1990년대 MLB (Major League Baseball)에서 득점과 실점을 이용한 승률에 관한 연구

유 승 현[1)]

요 약

야구에서 득점(Batter`s Runs, BR)과 실점(Pitcher`s Runs, PR)은 타점이나 타율 그리고 승리나 세이브에 비하여 상대적으로 가치를 인정받지 못하였다. 그러나 야구의 승리는 득점을 많이 하고 실점을 적게 해야 하는 기본적인 매커니즘을 무시할 수는 없다. 본 연구에서는 득점과 실점을 이용하여 승률과의 상관관계를 알아보고 승률을 추정하기 위하여 회귀분석을 실시하였으며 추정된 회귀식과 미국의 야구통계학자로 유명한 빌 제임스의 피타고리안 승률 등과 적합성을 비교하고자 한다.

주요용어: 득점, 실점, 회귀분석, 피타고리안 승률, 빌 제임스

Ⅰ. 서론

야구는 통계적으로 표현할 수 있는 가장 대표적이자 이상적인 스포츠이다. 투수가 던진 하나의 공에 따라 발생하는 사건들이 수치적인 정보로 기록될 수 있기 때문이다. 미국야구연구협회의(Society for American Baseball Research, SABR) 에서는 수많은 논문과 수식을 통하여 야구를 통계적으로 접근하고자 많은 노력을 하고 있다. 야구는 기본적으로 27번의 기회 가운데 누가 상대방보다 득점을 많이 내는가 하는 경기이다. 가장 기본적인 승리 공식은 득점은 많고 실점은 적은 팀일 것이다.

Ⅱ. 득점과 실점을 이용한 승률 추정의 이용실태

야구는 기본적으로 27번의 기회 가운데 누가 상대방보다 득점을 많이 내는가 하는 경기이다. 가장 기본적인 승리 공식은 득점(batter`s runs, BR)은 많고 실점(pitcher`s runs, PR)은 적은 팀일 것이다. 야구의 승률(winning average, WA)에 대한 정의는 승(Win, W)과 패(Lose, L)의 합에 대한 승의 비율로 나타내는데 다음과 같다.

1)712-749 경상북도 경산시 대동 214-1, 영남대학교 통계학과. E-mail : effort-result@naver.com

$$WA = \frac{W}{W+L}$$

피트 파머는 SABR의 연감 The National Pastime(1982)에서 득점과 실점을 고려한 승률에 대한 연구는 언쇼 쿡의 Percentage Baseball(1964)에서 시도되었다고 하였다. 언쇼 쿡은 1950년에서 1960년 사이의 메이저리그 경기 경기를 분석한 득점과 실점을 이용하여 승률을 다음과 같이 정의하였다.

$$WA = \frac{BR}{PR} \times .484$$

아놀드 술먼은 출판되지는 않았지만 일부 미디어의 관심을 모은 보고서에서 1901년에서 1970년까지의 경기 당 득점(batter`s runs per game, BRG)과 경기 당 실점(pitcher`s runs per game, PRG) 자료를 이용하여 승률을 다음과 같이 정의하였다.

$$WA = .102 \times BRG - .103 \times PRG + .505$$

빌 제임스는 The Bill James Baseball Abstract(1982)에서 승률에 대한 생각을 발전시켰다. 그는 승률을 다음과 같이 정의하였다.

$$WA = \frac{BR^2}{BR^2 + PR^2}$$

그리고 이후 연구를 통해 다음과 같은 식으로 승률을 정의하였다.

$$WA = \frac{BR^{1.83}}{BR^{1.83} + PR^{1.83}}$$

본 연구에서는 승률을 종속변수로 하고 경기당 득점과 경기당 실점을 종속변수로 하여 추정된 중선형 회귀식과 피타고리안 승률, 언쇼 쿡, 아놀드 술먼의 승률 추정식중 어떤 추정식이 1990년대 MLB에서 가장 잘 적합이 되는지 찾고자 한다.

Ⅲ. 1990년대 MLB에서 통계분석

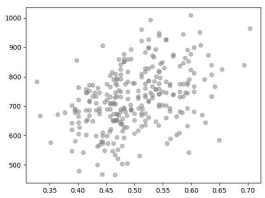
3-1.상관분석

WA**와** BRG, PRG, $\frac{BRG}{PRG}$의 상관분석을 실시한 결과는 [표 1]과 같다. WA와 BRG은 양의 상관관계($r = 0.4060$)가 있었고, PRG는 음의 관계($r = -0.4488$), $\frac{BRG}{PRG}$와는 양의 상관관계($r = 0.9316$)를 나타내었다. 이는 모두 통계적으로 유의한 차이가 있었으며($^{**}p < .01$) 산점도를 살펴 보아도 마찬가지 결과가 나타났음을 알 수 있다.

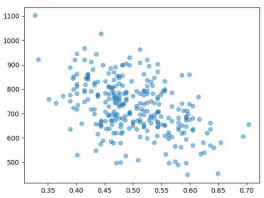
[표 1] 상관분석

항목	WA	BRG	PRG	$\frac{BRG}{PRG}$
WA	1	0.4060^{**}	-0.4488^{**}	0.9316^{**}
BRG		1	0.5745^{**}	0.4210^{**}
PRG			1	-0.4877^{**}
$\frac{BRG}{PRG}$				1

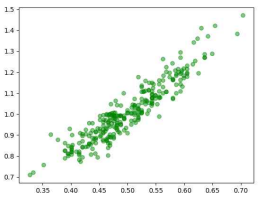
$^{**}p < .01$, $^{*}p < .05$



[그림 1] WA와 BRG의 상관관계



[그림 2] WA와 PRG의 상관관계



[그림 3] WA와 $\frac{BRG}{PRG}$의 상관관계

3-2.회귀분석

WA를 종속변수로 하고 BRG와 PRG를 독립변수로 하여 중선형 회귀분석을 한 결과는 [표 2]와 같다. 회귀식은 통계적으로 유의한 차이가 있었으며 PRG가 승률에 BRG에 비하여 상대적으로 영향을 더 주는 것으로 나타났으며 추정된 회귀식은 다음과 같다.

$$\widehat{y_1} = 0.538 + 0.081 \times BRG - 0.088 \times PRG$$

[표 2] BRG와 PRG를 독립변수로 한 회귀분석

독립변수	비표준화 계수		표준화 계수	t	유의확률
	B	표준오차	베타		
(상수)	.4949	.012		42.005	.000**
BRG	.0006	.0000178		35.853	.000**
PRG	-.0006	.0000171		-36.839	.000**
$F = 839.6 (p-value = .000^{**})$			$adj - R^2 = .858$		

$^{**}p < .01$, $^{*}p < .05$

WA를 종속변수로 하고 $\frac{BRG}{PRG}$를 독립변수로 하여 단순 선형 회귀분석을 한 결과는 [표 3]와 같다. 회귀식은 통계적으로 유의한 차이가 있었으며 추정된 회귀식은 다음과 같다.

$$\widehat{y_2} = 0.045 + 0.450 \times \frac{BRG}{PRG}$$

[표 3]BRG/PRG를 독립변수로 한 회귀분석

독립변수	비표준화 계수		표준화 계수	t	유의확률
	B	표준오차	베타		
(상수)	.0461	.011		4.284	.000**
BRG/PRG	.4494	.011		42.588	.000**
$F = 1814. (p-value = .000^{**})$			$adj - R^2 = .867$		

$^{**}p < .01$, $^{*}p < .05$

3-3. 분산분석

최종 군집 중심

	1	2	3	4
BRG	−.17007	1.37804	.64664	−.95322
PRG	.70792	1.15915	−.65222	−.76126

기술통계량

	N	평균	표준편차	F(p−value)
1	77	.45121	.037919	
2	48	.50683	.058279	
3	58	.57543	.046458	66.234 (.000)**
4	95	.49009	.059417	
전체	278	.50002	.067141	

Student-Newman-Keuls

케이스군집번호	N	1	2	3
1	77	.45121		
4	95		.49009	
2	48		.50683	
3	58			.57543
CTT 유의확률		1.000	.065	1.000

군집별 상위 10개

연도	팀	승률	군집	연도	팀	승률	군집
1990년	Atlanta Braves	.556	1	1994년	New York Yankees	.619	2
1990년	Chicago Cubs	.525	1	1995년	Boston Red Sox	.597	2
1990년	Detroit Tigers	.519	1	1995	Cleveland Indians	.584	2
1990년	Milwaukee Brewers	.519	1	1995년	California Angels	.538	2
1991년	Baltimore Orioles	.514	1	1991년	Texas Rangers	0.525	2
1991년	Detroit Tigers	.506	1	1993년	Detroit Tigers	.525	2
1991년	Oakland Athletics	.500	1	1994년	Minnesota Twins	.469	2
1992년	Detroit Tigers	.497	1	1994년	Detroit Tigers	.461	2
1992년	New York Yankees	.488	1	1994년	Texas Rangers	.456	2
1992년	Seattle Mariners	.488	1	1994년	Seattle Mariners	0.438	2

연도	팀	승률	군집	연도	팀	승률	군집
1990년	New York Mets	0.562	3	1990년	Chicago White Sox	0.58	4
1990년	Toronto Blue Jays	0.531	3	1990년	Cincinnati Reds	0.562	4
1991년	Atlanta Braves	0.58	3	1990년	Boston Red Sox	0.543	4

1991년	Chicago White Sox	0.537	3	1990년	Los Angeles Dodgers	0.531	4
1991년	Milwaukee Brewers	0.512	3	1990년	California Angels	0.494	4
1991년	Minnesota Twins	0.586	3	1990년	Cleveland Indians	0.475	4
1991년	Pittsburgh Pirates	0.605	3	1990년	Baltimore Orioles	0.472	4
1992년	Minnesota Twins	0.556	3	1990년	Kansas City Royals	0.466	4
1992년	Toronto Blue Jays	0.593	3	1990년	Houston Astros	0.463	4
1993년	Atlanta Braves	0.642	3	1990년	Minnesota Twins	0.457	4

Ⅳ. 승률 추정식의 비교 분석

모형확인의 가장 좋은 방법은 새로운 자료를 수집하여 선택된 모형을 검토하는 것이다. 새로운 자료에서 설명변수들의 특정한 값을 모형에 대입했을 때의 적합치가 새로운 자료의 반응치와 얼마나 차이가 나는지를 보기 위하여 MSPR(mean squared prediction error)을 사용하며, 그 공식(강근석 등, 1999)은 다음과 같이 정의된다. 여기서, y_i는 새로운 자료의 관측값, $\hat{y_i}$는 최종모형을 이용한 예측값, n^*는 자료의 개수를 말한다.

$$MSPR = \frac{\sum_{i=1}^{n^*} (y_i - \hat{y_i})^2}{n^*}$$

적합된 회귀식과 빌 제임스가 제시한 공식과의 $MSPR$을 비교하여 보면 [표 4]과 같다. BRG와 PRG를 독립변수로 한 회귀직선의 가장 적합성이 좋은 것으로 나타났다.($MSPR = 0.0004006587$)

[표 4] 추정된 승률과 $MSPR$

항목	추정된 승률 식에 따른 $MSPR$			
	피타고리안(2)	피타고리안(1.83)	$\widehat{y_1}$	$\widehat{y_2}$
$MSPR$	0.0004354635	0.0004159649	0.0004006587	0.0004740583

Ⅴ. 결론

WA는 설명변수인 BRG, $\frac{BRG}{PRG}$와 정비례, PRG와는 반비례의 예상과 동일한 결과를 확인할 수 있다. 특히 WA와 $\frac{BRG}{PRG}$의 R값은 0.9316^{**}로 굉장히 높다고 볼 수 있다. 그렇기에 승률은 승점 나누기 실점으로 설명할 수 있다고 말할 수 있다.