Journal of the Korean Data Analysis Society (August 2017)
Vol. 19, No. 4 (B), pp. 2071-2079

다중회귀분석을 이용한 메이저리그 승률의 모형구축과 예측

이석원[1], **천영진**[2]

요 약

본 연구에서는 미국 메이저리그 야구 경기에서 다중회귀분석을 이용하여 승률을 추정하는 통계 모형을 제시하고, 이를 바탕으로 2015년의 결과를 예측하고자 하였다. 이를 위하여 메이저리그 2009년부터 2013년까지 팀당 162경기의 경기 결과를 대상으로 타자 지표 27개, 투수 지표 37개, 수비 지표 15개 등을 선정하여 승률을 반응변수로, 나머지 78개의 변수들을 설명변수로 하여 분석하였다. 산점도를 통하여 35개의 변수를 선정한 후 stepwise 방식을 사용하여 최종적으로 6개의 변수를 선정하였다. 그 결과 득점, 실점, 세이브, 완봉경기, 삼진, 홈런의 변수가 승률에 영향을 미친 것으로 나타났다. 이 모형을 2015년 승률에 적용한 결과 각 리그의 순위는 정확하게 일치하였고, 승률은 1, 2위 예측은 실제 값과 3% 이내로 일치하였으며, 나머지 순위의 승률도 대부분 일치하는 것으로 나타났다. 승률을 높이기 위해서 투수 부문에서는 확실한 선발 투수와 경기를 마무리할 수 있는 전문 투수가 필요하고, 가장 쉽게 점수를 얻을 수 있는 방법인 홈런을 칠 수 있는 클러치 타자가 필요하며, 삼진을 당하지 않을수록 승률이 좋은 것을 알 수 있다.

주요용어 : 다중회귀분석, 승률, 모형구축, 메이저리그.

1. 서론

우리나라 프로 스포츠 중 야구 종목은 대중으로부터 가장 큰 관심을 받고 있고, 2012년에는 프로스포츠 사상 최초로 700만의 관중을 돌파한 종목이다(Jang, Moon, 2014). 2016년 들어 국내 프로야구 리그와 일본 프로야구 리그에서 활약하던 선수들이 대거 메이저리그로 진출하여 어느 해보다 메이저리그에 대한 관심이 높아졌다.

메이저리그는 1869년에 시작되어 약 150년에 가까운 역사를 가지고 있으며, 30개 팀이 양대 리그로 나뉘어 팀당 162경기를 한 후, 각 리그에서 지구별 1위 팀과 와일드카드 진출 1팀이 챔피언쉽 시리즈를 거친 후 양대 리그 우승팀끼리 월드시리즈를 통하여 최종 우승팀을 겨루는 약 8개월간의 장기 레이스로 펼쳐진다. 이로 인하여 경기 기록뿐만 아니라 경기 외적인 요인, 예를 들면 선수의 부상, 긴 이동거리로 인한 체력 저하 등으로 인하여 승률에 영향을 미칠 수 있다. 그러나 경기 외적인 요인으로 인하여 승률을 예측하는 데에는 한계가 있어, 메이저리그에서 사용하고 있는 경기 기록을 바탕으로 승률의 모형을 구축하고, 이를 바탕으로 승률을 예측하고자 하였다.

승률과 관련하여 기존 연구들을 살펴보면, 주로 한국 프로야구 경기를 대상으로 이루어졌는데, Cho, Cho(2003, 2004, 2005a, 2005b)는 한국 프로야구에서 Beane Count와 승률간의 관계를 살펴보고, 이닝당 주자 출루율이 방어율에 미치는 영향과 득점과 OPS(on-base percentage plus slugging)의 관계를 분석하였으며, 프로빗 모형을 이용하여 한국 프로야구의 승률을 추정하였다. 또한, Lee, Kim

[1]02504 서울시 동대문구 서울시립대로 163, 서울시립대학교 통계학과 학사졸업. E-mail : leesk0604@naver.com
[2](교신저자) 02504 서울시 동대문구 서울시립대로 163, 서울시립대학교 스포츠과학과 조교수.
E-mail : yj1000@uos.ac.kr
[접수 2017년 7월 17일; 수정 2017년 8월 14일; 게재확정 2017년 8월 17일]

(2007)은 한국 프로야구에서 승률 예측을 위한 효율적인 통계적 모형을 제시하였다. Jang et al.(2014)은 한국 프로야구의 자료를 사용하여 고정효과 패널회귀모형에다 고정효과벡터분해를 병행하여 시변변수뿐만 아니라 시간불변변수가 승률에 미치는 영향을 분석하였고, Kim, Jin(2014)은 베이스의 획득과 허용이 승리에 얼마나 영향을 주는지에 대하여 살펴보았다.

승률을 예측하는 문제는 세이버 메트릭션들의 가장 큰 관심사로, 야구의 승률을 계산하는 공식에 득점과 실점의 관계식을 이용하여 승률을 예측한 공식인 피타고라스 정리를 들 수 있다(James, 1982). 이 식을 간단히 설명하면 승률은 득점 제곱을 득점의 제곱과 실점 제곱의 합으로 나눈 것으로 설명할 수 있다. 이 공식에 대하여 Lee et al.(2007)은 위의 피타고라스 정리 공식이 주로 지나간 데이터를 이용하여 승률이 잘 맞는지를 알아보는데 사용된다는 단점을 지적하였다.

따라서, 본 연구에서는 메이저리그의 지난 5년간의 자료 중 투수, 타자 수비 부분의 자료를 바탕으로 승률을 예측할 수 있는 모형을 구축하고, 다음 년도와 그 이듬해의 승률을 예측하여, 구축한 모형이 실제 결과와 어느 정도 일치하는지 확인하는 데에 그 목적이 있다.

2. 연구방법

2.1. 자료수집

2009년부터 2013년까지 5년간 팀별 기록을 수집하기 위하여 메이저리그 공식 홈페이지(http://mlb.mlb.com)에서 타자부문, 투수부문, 수비부문으로 구분하여 각 분야에 해당하는 자료를 수집하였다. 타자 부문에서는 27개, 투수 부문에서는 36개, 수비 부분에서는 15개와 승률을 합한 총 79개의 변인을 조사하였으며 구체적인 변인 설명은 Table 1과 같다.

Table 1. Definition of predictor variables

Hitting		Pitching		Fielding
G(경기수)	SAC(희생번트)	W(승)	IBB(고의사구)	GS(선발)
AB(타수)	SF(희생타)	L(패)	GF(마무리)	INN(이닝)
R(득점)	TB(총루타)	ERA(방어율)	HLD(홀드)	TC(수비기회)
H(안타)	XBH(장타)	GS(선발)	GIDP(병살타)	PO(아웃)
2B(2루타)	GDP(병살타)	SV(세이브)	GO(땅볼)	A(보살)
3B(3루타)	GO(땅볼)	SVO(세이브기회)	AO(뜬공)	E(에러)
HR(홈런)	AO(뜬공)	IP(이닝)	WP(폭투)	DP(병살)
RBI(타점)	GO_AO(땅볼/뜬공)	H(피안타)	BK(보크)	SB(도루)
BB(사사구)	NP(피투구수)	R1(실점)	SB(도루)	CS(도루자)
SO(삼진)		ER(자책점)	CS(도루자)	SBPCT(도루성공률)
SB(도루)		HR(피홈런)	PK(견제)	PB(포일)
CS(도루자)		BB(사사구)	TBF(타자수)	C_WP(폭투캐치)
AVG(타율)		SO(삼진)	NP(총투구수)	FPCT(에러율)
OBP(출루율)		AVG(피안타율)	WPCT(승률)	DER(수비율)
SLG(장타율)		WHIP(이닝당출루허용률)	GO_AO(땅볼/뜬공)	
OPS(출루율+장타율)		CG(완투)	OBP(출루율)	
IBB(고의사구)		SHO(완봉)	SLG(장타율)	
HBP(사구)		HB(사구)	OPS(출루율+장타율)	

2.2. 분석 방법

분석을 위하여 79개의 변수에 대한 자료를 SAS 프로그램에 입력하였다. 이 중 승률을 반응변수

로 두고 나머지 78개의 변수들을 설명변수로 하였다. 분석방법은 다중회귀분석으로 $y=\beta_0+\beta_1 x_1+\cdots+\epsilon$라는 모델로 가정하였다.

3. 연구결과

설명변수의 수가 78개로 매우 많기 때문에 반응변수와 설명변수의 산점도를 통하여 연관성이 없어 보이는 변수들을 제거한 후 다시 중복되는 변수를 제거하였다. 최종적으로 모형에 유의한 변수들을 선택한 결과 다음과 같은 과정을 통하여 결과를 도출하였다.

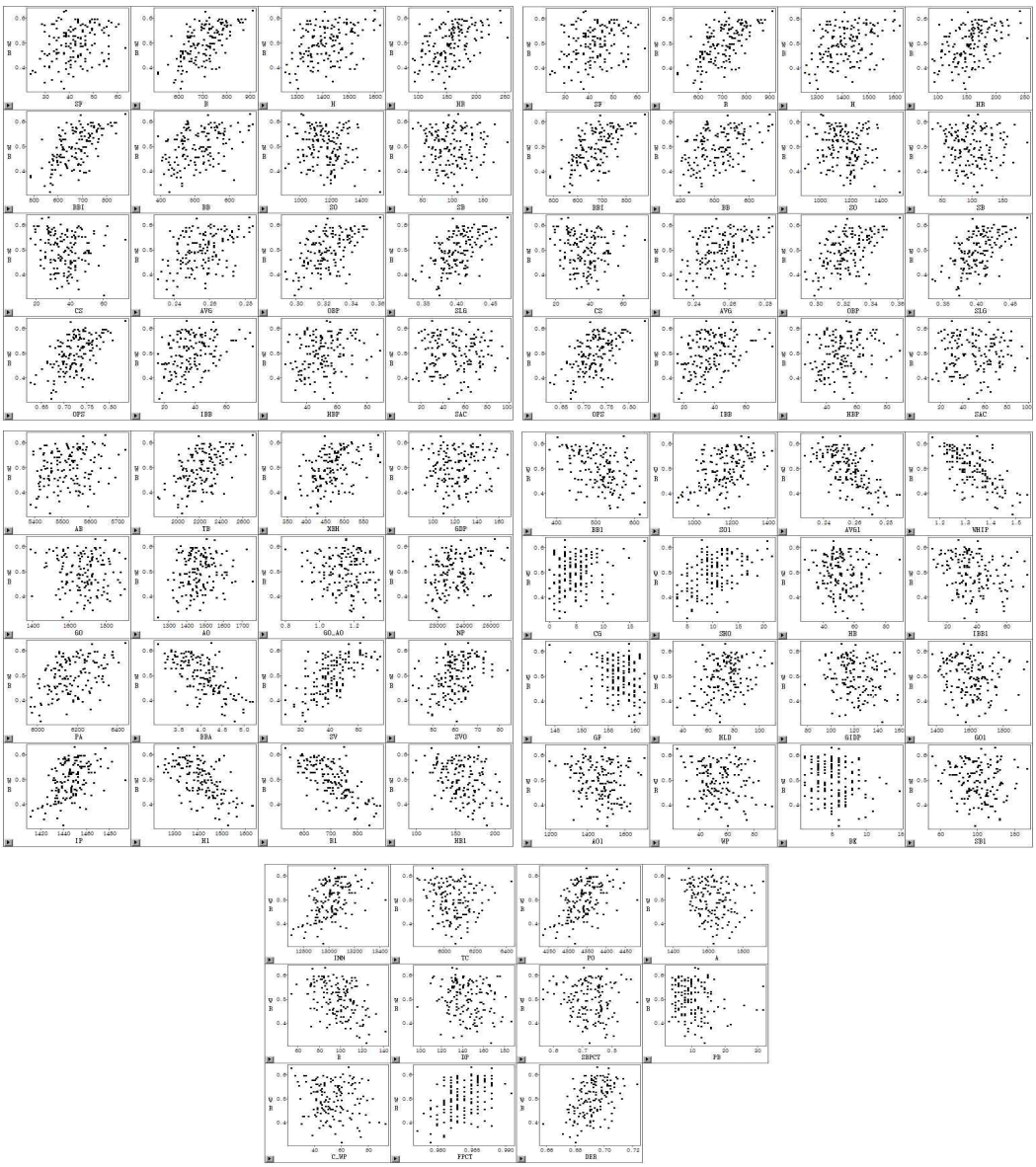

Figure 1. Scatter plot

3.1. 산점도를 통한 변수 선택

반응변수와 설명변수에 대한 산점도를 Figure 1과 같이 나타내었다. 산점도를 통하여 선형성이나 일정한 패턴이 존재할 경우 두 변수 간에는 관계가 있으며, 산점도에서 상관관계가 낮게 나타난 변수들은 설명변수에서 제거하였다. 그 결과 40개의 변수를 제거하여 타석, 득점, 안타, 홈런, 타점, 볼넷, 삼진, 타율, 출루율, 장타율, OPS, 고의사구, 루타수, 장타수, 투구수, 방어율, 세이브, 이닝, 피안타, 실점, 피타점, 탈삼진, 피안타율, 이닝당출루허용률, 완봉, 홀드, 피출루율, 피장타율, 피OPS, 이닝기회, 주자아웃, 에러, 수비성공률, 수비효율 등 35개의 변수를 선정하였다.

3.2. 35개의 변수에 대한 stepwise

35개의 변수 중 유의한 변수선택 기능인 stepwise를 실시한 결과 Table 2와 같이 나타났다. 이 모형에서 실점, 득점 세이브, 완봉, OPS, 타석 수, 투구이닝 수, 총루타 등 8개의 변수를 선택하였다. 모형적합을 한 결과 p값이 .00001 이하로 유의한 모델임을 나타내었고, R^2 역시 94%의 높은 설명력을 가지는 것으로 나타났다. 다만, 분산팽창인자의 값이 10 이상인 변수들이 존재함에 따라 모수들의 분산이 커지기 때문에 분석에 왜곡이 생길 수 있어 상관성을 확인하였다. 본 연구에서의 유의수준은 .05 이하로 하였다.

Table 2. Regression output for the 35 variables

Analysis of Variance					
Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	8	6996.77939	874.59742	279.01	<.0001
Error	141	441.98869	3.13467		
Corrected total	149	7438.76809			
	Root MSE	1.77050	R-square	0.9406	
Parameter Estimates					
Variable	DF	Parameter estimate	Standard error	t value	Pr>|t|
Intercept	1	-38.80556	26.80446	-1.45	0.1499
R1	1	-0.04554	0.00345	-13.20	<.0001
R	1	0.04501	0.00683	6.59	<.0001
SV	1	0.20512	0.02906	7.06	<.0001
SHO	1	0.11248	0.04771	2.36	0.0198
OPS	1	105.47125	27.28559	3.87	0.0002
PA	1	-0.01289	0.00355	-3.63	0.0004
IP	1	0.07688	0.02241	3.43	0.0008
TB	1	-0.01287	0.00464	-2.77	0.0063

3.3. 상관계수 확인 및 중복 변수 제거

Figure 2의 상관계수행렬에서 0.9 이상의 값을 지니는 변수들을 음영처리 하였고, 이러한 변수들에서 다중공선성이 발생한 것으로 보고 제거하였다. 이러한 변수들 중 OPS는 장타율과 출루율의 합으로 이루어진 변수이기 때문에 세 변수 중 OPS를 선택하였다. 또한, 상관계수가 0.9보다 큰 변수들을 승률에 대해 다중회귀분석 모형을 적합시킨 결과 분산팽창인자가 10 이상의 값이 발견되었다. 따라서 어느 한 변수는 제거해야하는데, 이때 변수의 모수추정치의 p값이 작은 값, 즉 승률에 더 유의한 변수를 선택하여 안타와 타율 중에는 타율을, 득점과 OPS 중에는 득점을 선택하였다.

이와 같은 방법으로 35개의 변수 중 득점, 홈런, 볼넷, 삼진, 고의사구, 장타수, 세이브, 실점, 방어율, 피탈삼진, 피안타율, 이닝당출루허용률, 완봉, 홀드, 이닝, 주자아웃, 에러, 수비성공률, 수비효율 등 20개의 변수를 선택하였다.

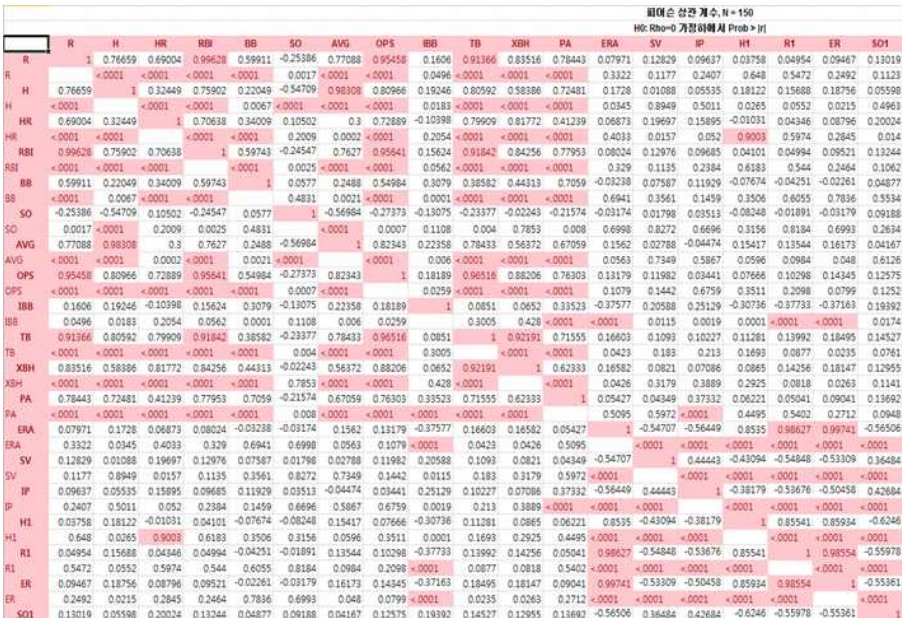

Figure 2. Correlation coefficient matrix

3.4. 20개 변수에 대한 stepwise

선택된 20개의 변수를 유의수준 .05 하에서 stepwise를 통해 모형 선택을 한 결과 Table 3과 같이 나타났다. 이 모형에서 실점, 득점, 세이브, 완봉, 홈런, 삼진 등 6개의 변수를 선택하였다. 이 변수들을 가지고 승률에 대한 회귀모형을 접합한 결과 p값이 .0001 이하로 모형이 유의한 것으로 나타

Table 3. Regression output for the 20 variables

Analysis of Variance					
Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	6	6955.46859	1159.24476	343.00	<.0001
Error	143	483.29950	3.37972		
Corrected total	149	7438.76809			
	Root MSE	1.83840	R-square	0.9350	
Parameter Estimates					
Variable	DF	Parameter estimate	Standard error	t value	Pr>|t|
Intercept	1	40.18424	4.14288	9.70	<.0001
R1	1	0.05108	0.00307	16.64	<.0001
R	1	-0.05146	0.00300	-17.13	<.0001
SV	1	0.24487	0.02790	8.78	<.0001
SHO	1	0.10074	0.04962	2.03	0.0442
HR	1	0.01539	0.00708	2.17	0.0315
SO	1	-0.00299	0.00148	-2.02	0.0452

났으며 R^2 또한 0.9350으로 높은 설명력을 나타내었다. 분산팽창인자 또한 모든 계수들에서 낮게 나오는 것을 알 수 있으며 이 모형을 승률을 예측하기 위한 모형으로 채택하였다.

3.5. 분석의 타당성 설명

1) 모형 선택의 통계량 확인

모형을 선택할 때 사용하는 통계량들인 R^2, adj-R^2, C(p), Root MSE 통계량을 Table 4와 같이 확인해 본 결과 회귀모형이 설명하는 능력인 R^2의 값이 상당히 높은 93.5%로 나타났으며 오차의 표준편차인 Root MSE가 1.8384로 나타났다.

Table 4. Summary of model

R^2	adj-R^2	C(p)	Root MSE
0.9350	0.9323	3.2021	1.8384

2) 잔차의 정규성 검정

Figure 3의 Q-Q PLOT에서 나타난 바와 같이 직선에 근접하고 중앙에 집중되고 있는 것을 보아 역시 잔차들의 정규성 증거라 할 수 있다.

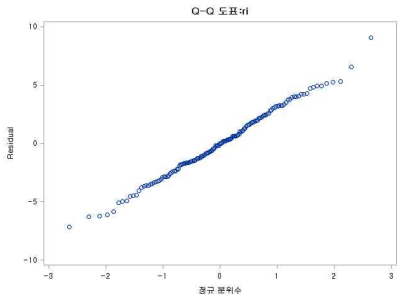

Figure 3. Q-Q plot for residuals

3.6. 최종 모형

2015년 결과 예측에 앞서 모형의 적합도를 확인하기 위하여 2014년의 기록으로 승률을 예측해 보았다. Table 5에서와 같이 예측의 1위부터 10위 내의 팀들 중에서 예상 승률과 실제 승률의 차이가 1~2% 내에 있음을 알 수 있다.

이를 바탕으로 한 최종모형은 다음과 같다.

$$\text{승률(WR, \%)} = 0.148+0.051\times\text{R}-0.051\times\text{R1}+0.245\times\text{SV}+0.101\times\text{SHO}-0.003\times\text{SO}+0.015\times\text{HR}$$

최종 모형에서 득점, 세이브, 완봉 경기수와 홈런은 승률과 양의 상관관계를 가지고 있고 실점과 삼진은 음의 상관관계를 가지고 있는 것으로 나타났다. 최종 모형을 통하여 투수 측면에서 살펴보면, 완봉 경기가 많다는 것은 팀 내에서 에이스급 투수를 보유하고 있다는 의미와 완봉을 통하여 불펜 투수에게 연투와 같은 무리한 운용을 할 필요가 없다는 것을 의미하기도 한다. 또한 세이브가 많다는 것은 확실한 마무리 투수를 보유함으로써 세이브 상황에서 팀의 승리를 지켜 승률을 높이는 데에 많은 도움이 되었다고 사료된다. 타자 측면에서 볼 때 홈런은 가장 쉽게 득점할

수 있고, 경기 분위기를 가져오는 데에 큰 도움이 된 것으로 판단되며, 삼진은 아웃카운트만 늘어날 뿐 주자의 진루나 득점할 수 있는 기회를 잃어버리기 때문에 음의 상관관계가 나타났다.

Table 5. Results of predicted WPCT and WPCT in 2014

Ranks	Team	Predicted WPCT	WPCT	Ranks	Team	Predicted WPCT	WPCT
1	BAL	.593	.593	6	OAK	.558	.543
2	LAA	.584	.605	7	STL	.548	.556
3	LAD	.580	.593	8	KC	.542	.549
4	WSH	.566	.580	9	TOR	.537	.512
5	SEA	.559	.537	10	PIT	.537	.543

3.7. 2015년 승률 예측

최종 모형을 통하여 양대 리그에서 각 지구별 팀에 대한 2015년도의 실제 승률과 모형으로 예측한 승률 예측을 한 결과 Table 6과 같이 나타났다. 각 리그 1, 2위의 예측은 실제 값과 3% 내에서 모두 일치하였고, 나머지 순위의 승률도 대부분 일치하는 것으로 나타났다.

Table 6. Results of WPCT and predicted WPCT in 2015

	Central			East			West		
	Team	WPCT	Predicted WPCT	Team	WPCT	Predicted WPCT	Team	WPCT	Predicted WPCT
A.L	KC	.586	.579	TOR	.574	.607	TEX	.543	.559
	MIN	.512	.506	NYY	.537	.550	HOU	.531	.517
	CLE	.503	.504	BAL	.500	.546	LAA	.525	.510
	CWS	.469	.443	TB	.494	.518	SEA	.469	.476
	DET	.460	.424	BOS	.481	.495	OAK	.420	.454
N.L	STL	.617	.612	NYM	.556	.560	LAD	.568	.564
	PIT	.605	.579	WSH	.512	.535	SF	.519	.540
	CHC	.599	.560	MIA	.438	.448	ARI	.488	.507
	MIL	.420	.446	ATL	.414	.404	SD	.457	.447
	CIN	.395	.423	PHI	.389	.380	COL	.420	.426

4. 결론

야구 경기는 단체 경기이면서 개인의 기록 또한 중요시 하는 스포츠이다. 이 모든 기록들은 가장 통계적이고 과학적인 자료로서 팀의 승률 예측, 개인의 연봉 협상 등에 사용되고 있다. 본 연구는 미국 메이저리그의 기록을 바탕으로 팀의 승률을 예측할 수 있는 모형을 제시하고자 하였다. 과거 5년 동안의 자료를 바탕으로 이듬해의 승률과 비교하였고, 그 이듬해의 승률을 예측하였다. 그 결과 승률에 영향을 미치는 요인은 공격 부문에서는 득점, 홈런이 양의 상관관계를, 삼진이 음의 상관관계를 나타내었고, 투수 부문에서는 실점, 세이브, 완봉경기수가 양의 상관관계를 나타내었다. Lee et al.(2007)은 한국프로야구를 대상으로 승률 예측을 위한 통계적 모형을 제안하였는데, 그 결과 타력 지표인 OPS와 투수력 지표인 WWHIP가 승률에 영향을 미쳤다고 하였고, Kim et al.(2014)에 의하면 출루율이 1할 증가하면 +29.6승, 이닝당 베이스 허용수가 1할 증가하면 -4.4승의 효과를 보인다고 하였다. 대상은 다르지만 본 연구에서처럼 메이저리그를 대상으로 한 결과와 일맥상통한다고 할 수 있다. 즉, 득점을 하기 위해서는 출루율과 장타율이 높아야 하고, 실점을 줄

이기 위해서는 베이스 허용을 최대한 줄여야 할 것이다. 한편, 본 연구에서는 수비 부문에서 영향을 미치는 변인이 나타나지 않았다. 이는 메이저리그에서 뛰고 있는 선수들의 수비 수준이 매우 높아 승률에는 영향을 크게 미치지 않은 것으로도 해석할 수 있으며, 나아가 수비 부문에 대하여 좀 더 분별력 있는 변인을 새로이 찾아야 할 것이다. 결론적으로 승률을 높이기 위해서 팀 내에서 많은 이닝을 소화할 수 있는 선발 투수와 이기고 있는 경기를 마무리 할 수 있는 투수를 보유해야 하고, 중요한 순간에 홈런을 칠 수 있는 클러치 히터를 보유해야 하며, 타자 입장에서는 삼진의 수를 줄이는 노력이 필요하다.

References

Cho, Y. S., Cho, Y. J. (2003). The research regarding a beane count application from Korean baseball league, *Journal of the Korean Data Analysis Society*, 5(3), 649-658. (in Korean).

Cho, Y. S., 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(5), 1415-1424. (in Korean).

Cho, Y. S., Cho, Y. J. (2005a). A study on OPS and runs from Korean baseball league, *Journal of the Korean Data Analysis Society,* 7(1), 221-231. (in Korean).

Cho, Y. S., Cho, Y. J. (2005b). A study on winning percentage using batter's runs and pitcher's runs in Korean professional baseball league, *Journal of the Korean Data Analysis Society*, 7(6), 2303-2312. (in Korean).

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

Jang, J. H., Moon, C. G. (2014). Determinants of team winning percentage in the Korean professional baseball league, *Korean Journal of Sport Management*, 19(3), 17-31. (in Korean).

Kim, C. K., Jin, S. (2014). Predicting wins through professional baseball statistics, *Journal of the Korean Data Analysis Society*, 16(1), 211-220. (in Korean).

Lee, J. T., Kim, Y. T. (2007). An effective statistical model that predicted winning percentage in Korean pro-baseball, *Journal of the Korean Data Analysis Society*, 9(2), 931-942. (in Korean).

Model Construction and Prediction of Major League Baseball WPCT Using Multiple Regression Analysis

Seok-Won Lee[1], *Young-Jin Chun*[2]

Abstract

In this study, we propose a statistical model that estimates the WPCT of a major league baseball game in the US using multiple regression analysis. To accomplish this, the results of 162 matches per team from 2009 to 2013 were selected, and 27 hitting ratios, 37 pitching ratios, and 15 fielding ratios were selected, respectively. After selecting 35 variables through scatter plot, six variables were finally selected using stepwise method. As a result, the variables of score, run, save, shut-out, strikeouts, and home-run affected the WPCT, and the result of applying this model to the 2015 WPCT was exactly the rank of each league. It is within 3% of the actual value, and the WPCT of the remaining ranks are almost the same. In order to increase the WPCT, the pitcher needs a reliable starting pitcher and professional pitcher to finish the game, and the batter needs a clutch hit to hit the home run, the easiest way to score points. Finally, the number of strikeouts should be reduced.

Keywords : multiple regression analysis, WPCT, model construction, major league baseball.

[1]Bachelor, Department of Statistics, University of Seoul, 163, Seoulsiripdae-ro, Dongdaemun-gu, Seoul 02504, Korea. E-mail : leesk0604@naver.com

[2](Corresponding Author) Assistant Professor, Department of Sport Science, University of Seoul, 163, Seoulsiripdae-ro, Dongdaemun-gu, Seoul 02504, Korea. E-mail : yj1000@uos.ac.kr

[Received 17 July 2017; Revised 14 August 2017; Accepted 17 August 2017]