한국산업기술대학교 빅데이터 분석, 2021
Korea Polytechnic University

선형회귀모형을 이용한 2020년 KBO 타격지표의 다각적 분석

장경준
[a)]한국산업기술대학교 IT경영학과, 2016314031

June, 2021

1. 서론

국내에서 KBO의 많은 인기로 사랑받고 있는 야구는 “기록의 스포츠”라고 불릴 정도로 경기 기록이 발달되어 있으며, 이를 바탕으로 한 다양한 통계 분석이 널리 사용되고 있다. 특히 야구는 몇 가지의 특정한 상황하에서 모든 경우의 수가 발생하여, 다른 스포츠 종목에 비해 정적인 특성이 있다. 따라서 기록화해 분석하기가 용이하다. 야구는 1860년대 미국에서 시작되어 150년 가량의 긴 역사를 갖고있는 만큼 누적된 기록이 풍부하고, 그만큼 많은 통계 지표들이 파생되어 널리 사용되고 있다. 가장 대표적으로 타율, 출루율, 장타율 등이 있고, 이에서 파생된 OPS,wOBA,WAR 등이 있다. 특히 WAR(Wins Above Replacement), 대체선수 승리 기여도는 가장 직관적이고 널리 쓰이는 지표로서 모든 선수(야수, 투수)의 타격, 수비, 주루, 투구를 종합하여 그 선수의 승리에 대한 기여도를 직관적으로 알 수 있어 널리 활용되고 있다.

본 연구에서는 2020년 KBO 타자 294명의 타격 기록을 바탕으로 회귀분석을 이용해 타격 지표들간의 상관관계에 대해 분석해보고자 한다. 데이터셋에서 기본적으로 제공된 OPS, wOBA, WAR이외에도 이들에서 파생된 wRAA(공격공헌도), wRC(득점창출력)의 변수를 추가하여, 타자가 직접적으로 결과에 기여할 수 있는 득점, 타점과의 상관관계를 증명한다. 또한 가장 널리 쓰이는 통계지표인 OPS,WAR와 다른 지표들간의 상관관계에 대해서도 시각화하여 보여주고자 한다.

2. 문헌조사

KBO 타자 데이터분석에 대한 선행연구에서 김혁주,김예형(2015)는 변수선택 기법을 이용한 한국 프로야구의 득점과 실점 설명 방안에 대해 연구하여, 전방선택법, 후방소거법, 단계별 회귀법 등 다양한 회귀모형을 통해 득점과 실점에 미치는 영향에 대해 탐구하였다.

조영석,조영주(2005)는 한국 프로야구에서 OPS와 득점에 관한 연구에서 득점과 OPS의 상관관계를 분석하기 위해, OPS를 이용한 경기당 득점 추정 회귀모형을 만들어 출루율과 장타율이 경기당 득점에 미치는 요소를 비교 분석 하였다.

또한 김혁주(2012)는 한국 프로야구에서 출루 능력과 장타력이 득점 생산성에 미치는 영향을 분석하여, 출루율,장타율과 득점 생산성의 상관관계에 대해 분석하였다.

3. 방법론

1) 탐색적 데이터 분석(EDA)

타율 = 안타수/타수

장타율 = (1루타 x1 + 2루타 x2 + 3루타 x3 + 홈런 x4) / 타수

출루율 = 안타+볼넷+사구 / 타수 + 볼넷 + 사구 + 희생플라이

OPS = 장타율 + 출루율

wOBA(가중출루율)

wOBA = (0.72*NIBB + 0.75*HBP + 0.90*1B + 0.92*RBOE + 1.24*2B + 1.56*3B + 1.95*HR) / (PA - IBB)

PA - 타석, IBB - 고의4구, NIBB - 고의4구 제외 볼넷, HBP - 몸에 맞는 공, RBOE - 실책으로 인한 출루

wRRA(Weighted Runs Above Average, 리그평균 대비 득점생산)

리그 평균과 비교해 한 타자가 팀에 기여하는 득점수를 나타내는 지표로, 타자가 득점에 기여하는 정도를 효율적으로 비교하기 위해 생성하였다.

wRRA = 선수 wOBA - 리그평균 wOBA / wOBA Scale(1.15) x PA(타석)

wRC(Weighted Runs Created, wOBA기반 득점생산)

타자의 전체 공격 가치를 정량화한 지표로, 가중출루율을 기반으로 한 득점 생산능력를 비교하기 위한 지표로 활용하기 위해 생성하였다.

wRC = {((wOBA - 리그평균wOBA) / wOBA Scale) + (리그득점/리그타석)} * 타석

WAR(Wins Above Replacement) : 대체 수준 대비 승리 기여도

WAR는 선수의 타격,투구,수비,주루 등 모든 능력은 하나의 단일화된 수치로 보여줄 수 있다는 점에서 유용하다. 타자에게 WAR는 대부분 타격 스텟이 크게 작용하므로 타격 기록과의 상관관계가 크지만, 현재 제공된 데이터셋 내에서는 타격기록만 제공되었으므로 수비, 주루에 대한 부분이 누락되었다. 따라서 현재 타격지표와 WAR를 직접적으로 비교하는 것은 정확성이 떨어진다.

표1. 기타 파생변수

lg.obp	리그 평균 출루율
lg.slg	리그 평균 장타율
lg.wOBA	리그 평균 wOBA
lg.R	리그 전체 득점합
lg.PA	리그 전체 타석 합
rg	경기별 득점(Runs for a game)
rbig	경기별 타점(Runs batted in for a game)
A_OPS	조정 OPS(리그평균대비 OPS)

※리그 평균 wOBA는 리그 전체 타자들의 데이터가 제공되지 않은 관계로, 다른 파생변수들의 정확한 비교를 위해 당해 리그 전체평균 수치(0.3315)를 대입하였다.

표2. 주요 DATA Summary

	안타	홈런	득점	타점	볼넷	삼진
최소	0.00	0.00	0.00	0.00	0.00	0.00
최대	199.00	47.000	116.00	135.00	91.00	154.00
표준편차	52.40	7.84	29.12	30.63	21.53	33.66
평균	46.24	4.652	25.38	24.03	18.14	33.64

	타율	출루율	장타율	OPS	wOBA	WAR
최소	0.00	0.00	0.00	0.00	0.00	0.00
최대	1.0000	1.0000	1.0000	2.0000	0.9070	8.7600
표준편차	0.108325	0.119961	0.163926	0.267514	0.115149	1.685146
평균	0.2281	0.3047	0.3245	0.6291	0.2895	0.6101

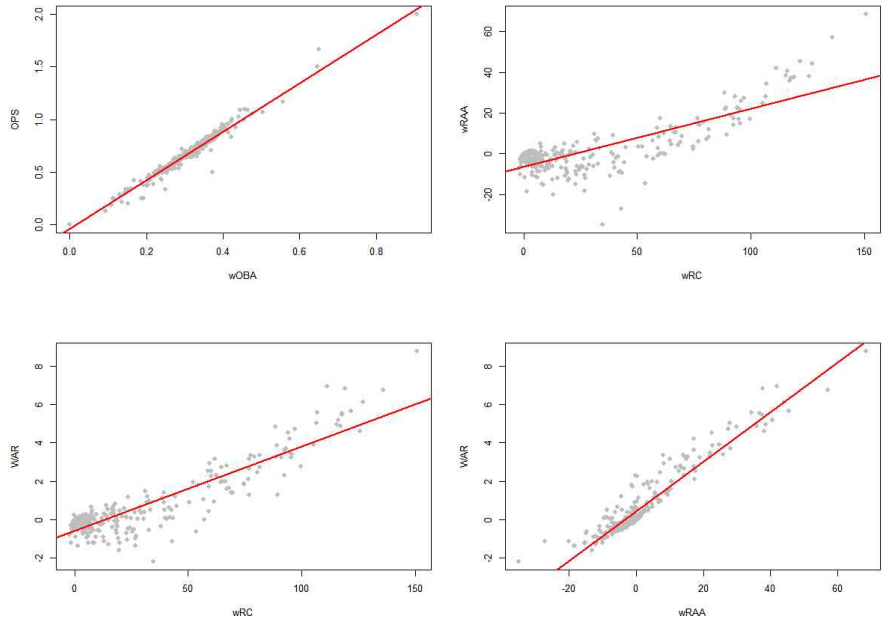

그림1. 파생 변수들간 상관관계 분석

비교대상	Multiple R-squared	Adjusted R-sqaured	상관계수
OPS-wOBA	0.9673	0.9671	0.9834
wRRA-wRC	0.6282	0.6269	0.7925
WAR-wRC	0.8092	0.8085	0.8995
WAR-wRRA	0.8973	0.8969	0.9472

2) 회귀분석

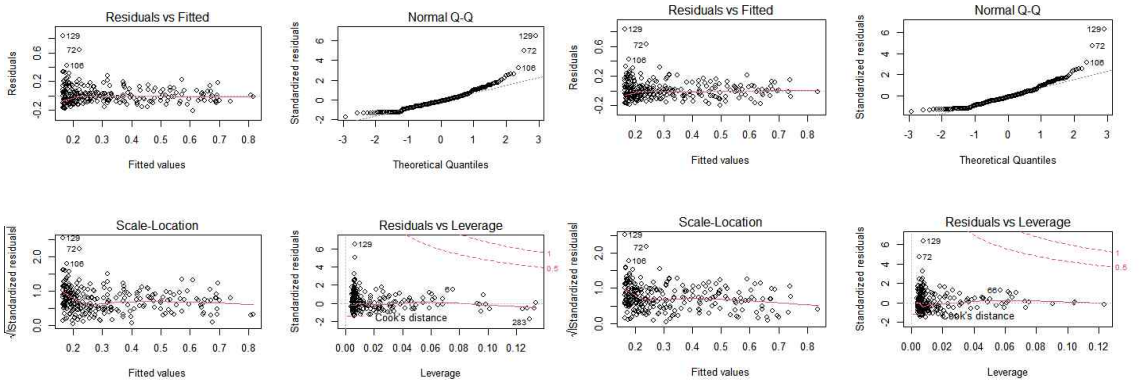

그림2. 득점 관련/타점 관련 상관관계 회귀분석 그래프

①득점 관련 상관관계 분석

득점에 영향을 미치는 변수의 상관관계를 파악하기 위해, 안타,홈런,볼넷,도루와 득점을 비교하는 모델을 생성하였고, 다중공선성 검정 결과 모두 10 이하의 수치를 나타내어 회귀계수 추정에 신뢰성 문제는 발생하지 않았다.

전방 선택법을 적용한 결과 안타,볼넷,홈런,도루 순이었고 볼넷이 가장 상관관계가 높은 것으로 나타났다.

후방소거법을 적용한 결과, 볼넷 홈런 도루 안타 순으로 모형에서 제거되었다.

단계별 회귀를 적용한 결과, 안타,홈런,도루,볼넷 순으로 모형에 들어갔다.

②타점 관련 상관관계 분석

타점에 영향을 미치는 변수의 상관관계를 파악하기 위해, 안타,홈런,삼진과 타점을 비교하는 모델을 생성하였고, 다중공선성 검정 결과 모두 10 이하의 수치를 나타내어 회귀계수 추정에 신뢰성 문제는 발생하지 않았다.

전방 선택법을 적용한 결과 안타,홈런,삼진 순이었고 삼진이 가장 상관관계가 높은 것으로 나타났다.

후방소거법을 적용한 결과, 삼진,홈런,안타 순으로 모형에서 제거되었다.

단계별 회귀를 적용한 결과, 안타,홈런,삼진 순으로 모형에 들어갔다.

4. 결과

야구에서 점수를 결정하는 득점과 타점에 영향을 미치는 요인들을 분석하기위해 wRRA,wRC등 다양한 파생변수를 이용해 분석하였다. wOBA와 WRC는 높은 상관관계를 가지며 타자의 공격력을 평가하는데 적합한 지표로서 평가되었다. 리그 평균을 적용한 RRA와 wRC를 이용하면 더 정확한 타점, 득점에 대한 접근이 가능하다.

회귀분석기법, 전방선택, 후방소거, 단계별 회귀를 통한 분석결과 득점에 영향을 미치는 요소는 볼넷, 안타, 홈런 순으로 나타났다. 같은 방법으로 분석한 타점에 영향을 미치는 요소는 삼진, 홈런, 안타 순이었다.