The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.7.1271
29(7), 1271–1281

Suggestion of batter ability index in Korea baseball - focusing on the sabermetrics statistics WAR

Jea-Young Lee[a,1] · Hyeon-Gyu Kim[a]

[a]Department of Statistics, Yeungnam University

(Received August 8, 2016; Revised August 30, 2016; Accepted August 30, 2016)

Abstract

Wins above replacement (WAR) is one of the most widely used statistic among sabermatrics statistics that measure the ability of a batter in baseball. WAR has a great advantage that is to represent the attack power of the player and the base running ability, defensive ability as a single value. In this study, we proposed a hitter ability index using the sabermetrics statistics that can replace WAR based on Korea Baseball Record Data of the last three years (2013–2015). First, we calculated Batter ability index through the arithmetic mean method, the weighted average method, principal component regression and selected the method that had high correlation with WAR.

Keywords: principal component analysis, principal component regression, sabermatrics, wins above replacement

1. 서론

야구에서 타자의 타격능력을 쉽게 계산하면서 평가할 수 있는 통계량을 개발하기 위한 연구는 세이버메트릭스(sabermatrics)를 통해서 계속 진행 중이다. 세이버메트릭스는 누적된 자료를 토대로 통계적인 관점에서 야구에 관한 분석을 하는 연구분야이며, 이와 같은 방법으로 자료 분석하는 사람을 세이버메트릭션(sabermatrician)이라고 부른다 (Hong 등, 2016). 한국프로야구(Korea Baseball Organization; KBO)에서 타자 능력에 관한 연구는 Kim (2012), Lee와 Cho (2009), Lee (2014) 등이 있으며 특히, 기존의 단순한 통계량을 가공하여 야구기록을 보다 수학적·과학적으로 분석하는 세이버메트릭스 분야의 중요성은 점차 강조되고 있다 (Kang 등, 2014; Cho 등, 2007). 이는 타율, 타점, 득점 등과 같이 일차적인 방법으로 선수의 능력을 분석하는 것에서 벗어나 더 객관적이고 구체적이며 조금 더 고차원적인 방법으로 선수의 경기력을 평가할 수 있다는 점에서 야구경기 분석의 주류로 자리 잡아 가고 있다 (Seung과 Kang, 2012). 특히, 대체선수대비승수를 나타내는 wins above replacement(WAR)은 특정 선수 대신 투입됐을 때 얼마나 많은 승리에 기여했는가를 나타내는 수치로 지금까지 나와있는 많은 세이버메트릭스 통계량 중 미국프로야구(Major League Baseball; MLB)와 KBO에서 가장 공신력있는 통계량으로 이용되고 있다. 그 중 KBO에서 사용하고 있는 WAR은 다음과 같이 구할 수 있다 (kbreport, http://www.kbreport.com/statDic/detail?seq=22&contentsType=a304).

$$\text{WAR} = \frac{\text{Batting Runs} + \text{Base Running Runs} + \text{Fielding Runs} + \text{Positional Adjustment}}{\text{Runs Per Win}} \quad (1.1)$$

[1]Corresponding author: Department of Statistics, Yeungnam University, 280, Daehak-ro, Gyeongsan-si, Gyeongsangbuk-do 38541, Korea. E-mail: jlee@yu.ac.kr

KBO에서의 WAR은 선수의 공격능력(batting runs)과 주루능력(base running runs), 수비능력(fielding runs), 포지션조정(positional adjustment)의 합을 1승에 해당하는 득점(runs per win)으로 나눈 값으로 선수의 능력을 종합적으로 나타낸다는 점에서 큰 장점을 가진다. MLB에서도 WAR을 이용하고 있으나, 식 (1.1)에서 리그 조정(league runs)과 대체선수대비 타석수 보정(replacement runs)이 추가되어 계산된다는 점에서 KBO에서 사용되는 WAR과 조금 다르다. 따라서 본 연구에서는 먼저 KBO 자료로부터 구한 WAR과 세이버메트릭스 통계량들을 산술평균방법, 가중평균방법, 주성분회귀분석방법에 적용한 뒤 비교분석하여 상관계수가 가장 높은 분석방법을 채택하고, 최종 타자능력지수(batter ability index; BAI)를 제안한다. 데이터는 KB Report(kbreport.com) 기록실에 게시되어있는 2013년부터 2015년까지 지난 3년간의 데이터를 이용하였다.

본 연구의 구성은 다음과 같다. 2절에서는 세이버메트릭스 변수에 대한 설명과 연구에 사용된 데이터 및 분석방법을 소개한다. 3절에서는 산술평균방법과 상관계수를 이용한 가중평균방법, 주성분회귀분석방법을 통해 타자의 능력을 파악할 수 있는 지수를 개발하고 WAR과 비교하여 가장 근접한 지수를 최종으로 선택하여 타자능력지수로 제안한다. 마지막 4절에서는 연구의 결과를 요약하고 결론을 맺는다.

2. 연구 방법

2.1. 데이터 소개

본 연구는 타자력 요인에 관한 연구를 하기 위하여 2013년부터 2015년까지 한국프로야구의 규정타석을 만족한 153명의 타자들 중 동일한 타자의 경우 각각 서로 연관이 있을 것이라 생각하여 평균값으로 데이터를 종합해 총 83명의 선수들에 대한 데이터로 분석하였다. 데이터는 케이비리포트(kbreport.com) 기록실에 게시되어있는 데이터를 이용하였다. 변수는 타자의 경기력을 분석하는데 필요한 지수로 제한하였고, 체육전공 교수 1인, 야구전문가 2인의 도움을 받아 다음과 같은 13개의 세이버메트릭스 통계량을 이용하였다 (Yang 등, 2015).

- **공격공헌도(On Base plus Slugging; OPS)**
  가장 보편화되고 잘 알려진 세이버메트릭스 통계량으로서 출루율과 장타율의 합으로 구할 수 있어 쉽게 계산할 수 있는 장점이 있다. 단점으로는 단순히 출루율(OBP)과 장타율(SLG)을 더한다는 점인데, 이는 장타율에 치중된 값이라고 볼 수 있다. 하지만 많이 사용되는 이유는 아무래도 간편한 식 때문이다.

$$\mathrm{OPS} = \mathrm{OBP} + \mathrm{SLG}.$$

- **총생산평균(Gross Production Average; GPA)**
  OPS의 단점을 보완한 통계량으로 출루율에 1.8의 가중치를 두고 계산한다. 따라서 장타율이 과대평가되는 단점을 보완할 수 있다.

$$\mathrm{GPA} = \frac{1.8\mathrm{OBP} + \mathrm{SLG}}{4}.$$

- **수정타율(SECondary Average; SECA)**
  안타를 타수로 나누는 타율의 공식이 갖는 가장 큰 단점은 장타와 단타의 가치를 동일 시 하면서 볼넷은 인정하지 않으므로 SECA는 장타율의 가중치에 4사구와 도루의 가치를 고려해 만든 지수이다.

$$\mathrm{SECA} = \frac{(2\mathrm{B}) + 2(3\mathrm{B}) + 3(\mathrm{HR}) + (\mathrm{UBB}) + (\mathrm{SB}) - (\mathrm{CS})}{\mathrm{AB}},$$

여기서 2B는 2루타, 3B는 3루타, HR는 홈런, UBB는 고의4구 제외 4사구, SB는 도루성공, 그리고 CS는 도루실패이다.

- **종합공격력(Total Average; TA)**

  총 진루수를 총 아웃수로 나눈 것으로 한 시즌동안 타자가 한 번의 공격기회에서 얼마만큼 진루했는가를 나타내는 지수이다.

$$\mathrm{TA} = \frac{\mathrm{TB} + \mathrm{HBP} + \mathrm{BB} + \mathrm{SB}}{\mathrm{AB} - \mathrm{H} + \mathrm{CS} + \mathrm{GDP}},$$

  여기서 TB는 총루타수, HBP는 몸에 맞는 공, BB는 볼넷, AB는 타수, H는 안타수, 그리고 GDP는 병살타이다.

- **득점생산력(Run Created; RC)**

  세이버메트릭스의 창시자인 빌 제임스가 고안한 것으로 타자의 출루능력(A)과 주자를 진루시키는 능력(B)을 타자의 득점에 관한 창출능력으로 보고 한 타자가 팀이 득점을 올리는데 있어서 어느 정도 기여했는지를 나타낸다.

$$\begin{aligned}
\mathrm{RC} &= \frac{A \times B}{C},\\
A &= (\mathrm{H} + \mathrm{BB} + \mathrm{HBP} - \mathrm{CS} - \mathrm{GDP}),\\
B &= (\mathrm{TB}) + 0.52(\mathrm{SB} + \mathrm{SF} + \mathrm{SH}) + 0.26(\mathrm{BB} + \mathrm{HBP} - \mathrm{IBB}),\\
C &= (\mathrm{AB} + \mathrm{BB} + \mathrm{HBP} + \mathrm{SF} + \mathrm{SH}),
\end{aligned}$$

  여기서 SF는 희생플라이, SH는 희생타, IBB는 고의4구이다.

- **경기당 득점기여도(Run Created per 27 outs; RC/27)**

  RC가 타자가 몇 점을 만들어냈는지를 의미하는 것이라면 RC/27은 한 타자로 9명이 구성된 팀이 한 경기를 치르면 몇 점이나 뽑을 것인가를 평균수치화 한 것이다.

$$\begin{aligned}
\frac{\mathrm{RC}}{27} &= \frac{\mathrm{RC} \times 27}{A},\\
A &= (\mathrm{AB} - \mathrm{H} + \mathrm{SF} + \mathrm{SH} + \mathrm{CS} + \mathrm{GDP}).
\end{aligned}$$

- **추정득점력(eXtrapolated Runs; XR)**

  RC와 유사한 개념으로 팀의 득점에 얼마나 공헌했는지를 타나내며 RC와의 차이점으로는 1955년부터 1997년까지 메이저리그 공식기록을 회귀분석하여 도출한 선형공식이라는 점으로 정확도 면에서는 RC보다 낫다고 알려져 있다.

$$\begin{aligned}
\mathrm{XR} = {} & 0.5(1\mathrm{B}) + 0.72(2\mathrm{B}) + 1.04(3\mathrm{B}) + 1.44(\mathrm{HR}) + 0.34(\mathrm{BB} + \mathrm{HBP} - \mathrm{IBB})\\
& + 0.25(\mathrm{IBB}) + 0.18(\mathrm{SB}) - 0.32(\mathrm{CS}) - 0.09(\mathrm{AB} - \mathrm{H} - \mathrm{SO}) - 0.098(\mathrm{SO})\\
& - 0.37(\mathrm{GDP}) + 0.37(\mathrm{SF}) + 0.04(\mathrm{SH}),
\end{aligned}$$

  여기서 1B는 1루타, SO는 삼진아웃이다.

- **득점공헌도(Batting Runs; BR)**

  리그 타자들의 평균공격력을 0으로 놓은 상태에서 상대적으로 타자의 팀 공격기여도를 수치화 한 지

수다.

$$\begin{aligned}\mathrm{BR} = {} & 0.47(1\mathrm{B}) + 0.78(2\mathrm{B}) + 1.09(3\mathrm{B}) + 1.4(\mathrm{HR}) + 0.33(\mathrm{BB} + \mathrm{HBP}) + 0.3(\mathrm{SB}) \\ & - 0.6(\mathrm{CS}) - 0.25(\mathrm{AB} - \mathrm{H}) - 0.5(\mathrm{OOB}),\end{aligned}$$

여기서 OOB는 주루사이다.

- **순수장타율(ISOlated power; ISO)**
  타자의 파워히팅 능력을 나타낸 것으로 장타율에 타율(AVG)이 포함되어 있는 것을 고려해서 고안된 지수이다.

$$\mathrm{ISO} = \mathrm{SLG} - \mathrm{AVG}.$$

- **호타준족(Power Speed Number; PSN)**
  타자의 호타준족 정도를 나타낸 지수로 도루실패를 감안하지 않는다는 단점을 지니고 있다.

$$\mathrm{PSN} = \frac{\mathrm{HR} \times \mathrm{SB} \times 2}{\mathrm{HR} + \mathrm{SB}}.$$

- **타석 당 득점기대(weight On Base Average; wOBA)**
  타자가 타석에 들어섰을 때의 여러 가지 상황에 따른 득점 가치를 고려하여 타자의 생산력을 나타낸 것으로 득점과의 상관관계가 매우 높고, 출루율의 가치가 저평가 받는 문제점을 개선하였지만 구장효과를 반영하지 못하며 주루와 타격을 분리하지 못했다는 단점을 지닌다.

$$\mathrm{wOBA} = \frac{0.7(\mathrm{BB}-\mathrm{IBB})+0.73(\mathrm{HBP})+0.89(1\mathrm{B})+1.27(2\mathrm{B})+1.61(3\mathrm{B})+2.07(\mathrm{HR})+0.25(\mathrm{SB})+0.5(\mathrm{CS})}{\mathrm{AB} - \mathrm{IBB}}.$$

- **공격기대승률(Offensive Winning Percentage; OW%)**
  리그의 평균 득점과 한 타자의 RC/27을 고려해 한 타자만으로 이루어진 타선이면 몇 %의 승률을 가지는 지를 나타내는 지수로 리그평균 변화를 고려하여 한 타자의 한 시즌 리그 지배력을 알 수 있다.

$$\mathrm{OW}\% = \frac{(\mathrm{RC}/27)^2}{(\mathrm{RC}/27)^2 + \text{리그}(\mathrm{RC}/27)^2}.$$

- **인플레이타구비율(Batting Average on Balls In Play, BABIP)**
  타자가 친 공이 페어지역 안에 떨어진 경우만을 나타내는 지수로 타자와 투수에게 모두 적용이 가능하다. 본인의 타격 스타일에 따라 자신만의 고유한 BABIP를 가지게 되며, 라인드라이브 > 그라운드볼 > 플라이볼 순으로 BABIP 값이 높게 형성된다.

$$\mathrm{BABIP} = \frac{\mathrm{H} - \mathrm{HR}}{\mathrm{AB} - \mathrm{SO} - \mathrm{HR} + \mathrm{SF}}.$$

2.2. 분석 방법

본 연구에서는 한국프로야구 타자의 능력을 파악하는 지수를 개발 및 제안하기 위해 13개의 세이버메트릭스 통계량을 이용해서 산술평균방법과 가중평균방법, 주성분 분석방법을 적용하였다. 먼저 산술평균방법은 총 13개 변수를 표준화하여 산술평균을 구한 뒤, 각 타자들의 능력을 평가하였다. 여기에서는 모든 변수들이 동일한 가중치$(1/n)$로 반영이 되었으므로, OPS와 GPA, wOBA와 같은 비슷한 능력을 측정하는 경우 이 부분의 값이 큰 타자가 높은 점수를 받을 것이다. 이러한 단점을 보완하기 위해서 두

Table 3.1. Simple statistics for 13 variable

	N	Mean	Std	Min	Max
OPS	83	.839	.104	.621	1.199
GPA	83	.285	.031	.217	.392
SECA	83	.311	.088	.171	.617
TA	83	.852	.159	.567	1.499
RC	83	79.123	23.747	36.960	166.810
RC/27	83	6.557	1.884	3.340	14.510
XR	83	71.180	19.763	29.566	135.955
BR	83	25.266	17.504	−8.350	89.540
ISO	83	.163	.065	.057	.377
PSN	83	8.204	6.429	.000	30.088
wOBA	83	.370	.037	.291	.488
OW%	83	.469	.125	.199	.813
BABIP	83	.337	.031	.214	.407

OPS = on base plus slugging; GPA = gross production average; SECA = SECondary average; TA = total average; RC = run created; RC/27 = run created per 27 outs; XR = eXtrapolated Runs; BR = batting runs; ISO = ISOlated power; PSN = power speed number; wOBA = weight on base average; OW% = offensive winning percentage; BABIP = batting average on balls in play.

번째로 가중평균방법을 이용하였다. 모든 변수의 상관계수를 구하고, 이를 이용해서 구한 가중평균으로 타자들의 능력을 평가하였다. 상관계수가 높은 세이버메트릭스 통계량끼리 그룹으로 묶은 후, 각 다른 가중치를 부여함으로써 타자의 능력을 살펴 볼 수 있다. 그러나 13개의 변수를 모두 사용하여 다중회귀분석을 하는 경우 설명변수들 사이의 높은 상관관계에 의해 다중공선성(multicollinearty) 문제를 야기시킬 수 있다 (Kwon, 2008).

따라서, 이러한 문제를 해결하기 위해서, 본 논문에서는 주성분분석을 통해 주성분변수를 얻어 이를 설명변수로 이용함으로써 다중공선성 문제를 해결 하였다 (Oh 등, 2012). 주성분분석에서는 주성분의 개수를 선택할 때, 상관계수행렬을 이용할 시 일반적으로 고유치 값이 1이상인 주성분과 총 변동의 설명력이 80% 이상인 주성분 변수를 선택할 수 있다. 성분 부하 값이 크다는 것은 그에 대응하는 원 변수의 영향이 크다는 것을 의미하므로 성분 부하 값이 큰 변수를 파악하여 주성분의 이름을 부여하면 된다. 여기서 선택된 주성분이 새로운 회귀모형의 설명변수로 이용되고 주성분 점수가 설명변수의 측정치가 된다. 새로운 회귀모형은 다음과 같다.

$$y_i = \beta_0 + \beta_1 \mathrm{Prin}_1 + \beta_2 \mathrm{Prin}_2 + \cdots + \beta_p \mathrm{Prin}_p + \epsilon_i, \quad i = 1, \ldots, k, \tag{2.1}$$

여기서 $\mathrm{Prin}_1, \mathrm{Prin}_2, \ldots, \mathrm{Prin}_p$는 주성분 변수가 되고, $\beta_0, \beta_1, \ldots, \beta_p$는 회귀계수 추정치이며, ϵ_i는 평균벡터가 0, 공분산행렬이 $\mathrm{cov}(\epsilon_i) = I$인 확률오차벡터이다 (Bae 등, 2012). 식 (2.1)에서 추정된 y값을 타자능력지수로 두고 WAR과 비교한다.

3. 타자능력지수 제안

분석에 앞서 사용될 지수들의 기초통계량은 Table 3.1과 같다. Table 3.1에서 13개의 변수들의 평균값을 보면 RC, RC/27, XR, BR, PSN 값의 단위가 차이나는 것을 확인 할 수 있다. 따라서 우리는 BAI를 제안하기 위해, 3.1절에서는 변수들의 측정단위가 차이 나기 때문에 표준화 시킨 변수들을 이용하여 산술평균, 가중평균, 주성분회귀분석을 실시한다. 3.2절에서는 분석된 결과들을 가지고 WAR과 비교하여 가장 근접한 방법을 찾고 3.3절에서 WAR과 가장 근접한 방법으로 구한 타자능력지수를 제안한다.

Table 3.2. Correlation coefficient matrix of sabermetrics statistics

	OPS	GPA	SECA	TA	RC	RC/27	XR	BR	ISO	PSN	wOBA	OW%	BABIP
OPS	1	**.993****	.838**	.961**	.914**	.973**	.914**	.966**	.855**	.472**	**.983****	**.973****	.464**
GPA		1	.822**	.966**	.911**	.978**	.905**	.969**	.797**	.443**	**.994****	**.983****	.518**
SECA			1	.895**	.802**	.826**	.818**	.857**	**.881****	.518**	.815**	.792**	.137
TA				1	.940**	**.982****	.921**	**.982****	.802**	.539**	.960**	.960**	.477**
RC					1	.937**	**.989****	.969**	.761**	.565**	.893**	.919**	.457**
RC/27						1	.918**	**.977****	.783**	.482**	.973**	.973**	.529**
XR							1	.961**	.798**	.540**	.882**	.913**	.391**
BR								1	.794**	.537**	.962**	.964**	.481**
ISO									1	.506**	.772**	.755**	.046
PSN										1	.433**	.445**	.099
wOBA											1	**.979****	.532**
OW%												1	.572**
BABIP													1

OPS = on base plus slugging; GPA = gross production average; SECA = SECondary average; TA = total average; RC = run created; RC/27 = run created per 27 outs; XR = eXtrapolated Runs; BR = batting runs; ISO = ISOlated power; PSN = power speed number; wOBA = weight on base average; OW% = offensive winning percentage; BABIP = batting average on balls in play.

3.1. 데이터 분석 및 결과

3.1.1. 산술평균결과 세이버메트릭스 통계량들 간에 값의 차이가 크기 때문에 변수를 표준화하여 분석하였다. 표준화한 변수의 값을 산술평균(AVG_{H1})으로 계산하면, 식 (3.1)과 같다.

$$\mathrm{AVG}_{H1} = \frac{(Z_1 + Z_2 + Z_3 + \cdots + Z_n)}{n}, \qquad Z_i = \frac{(X_i - \mu_i)}{\sigma_i}, \quad i = 1, \ldots, n, \tag{3.1}$$

여기에서 변수는 총 13개가 쓰였기 때문에 $n = 13$이 된다. 산술평균을 계산하여 상위 10명의 순위를 나타낸 결과와 WAR과의 비교는 3.4절에서 다루도록 한다.

3.1.2. 상관계수를 활용한 가중평균결과 산술평균을 이용하는 경우 모든 변수들이 같은 가중치를 가지기 때문에 비슷한 성향의 변수인 OPS와 GPA 등이 높은 사람이 높은 점수를 받을 것이다. 이러한 문제점을 보완하기 위해 가중평균을 이용하였다. 세이버메트릭스 변수에 가중치를 부여할 때, 주관적으로 부여하는 방법보다 객관적인 방법으로 상관계수가 높은 변수를 그룹화 하여 가중치를 설정하였다. Table 3.2는 세이버메트릭스 통계량들의 상관계수를 나타낸 표이다. Table 3.2를 보면 OPS, GPA, wOBA 등은 상관계수가 매우 높은 것을 확인 할 수 있다. 그리고 타자의 능력을 계산한 세이버메트릭스 통계량이기 때문에 전반적으로 상관계수 값들이 높은 것을 알 수 있었다. 상관계수가 높다는 것은 유사한 능력을 가지고 있는 것이기 때문에 같은 그룹이 된다. 따라서 13개의 변수를 6개의 그룹으로 분류하여 분석을 진행하였다. 가중평균(wAVG_{H2})의 가중치는 아래와 같이 구할 수 있다.

$$\mathrm{wAVG}_{H2} = \left[\frac{\mathrm{OPS} + \mathrm{GPA} + \mathrm{wOBA} + \mathrm{OW\%}}{4} + \frac{(\mathrm{SECA} + \mathrm{ISO})}{2} + \frac{\mathrm{TA} + \mathrm{RC/27} + \mathrm{BR}}{3} + \frac{\mathrm{RC} + \mathrm{XR}}{2} + \mathrm{PSN} + \mathrm{BABIP}\right] \Big/ 6. \tag{3.2}$$

위 식으로 가중평균을 계산하여 상위 10명의 순위를 나타낸 결과와 WAR과의 비교는 3.4절에서 다루도록 한다. 가중평균방법에서 생길 수 있는 문제점은 데이터 숫자에 비해 변수의 개수가 많아 변수들 간의

Table 3.3. Eigenvectors of sabermetrics statistics

	HA (Prin_1)	QB (Prin_2)	SASR (Prin_3)
OPS	.984	.028	−.098
GPA	.981	.103	−.092
SECA	.875	−.336	−.144
TA	.988	.010	−.002
RC	.959	.001	.090
RC/27	.984	.096	−.030
XR	.952	−.057	.028
BR	.991	.026	.013
ISO	.838	−.418	−.196
PSN	.551	−.416	.716
wOBA	.972	.127	−.089
OW%	.974	.161	−.043
BABIP	.473	.821	.246

OPS = on base plus slugging; GPA = gross production average; SECA = SECondary average; TA = total average; RC = run created; RC/27 = run created per 27 outs; XR = eXtrapolated Runs; BR = batting runs; ISO = ISOlated power; PSN = power speed number; wOBA = weight on base average; OW% = offensive winning percentage; BABIP = batting average on balls in play.

다중공선성이 발생할 수 있으며, 상관계수의 크기만으로 변수들을 분류하는 게 쉽지 않다는 것이다. 따라서 변수를 정량적으로 축약하는 주성분분석을 다음 절에서 활용하였다.

3.1.3. 주성분 회귀분석결과 원자료의 모든 변수 13개를 이용해서 상관계수의 크기를 분류하는 것은 쉽지 않다. 이러한 문제를 해결하기 위해서 주성분 분석을 통해 변수를 축약하였다. 주성분 분석을 통해서 나온 고유치와 누적 설명력을 이용해서, 그에 맞는 합당한 변수들로 축약을 할 수 있는데, 일반적으로 고유치 1이상이고 누적 설명력이 80% 이상인 주성분을 선택하는 것이 기본이다. 하지만 본 연구에서는 13개 변수의 효과를 모두 포함하기 위해 총 3개의 변수로 축약을 했다. 제1주성분의 고유치는 10.572이고 제2주성분은 1.201, 제3주성분은 0.67의 값을 가졌다. 축약 된 3개의 주성분 변수가 96%의 누적설명력을 가지고 있다. Table 3.3은 선택된 주성분 변수에 의해 얻어진 고유벡터를 나타낸 표이다. 이 고유벡터를 바탕으로 변수를 축약을 할 수 있었다.

각각 주성분 내에서 고유벡터 값을 큰 변수들끼리 묶은 후 이를 이용해서 주성분에 이름을 부여할 수 있다. Table 3.3에서 제1주성분(prin_1)의 계수 크기를 보면 PSN과 BABIP를 제외하고 변수의 부하 값이 크므로 제1주성분은 타격능력(hitting ability; HA)이라 할 수 있다. 제2주성분(prin_2)에서는 BABIP의 부하값이 크므로 타구의 질(quality of batting; QB)이라 할 수 있다. 제3주성분(prin_3)에서는 PSN의 부하값이 크므로 호타준족(slugger and swift runner; SaSR)이라 할 수 있다. 각 주성분의 이름을 정한 뒤, 주성분 점수를 구할 수 있다. 아래 식의 각 부분은 얻어진 주성분의 점수 산출식이다.

$$\text{HA}(\text{prin}_1) = 0.984Z_1 + 0.981Z_2 + \cdots + 0.473Z_{13}, \tag{3.3}$$

$$\text{QB}(\text{prin}_2) = 0.028Z_1 + 0.103Z_2 + \cdots + 0.821Z_{13}, \tag{3.4}$$

$$\text{SaSR}(\text{prin}_3) = -0.098Z_1 - 0.092Z_2 - \cdots + 0.246Z_{13}. \tag{3.5}$$

위 식에서 Z_i는 각 변수를 표준화한 값이며, 총 13개의 세이버메트릭스 변수를 표준화하여 주성분

Table 3.4. Comparison of top 10 rank and scores result in three ways and WAR

Name	Team	WAR index (rank)	AVG_{H1} scores (rank)	$wAVG_{H2}$ scores (rank)	$PRIN_{H3}$ scores (rank)
Eric Allyn Thames	NC	9.045 (1)	2.990 (1)	2.837 (1)	67.919 (1)
Park, Byeong-ho	Nexen	7.580 (2)	2.104 (2)	1.846 (2)	49.701 (2)
Seo, Geon-chang	Nexen	7.510 (3)	1.363 (4)	1.284 (4)	32.599 (4)
Kang, Jeong-ho	Nexen	7.210 (4)	1.401 (3)	1.239 (5)	33.693 (3)
Choi, Jeong	SK	6.690 (5)	1.321 (5)	1.388 (3)	30.565 (6)
Tamaico Navarro	Samsung	6.485 (6)	1.297 (6)	1.211 (6)	31.840 (5)
Park, Seok-min	Samsung	5.797 (7)	0.884 (10)	0.608 (13)	23.413 (8)
Yang, Eui-ji	Doosan	5.510 (8)	0.510 (19)	0.385 (21)	14.557 (18)
Andy Marte	KT	5.250 (9)	0.693 (14)	0.452 (17)	19.020 (14)
Choi, Hyeong-woo	Samsung	5.133 (10)	0.911 (8)	0.667 (10)	24.062 (7)

점수를 구하였다. 축약된 변수를 이용해서 회귀분석을 진행한 결과 얻어진 주성분 회귀분석의 모형($PRIN_{H3}$)은 다음과 같다.

$$PRIN_{H3} = 2.738 + 1.843 \times HA - 0.008 \times QB + 0.077 \times SaSR. \tag{3.6}$$

위 주성분 회귀모형($PRIN_{H3}$)에서 계수들의 유의성을 확인해 본 결과 HA($prin_1$)의 p-값은 < 0.0001으로 매우 유의하게 나왔으나 QB($prin_2$)의 p-값은 0.914, SaSR($prin_3$)의 p-값은 0.299으로 유의하지 않게 나왔다. 따라서 제1주성분(HA)만을 이용한 회귀모형과 $PRIN_{H3}$를 WAR과 각각 비교한 결과, HA만을 이용한 회귀모형과의 상관계수 값은 0.94($p < 0.0001$)였고 $PRIN_{H3}$와의 상관계수 값은 0.943($p < 0.0001$)였다. 두 회귀모형은 미세한 차이를 보였으나 $PRIN_{H3}$의 상관계수 값이 더 크므로 $PRIN_{H3}$와 AVG_{H1}, $wAVG_{H2}$를 다음 절에서 비교하였다.

3.2. WAR과 분석방법에 따른 결과 비교

3.1절에서 산술평균과 가중평균, 주성분회귀분석을 통해 타자를 평가할 수 있는 지수를 만들었다. 이 결과들을 바탕으로 KBO에서 2013년부터 2015년까지 규정타석을 만족하고 동일한 선수의 경우 평균값을 계산하여 총 83명의 타자들로부터 WAR과 세 가지 방법에 따른 상위 10명의 지수 값과 순위를 비교한 결과는 Table 3.4와 같다.

Table 3.4를 보면 테임즈(Eric Allyn Thames)와 박병호(Park, Byeong-ho)는 항상 같은 결과가 나왔고, 3등부터 6등까지는 순서가 섞이며 7등 이후로는 바뀌는 폭이 큰 것으로 나왔다. $PRIN_{H3}$의 9등과 10등은 김태균(Kim, Tae-kyun)과 아두치(Jim Charles Adduci)로 나타났다. 정확한 비교를 위해 상관분석을 실시한 결과 3개 방법 모두 유사한 결과가 나왔으나 면밀히 살펴보면 WAR과 산술평균 사이에는 상관계수 값이 0.941($p < 0.0001$)로 나왔고 가중평균과의 비교에서는 0.929($p < 0.0001$)가 나왔다. 마지막 $PRIN_{H3}$와의 비교에서는 상관계수 값이 0.943($p < 0.0001$)으로 가장 높게 나온 것을 알 수 있었다. 각 점수별로 값의 단위가 차이 나기 때문에 변수를 표준화시켜 WAR과 각 분석방법으로 얻은 값들을 비교한 그래프는 Figure 3.1과 같다.

Figure 3.1을 보면 3개의 방법 모두 WAR과 높은 상관관계를 띄고 있어 그래프만으로는 구분하기가 어렵다. 따라서 WAR과 각 분석방법에 따른 결과 값과의 상관계수를 이용해서 가장 큰 상관계수를 가지는 $PRIN_{H3}$이 최종 타자능력지수로 적합하다고 판단하였다.

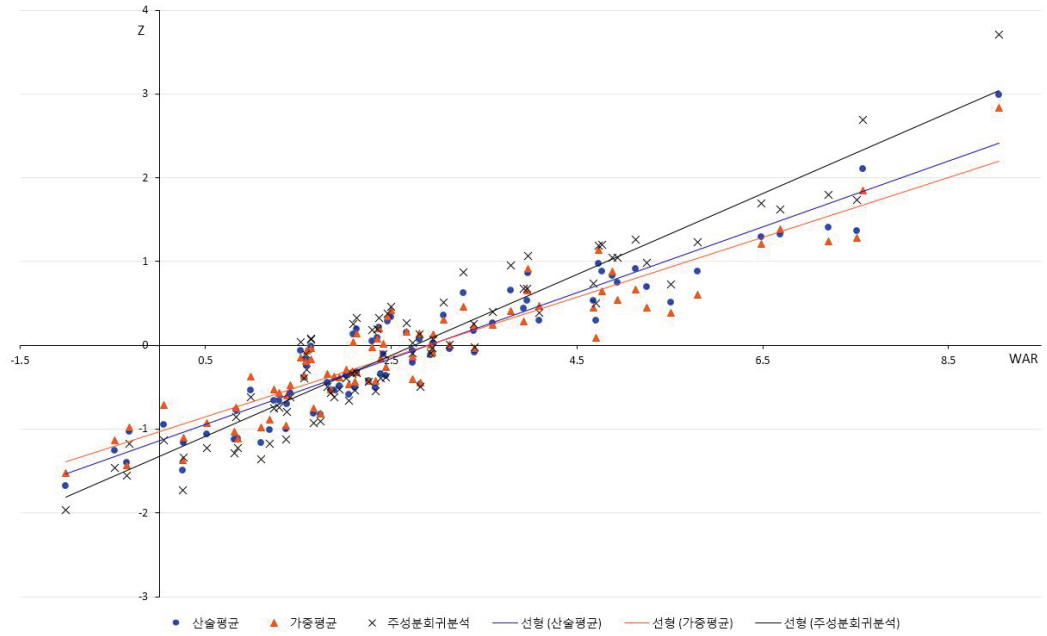

Figure 3.1. Scatter plot.

3.3. 타자능력지수(Batter Ability Index; BAI) 제안

본 논문은 타자의 능력을 평가하는 지수 중 WAR과 가장 근접한 방법을 제안하기 위해 산술평균방법, 가중평균방법, 주성분 회귀분석방법을 이용하였다. 각 방법으로부터 얻은 값과 WAR을 비교한 결과, 주성분 회귀모형의 상관계수가 0.943($p < 0.0001$)으로 WAR과 가장 근접하며 효율적인 분석 방법으로 나타났다. 따라서 주성분분석을 이용하여 총 13개의 세이버메트릭스 변수를 3개의 주성분 변수(HA, QB, SaSR)로 축약하고 3개의 주성분 변수를 통해 최종 타자능력지수(BAI)를 제안한다. 식 (3.6)을 최종 타자능력지수로 선택하여 BAI로 명명하였다.

$$\text{BAI}(= \text{PRIN}_{H3}) = 2.738 + 1.843 \times \text{HA} - 0.008 \times \text{QB} + 0.077 \times \text{SaSR}. \tag{3.7}$$

식 (3.7)인 BAI와 WAR의 R^2값은 0.889($r = 0.943$)로 뛰어난 설명력을 가지고 있으며, WAR에 가장 근접한 모형이기 때문에 타자의 능력을 파악하는 지수로 타당하다고 평가하였다.

4. 결론 및 토의

본 논문은 MLB 뿐만아니라 KBO에서 타자의 능력을 평가하는데 가장 공신력 있는 통계량으로 사용되는 WAR에 가장 근접한 타자능력지수를 제안하기 위해 타자능력에 영향을 미치는 세이버메트릭스 변수 13개를 사용하여 산술평균, 가중평균, 주성분 분석, 주성분분석에 의한 회귀분석을 적용하였다. 데이터는 케이비리포트(kbreport.com) 기록실에 게시되어있는 2013년부터 2015년까지 지난 3년간의 데이터를 이용하였다. 먼저 변수의 단위가 다르기 때문에 변수를 표준화 하여 산술평균을 구하고, 두 번째로 13개 세이버메트릭스 통계량의 상관관계를 이용하여 6개의 그룹으로 나눈 뒤 가중평균을 계산하였다. 그러나 앞의 두 방법의 경우 유사한 항목이 많고 가중치 문제와 다중공선성 문제가 발생하기 때

문에 세 번째로 이를 보완하는 주성분 분석을 실시하였다. 주성분 분석을 통해 13개의 변수를 3개의 주성분 변수(HA, QB, SaSR)로 축약하고, 주성분 점수를 계산하여 회귀모형을 구하였다. 마지막으로 이 세 가지 방법을 통해 구한 값들을 WAR과 비교하여 상관계수가 가장 높은 주성분분석방법($r = 0.943$, $p < 0.0001$)을 채택하고 최종 BAI로 제안했다.

상황에 따라 타자의 심리도 타자능력에 영향을 끼치는 만큼 타격지표로 타자의 모든 능력을 파악할 수는 없기 때문에 본 연구에서 제안한 BAI 또한 완벽한 타자능력지수는 아니다. 하지만 BAI 지수는 크게 HA, QB, SaSR 3가지를 반영하여 타자의 능력을 평가하는 지수로써, 많은 세이버메트릭스 통계량들을 포함하여 복잡하게 계산하여야 하는 지수들보다 좀 더 간편하고 쉽게 계산할 수 있어서 타자를 객관적으로 평가하고 경기 전략을 짜는 데 도움이 될 것이다.

References

Bae, J. Y., Lee, J. M., and Lee, J. Y. (2012). Predicting Korea Pro-Baseball Rankings by principal component regression analysis. *The Journal of Korean Statistical Society*, **19**, 367–379.

Cho, Y. S., Cho, Y. J., and Sin, S. G. (2007). A study on winning and losing in Korean Professional Baseball League, *Journal of the Korean Data & Information Science Society*, **9**, 501–510.

Hong, J. S., Kim, J. Y., and Sin, D. S. (2016). Alternative hitting ability index for KBO, *Journal of the Korean Data & Information Science Society*, **27**, 677–687.

Kang, J. G., Park, S. C., and Kim, J. H. (2014). Suggestion of Korea professional baseball record system using Saber-Metrics, *Korean Society for Internet Information*, **15**, 143–144.

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.

KB Report (2013–2015). http://www.kbreport.com/leader /main.

Kwon, S. H. (2008). *Utilizing and Analysis of Multivariate Data*, Freeacademy, Seoul.

Lee, J. T. (2014). Measurements for hitting ability in the Korean pro-baseball, *Journal of the Korean Data & Information Science Society*, **25**, 349–356.

Lee, J. T. and Cho, H. S. (2009). Estimation of OBP coefficient in Korean professional baseball, *Journal of the Korean Data & Information Science Society*, **25**, 357–363.

Oh, G. J., An, J. J., and Sim, G. S. (2012). Multi-currencies portfolio strategy using principal component analysis and Logistic regression, *Journal of the Korean Data & Information Science Society*, **23**, 151–159.

Seung, H. B. and Kang, G. 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.

Yang, D. E., Cho, E. H., Bae, S. W., and Jung, S. W. (2015). Analysis of professional Korean baseball batter's performances factors, *Journal of Sport and Leisure Studies*, **60**, 305–313.

한국프로야구에서 타자능력지수 제안 - 대체선수대비승수(WAR)을 중심으로

이제영[a,1] · 김현규[a]

[a]영남대학교 통계학과

(2016년 8월 8일 접수, 2016년 8월 30일 수정, 2016년 8월 30일 채택)

요 약

야구에서 타자의 능력을 측정하는 많은 세이버메트릭스 통계량들 중에서 대체선수대비승수(wins above replacement; WAR)은 가장 많이 쓰이는 통계량이다. WAR은 선수의 공격능력과 주루능력, 수비능력 등을 하나의 수치로 표현하는 방법이란 점에서 큰 장점을 가지고 있다. 본 논문에서는 지난 3년간(2013–2015년) 한국프로야구 기록 자료를 바탕으로 세이버메트릭스 변수들의 값을 구한 뒤, 이를 이용하여 WAR을 대체할 수 있는 타자능력지수를 제안하였다. 타자능력지수는 산술평균방법, 가중평균방법, 주성분회귀분석 등을 통해 산출하고 WAR과 비교하여 가장 관계가 높은 방법을 선택하였다.

주요용어: 대체선수대비승수, 세이버메트릭스, 주성분분석, 주성분회귀분석

[1]교신저자: (38541) 경상북도 경산시 대학로 280, 영남대학교 통계학과. E-mail: jlee@yu.ac.kr