Journal of the Korean Data & Information Science Society
2016, **27**(3), 677–687

http://dx.doi.org/10.7465/jkdi.2016.27.3.677
한국데이터정보과학회지

한국프로야구에서 타자력 지수 제안

홍종선[1] · 김재영[2] · 신동식[3]

[123]성균관대학교 통계학과

접수 2016년 4월 15일, 수정 2016년 5월 13일, 게재확정 2016년 5월 17일

요 약

야구 타자의 능력을 측정하는 많은 세이버메트릭스 통계량들 중에서 WAR은 미국프로야구에서 가장 많이 사용하는 통계량이다. 그러나 한국프로야구 자료에는 수비에 관련된 변수에 포함된 야구장 요인, 포지션조정 그리고 리그 조정 통계량들이 존재하지 않으므로 WAR을 한국프로야구에 적용하는 데에는 문제가 있다. 본 연구에서는 타자의 능력을 측정하는 대안적인 세이버메트릭스 통계량을 제안하여 미국프로야구 뿐만 아니라 한국프로야구에서도 동시에 사용할 수 있도록 한다. 본 연구에서 제안한 타자력 지수 모형은 한국프로야구와 미국프로야구 타자들에 대한 다섯 종류의 통계량을 사용하여 개발한다. 우선 2015년도 최소 규정 타석을 만족한 미국프로야구 타자들의 자료를 바탕으로 타자력 지수 모형을 개발한다. 미국프로야구 타자들의 WAR과 비교하면서 본 연구에서 제안한 타자력 지수의 능력의 타당성을 검토한다. 다음으로 이 모형을 2015년도 한국프로야구 자료에 적용하여 한국형 타자력 지수를 제안한다. 한국프로야구 타자력 지수를 서로 다른 팀별, 나이별, 포지션별로 통계적으로 분석하고, 타자력 지수와 그들의 연봉과의 선형관계성을 토론한다. 연봉에 관한 회귀모형의 신뢰영역을 바탕으로 연봉책정의 적절함에 따라 46명의 타자를 세 그룹으로 할당하고, 세 그룹에 속한 연봉을 다양한 인자에 대하여 통계적으로 탐색한다.

주요용어: 세이버메트릭스, 주성분, 타자능력, 회귀모형.

1. 서론

야구에서 타자의 공격능력을 쉽게 계산하면서 평가할 수 있는 통계량을 개발하기 위한 연구는 세이버메트릭스 (sabermetrics)를 중심으로 계속해서 진행되고 있다. 세이버메트릭스는 누적된 자료를 이용하여 통계적인 관점에서 야구에 관한 분석을 하는 연구 분야이며, 세이버메트릭스 방법으로 자료 분석하는 사람을 세이버메트리션 (sabermetrician)이라고 부른다.

미국 프로야구 (Major League Baseball; MLB)에서 세이버메트리션인 James (1982)를 필두로 통계적인 연구가 진행 중이며 한국 프로야구 (Korean Baseball Organization; KBO)에서 타자 능력에 관한 연구는 Cho 등 (2003, 2004, 2005), Lee와 Kim (2005, 2006a, 2006b), Lee와 Cho (2009), Choi와 Kim (2011) Kim (2012) Lee (2014a, 2014b) 등이 있다. 타자에 관한 많은 연구 중에서 Lee (2014c)는 2000년 부터 2013년의 자료를 바탕으로 BGI (batting grade index)를 제안하였다. BGI는 타자의 능력을 측정하는 세이버메트릭스 통계량들을 주성분분석을 이용하여 8개의 인자를 추출하여 타자의 능력을 비교하는 통계량이다. 그리고 BGI를 기반으로 군집분석을 실시하여 선수들을 네 개의 군집으로 분할하

[1] 교신저자: (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 교수.
E-mail: cshong@skku.edu
[2] (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 대학원생.
[3] (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 대학원생.

여 분석하였다. 또한, Yang 등 (2015)은 Lee (2014c)의 연구를 확장하여 타고투저 현상의 발생 원인을 분석에 포함하였다.

WAR은 야구선수의 모든 요소를 계산해서 종합한 통계량이기 때문에 계산하기가 상당히 까다롭지만 기본 정의는 다음과 같다.

$$WAR = (Batting\ Runs + Base\ Running\ Runs + Fielding\ Runs + Positional\ Adjustment + League\ Adjustment + Replacement\ Runs) / (Runs\ Per\ Win).$$

WAR 계산에 쓰이는 각 변수를 구하는 식이 복잡하며 여러 방법이 존재하기 때문에 세이버메트릭션마다 해석을 다르게 할 수 있다 (2.1절에서 언급한 bWAR과 fWAR). 그러나 선수의 가치를 하나로 표현하는 방법이란 공통된 점에서 매우 큰 장점을 가지고 있다.

MLB에서 사용하는 WAR식을 KBO에 적용하기에는 한계가 있다. MLB에서는 우익수와 좌익수에 대한 포지션 조정 (position adjustments)변수, 리그 조정 (league adjustment)변수 그리고 서로 다른 야구장의 크기를 조정하기 위한 야구장 요인 (park factor)이 포함된 수비에 관련된 변수 (fielding runs)를 사용하는데, KBO에는 이러한 세 종류의 변수계산에 필요한 충분한 자료가 존재하지 않는다. 한국 야구장은 지역별 구장의 편차가 크고 잦은 개보수로 인해 야구장 요인 변수를 계산하기에 어려움이 많으며 MLB 선수와 KBO 선수의 신체 조건이 서로 다르기 때문에 이런 변수를 KBO 자료에 포함시키는데 많은 문제점이 존재한다. 따라서 본 연구에서는 타자의 능력을 측정하는 대안적인 세이버메트릭스 통계 모형을 개발하여, 미국프로야구 뿐만 아니라 한국프로야구에서도 동시에 사용하면서 비교 분석할 수 있는 타자력 지수를 개발한다. KBO와 동일한 방식으로 측정되는 MLB의 자료를 바탕으로 타자력 지수를 개발하고 이 자료를 MLB 선수들의 WAR과 비교 분석하여 타당성을 검정한 후에 KBO의 자료에 대하여 한국형 타자력 지수를 제안한다. 그리고 한국형 타자력 지수를 바탕으로 KBO에서 활동하는 타자들에 대한 다양한 통계분석을 실시하고자 한다. 우선 본 논문에서 제안한 타자력 지수와 연봉과의 관계를 심도있게 분석하기 위하여 책정된 연봉은 2016년도를 기반으로 하였고, 2016년 연봉책정에 영향을 주는 타자력 지수 모형을 개발하기 위하여 2015년도의 KBO 자료를 이용하였다. 따라서 한국형 타자력 지수를 제안하고, 2015년도의 타자력 지수와 책정된 2016년의 연봉과의 관계를 연구하는 것이 본 논문의 연구 목적이다. 그리고 이 관계를 통하여 연봉책정의 적절성에 따라 선수들을 군집화하여 다양한 인자에 대하여 분석하면서 토론한다.

본 연구의 구성은 다음과 같다. 2절에서는 타자에 관한 통계량 중에서 다섯 가지의 통계량을 소개하면서 2015년도 MLB 자료를 바탕으로 다섯 가지의 통계량을 이용한 타자력 지수 모형을 개발하고, 미국 프로선수들의 WAR 과 비교하여 적절한 지수임을 보인다. 개발한 모형을 이용하여 2015년 KBO에서 프로야구 선수 379명 중 규정타석 (446타석)을 만족한 타자 46명 (총 51명 중 KBO를 떠난 5명은 제외)의 자료를 바탕으로 한국형 타자력 지수를 제안한다. 3절에서는 한국형 타자력 지수를 적용한 선수들의 자료를 바탕으로 10개 프로팀과 선수들의 나이 그리고 포지션별로 통계분석을 실시하여 설명한다. 4절에서는 2015년도의 타자력 지수와 책정된 2016년의 연봉과의 관계를 회귀분석을 이용하여 모형을 설정한다. 이 모형을 바탕으로 연봉 책정의 적절성에 따라 선수들을 세 그룹으로 나누어 다양한 분석을 실시하여 탐색한다. 팀, 나이, 내·외국인, 포지션 그리고 연봉변동에 따라서 분석하고 설명한다. 마지막 5절에서는 결론을 유도한다.

2. 타자력 지수 제안

2.1. MLB에서의 HAI 설정

기존의 연구 문헌에서는 타자들의 능력을 평가하는 80여 개의 세이버메트릭스 통계량 중에서 주성분

분석 (principal component analysis)을 이용하여 각각 8개와 13개의 통계량을 선정하여 인자를 설정하였는데, 선정된 통계량의 계산식에 중복되어 정의되는 통계량들이 존재하며 나아가 통계량들 사이에 매우 유의한 상관관계가 존재하기 때문에 통계적인 분석이지만 선정된 통계량들을 설명하고 이해하기에는 문제점이 존재한다. 이를 극복하기 위하여 본 연구에서는 타자들의 능력을 평가하는 통계량들 중에서 가장 핵심인 출루와 득점에 연관된 중요한 통계량만을 선정하여 사용하고자 한다. 선정 조건으로는 첫 번째로 타자의 경기력을 잘 설명해야 하고, 두 번째로는 편의 (bias)가 작아야 하며, 세 번째로 MLB의 세이버메트릭스 통계량이 KBO에서도 제공하는 통계량으로 구성할 수 있어야 한다. 네 번째로는 선정된 통계량이 역사적으로 좋은 평가를 받아온 통계량이라는 조건 하에서 선정된 다섯 가지 통계량만을 사용한다. 각 통계량들에 대한 설명과 계산식은 다음과 같다.

GPA (Gross Production Average): 가장 보편화된 타격 퍼포먼스를 평가하는 지수인 OPS (출루율과 장타율)를 보완하기 위해서 개발된 통계량이다. OPS의 장점인 계산의 편의성과 타자의 출루율 (OBP)과 장타율 (SLG)을 함께 평가하는 점은 그대로 살리고, 단점인 장타율을 과대평가하는 부분을 개선한 통계량이다.

$$GPA = \ (1.8OBP + SLG) \ / \ 4.$$

BB/K (Base on Balls / striKe out): 타자가 얻은 볼넷의 수를 삼진 수로 나눈 값으로, 타자의 출루능력을 평가하는 통계량이다.

$$BB/K = \ Base\ on\ Balls \ / \ Strike\ Out.$$

wRC+ (weighted Runs Created+): wOBA (weight On Base Average, 가중출루율)를 바탕으로 타자의 득점창출력을 측정하는 통계량으로 타자가 타석에 들어섰을 때 일어나는 각각의 사건의 가치를 분리하여 생각해 각기 다른 가중치를 적용하여 계산하는 통계량이다.

$$wRC+ = (\ \frac{(wOBA - league\ wOBA) \ / \ wOBAscale}{League\ R/League\ PA} + 1 \ ) \times \ 100,$$

여기서 $wOBA = (0.72NIBB + 0.75HBP + 0.90(1B) + 0.92ROBE + 1.24(2B) + 1.56(3B) + 1.95HR)/PA$,
$league\ wOBA$= 리그 평균 $wOBA$, $wOBA\ scale$= 해당연도 wOBA 조정지수,
$League\ R$= 리그 전체 득점, $League\ PA$= 리그 전체 타석, PA= 타자의 총 타석.

RC27 (Runs Created 27): Bill James가 고안한 통계량으로, 특정 한 선수가 한 경기 모든 타석에 설 경우 팀이 몇 점을 득점할 것인가를 평가하는 통계량이다.

$$RC27 = \frac{RC}{PA - H + SH + SF + CS + GIDP} \times \ 27,$$

여기서 $RC = \ (2.4C + A)(3C + B)/9C - 0.9C$,
$A = H + BB - CS + HBP - GIDP$, $C = AB + BB + HBP + SH + SF$,
$B = 1.125(1B) + 1.69(2B) + 3.02(3B) + 3.73HR + 0.29(BB - IBB + HBP) + 0.492(SH + SF + SB) - 0.04K$.

XR27 (eXtrapolated Runs 27): 선수 혼자 만들어낸 안타, 홈런, 번트등과 같은 기본 통계량을 이용하여, 그 선수의 득점생산성을 측정하기 위한 세이버메트릭스 타격 통계량이다.

$$XR27 = \frac{XR}{PA - H + SH + SF + CS + GIDP} \times \ 27,$$

여기서 $XR = 0.50(1B) + 0.72(2B) + 1.04(3B) + 1.44HR + 0.34(HBP + TBB - IBB) + 0.25IBB + 0.18SB$
$-0.32CS - 0.09(AB - H - K) - 0.098K - 0.37GIDP + 0.37SF + 0.04SH$.

본 연구에서는 2015년도 MLB의 타자에 관한 자료를 가장 잘 설명하는 위에서 선정한 다섯 개의 통계량들의 선형 결합으로 나타나는 첫 번째 주성분 (principal component)으로 타자력 지수를 설정하고자 한다. 우선 2015년 MLB에서 규정타석 조건을 만족한 142명의 타자의 통계자료를 가지 주성분분석을 이용하여 타자력 지수를 개발한다 (MLB 2015년 자료 참조). 주성분분석의 기본 가정을 만족하는 방법으로 KMO와 Bartlett 검정을 사용하였다. 표본적합도 KMO는 0.806으로 비교적 큰 값으로 나타났으며, Bartlett 검정통계량값은 1041.062 (p-값$<$ 0.0001)이므로 변수들 간의 상관관계가 존재한다고 판단되어 주성분분석에 적합한 가정을 만족한다고 확인할 수 있다. 각각의 표준화된 통계량들의 선형결합으로 나타나는 첫 번째 주성분인 MLB의 타자력 지수 (Hitting Ability Index; HAI)는 다음과 같이 설정한다. 앞으로, KBO에 대한 타자력 지수와 구별하기 위하여 MLB와 KBO의 타자력 지수를 각각 HAI_M와 HAI_K로 표기한다.

$$HAI_M = 0.248Z_{GPA} + 0.163Z_{BB/K} + 0.245Z_{wRC+} + 0.215Z_{RC27} + 0.242Z_{XR27}. \tag{2.1}$$

이 지수는 전체변동의 79.148%를 설명하며, 다섯 개 통계량에 대한 신뢰도 cronbach's α값은 0.926이다. 본 연구에서 설정한HAI_M의 신뢰성을 살펴보기 위하여 bWAR, fWAR과 상관분석을 실시하였다. bWAR은 베이스볼레퍼런스 (`www.baseball-reference.com`)에서 기존의 고전적인 통계량을 기반으로 계산한 WAR이고, fWAR은 팬그래프 (`www.fangraphs.com`)에서 세이버메트릭스 통계량을 기반으로 계산한 WAR이다. 상관분석 결과, HAI_M와 bWAR의 상관계수는 0.758 그리고 HAI_M와 fWAR과는 0.771로 둘 다 매우 유의한 상관관계를 갖는다. 따라서 (2.1)식에서 설정한 타자력 지수 HAI_M는 두 종류의 bWAR, fWAR과의 관계분석을 통하여 MLB에서의 타자력 지수로서 타당하다고 평가할 수 있다.

2.2. KBO에서의 HAI 설정

MLB 자료를 바탕으로 타자력 지수를 설정하는 방법을 KBO에 적용해 본다. 2015년 규정타석을 만족하고, 2016년 재계약을 마친 타자 46명에 대한 자료를 분석한다 (KBO, `http://www.koreabaseball.com/Record/Player/HitterBasic/BasicOld.aspx`). 우선, KBO 타자력 지수에 포함된 변수들의 기초 통계량은 Table 2.1과 같다.

Table 2.1 Simple statistics for 5 variables

	N	MIN	MAX	MEAN	STD
GPA	46	.227	.421	.29	.035
BB/K	46	.25	1.51	.65	.257
wRC+	46	64.9	222.4	119.40	28.399
RC27	46	3.367	17.331	7.74	2.354
XR27	46	3.998	22.118	7.99	3.030

이 변수들에 대한 KMO 검정값은 0.802으로 적합하고, Bartlett 검정통계량값이 331.147 (p-값$<$ 0.0001)이므로 주성분분석의 가정을 만족한다. 표준화된 통계량들의 선형 결합으로 나타나는 첫 번째 주성분인 KBO의 타자력 지수 HAI_K는 다음과 같다.

$$HAI_K = 0.249Z_{GPA} + 0.147Z_{BB/K} + 0.246Z_{wRC+} + 0.225Z_{RC27} + 0.250Z_{XR27}. \tag{2.2}$$

이 지수는 전체변동의 77.759%를 설명하며, 다섯 개 통계량에 대한 신뢰도인 크론바 알파 (cronbach's α)값은 0.878로 HAI_M보다 조금 작지만 여전히 매우 높은 신뢰성을 갖는다.

그러므로 2.1절에서 논의한 방법으로 개발한 MLB의 타자력지수 HAI_M (2.1)식은 WAR을 대체할 수 있는 타당성 있는 적절한 방법임을 보였다. 그리고 이 모형을 KBO 자료에 적용하여 구

한 KBO의 타자력지수 HAI_K (2.2)식도 HAI_M와 유사하게 높은 신뢰성을 갖고 있음을 탐색하였다. 2016년 KBO타자 46명의 HAI와 2016년 MLB타자 중 상위 46명에 대한 HAI는 부록의 Table A.1과 A.2에 각각 정리하였다.

3. HAI의 자료 분석

3.1. 팀별 HAI

KBO 자료를 바탕으로 제안한 타자의 경기력을 나타내는 HAI_K를 이용하여 다양한 분석결과를 살펴본다. 여기부터는 KBO의 HAI에 대하여만 분석하므로HAI_K를 HAI로 표기한다. 우선 10개의 프로야구 소속팀에 따른 HAI의 평균분석은 Table 3.1과 같다. Levene 통계량값이 1.263 (p-값=0.290)으로 등분산을 가정하여 진행한 ANOVA 분석에서의 F 통계량값이 0.738 (p-값=0.0672)로 유의수준 5%하에서 유의하지 않다. 즉 소속팀별 간에 HAI는 통계적으로 차이가 없다고 판단할 수 있다.

Table 3.1 Means of HAI with respect to teams

	Doosan	Hanwha	KIA	KT	LG	Lotte	NC	Nexen	Samsung	SK
N	6	3	2	5	2	6	9	3	6	4
MEAN	-0.28	0.91	-0.06	-0.21	-0.2	0.37	-0.15	0.17	0.33	-0.7
STD	0.46	0.657	0.005	0.771	0.242	0.705	1.785	0.968	0.864	0.129

3.2. 나이별 HAI

20대 선수와 30대 초반 그리고 그 이후의 선수 나이로 집단화하여 선수 나이에 따른 HAI의 평균분석은 Table 3.2에 정리하였다. Levene 통계량값이 2.375 (p-값=0.105)로 등분산을 가정한 ANOVA 분석에서의 F 통계량값이 5.719 (p-값=0.010)로 유의수준 1%에서도 유의하다. 즉 나이별 HAI는 차이가 있으며 특히 20명의 30대 전반의 선수들이 매우 큰 HAI값을 갖고 있음을 탐색할 수 있다.

Table 3.2 Means of HAI with respect to age

		N	MEAN	STD
age	20 - 29	20	-.379	.618
	30 - 34	20	.497	1.192
	35 -	6	-.393	.26

3.3. 포지션별 HAI

9개의 포지션에 따른 HAI의 평균분석은 Table 3.3에 요약하였다. Levene 통계량 값이 1.998 (p-값=0.074)로 등분산을 가정한 ANOVA 분석에서의 F 통계량값이 1.182 (p-값=0.0336)로 유의수준 5%하에서 유의하지 않다. 따라서 포지션별 간에 HAI는 통계적으로 유의한 차이가 없다고 판단된다.

Table 3.3 Means of HAI with respect to positions

	1B	2B	3B	C	CF	DH	LF	RF	SS
N	6	5	7	5	6	3	4	4	6
MEAN	.92	-.10	-.14	-.29	-.27	.45	-.12	.37	-.55
STD	1.811	.439	.936	1.230	.813	.501	.616	.663	.470

4. HAI와 연봉 분석

2015년 규정타석을 만족하고, 2016년 재계약을 마친 타자 46명의 HAI 값과 2016 연봉 (salary)과의 관계를 살펴보기 위하여 회귀분석을 하였다. 2016년에 책정된 연봉은 KBO 보도자료 (`http://www.koreabaseball.com/News/Notice/View.aspx?bdSe=6181`)를 참조하였으며, 1억원 미만의 타자의 연봉은 각 구단 홈페이지 자료를 사용하였다. 회귀모형의 기본 가정인 잔차의 독립성, 동집성 그리고 정규성을 검토한 결과 회귀모형의 가정이 적절하고, 이를 바탕으로 분석한 HAI와 2016 KBO 책정된 연봉 (단위: 억원)과의 회귀모형식은 (4.1)과 같으며, 이를 Figure 4.1에 구현하였다.

$$\widehat{SALARY} = 4.81 + 2.21\ HAI. \tag{4.1}$$

회귀모형식에서 기울기는 2.21이므로 HAI가 1.0 증가할 때 연봉은 약 2억2천만원 이상 증가한다고 예측할 수 있다. 이 자료에 대하여 Pearson 상관계수는 0.652로 유의한 값이며, 회귀모형에 대한 F 통계량값은 32.568 (p-값$<$ 0.0001)이므로 회귀모형은 유의하다. 그리고 결정계수와 수정된 결정계수는 각각 0.652와 0.425이다.

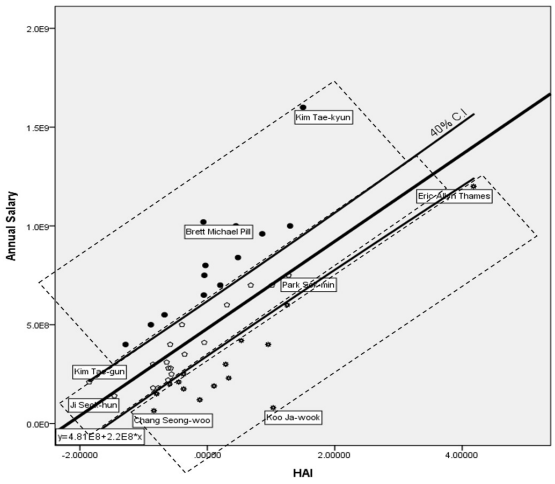

Figure 4.1 Relation between HAI and annual salary

본 연구의 분석 대상인 타자 46명의 2016년도 책정된 연봉이 적절한지를 파악하기 위하여 세 집단으로 나누어 보았다. 세 집단에 적절한 표본수를 할당시키기 위하여, 회귀모형의 신뢰구간 40%를 기준으로 하여 상한과 하한을 설정하였다. 이 상한선을 넘어있는 13명의 선수는 타자의 경기력이 고평가 (over evaluated)되어 2016년 선수의 실적보다 높은 연봉을 받았으며, 하한 아래에에 위치한 15명의 선수는 타자의 경기력이 저평가 (under evaluated)되어 낮은 연봉을 받았다고 판단하고, 회귀직선의 신뢰구간 40% 안에 분포되어있는 18명의 선수는 HAI에 비교하여 2016년 연봉책정이 적절하게 평가 (fully evaluated)되었다고 할 수 있다 (Table 4.1 참조). 예를 들어 회귀직선의 99% 신뢰구간 밖에 위치한 김태균 선수는 구단 내 간판스타로서 외적인 요소가 포함돼 연봉책정이 매우 잘 되었으며, 반면에 구자욱 선수는 회귀직선의 95% 신뢰구간 아래에 위치하고 있는데 데뷔 1년 차 신인으로서 자신의 경기력보다 현저히 못 미치는 연봉으로 책정되었다고 설명할 수 있다. 회귀모형 분석을 바탕으로 책정된 연봉의 적

절성에 따라 세 그룹으로 선수들을 나누어 팀별, 나이별, 내·외국인별, 연봉변동별, 포지션별로 연봉에 대하여 분석한다.

Table 4.1 Evaluated batters list

Over evaluated	Andy Manuel Marte, Brett Michael Pill, James Charles Adduci, Jeong Keun-woo, Kang Min-ho, Kim Tae-kyun, Lee Bum-ho, Lee Ho-jun, Lee Jong-wook, Lee Seung-yuop, Oh Jae-won, Park Yong-taik, Son Si-heon	13
Fully evaluated	Choi Hyung-woo, Chung Soo-bin, Hwang Jae-kyun, Ji Seok-hun, Kim Jae-ho, Kim Min-seong, Kim Sang-hyeon, Kim Sang-soo, Kim Seong-hyeon, Kim Tae-gun,Lee Dae-hyeong, Lee Jae-won, Lee Myoung-ki, Lee Yong-gyu, Min Byeong-heon, Park Jeong-kwon, Park Sok-min, Son Ah-seop	18
Under evaluated	Chang Seong-woo, Choi Jun-seok, Eric Allyn Thames, Hur Kyeong-min, Jung Hoon, Kim Ha-seong, Kim Jong-ho, Koo Ja-wook, Na Sung-bum, Oh Ji-hwan, Park Hae-min, Park Kyung-soo, Park Min-woo, Yang Eui-ji, Yoo Han-jun	15

4.1. 팀별과 나이별 연봉 분석

소속팀별과 나이별로 살펴본 연봉과의 관계는 Table 4.2에 정리하였다. 여기서 주목할 점은 NC는 모든 주전 선수가 규정타석을 만족하였다. 그리고 SK의 경우 모두 타자의 경기력을 정평가하여 연봉책정을 했다고 확인할 수 있다. 그러나 평가된 연봉책정의 적절성과 소속팀별간의 HAI의 평균차이는 통계적으로 유의하지 않다 (F 통계량값=1.789, p-값= 0.105).

연차가 높은 나이일수록 고평가의 비율이 높아지는 것으로 판단할 수 있다. 따라서 한국프로야구의 연봉은 현재의 타자 실력보다는 누적된 선수 경력 등이 중요하다는 것을 인지할 수 있다 (F 통계량값=11.778, p-값$<$ 0.0001).

Table 4.2 Evaluated batters with respect to teams and ages

	team										age			sum
	Doosan	Hanwha	KIA	KT	LG	Lotte	NC	Nexen	Samsung	SK	20-29	30-34	35-	
batters	6	3	2	5	2	6	9	3	6	4	20	20	6	46
Over evaluated	1	2	2	1	1	2	3	-	1	-	-	8	5	13
Fully evaluated	3	1	-	2	-	2	2	1	3	4	10	7	1	18
Under evaluated	2	-	-	2	1	2	4	2	2	-	10	5	-	15

4.2. 내·외국인별, 포지션별 그리고 연봉변동별 연봉 분석

국내·외의 선수들의 KBO 연봉과의 관계 그리고 포지션별과 연봉변동별과의 연봉책정과의 관계에 대한 결과를 Table 4.3에 정리하였다. 외국인 선수는 연봉에 대해서 고평가되는 경향이 있다는 것을 파악할 수 있지만 외국인 선수의 수가 국내 선수의 수에 비하여 매우 적기 때문이라고 판단된다. 그리고 KBO의 연봉책정에서 대부분의 선수들은 9개의 포지션별에 큰 영향을 받지 않는다고 탐색할 수 있다. 또한 규정타석을 만족한 타자의 경우에는 연봉은 대부분이 증가한 것으로 분석된다. 따라서 4.2절에서 논의한 분석에서 내·외국인간의 평균차이에 대한 분석 (t 통계량값=3.666, p-값$<$ 0.0001), 포지션간의 평균차이에 대한 분석 (F 통계량값=1.07, p-값=0.400), 연봉변동간의 평균차이에 대한 분석 (F 통계량값=0.139, p-값=0.871) 모두 통계적으로 유의하지는 않다고 결론내릴 수 있다.

Table 4.3 Evaluated batters with respect to teams and ages

	team										age			sum
	Doosan	Hanwha	KIA	KT	LG	Lotte	NC	Nexen	Samsung	SK	20-29	30-34	35-	
batters	6	3	2	5	2	6	9	3	6	4	20	20	6	46
Over evaluated	1	2	2	1	1	2	3	-	1	-	-	8	5	13
Fully evaluated	3	1	-	2	-	2	2	1	3	4	10	7	1	18
Under evaluated	2	-	-	2	1	2	4	2	2	-	10	5	-	15

5. 결론

본 연구는 2015년 KBO에서 규정타석 (446타석)을 만족한 타자 46명 (총 51명 중 KBO를 떠난 5명은 제외)을 대상으로 KBO에서 제공하는 타자의 모든 기록 통계량 중 타자의 능력을 가장 잘 나타내는 다섯 개의 세이버메트릭스 통계량을 이용하여 새로운 세이버메트릭스 통계량인 타자력 지수 (HAI)를 제안하였다. 그리고 2015년 KBO선수들의 경기력을 바탕으로 산출된 2016년 연봉책정과 관련하여 HAI와 2016년 연봉책정과의 관계 속에서 팀별, 나이별, 국내·외별, 연봉변동별, 포지션별로 통계분석을 실시하였다. 나이별 HAI의 평균을 살펴보면 유의한 차이가 있으며, 특히 20명의 30대 전반의 선수들이 매우 큰 HAI값을 갖고 있음을 발견하였다. 그리고 소속팀에 따른 HAI의 평균은 유의한 차이가 없으며 또한 선수의 포지션별에도 유의한 차이가 없다.

본 연구의 분석 대상인 타자 46명을 회귀직선의 신뢰구간을 바탕으로 연봉책정의 고평가, 정평가, 저평가인 세 그룹으로 할당하여 팀별, 나이별, 내·외국인별, 연봉변동별, 포지션별로 연봉에 대하여 분석하였다.

우선 소속팀과 연봉과 관계를 살펴보면, NC는 모든 주전 선수가 규정타석을 만족하였으며 SK의 경우 모두 타자의 경기력을 정평가하여 연봉책정을 했다고 확인할 수 있다. 그리고 선수들의 나이가 많을수록 고평가의 비율이 높아지므로 KBO의 연봉은 현재의 타자 실력보다는 누적된 선수 경력 등이 중요하다는 것을 인지할 수 있다. 외국인 선수와 연봉과의 관계는 외국인선수의 수가 매우 적지만 국내 선수보다 고평가하는 경향이 있다는 것을 파악할 수 있다. 그리고 KBO의 연봉책정에 포지션의 난이도는 반영되지 않는다고 탐색한다. 그러나 이와 같은 분석은 나이를 제외하고 나머지는 통계적으로 유의하지는 않다고 결론내릴 수 있다.

MLB 야구 통계자료를 세이버메트릭스 방법으로 분석하는 데 있어 가장 널리 알려졌고 공신력 있다고 판단하는 WAR 통계량을 KBO에서도 사용하여야 하는데 1절에서 언급한 바와 같이 이를 사용하기에는 극복해야할 많은 어려움이 있다. 한국에서도 사용할 수 있는 여러 가지 지수들을 개발하는 것도 중요하지만, MLB와 KBO를 비교 분석하기 위해서는 MLB에서 사용하는 WAR 통계량을 사용할 수 있도록 KBO의 노력이 필요하다고 제안한다. 즉 KBO의 발전이 한 걸음 더 나아가기 위해서는 KBO와 각 구단에서 MLB의 WAR 통계량을 사용하기 위한 노력이 필요하다고 제안한다.

야구를 통계학을 비롯한 과학적인 방법으로 접근한다 하더라도 2015년 KBO 깜짝 우승팀인 두산의 경우처럼 승패를 예측하는 것은 어렵다. 예측이 다양하고 변수가 많은 야구에서 여느 세이버메트릭스 통계량과도 마찬가지로 본 연구에서 제안한 HAI도 승패를 가르기 위함이 아닌 선수들의 활약 정도에 따른 평가 측도로만 인지하고 선수들의 발전에 기여를 하는 정도로 판단하여 사용하여야 하겠다.

부록

Appendix A: Betters and HAI for KBO and MLB in 2015

Table A.1 HAI for KBO (46 people)

name	KHAI
Eric Allyn Thames	4.17827
Kim Tae-kyun	1.50242
Kang Min-ho	1.29866
Park Sok-min	1.276
Yoo Han-jun	1.25274
Koo Ja-wook	1.03303
Lee Yong-gyu	1.01529
Choi Jun-seok	0.95503
Andy Manuel Marte	0.86228
Choi Hyung-woo	0.68129
Yang Eui-ji	0.53273
James Charles Adduci	0.48059
Lee Seung-yuop	0.44825
Park Kyung-soo	0.33509
Son Ah-seop	0.3071
Na Sung-bum	0.28728
Jeong Keun-woo	0.20266
Park Min-woo	0.10506
Park Yong-taik	-0.0294
Lee Ho-jun	-0.04676
Kim Jae-ho	-0.04824
Lee Bum-ho	-0.05506
Brett Michael Pill	-0.06218
Kim Ha-seong	-0.1191
Min Byeong-heon	-0.35665
Oh Ji-hwan	-0.37165
Kim Jong-ho	-0.37249
Hwang Jae-kyun	-0.39819
Jung Hoon	-0.4501
Kim Sang-hyeon	-0.56301
Chung Soo-bin	-0.57552
Park Jeong-kwon	-0.58347
Hur Kyeong-min	-0.59008
Lee Jae-won	-0.61032
Kim Min-seong	-0.61603
Kim Sang-soo	-0.63841
Oh Jae-won	-0.67201
Lee Myoung-ki	-0.77258
Park Hae-min	-0.79607
Chang Seong-woo	-0.83912
Kim Seong-hyeon	-0.85193
Lee Dae-hyeong	-0.8529
Lee Jong-wook	-0.88577
Son Si-heon	-1.28332
Ji Seok-hun	-1.45974
Kim Tae-gun	-1.85368

Table A.2 HAI for MLB (Top 46 people)

name	MHAI
Bryce Harper	3.901445
Paul Goldschmidt	2.678461
Miguel Cabrera	2.534051
Joey Votto	2.410598
Mike Trout	2.36835
Nelson Cruz	1.752319
Michael Brantley	1.612309
Andrew McCutchen	1.608335
Anthony Rizzo	1.598884
Edwin Encarnacion	1.579276
Josh Donaldson	1.505694
Buster Posey	1.458279
Jose Bautista	1.331017
David Peralta	1.268521
A.J. Pollock	1.138084
Yunel Escobar	1.055289
David Ortiz	1.030714
Manny Machado	0.995963
Chris Davis	0.937327
Kris Bryant	0.933862
Matt Carpenter	0.933369
Prince Fielder	0.906865
J.D. Martinez	0.893169
Ben Zobrist	0.890335
Eric Hosmer	0.868144
Adam Lind	0.846854
Brandon Belt	0.776349
Ryan Braun	0.731705
Kendrys Morales	0.718014
Shin-Soo Choo	0.716752
Yoenis Cespedes	0.708969
Lorenzo Cain	0.695518
Nolan Arenado	0.676959
Mookie Betts	0.663637
Josh Reddick	0.632314
Curtis Granderson	0.609443
Alex Rodriguez	0.604482
Jose Abreu	0.580795
Adrian Gonzalez	0.56981
Jason Heyward	0.553523
Carlos Gonzalez	0.538178
Jason Kipnis	0.526586
Mike Moustakas	0.511623
Carlos Beltran	0.507541
Logan Forsythe	0.471444
Jose Altuve	0.467148

Appendix B: Statistic List in sabermetrics

Table B.1 Selected statistic List

abbreviation		meaning
1B		Single Base
2B		Double Base
3B		Triple base
AB		At bat
BB		Base on balls (also called a “Walks”)
CS		Caught stealing
GIDP		Ground into double play (also called a “GDP”)
H		Hit
HBP		Hit by pitch
HR		Home runs
IBB		Intentional base on balls
K		Strike out (also abbreviated SO)
NIBB		Not intentional base on balls
OBP		On-base percentage
PA		Plate appearances
R		Runs scored
RC		Runs created
ROBE		Reached base on error
SB		Stolen base
SF		Sacrifice fly
SH		Sacrifice hit
SLG		Slugging average
TBB		Total base on balls
wOBA		weight on base average
position	1B	First baseman
	2B	Second baseman
	3B	Third Baseman
	C	Catcher
	CF	Center Fielder
	DH	Designated Hitter
	LF	Left Fielder
	RF	Right Fielder
	SS	Short Stop

References

Cho, Y. S. and Cho, Y. J. (2003). The research regarding a Beane Count application from Korean baseball league. *Journal of the Korean Data Analysis Society*, **5**, 649-658.

Cho, Y. S. and Cho, Y. J. (2004). Study about the influence that WHIP has on ERA in 2003 season Korean professional baseball. *Journal of the Korean Data Analysis Society*, **6**, 1415-1424.

Cho, Y. S. and Cho, Y. J. (2005). A study on OPS and runs from Korean baseball league. *Journal of the Korean Data Analysis Society*, **7**, 221-231.

Choi, Y. G. and Kim, H. M. (2011). A statistical study on Korean baseball league games. *The Korean Journal of Applied Statistics*, **24**, 915-930.

James, B. (1982). *The Bill James Baseball Abstract*. Ballantine Books, New York.

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 (2015). `http://www.koreabaseball.com/Record/Player/HitterBasic/BasicOld.aspx`.

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

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

Lee, J. T. (2014c). 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). Win-lose models when two teams meet using data mining in the Korean pro-baseball. *Journal of the Korean Data Analysis Society*, **11**, 3417-3426.

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.

Major League Baseball (2015). `http://mlb.mlb.com/stats`.

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.

Journal of the Korean Data & Information Science Society
2016, **27**(3), 677–687

http://dx.doi.org/10.7465/jkdi.2016.27.3.677
한국데이터정보과학회지

Alternative hitting ability index for KBO

Chong Sun Hong[1] · Jae Young Kim[2] · Dong Sik Shin[3]

[123]Department of Statistics, Sungkyunkwan University

Received 15 April 2016, revised 13 May 2016, accepted 17 May 2016

Abstract

Among lots of sabermetric statistics for baseball batters' ability, the wins above replacement (WAR) is the most popular statistic in MLB. However, there exists a difficulty applying WAR to KBO, since KBO data do not have position adjustment, league adjustment and park factor which are essential in calculating WAR. In this paper, using five statistics for both KBO and MLB qualified batters, we propose hitting ability index (HAI), an alternative sabermetric indices to represent batters' ability. Comparing HAI with WAR of MLB batters, we evaluate the validity of HAI and then applied HAI to 2015 KBO data in which HAI is analyzed statistically with respect to different teams, ages, and positions. Moreover, the linear relationship between KBO batter's HAI and their annual salary is discussed. Grouping 46 KBO batters based on confidence region of the regression model for annual salary, we also statistically investigate batter's annual salary in these groups with respect to several factors.

Keywords: Hitting ability, principal component, regression model, sabermetrics.

[1] Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, Seoul 03063, Korea. E-mail: cshong@skku.edu
[2] Graduate student, Department of Statistics, Sungkyunkwan University, Seoul 03063, Korea.
[3] Graduate student, Department of Statistics, Sungkyunkwan University, Seoul 03063, Korea.