Journal of the Korean Data & Information Science Society
2017, **28**(4), 863–874

http://dx.doi.org/10.7465/jkdi.2017.28.4.863
한국데이터정보과학회지

한국프로야구에서 선발투수의 투수능력지수 제안 - 대체선수대비승수 (WAR)을 중심으로

김현규[1] · 이제영[2]

[12]영남대학교 통계학과

접수 2017년 5월 16일, 수정 2017년 7월 5일, 게재확정 2017년 7월 5일

요 약

야구선수들의 능력을 측정하는 많은 세이버메트릭스 통계량들 중에서 대체선수대비승수 (WAR)은 가장 많이 사용되는 통계량이다. WAR의 장점은 투수와 타자처럼 서로 다른 포지션임에도 불구하고 선수들의 WAR을 비교할 수 있다는 점이다. 하지만 WAR은 복잡한 형태로 일반적으로 제공되는 기록만으로 구하기 어렵다. 따라서 본 논문에서는 지난 3년간 (2014-2016년) 한국프로야구 기록 자료를 바탕으로 세이버메트릭스 변수를 계산한 뒤, 이를 이용하여 WAR을 대체할 수 있는 선발투수능력지수를 제안한다. 선발투수능력지수는 산술평균방법, 가중평균방법, 주성분회귀분석 등을 통해 산출한 뒤, WAR과 비교하여 가장 관계가 높은 방법을 선택하였다. 이는 선발투수의 능력을 파악하는데 유용하게 사용될 것이다.

주요용어: 대체선수대비승수, 세이버메트릭스, 주성분분석, 주성분회귀분석.

1. 서론

한국프로야구 (Korea baseball organization; KBO)의 연 관중 수는 꾸준히 증가 추세를 보이고 있다. 지난해에는 한국프로야구 역사상 가장 최대인 800만 관중을 넘어섰다. 프로야구에 대한 관심이 높아지면서 야구기록에 대한 중요성과 세이버메트릭스 (Sabermatrics)에 대한 관심도 점점 더 커지고 있다. 세이버메트릭스는 야구에 대한 실증적인 분석을 하는 것을 말한다. 야구의 통계적인 분석, 수학적 분석 또한 세이버메트릭스라고 볼 수 있다. 이와 같은 방법으로 자료 분석하는 사람을 세이버메트릭션 (Sabermatrician)이라고 부른다 (Hong 등, 2016). 한국프로야구에서도 세이버메트릭스를 통한 연구는 Kim (2012), Lee와 Cho (2009), Lee (2014) 등이 있다. 이는 단순히 스트라이크 수, 볼넷 수 등으로 선수의 능력을 분석하는 것에서 벗어나 보다 실증적인 연구로 선수의 능력을 평가한다는 점에서 야구 경기 분석의 주류로 자리 잡고 있다 (Seung과 Kang, 2012). 세이버메트릭스 통계량은 야구 경기 중에도 흔히 볼 수 있는데 KBO 기록실에서 나타내는 통계량은 타자의 경우 공격공헌도 (OPS), 타율 (AVG), 출루율 (OBP), 순수장타율 (ISOP) 등 11개가 있고, 투수의 경우 평균자책점 (ERA), 이닝당 출루율 (WHIP), 피안타율 (oAVG) 등 12개가 있다. 특히, 세이버메트릭스 통계량 중에서 대체선수대비승수인 WAR (wins above replacement)은 가장 공신력 있는 통계량이고, 미국프로야구 (major league baseball; MLB)와 KBO에서 선수를 평가함에 있어 가장 많이 사용된다 (`http://m.`

[1] (38541) 경상북도 경산시 대학로 280, 영남대학교 통계학과, 석사과정.
[2] (38541) 경상북도 경산시 대학로 280, 영남대학교 통계학과, 교수. E-mail: jlee@yu.ac.kr

mlb.com/news/article/182980276/best-late-round-picks-in-draft-history/, http://osen.mt.co.kr/article/G1110627260). 실제로 야구 경기 중계 방송 중에도 포지션 별 WAR 수치를 보여주고 있으며, WAR을 포함한 세이버메트릭스 통계량을 활용하여 선수들의 연봉추정 모형을 제시한 연구도 진행된 바 있다 (Chang과 Zenilman; 2013). 이러한 WAR의 가장 큰 장점은 투수와 타자 등 모든 포지션의 선수의 어떤 행위도 철저하게 득점과 승리를 위한 수단으로 바라보아 승리기여도라는 단 하나의 숫자로 표현해 주는 것으로 서로 다른 구단, 다른 포지션의 선수들과 비교가 가능하다.

WAR을 산출하는 방식은 조정 실점 (Adjusted RA) 산출, 1승당 필요한 점수 (Runs to Win), 기대승률 산출 등의 복잡한 과정을 거쳐서 정의된다. 식은 다음과 같다.

$$\text{WAR} = (\text{Expected\%} - 0.38) \times \text{IP}/9. \tag{1.1}$$

식 (1.1)에서 투수의 기대승률 (Expected%)은 다음과 같이 정의된다.

$$\text{Expected\%} = (\text{Adjusted RA} - \text{RA}/9)/\text{Runs per Win} + 0.5. \tag{1.2}$$

기대승률에서 조정 실점 값 (adjusted RA)은 구장효과 (park factor), 팀 수비능력 등을 고려한 중립적인 평균 실점 값과 리그평균 자책점과 리그평균 실점의 비를 나눈 값으로 아래의 식과 같다.

$$\text{Adjusted RA} = \text{FIP}/(\text{league ERA} : \text{league RA}). \tag{1.3}$$

케이비레포트에서 조정 실점 값은 수비무관 평균자책점 (fielding independent pitching; FIP)에 기반을 두고 있다. Runs per win은 특정 투수가 등판했을 때 1승당 필요한 점수로 아래의 식과 같다.

$$\text{Runs per win} = (((18 - \text{IP}) \times \text{Adjusted RA}/9 + \text{IP} \times \text{RA}/9) + 2) \times 1.5. \tag{1.4}$$

Runs per win은 한 경기에서 양 팀의 공격이 각각 9이닝씩 총 18이닝이 이루어진다고 보았을 때, 특정 투수가 등판한 이닝에서는 그 선수의 이닝당 실점 값을 적용하고 그 외의 이닝에서는 리그의 평균적인 점수, 즉 앞서 산출한 조정 실점만큼의 점수가 발생했다고 가정하고 1승을 올리기 위한 값을 구한다는 개념이다. 마지막으로 기대승률에 더해지는 0.5는 리그 전체의 평균 승률을 의미한다. 이 과정을 통해 산출한 1승당 점수를 바탕으로 투수의 기대 승률 (Expected%)이 정의된다. 따라서 WAR은 기대승률과 투수의 시즌 투구이닝 (IP)을 종합하여 정의된다. 이와 같이 WAR (1.1)식은 식 (1.3)의 조정실점 (adjusted RA)과 식 (1.4)의 1승당 필요한 점수 (runs per win)에 의해 복잡한 과정을 거쳐 산출된다. 특히 식 (1.3)의 조정실점은 구장효과 부분도 포함되기 때문에 기본적으로 제공되는 자료만으로는 일반인이 구하기 어렵다. 따라서 본 연구에서는 케이비레포트에 제시되어 있는 세이버메트릭스를 사용하여 WAR에 가장 근접한 선발투수의 투수능력지수를 제안하는 것이 목적이다.

본 연구에서는 KBO 자료로부터 구한 세이버메트릭스 통계량들을 산술평균, 가중평균, 주성분회귀분석 방법을 적용한 뒤 WAR과 비교하여 상관계수가 가장 높은 방법을 채택하고, 최종적으로 선발투수능력지수 (Starting pitcher ability index; SPAI)로 제안한다. 데이터는 케이비레포트 (www.kbreport.com) 기록실에 있는 자료를 이용하였고, 최근 선발투수의 능력을 분석하기 위한 일환으로 지난 3년간 (2014-2016)의 데이터를 사용하였다.

본 연구의 구성은 다음과 같다. 2절에서는 분석에 사용된 세이버메트릭스 통계량에 대한 설명과 분석방법을 소개한다. 3절에서는 산술평균방법과 상관계수를 이용한 가중평균방법, 주성분회귀분석 방법을

통해 선발투수의 능력을 평가할 수 있는 지수를 개발하고 WAR과 비교하여 가장 근접한 지수를 최종으로 선택하여 선발투수능력지수로 제안한다. 마지막으로 4절에서는 연구의 결과를 요약하고 결론을 맺는다.

2. 연구방법

2.1. 데이터 소개

본 연구는 선발투수의 능력에 관한 연구를 하기 위하여 2014년부터 2016년 까지 한국프로야구의 규정이닝을 만족한 투수 60명 중 동일한 선수의 경우 년도별로 서로 연관이 있을 것이라 생각하여 평균값으로 데이터를 종합해 총 39명의 선수들에 대한 데이터로 분석하였다. 데이터는 케이비레포트 (www.kbreport.com) 기록실에 게시되어있는 데이터를 이용하였다. 변수는 자주 사용되고 있는 경기력 지수들에 활용되고 있는 개인기록들을 참고하고 (Lee, 2014), 기록 수집 가능 여부를 고려하여 선정하였다 (Lee, 2014). 분석에 사용된 세이버메트릭스 변수는 다음과 같다.

P/IP (이닝당 투구수) 이닝당 던진 공의 개수를 말한다.

K/9 (9이닝당 삼진) 9이닝 동안의 탈삼진 개수를 뜻한다.

BB/9 (9이닝당 볼넷) 9이닝 동안의 허용한 볼넷의 개수를 뜻한다.

HR/9 (9이닝당 홈런) 9이닝 동안의 허용한 홈런의 개수를 뜻한다.

H/9 (9이닝당 안타) 9이닝 동안의 허용한 안타의 개수를 뜻한다.

K/BB (삼진-볼넷 비율) 삼진과 볼넷의 비율을 말한다.

GO/AO (땅볼-뜬공 비율) 땅볼과 뜬공의 비율을 말한다. GO는 땅볼 AO는 뜬공이다.

oAVG (피안타율) 상대한 모든 타자들과의 대결에서 안타를 허용한 비율을 의미한다.

oSLG (피장타율) 본래는 타자에게 적용하는 지표인 장타율을 투수를 기준으로 적용한 지표이다.

oOBP (피출루율) 타율, 장타율과 함께 타자에게 적용하는 지표지만 투수를 기준으로도 적용이 가능하다.

PFR (파워-기교 비율) 스트라이크와 볼넷의 합을 이닝수로 나눈 값이다. 이 값이 높을수록 투수가 던진 공이 페어그라운드 안으로 떨어져서 안타나 땅볼 혹은 뜬공이 되는 경우가 적음을 알 수 있다.

Ground% (땅볼 비율) Ground balls%. 땅볼의 비율을 말한다. 땅볼이 많은 유형의 투수들은 내야수비진에 대한 의존도가 높은 편이다.

Fly% (뜬공 비율) Fly balls%. 뜬 공의 비율을 말한다. 뜬 공의 비율이 높은 선수들은 땅볼 투수들에 비해 수비진에 대한 의존도가 낮은 편이다.

WHIP (이닝 당 출루 허용) 투수가 한 이닝당 얼마나 많은 주자를 출루시키는지를 나타내는 지표로 1.1이하의 WHIP 수치를 보이는 투수를 특급 투수로 간주할 수 있다.

BABIP (인플레이 타구 피안타율) 'Batting average on balls in play'의 약자로 인플레이로 이어진 타구에 대한 타율을 계산하는 용어로 타자와 투수 모두에 적용할 수 있다.

ERA (평균자책점) 투수의 9 이닝당 자책점으로 나타내며 비교적 계산하기 쉽기 때문에 예전부터 투수를 평가하는 지표로 사용되고 있다.

FIP (수비 무관 평균자책점) 투수가 전적으로 책임지는 지표만을 대상으로 고안한 스탯으로, ERA보다 연도별 변동성도 적고 따라서 예측력도 높은 편이다. cFIP는 FIP값이 ERA와 유사한 범위를 갖도록 고안된 상수값이다.

2.2. 분석 방법

본 연구에서는 한국프로야구 선발투수의 능력을 파악하는 지수를 개발 및 제안하기 위해 17개의 세이버메트릭스 통계량을 이용해서 산술평균방법과 가중평균방법, 주성분 분석방법을 적용하였다.

먼저 산술평균방법은 총 17개 변수를 표준화하고 낮을수록 좋은 값인 경우 -1을 곱해 산술평균을 구한 뒤, 각 투수들의 능력을 평가하였다. 여기에서는 모든 변수들이 동일한 가중치 ($1/n$)로 반영이 되었으므로, oAVG와 H/9와 같은 비슷한 능력을 측정하는 경우 이 부분의 값이 큰 투수가 높은 점수를 받을 것이다. 이러한 단점을 보완하기 위해서 두 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 투수의 능력을 살펴 볼 수 있다. 그러나 17개의 변수를 모두 사용하여 다중회귀 분석을 하는 경우 설명변수들 사이의 높은 상관관계에 의해 다중공선성 (multicollinearty) 문제를 야기 시킬 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.

$$y_i = \beta_0 + \beta_1 \mathrm{Prin}_1 + \beta_2 \mathrm{Prin}_2 + \cdots + \beta_p \mathrm{Prin}_p + \epsilon_i.$$

Prin_1, Prin_2, $\cdots$, Prin_p는 주성분 변수가 되고, β_0, β_1, $\cdots$, β_p는 회귀계수 추정치이며, ϵ_i는 평균벡터가 0, 공분산행렬이 $cov(\epsilon) = \sigma^2 I$인 확률오차벡터이다 (Bae 등, 2012). 식 (2.1)에서 추정된 y값을 선발투수능력지수로 두고 WAR과 비교한다.

3. 선발투수능력지수 제안

분석하기에 앞서 먼저 사용될 변수들의 기초통계량은 다음과 같다.

Table 3.1 Simple statistics for 17 variable

	N	Mean	Std	Min	Max
ERA	39	4.417	0.818	3.180	6.370
FIP	39	4.727	0.615	3.520	6.480
WHIP	39	1.407	0.145	1.110	1.710
P/IP	39	16.786	1.037	13.670	19.050
H/9	39	9.689	1.060	7.390	11.910
K/9	39	6.675	1.353	3.720	10.610
BB/9	39	2.987	0.941	1.120	5.660
HR/9	39	0.920	0.297	0.300	1.870
BABIP	39	0.319	0.021	0.280	0.370
K/BB	39	2.485	0.903	0.810	5.020
oAVG	39	0.278	0.023	0.220	0.330
oOBP	39	0.340	0.024	0.290	0.400
oSLG	39	0.415	0.046	0.350	0.520
PFR	39	1.075	0.199	0.680	1.510
GO/AO	39	1.229	0.359	0.650	2.070
Ground%	39	0.286	0.040	0.210	0.380
Fly%	39	0.245	0.043	0.170	0.350

Table 3.1에서 평균값을 보면 P/IP, K/9, BB/9 등 변수들의 값 단위가 차이 나는 것을 확인 할 수 있었다. 그리고 P/IP, BB/9, HR/9 등의 값들은 의미하는 것이 이닝 당 투구수, 9이닝당 볼넷 수, 9이닝당 홈런 수 인데 이는 작을수록 뛰어난 투수임을 나타내는 변수들이다. 따라서 우리는 선발투수능력지수를 제안하기 위해, 3.1절에서는 변수들의 단위가 차이나기 때문에 표준화를 시키고 작을수록 뛰어난 능력을 나타내는 변수 (ERA, FIP, WHIP, BB/9, HR/9, BABIP, oAVG, oOBP, oSLG, H/9, Ground%, Fly%)들은 -1을 곱하여 산술평균, 가중평균, 주성분회귀분석을 실시한다. 3.2절에서는 분석된 결과를 가지고 WAR과 비교하여 가장 근접한 방법을 찾는다. 3.3절에서는 WAR과 가장 근접한 지수를 선발투수능력지수로 제안한다.

3.1. 데이터 분석 및 결과

3.1.1. 산술평균결과

세이버메트릭스 변수들 간에 값의 차이가 크기 때문에 표준화를 시킨 후, 작을수록 뛰어난 능력을 나타내는 변수 (ERA, FIP, WHIP, BB/9, HR/9, BABIP, oAVG, oOBP, oSLG, H/9, Ground%, Fly%)들의 경우 -1을 곱해 분석 하였다. 따라서 산술평균에 의해 얻어진 선발투수능력지수 (AVG_{P1})는 다음과 같다.

$$\mathrm{AVG}_{P1} = \frac{(-Z_1(\mathrm{ERA}) - Z_2(\mathrm{FIP}) - Z_3(\mathrm{WHIP}) + \cdots - Z_{17}(\mathrm{Fly\%}))}{17}, \tag{3.1}$$
$$Z_i = \frac{(X_i - \mu_i)}{\sigma_i}, \quad i = 1, \cdots, 17.$$

식 (3.1)을 계산하여 상위 10명의 순위를 나타낸 결과와 WAR과의 비교는 3.4절에서 다루도록 한다.

3.1.2. 상관계수를 활용한 가중평균결과

산술평균을 사용하는 경우에는 모든 변수들이 같은 가중치를 가지기 때문에 비슷한 성향의 변수인 oSLG, HR/9나, oAVG, H/9 등의 변수 값이 높은 경우 높은 점수를 받을 것이다. 따라서 이러한 문제점을 보완하기 위해 상관계수를 활용한 가중평균을 이용하였다. 가중치를 부여할 때 주관적인 방법보다 객관적인 방법인 상관분석 결과를 토대로 가중치를 설정하였다. Table 3.2는 세이버메트릭스 변수들의 상관분석결과이다.

Table 3.2 Correlation coefficient matrix of sabermetrics statistics

	ERA	FIP	WHIP	K/9	BB/9	HR/9	K/BB	PFR	BABIP	oAVG	oOBP	oSLG	P/IP	GO/AO	H/9	Ground%	Fly%
ERA	1	.726**	.832**	.385*	.753**	-.324*	0.306	.565**	.529**	-.410**	.743**	.803**	.754**	-0.084	0.067	0.087	-0.033
FIP		1	.647**	0.274	.492**	-.531**	.345*	.762**	0.01	-.587**	.517**	.661**	.714**	-0.221	-0.083	0.088	0.252
WHIP			1	.474**	.704**	-0.308	.596**	0.294	.647**	-.658**	.740**	.983**	.564**	0.081	.321*	0.288	-0.296
P/IP				1	0.127	0.049	.513**	0.026	0.224	-.347*	0.152	.469**	0.084	0.307	0.168	0.081	-0.193
H/9					1	-.546**	-0.15	.414**	.743**	-0.116	.991**	.665**	.831**	-.492**	-0.024	0.045	0.082
K/9						1	0.189	-0.149	0.035	0.294	-.580**	-.342*	-.423**	.856**	-0.01	-.357*	-.324*
BB/9							1	-0.058	0.061	-.783**	-0.09	.617**	-0.153	.669**	.471**	.348*	-.502**
HR/9								1	0.029	-0.085	.387*	0.264	.813**	-0.144	-.429**	-.366*	.443**
BABIP									1	0.001	.743**	.603**	.457**	0.06	0.201	0.019	-.364*
K/BB										1	-0.179	-.667**	-0.104	-0.189	-.361*	-.446**	0.26
oAVG											1	.714**	.813**	-.486**	0.022	0.099	0.046
oOBP												1	.526**	0.066	.406*	.388*	-.354*
oSLG													1	-.401*	-0.293	-0.203	.333*
PFR														1	0.241	-0.086	-.510**
GO/AO															1	.910**	-.910**
Ground%																1	-.710**
Fly%																	1

Table 3.2를 보면 변수들 간에 상관관계가 있는 것이 있고 없는 것도 있다. 이를 바탕으로 유사한 관계에 있는 변수들을 같은 그룹으로 분류하였다. 따라서 17개의 변수를 표준화 시키고 작을수록 뛰어난 능력을 나타내는 변수 (ERA, FIP, WHIP, BB/9, HR/9, BABIP, oAVG, oOBP, oSLG, H/9, Ground%, Fly%)들의 경우 -1을 곱해 6개의 그룹으로 나누었다. 첫 번째 그룹은 실점에 관련된 변수 (ERA, WHIP, BABIP, oAVG, oOBP, H/9)로 묶었고, 두 번째 그룹은 장타 (FIP, HR/9, oSLG), 세 번째 그룹은 제구 (K/BB, BB/9), 네 번째 그룹은 땅볼유도 (Fly%, Ground%, GO/AO), 다섯 번째 그룹은 스트라이크 능력 (K/9, PFR), 여섯 번째 그룹은 이닝당 투구수를 나타내는 P/IP이다. 따라서 가중평균에 의해 얻어진 선발투수능력지수 ($wAVG_{P2}$)는 다음과 같다.

$$\begin{aligned}\mathrm{wAVG}_{P2} =&[(-Z_1(\mathrm{ERA}) - Z_3(\mathrm{WHIP}) - Z_9(\mathrm{BABIP}) - Z_{10}(\mathrm{oAVG})\\ &- Z_{11}(\mathrm{oOBP}) - Z_{15}(\mathrm{H/9}))/6 + (-Z_2(\mathrm{FIP}) - Z_6(\mathrm{HR/9})\\ &- Z_{12}(\mathrm{oSLG}))/3 + (Z_7(\mathrm{K/BB}) - Z_5(\mathrm{BB/9}))/2\\ &+ (-Z_{17}(\mathrm{Fly\%}) - Z_{16}(\mathrm{Ground\%}) + Z_{14}(\mathrm{GO/AO}))/3\\ &+ (Z_4(\mathrm{K/9}) + Z_8(\mathrm{PFR}))/2 + Z_{13}(\mathrm{P/IP})]/6.\end{aligned} \tag{3.2}$$

식 (3.2)를 계산하여 상위 10명의 순위를 나타낸 결과와 WAR과의 비교는 3.4절에서 다루도록 한다. 가중평균을 사용 시 발생할 수 있는 문제점은 데이터의 개수에 비해 변수의 개수가 많아 변수들 간의 다중공선성이 발생할 수 있다. 따라서 변수를 정량적으로 축약하는 주성분 분석을 다음 절에서 활용하였다.

3.1.3. 주성분 회귀분석결과

원자료의 모든 변수 17개를 이용해서 상관계수의 크기에 따라 분류하는 것은 쉽지 않다. 이러한 문제를 해결하기 위해 주성분 분석을 통해 변수를 축약하였다. 주성분법 (요인분석, varimax)을 통해 나온 고유치와 누적 설명력을 활용하여 변수를 축약할 수 있는데, 일반적으로 고유치가 1이상이고 누적 설명력이 80%이상인 주성분을 선택하는 것이 기본이다. 이를 바탕으로 본 연구에서는 고유치가 1이상의 값을 가지는 총 4개의 변수로 축약하였다. 제1주성분의 고유치는 6.809이고 제2주성분은 4.404, 제3주성분은 2.228, 제4주성분은 1.871의 값을 가졌다. 축약된 4개의 변수가 가지는 누적 설명력은 90.074%로 나타났다. Table 3.3은 주성분 변수에 의해 얻어진 회전된 고유벡터를 나타낸 표이다. 여기서 회전을 시킨 이유는 하나의 원 변수에 부하값이 큰 요인이 2개 이상 존재하여 VARIMAX 방법을 이용하여 요인 회전을 하였다.

Table 3.3 Simple statistics for 17 variable

	Prin1	Prin2	Prin3	Prin4
H/9	**.927**	.061	.094	.343
BABIP	**.921**	-.094	-.206	-.269
oAVG	**.917**	.112	.047	.362
oSLG	**.736**	.206	.495	.298
ERA	**.734**	.531	.165	.094
WHIP	**.718**	.658	-.143	.006
K/BB	.019	**.879**	-.222	.131
BB/9	-.049	**.846**	-.304	-.381
FIP	.307	**.755**	.373	.396
oOBP	.676	**.680**	-.217	.052
P/IP	-.215	**-.524**	.049	.313
GO/AO	-.074	-.264	**.915**	.024
Fly%	-.140	-.170	**.856**	.372
Ground%	-.001	.307	**-.851**	.350
HR/9	.321	.341	**.728**	.140
K/9	.228	.173	-.022	**.932**
PFR	.197	-.314	.144	**.906**

각각의 주성분 내에서 고유벡터 값이 큰 변수들 끼리 묶은 후 이를 이용하여 주성분에 이름을 부여할 수 있다. Table 3.3에서 제1주성분 (prin1)의 계수 크기를 보면 H/9, BABIP, oAVG, oSLG, ERA, WHIP까지 고유벡터 값이 크므로 제1주성분은 안타억제능력 (hit suppression ability; HSA)이라 할 수 있다. 제2주성분 (prin2)은 K/BB, BB/9, FIP, oOBP, P/IP의 고유벡터 값이 크므로 제구능력 (control ability; CA)이라 할 수 있다. 제3주성분 (prin3)은 GO/AO, Fly%, Ground%, HR/9의 값이 크므로 장타억제능력 (long hit suppression ability; LSA)이라 할 수 있다. 제4주성분 (prin4)은 K/9, PFR의 값이 크므로 투구의 구질 (quality of pitching; QP)이라 할 수 있다. 이렇게 각 주성분의 이름을 정한 뒤, 다음과 같이 주성분 점수를 구할 수 있다.

$$\text{HSA (prin1)} = 0.927Z_1 + 0.921Z_2 + 0.917Z_3 + \cdots + 0.197Z_{17} \tag{3.3}$$

$$\text{CA (prin2)} = 0.061Z_1 - 0.094Z_2 + 0.112Z_3 + \cdots - 0.314Z_{17} \tag{3.4}$$

$$\text{LSA (prin3)} = 0.094Z_1 - 0.206Z_2 + 0.047Z_3 + \cdots + 0.144Z_{17} \tag{3.5}$$

$$\text{QP (prin4)} = 0.343Z_1 - 0.269Z_2 + 0.362Z_3 + \cdots + 0.906Z_{17} \tag{3.6}$$

위 식에서 Z는 각 변수를 표준화한 값이며, 총 17개의 세이버메트릭스 변수를 표준화하여 주성분 점수를 구하였다. 축약된 변수를 이용하여 회귀분석을 진행한 결과 얻어진 선발투수의 회귀모형

($PRIN_{P3}$)은 다음과 같다.

$$\text{PRIN}_{P3} = 3.429 + 0.560 \times \text{HSA} + 1.089 \times \text{CA} + 0.239 \times \text{LSA} + 0.705 \times \text{QP}. \tag{3.7}$$

식 (3.7)에서 계수들의 유의성을 확인해 본 결과 HSA와 CA, QP의 p값이 0.000으로 매우 유의하게 나왔고, LSA의 p값은 0.011로 모든 회귀계수값이 유의하게 나타났다. 따라서 우리는 선발투수의 능력을 나타내는 지수 PRIN_{P3}와 앞서 구한 AVG_{P1}과 wAVG_{P2}를 WAR과 다음절에서 비교하였다.

3.2. WAR과 분석방법에 따른 결과 비교

3.1절에서 산술평균, 가중평균 그리고 주성분회귀분석을 통해 선발투수의 능력을 평가 할 수 있는 지수를 만들었다. 이 결과들을 바탕으로 총 39명의 투수들로부터 WAR과 세가지 방법에 따른 상위 10명의 점수 값과 순위를 비교한 결과는 다음과 같다.

Table 3.4 Top 10 rank and scores result in three ways and WAR

Name	Team	WAR index (rank)	AVG_{P1} scores (rank)	wAVG_{P2} scores (rank)	PRIN_{P3} scores (rank)
Andy Van Hekken	Nexen	6.880 (1)	0.736 (3)	0.864 (2)	14.777 (3)
Hector Noesi	KIA	6.610 (2)	0.485 (7)	0.226 (12)	13.530 (5)
Rick VandenHurk	Samsung	6.380 (3)	1.309 (1)	1.384 (1)	23.873 (1)
Henry Sosa	LG	5.975 (4)	0.077 (20)	0.087 (15)	7.812 (12)
Michael Matthew Bowden	Doosan	5.400 (5)	0.769 (2)	0.658 (4)	15.556 (2)
Merrill Kelly	SK	4.915 (6)	0.339 (10)	0.151 (14)	8.872 (10)
Dustin Nippert	Doosan	4.815 (7)	0.493 (6)	0.331 (8)	13.257 (6)
Eric Lynn Hacker	NC	4.540 (8)	0.638 (4)	0.264 (11)	14.383 (4)
Yang, Hyeon-jong	KIA	4.413 (9)	0.451 (8)	0.287 (9)	10.919 (7)
Kim, Gwang-hyeon	SK	4.305 (10)	0.219 (12)	0.026 (23)	8.614 (11)

Table 3.4를 보면 WAR과 다소 차이를 보인다. 먼저 산술평균의 상위 10명과 WAR의 상위 10명을 비교했을 때 8명이 동일한 상위권으로 나타났다. 가중평균의 상위 10명과 WAR의 상위 10명을 비교했을 때는 5명의 선수가 상위권으로 나타났다. 마지막 주성분회귀분석의 경우 상위 10명이 WAR의 상위 10명과 비교했을 때 8명의 선수가 상위권으로 나왔으며 1등이 산술평균과 가중평균 결과와 동일하게 벤덴헐크 (Rick VandenHurk)으로 나타났다. 정확한 비교를 위해 상관분석을 실시한 결과 WAR과 산술평균 사이에는 상관계수 값이 0.838 (p = 0.000)로 나왔고 가중평균과의 비교에서는 상관계수 값이 0.786 (p = 0.000)이 나왔다. 마지막 PRIN_{P3}와의 비교에서는 0.941 (p = 0.000)으로 가장 높게 나온 것을 확인 할 수 있었다. 따라서 WAR과 세가지방법을 비교한 결과 가장 큰 상관계수를 가지는 PRIN_{P3}가 최종 선발투수능력지수로 적합하다고 판단하였다.

3.3. 선발투수능력지수 (starting pitcher ability index; SPAI) 제안

본 논문은 선발투수의 능력을 평가하는데 있어 WAR과 가장 근접한 방법을 찾기 위해 산술평균방법, 가중평균방법, 주성분 회귀분석 방법을 이용하였다. 각 방법으로부터 얻은 값과 WAR을 비교한 결과 주성분 회귀모형 ($PRIN_{P3}$)의 상관계수가 0.941 (p = 0.000)로 가장 근접하고 효율적인 분석방법으로 나타났다. 따라서 주성분분석을 이용하여 총 17개의 변수를 4개의 주성분 변수 (HSA, CA, LSA, QP)로 축약하고 이를 통해 최종 선발투수능력지수를 제안한다. 식 (3.7)을 최종 선발투수능력지수로

선택하여 SPAI (starting pitcher ability index)로 명명하였다.

$$\mathrm{SPAI}(\mathrm{PRIN}_{P3}) = 3.429 + 0.560 \times \mathrm{HSA} + 1.089 \times \mathrm{CA} + 0.239 \times \mathrm{LSA} + 0.705 \times \mathrm{QP}. \tag{3.8}$$

식 (3.8)과 WAR의 R^2값은 0.885로 뛰어난 설명력을 가지고 있으며, 투수의 WAR과 가장 근접한 모형이기에 선발투수의 능력을 파악하는데 있어 부족함이 없다고 판단하였다.

4. 결론 및 토의

본 논문은 선발투수의 능력을 파악하기 위한 선발투수능력지수를 제안하였다. WAR은 MLB 뿐만 아니라 KBO에서 투수의 능력을 파악하는데 가장 공신력 있는 통계량이다. 하지만 WAR은 구하기 어려운 세이버메트릭스 통계량이 포함되어 있어 기본적으로 제공되는 기록만으로는 산출하기가 어렵다는 문제점을 가진다. 따라서 우리는 케이비레포트 기록실에 게시되어있는 2014년부터 2016년까지 규정이닝을 만족한 선발투수 데이터를 사용하여 WAR에 가장 근접한 선발투수능력지수를 제안하고자 하였다. 선발투수의 능력을 살펴볼 수 있는 기본적인 17개의 세이버메트릭스 변수들을 사용하여 산술평균방법과 가중평균방법 그리고 주성분 분석에 의한 회귀분석을 진행하였다. 먼저 기초통계량을 확인하여 변수의 단위가 다름을 인지하고 표준화를 통해 분석을 진행하였고, 변수의 값이 작을수록 좋은 값인 경우 표준화시킨 값에 -1을 곱해 클수록 좋은 값으로 바꾼 뒤, 산술평균을 구하고 두 번째로 17개의 변수들의 상관관계를 통해 6개의 그룹으로 나눈 뒤 가중평균을 계산하였다. 그러나 몇몇 변수들 간의 상관관계가 크고 데이터에 비해 변수가 많아 다중공선성의 문제가 발생하기 때문에 세 번째로 주성분 분석을 실시하였다. 주성분 분석을 통해 17개의 변수를 4개의 주성분 변수 (HSA, CA, LSA, QP)로 축약하고 주성분 점수를 계산하여 회귀모형을 구하였다. 마지막으로 각 분석 방법을 통해 얻어진 값과 WAR을 비교하여 상관계수가 가장 높은 주성분분석방법 ($r = 0.941, p = 0.000$)을 채택하여 최종 선발투수능력지수 (SPAI)로 제안했다. SPAI지수에 따른 선수들의 결과는 Appendix A의 Table A.1에 나타내었다.

투수의 세이버메트릭스 통계량들은 온전히 자기 자신만의 수치라고 보기 힘들다. 수비의 도움도 필요하고 승리를 따내기 위해서는 타자의 도움도 필요하다. 타고투저 현상은 전체적으로 타격이 우세하고 투수가 열세라는 뜻으로 여러 가지 이유가 있지만 투수의 기록 중 ERA에서만 보더라도 평균값이 4.417에 달한다. 이는 투고타저 시즌인 2005년 ERA (3.744)에 비해 상당히 높게 나타난 것을 알 수 있다. 결국 투수가 일정하게 던지더라도 상대편의 타자가 얼마나 잘 치느냐에 따라 값이 크게 달라질 수도 있다는 것을 의미한다. 따라서 본 논문에서 제안한 SPAI 또한 완벽하게 선발투수의 능력을 파악하기는 힘들다. 하지만 SPAI는 투수의 안타억제능력 (HSA), 제구능력 (CA), 장타억제능력 (LSA), 투구의 구질 (QP) 4가지를 반영하여 투수의 능력을 파악하는 지수로써 투수가 가져야할 역량을 보다 쉽게 계산할 수 있다. 이를 통해 투수의 능력을 쉽게 파악하여 경기 전략을 짜고, 연봉 협상시 하나의 지표로 적절한 연봉을 매기는데 도움이 될 것이다.

부록 A: Starting pitcher ability index for KBO in 2014-2016

Table A.1 A player list of the SPAI

Rank	Name	Team	SPAI
1	Rick Vandenhurk	Samsung	23.873
2	Michael Matthew Bowden	Doosan	15.556
3	Andy Van Hekken	Nexen	14.777
4	Eric Lynn Hacker	NC	14.383
5	Hector Noesi	KIA	13.530
6	Dustin Nippert	Doosan	13.257
7	Yang, Hyeon-jong	KIA	10.919
8	U, Gyu-min	LG	10.907
9	Alfredo Figaro	Samsung	10.782
10	Merrill Kelly	SK	8.872
11	Kim, Gwang-hyeon	SK	8.614
12	Henry Sosa	LG	7.812
13	Sin, Jae-yeoung	Nexen	7.549
14	Ryu, Je-Guk	LG	6.186
15	Yun, Seong-hwan	Samsung	5.861
16	Zach Stewart	NC	5.062
17	Cory Riordan	LG	4.955
18	Lee, Jae-Hak	NC	4.782
19	Brooks Raley	Lotte	4.371
20	Josh Lindblom	Lotte	3.811
21	Chang, Won-jun	Doosan	3.352
22	Cha, U-chan	Samsung	3.168
23	John Dale Martin	Samsung	3.099
24	Chris Oxspring	kt	2.797
25	Yu, Hee-gwan	Doosan	1.388
26	Mitch Talbot	Hanwha	1.285
27	Tyler Cloyd	Samsung	0.135
28	Lucas Harrell	LG	-0.371
29	Charlie Shirek	NC	-0.635
30	Ryan Robert Feierabend	KT	-2.196
31	Bae, Yeong-su	Samsung	-3.410
32	Chang, Won-sam	Samsung	-3.638
33	Josh Stinson	KIA	-4.267
34	Zeke Spruill	KIA	-5.731
35	Lee, Tae-yang	Hanwha	-6.785
36	Andrew Albers	Hanwha	-7.124
37	Che, Byeong-yong	SK	-12.716
38	Shane Youman	Lotte	-13.914
39	Im, Jun-seop	KIA	-16.565

References

Bae, J. Y., Lee, J. M. and Lee, J. Y. (2012). Predicting Korea pro-baseball rankings by principal component regresstion analysis. *The Journal of Korean Statistical Society*, **19**, 367-379.

Chang, J. and Zenilman, J. (2013). *A study of sabermetrics in major league baseball: The impact of moneyball on free agent salaries*, Washington University, Saint Louis.

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.

KBO (2017). http://osen.mt.co.kr/article/G1110627260

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**, 1065-1174.

KBreport (2013-2016), http://www.kbreport.com

Kwon, S. H. (2008). *Utilizing and analysis of multivariate data*, Freeacademy, Seoul.

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.

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, S. I. (2014). *Development of pitcher's performance index in the Korean professional baseball games*, Master's Thesis, Myoungji University, Seoul.

MLB (2017). http://m.mlb.com/news/article/182980276/best-late-round-picks-in-draft-history/

Oh, G. J., An, J. J. and Sim, G. S. (2012). Multicurrencies 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.

Journal of the Korean Data & Information Science Society
2017, **28**(4), 863–874

http://dx.doi.org/10.7465/jkdi.2017.28.4.863
한국데이터정보과학회지

Suggestion of starting pitcher ability index in Korea baseball - Focusing on the sabermetrics statistics WAR

Hyeon-Gyu Kim[1] · Jea-Young Lee[2]

[12]Department of Statistics, Yeungnam University

Received 16 May 2017, revised 5 July 2017, accepted 5 July 2017

Abstract

Wins above replacement (WAR) is the most commonly used statistics of the many sabermetrics that measure baseball players' abilities. The advantage of a WAR is that it enables to compare performances of players even though they have different roles such as pitcher and hitter. However, WAR is difficult to obtain with common records. Thus, in this paper, we have calculated the sabermetrics variable based on Korean professional baseball records for the past three years (2014-2016). Using these variables, we suggest starting pitcher ability index that can replace WAR. Starting pitcher ability index was calculated by means of arithmetic mean, weighted average and principal component regression. Then, compared to the WAR, the most relevant method was selected, which would be useful to identify for the starting pitcher ability.

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

[1] Graduate student, Department of Statistics, Yeungnam University, Kyungsan 38541, Korea.
[2] Professor, Department of Statistics, Yeungnam University, Kyungsan 38541, Korea.
E-mail: jlee@yu.ac.kr