Journal of the Korean Data & Information Science Society
2014, **25**(2), 349–356

http://dx.doi.org/10.7465/jkdi.2014.25.2.349
한국데이터정보과학회지

한국프로야구에서 타자능력의 측정[†]

이장택[1]

[1]단국대학교 응용통계학과

접수 2014년 1월 18일, 수정 2014년 2월 18일, 게재확정 2014년 2월 24일

요 약

타자들의 평가에 대한 불완전한 부분을 보완하기 위하여 세이버메트리션들이 세워놓은 기준들이 선수 평가에 중요한 잣대가 되고 있다. 하지만 평가지표들은 개수가 많고 형태가 일정하지 않아서 팬들을 혼동에 빠지게 한다. 본 연구에서는 대표적인 타자평가지표를 이용하여 지표들의 특성을 골고루 반영하는 주성분을 찾아보고 한국프로야구에 적합한 타자지표를 제안한다. 제안된 지표는 타자들의 능력을 그룹화하여 객관적으로 설명할 수 있기 때문에 선수들의 연봉을 합리적으로 결정할 수 있다.

주요용어: 세이버메트리션, 주성분분석, 케이-평균 군집, 타자평가지표.

1. 머리말

야구에서 누적된 수많은 기록들은 통계적으로 신뢰할 수 있는 영향력을 행사하기 시작한다. 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. 타자들은 야구경기에서 많은 결과들을 만드는 데, 대표적인 야구 통계량들은 단타 (1B), 2루타 (2B), 3루타 (3B), 타수 (AB), 볼넷 (BB), 도루실패 (CS), 참가경기 수 (G), 병살타 (GIDP), 안타 (H), 사구 (HBP), 홈런 (HR), 잔루 (LOB), 출루율 (OBP), 타점 (RBI), 도루 (SB), 희생번트 (SH), 희생플라이 (SF), 장타율 (SLG), 삼진 (SO), 루타 (TB), 고의사구 (IBB), 그리고 고의사구를 제외한 볼넷 (UBB) 등이 있다.

그러나 특정 개인이 야구의 모든 부분에서 뛰어날 수가 없기 때문에 이런 단편적인 양적 수치는 타자의 능력을 나타내는 데 도움은 되지만 각 타자들을 평가하는 통합지표의 역할을 할 수는 없다. 또한 안타의 개수가 많은 선수가 잘 하는 선수이지만 타석의 수와 안타의 질도 매우 중요한 의미를 지니기 때문에 타자의 비교 우월성을 간결하게 나타내지는 못한다. 그래서 타자들의 능력을 평가하는 데 사용되는 좀 더 진일보된 통계량들이 바로 타율 (BA), 홈런 (HR), 타점 (RBI) 등이다. 하지만 타율은 상대팀의 수비 능력과 같은 외부 요인들의 영향을 많이 받고, 타점은 팀 동료들이 해당 타자 앞에서 얼마나 출루를 해 주느냐에 크게 좌우된다. 이와 같은 이유로 오늘날 타자들의 평가에 대한 불완전한 부분을 보완하기 위하여 세이버메트리션들이 세워놓은 기준들이 선수 평가에 중요한 잣대가 되고 있다.

세이버메트리션들이 고안한 타자평가지표들은 출루율 (OBP), 장타율 (SLG), OPS (on base plus slugging), 루타수 (TB), ISO (isolated power), SECA (secondary average), 종합공격력 (TA), RC (runs created)와 RC/27 (runs created per game) 등이 있는데, 특정 지표를 사용하는 것에 따라 타자

[†] 이 연구는 2013학년도 단국대학교 대학연구비 지원으로 연구되었음
[1] (448-701) 경기도 용인시 죽전동 126번지, 단국대학교 응용통계학과, 교수. E-mail: jtlee@dankook.ac.kr

들의 평가가 서로 상이하게 될 수 있기 때문에 지표들에 대한 종합적인 판단을 할 필요가 있다. 따라서 변수 개수가 많을 때 변수의 성질에 따라 묶인 소수의 변수로 만들어 주는 주성분분석과 같은 분석이 필요할 수 있으며, 따라서 본 연구에서는 여러 가지 타자지표를 1개의 요인으로 요약하고 만들어진 지표를 이용하여 우리나라 대표타자들의 능력을 비교하고, 또한 몇 개의 군집으로 나누어 선수들의 수준을 평가하여 본다.

한편 한국프로야구 데이터는 학자들에 의해 많이 연구되었는데, 최근 연구들을 살펴보면 시계열모형을 이용하여 관중 수 예측을 다룬 Lee와 Bang (2010), 인공신경망을 이용하여 포스트시즌 진출 예측을 살펴본 Chea 등 (2010), 한국프로야구 타자들에 대한 세이버메트릭스 지수 값을 이용하여 선수들의 경기력과 연봉간의 패턴을 분석한 Seung과 Kang (2012), 출루율과 장타율이 득점에 미치는 연구를 한 Kim (2012) 등이 있다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 데이터의 구성, 타자평가지표 및 주성분분석에 대해 언급하며, 3절에서는 주성분분석을 이용하여 제안된 타자지표를 설명하고 K-평균 군집분석의 결과를 설명하였다. 끝으로 4절에서는 본 연구의 결론에 대해 언급하였다.

2. 연구방법

2.1. 데이터의 구성

본 연구에 사용된 데이터는 한국야구위원회 (KBO)에 기록되어 있는 2000년부터 2013년 사이에 있었던 경기 중 규정타석 수를 채운 594명의 경기결과이다. 통계패키지 SPSS 21K와 수집된 데이터를 이용하여 세이버메트리션 타자통계량을 계산하였다. 조사된 선수들의 수는 팀명이 우리, 히어로즈, 현대를 넥센에, 해태를 KIA에 포함시키고 신생팀 NC를 제외하면 삼성과 넥센이 각각 83명으로 가장 많았으며, KIA가 61명으로 가장 작았다.

2.2. 타자평가지표들의 정의

타자들의 평가를 위한 세이버메트리션들이 고안해 놓은 각종 지표들은 매우 많지만 그중에서도 보편화되어있고 많이 사용하는 다음과 같은 평가지표를 본 연구에서 사용하였으며 각 지표들에 대한 자세한 계산 방법은 Table 2.1에 정리되어 있다.

(1) OPS (on base percentage plus slugging percentage)

세이버메트릭스에서 가장 유명하고 보편화된 타자지표로 출루율 (OBP)과 장타율 (SLG)의 단순한 합으로 계산되는 지수이다. 타자의 출루능력과 장타력을 골고루 평가하며 누구나 쉽게 이해하고 계산할 수 있다는 장점도 있으나 타자의 주루능력, 병살타 및 희생타 생산능력 등이 직접 반영되지 않는 단점도 있다.

(2) ISO (isolated power)

장타율에는 단타가 포함되어 있어 타율이 높으면 장타율도 덩달아 높아지는 결과가 발생하기 때문에 이를 보정한 순수한 장타력만을 측정하는 지표이다.

(3) SECA (secondary average)

장타율과 출루율의 문제점을 극복하기 위해 장타율의 가중치에 볼넷과 도루의 가치를 고려해서 만든 수정타율이다.

(4) TA (total average)

종합공격력으로 지칭되는데 타자가 한 시즌 동안 한 번의 공격 기회에서 어느 정도 진루했는지를 나타낸다. TA와 SECA는 비슷하지만 TA는 출루와 도루에 비중을 두고 있으며, SECA가 좀 더 장타에 비

중을 두고 있는 차이가 있다.

(5) RC (runs created)

타자의 득점 공헌도에 관한 지수로 기본 생각은 출루율과 루타수의 곱으로 계산되는데, 전체 시즌을 통해 타자가 실제로 득점에 공헌한 바를 나타낸다.

(6) RC/27 (runs created per game)

한 경기를 통해 타자가 득점에 공헌한 바를 계산하려는 지표로 선수 A의 RC/27은 한 경기에 타순이 A로만 구성되는 경우에 한 경기에 몇 득점을 만들어 낼 수 있는 지를 설명하는 수치이다.

(7) wOBA (weighted on base average)

OPS의 단점들을 극복하고 좀 더 정확히 타자의 능력을 계량하기 위해 고안된 통계량으로 타석 당 득점 기대치를 의미하며, 볼넷, 사구, 단타, 2루타, 3루타, 홈런, 에러로 인한 출루 등에 가중치가 부여되고 이들의 합을 타석의 수로 나눈 값이다.

(8) XR (extrapolated runs)

타자의 진루 및 주루와 관련된 사항을 선형가중식으로 구한 득점공헌도로써 RC보다 더 우수하다고 알려져 있다. RC가 비율통계량과 누적통계량을 혼합했다면 XR은 순수한 누적통계량에 가깝다.

일반적으로 소수점을 포함한 계수가 등장하는 지표들은 모두 과거의 기록에 기초한 것이므로 매년 업데이트되어 변할 수 있으며 따라서 인터넷에 제시되는 공식들은 조금씩 차이가 있다. 또한 이 수치들은 메이저리그 야구를 기반으로 하기 때문에 한국프로야구를 이용하여 재추정하면 좀 더 다르게 나타날 수 있다.

Table 2.1 Formulae for sabermetric batting statistics

batting statistics	formulae
TB	$TB = 1B + 2(2B) + 3(3B) + 4(HR)$
OBP	$OBP = (H + BB + HBP)/(AB + BB + HBP + SF)$
SLG	$SLG = TB/AB$
OPS	$OPS = OBP + SLG$
ISO	$ISO = (TB - H)/AB$
SECA	$SECA = (TB - H + BB + SB - CS)/AB$
TA	$TA = (TB + BB + HBP + SB - CS)/(AB - H + CS + GIDP)$
RC	$RC = A \times B/C, A = H + BB + HBP - CS - GIDP,$ $B = TB + 0.26(BB - IBB + HBP) + 0.52(SH + SF + SB),$ $C = AB + BB + HBP + SH + SF$
RC/27	$RC/27 = RC/D, D = (AB - H + SH + SF + CS + GIDP)/27$
wOBA	$wOBA = E/F, E = 0.691(UBB) + 0.722(HBP) + 0.884(1B) + 1.257(2B) +$ $1.593(3B) + 2.058(HR), F = AB + BB - IBB + SF + HBP$
XR	$XR = 0.50(1B) + 0.72(2B) + 1.04(3B) + 1.44(HR) - 0.32(CS) + 0.34(HBP +$ $BB - IBB) + 0.25(IBB) - 0.09(AB - H - SO) + 0.18(SB) - 0.098(SO) -$ $0.37(GIDP) + 0.37(SF) + 0.04(SH)$

2.3. 주성분 분석

다변량 기법에서 가장 오래된 역사를 가지고 있으며 폭넓게 사용되는 주성분 분석 (principle component analysis)은 차원축소를 통하여 저차원상에서 변수의 관계를 규명하는 데이터 분석기법이다. 전체의 변동을 변수들의 선형결합들로 이루어진 주성분이라 하는 새로운 변수에 의해서 나타내고, 가급적 적은 주성분으로 설명하려는 점이 이 분석의 핵심이다. 주성분 분석에서는 주성분을 찾아 원래의 변수를 정보의 손실 없이 대신하려고 하며 용도는 다변량 자료의 탐색적 조사, 차원축소를 통한 자료의 단순화, 중회귀분석, 군집분석에서 사용된다. 주성분분석은 변수들의 상관행렬의 고유값과 고유벡터를 이용한

스펙트랄 분해를 이용하여 주성분을 생성하는 방법으로 요인분석에서 요인이라는 새로운 변수를 생성하는 방법으로 사용된다.

3. 데이터분석

3.1. 변수 개수의 축약

본 연구에서는 상관관계가 큰 변수들을 이용하여 변수 개수의 축약과 측정변수들의 분산을 최대로 설명하는 것이 목적이므로 요인분석 중에서 주성분분석을 선택하였다. 사용된 데이터가 요인분석의 기본 가정을 만족하는 것인지를 확인하기 위하여 KMO와 Bartlett 검정을 사용하였는데, 표본적합도 KMO 값은 0.78로 권장치인 0.5를 훨씬 상회하였으며, Bartlett의 구형성 검정은 유의확률이 0.000으로 나타나 요인분석에 적합함을 확인할 수 있었다. 그리고 8개의 세이버메트릭스 지수 (OPS, ISO, SECA, TA, RC, RC/27, wOBA, XR)에 대해 주성분분석을 실시하였고 설명된 총분산은 Table 3.1과 같다.

Table 3.1 Total variance explained

component	initial eigenvalues		
	Total	% of variance	cumulative %
1	7.239	90.490	90.490
2	0.405	5.062	95.551
3	0.180	2.245	97.797
4	0.145	1.812	99.609
5	0.024	0.300	99.909
6	0.004	0.056	99.965
7	0.001	0.019	99.983
8	0.001	0.017	100.000

Table 3.1의 고유값들 중에 한 개만이 1보다 크며, 이에 대응하는 주성분이 전체 변동의 90.49%를 설명함으로써 8개의 변수에 대한 정보는 하나의 주성분이면 충분하다고 간주되며 이는 스크리 플롯 (scree plot)을 이용하여도 마찬가지 결론이다.

Table 3.2 Factor loading for component 1

OPS	ISO	SECA	TA	RC	RC/27	WOBA	XR
0.984	0.865	0.914	0.986	0.956	0.972	0.979	0.948

Table 3.2의 요인적재값을 보면 주성분 1에 대하여 ISO를 제외한 7개 변수의 공통성은 거의 비슷하게 나타나며 수치가 해당 변수와 주성분 사이의 상관계수를 의미하므로 모두 양의 상관관계가 있으며 따라서 주성분 1은 타자의 능력을 나타내는 성분으로 이해할 수 있다.

3.2. 타자등급지표 (batting grade index)

본 연구에서 사용하는 주성분은 1개이므로 요인의 회전은 필요 없으나 요인분석의 결과를 이용하기 위하여 요인점수를 산출할 필요가 있다. 요인점수는 각각의 표본이 갖고 있는 각 요인에 대한 복합적인 측정치이다. SPSS를 이용하여 요인을 생성한 후에 각 케이스의 요인점수를 산출하였는데, 요인점수의 추정방법은 디폴트 방법인 회귀분석에 의한 추정을 사용하였다. Table 3.3은 각 표본의 변수 값들을 구해진 요인들의 값으로 바꾸어 준 요인점수의 계수로서 Z_1 부터 Z_8을 각각 OPS, ISO, SECA, TA, RC, RC/27, wOBA, XR의 표준화된 변수라고 할 때 어떤 선수의 요인점수 S는 다음과 같이 계산된다.

$$S = 0.136Z_1 + 0.120Z_2 + 0.126Z_3 + 0.136Z_4 + 0.132Z_5 + 0.134Z_6 + 0.135Z_7 + 0.131Z_8$$

그리고 Figure 3.1은 계산된 각 선수들의 요인점수를 오름차순으로 정렬하여 그림으로 나타낸 결과이다.

Table 3.3 Component score coefficient

OPS	ISO	SECA	TA	RC	RC/27	WOBA	XR
0.136	0.120	0.126	0.136	0.132	0.134	0.135	0.131

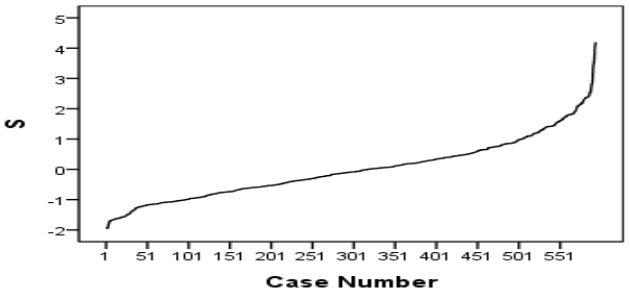

Figure 3.1 Factor score diagram

Figure 3.1을 이용하면 규정타석을 채운 한국프로야구 타자들의 능력을 몇 개의 그룹으로 나눌 수 있는 데, 대략 3개에서 5개의 그룹으로 나눌 수 있겠다. 우수한 요인점수 순서로 그룹을 G_1, G_2, G_3, G_4, G_5로 나타낼 때 다음 Table 3.4는 각 그룹에 속한 선수들의 빈도수와 최종 군집중심값을 나타낸다. 그룹의 분할은 유클리드 거리를 이용한 K-평균 군집분석을 이용하였으며, 이 경우 K는 군집의 개수를 나타낸다.

Table 3.4 Number of players per group (final cluster centers in parenthesis)

K	G_1	G_2	G_3	G_4	G_5
3	83 (1.79)	255 (0.30)	256 (-0.88)		
4	15 (2.83)	85 (1.45)	250 (0.22)	244 (-0.91)	
5	4 (3.80)	37 (2.08)	97 (1.06)	228 (0.09)	228 (-0.95)

만일 군집의 개수를 3개로 하면 군집의 특색이 별로 없고 5개로 하면 최우수 군집에 속하는 선수들의 수가 너무 적어서 4개인 경우가 가장 적절하다고 간주된다. 4개로 나누는 경우에 군집의 이름을 매우우수 (G_1), 우수 (G_2), 보통 (G_3), 미흡 (G_4)으로 하고 유의수준 5%에서 일원배치분산분석과 던칸의 다중비교를 통하여 집단 간 차이의 유의성을 요인점수 S와 중요한 야구 통계량들인 안타, 단타, 2루타, 3루타, 홈런, 볼넷, 병살타, 사구, 삼진, 타점, 타율, 출루율, 장타율에 대하여 살펴보면, 요인점수 S는 4개의 그룹 간 차이가 모두 유의한 $G_4 < G_3 < G_2 < G_1$으로 나타났다. 요인점수 S와 똑같은 집단 간의 유의성이 나타나는 통계량은 홈런, 볼넷, 타점, 출루율, 장타율로 나타났으며 단타인 경우는 집단 간에 차이가 없는 것으로 나타났다. Table 3.5는 주요타격 통계량의 평균값을 4개의 군집별로 나타낸 결과이다. 2루타와 3루타를 제외하고 모든 통계량들의 순위가 요인점수와 같게 나타남을 확인할 수 있다. 2루타는 G_2그룹의 평균이 가장 높고 3루타는 G_4그룹의 평균이 가장 높지만 타 그룹과의 차이는 크지 않다. 지금부터 8개의 주요 세이버메트릭스 지수를 이용하여 만든 요인점수 S를 타자등급지표(batting grade index)라고 명명하고 BGI로 표기하기로 한다. Table 3.6은 BGI를 이용하여 구한 상위 10위에 속하는 선수들이다.

Table 3.5 Mean of batting statistics by groups

name	H	2B	HR	3B	BB	BA	RBI	OBP	SLG
G_1	143.27	22.53	38.27	1.00	92.93	0.322	110.00	0.445	0.635
G_2	141.56	26.19	24.54	1.32	66.82	0.317	89.08	0.412	0.546
G_3	124.17	22.30	15.26	1.80	51.71	0.290	66.90	0.373	0.459
G_4	110.55	17.88	6.20	2.02	39.64	0.269	45.17	0.340	0.369

Table 3.6 Player listing of the top 10 BGI scores

rank	name	year	team	BGI
01	Sim, jeongsu	2003	Hyundai	4.19685
02	Jose	2001	Lotte	4.03104
03	Lee, seungyeop	2003	Samsung	3.55553
04	Lee, seungyeop	2002	Samsung	3.40747
05	Lee, daeho	2010	Lotte	2.83538
06	Brumbaugh	2004	Hyundai	2.80970
07	Park, ByeongHo	2013	Nexen	2.54759
08	Sim, jaehak	2001	Doosan	2.54640
09	Sim, jeongsu	2002	Hyundai	2.43158
10	Lee, seungyeop	2001	Samsung	2.39747

BGI도 결국 투수의 수준에 따라 영향을 많이 받지만 2003년 현대의 심정수 선수의 기록이 4.196으로 가장 높게 나타났다. BGI 값이 4를 초과할 확률은 규정타석을 채운 선수 중에서 0.3% 정도로 나타났다. 또한 삼성의 이승엽 선수는 10위 내에 3번이나 속해 있는데, 2004년부터 2011년까지 일본에서 활약한 것을 고려하면 정말 탁월한 기록이다. 하지만 상위 수준 대부분의 BGI 기록은 2000년 초반에 작성되었으며 2010년 이후에는 롯데의 이대호와 넥센의 박병호 선수 2명뿐이어서 거포부재의 야구가 국내에서 진행되고 있음을 알 수 있다.

4. 결론

야구는 기록의 스포츠이며 기록은 야구를 더욱 풍성하고 흥미롭게 만든다. 그런데 선수들의 공격력에 관해서 오래전부터 타율은 선수를 평가하는 절대적 진리라고 생각되어왔지만 최근에는 타자의 생산력을 측정하는데 있어서 타율보다 OPS, RC, RC /27, wOBA, XR 등과 같은 새로운 평가기준이 더 합리적이라는 근거가 등장했으며 매스컴에서도 이와 같은 측도들을 언급한다. 하지만 이런 측도들이 너무 많아서 많이 활용하면 야구팬들은 야구가 어렵게 생각되기 때문에 흥미가 감소될 수가 있다. 본 연구에서는 지금까지 세이버메트리션들이 만든 유명한 공격지표들을 이용하여 통합된 공격지표 BGI를 만들었다. 제안된 지표는 주성분분석을 이용하여 작성되었으며, K-평균 군집분석을 활용하여 한국프로야구 선수들의 공격력을 4개의 군집으로 나누었다. 본 연구의 BGI는 2000년 이후의 모든 한국프로야구 데이터를 사용하여 작성하였기 때문에 적절한 이상치나 영향점을 제거하여 주성분분석을 사용하면 좀 더 안정적인 결과를 제공할 수 있을 것이다.

References

Chea, J. S., Cho, E. H. and Eom, H. J. (2010). Comparisons of the outcomes of statistical models applied to the prediction of post-season entry in korean professional baseball. *The Korean Journal of Measurement and Evaluation in Physical Education and Sport Science*, **12**, 33-48.

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.

Korea Baseball Organization (2001-2006). *2000-2005 official baseball guide*, Korea Baseball Organization, Seoul.

Korea Baseball Organization (2014). `http://www.koreabaseball.com/Record`.

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.

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
2014, **25**(2), 349–356

http://dx.doi.org/10.7465/jkdi.2014.25.2.349
한국데이터정보과학회지

Measurements for hitting ability in the Korean pro-baseball[†]

Jang Taek Lee[1]

[1]Department of Applied Statistics, Dankook University

Received 18 January 2014, revised 18 February 2014, accepted 24 February 2014

Abstract

In baseball, sabermetric batting statistics are used to compare an offensive performance of players. There exist dozens of sabermetric statistics, but baseball fans don't like the complexity of an abundance of measures. This paper provides a batting grade index (BGI) using principal component based on eight batting statistics. These are OPS, ISO, SECA, TA, RC, RC/27, wOBA and XR. We show that how standardized batting statistics are aggregated and weighted to arrive at a single composite measure of BGI. Also our result allows for segmentation of players into groups using the K-means clustering algorithm.

Keywords: Batting grade index, K-means clustering, principal components analysis, sabermetrics.

[†] The present research was conducted by the research fund of Dankook University in 2013.
[1] Professor, Department of Applied Statistics, Dankook University, Yongin 448-701, Korea. E-mail: jtlee@dankook.ac.kr