Journal of the Korean Data & Information Science Society
2015, **26**(1), 1–10

http://dx.doi.org/10.7465/jkdi.2015.26.1.1
한국데이터정보과학회지

한국프로야구 기록들의 장기추세

이장택[1]

[1]단국대학교 응용통계학과

접수 2014년 8월 16일, 수정 2014년 10월 2일, 게재확정 2014년 10월 29일

요 약

본 연구에서는 한국프로야구 변천사를 야구 통계량들을 중심으로 살펴보았다. 분석방법으로는 1982년부터 2013년까지의 한국프로야구 데이터를 이용하여 야구 통계량들의 시계열 그래프와 상관계수를 이용하였다. 그 결과 유의수준 1%에서 연도와 유의한 양의 상관관계를 보인 통계량은 2루타, 타점, 4구, 삼진, 병살타, 사구, 출루율, OPS, 방어율, 폭투, WHIP이고, 유의한 음의 상관관계를 보인 통계량은 3루타, 도루자, 실책, 완투, 완봉, 보크였다. 상관계수가 유의한 야구통계량의 예측을 위해서는 Box-Jenkins의 ARIMA 모형을 이용하였다. 결론적으로 세월의 흐름과 가장 상관이 큰 것은 완투 횟수의 감소이며, 그 다음으로 삼진 개수의 증가를 들 수 있었다.

주요용어: 삼진, 상관, 아리마 모형, 완투, 한국프로야구.

1. 서론

1982년 시작된 프로야구는 한국에서 가장 인기가 많은 프로스포츠이며, 남녀노소를 불문하고 많은 팬을 확보하고 있다. 2004년부터 2009까지 매년 10% 이상 관중수가 증가하는 등 세월이 흘러도 지속적으로 그 인기가 유지되고 있는데, 지난 30년 동안 TV, 인터넷, 휴대폰 등에서도 실시간으로 야구를 즐길 수 있는 수단이 늘어나고 2006년 월드베이스볼클래식 4강, 2008년 베이징 올림픽 우승, 2009년 월드베이스볼클래식 준우승과 같은 세계 대회에서 좋은 성적을 거둔 것도 야구의 인기몰이에 기여했다고 판단되어진다.

모든 스포츠가 그렇듯이 프로야구를 관람하는 사람들도 일차적으로 응원하는 팀의 승패에 관심이 있겠지만, 야구는 통계의 경기이기 때문에 많은 팬들은 승패를 추월하여 팀의 전력을 비교하고 투수의 다승, 방어율 등과 타자들의 타율, 타점, 득점 등 경기력에 관심을 가지고 분석하기에 이른다. 그리고 방대한 야구데이터가 어느 스포츠보다 데이터베이스로 구축이 잘 되어 있어서 프로야구가 많은 연구자들의 분석 대상으로 선호되어 왔는데, 한국 프로야구에 관한 최근 연구들을 살펴보면 시계열모형을 이용하여 관중 수 예측을 다룬 Lee와 Bang (2010), 한국프로야구 타자들에 대한 세이버메트릭스 지수 값을 이용하여 선수들의 경기력과 연봉간의 패턴을 분석한 Seung과 Kang (2012), 출루율과 장타율이 득점에 미치는 연구를 한 Kim (2012), 한국프로야구에서의 투수평가지표에 관한 Lee (2014a), 한국프로야구에서 피타고라스 지수의 추정에 관한 Lee (2014b) 등이 있다. 하지만 지금까지의 프로야구 연구는 대부분 승패와 득점 등의 추정문제에 국한되어왔는데, 본 연구에서는 한국프로야구 30여 년 동안의 많은 기록 중에서 특히 어떤 부분들이 변화를 많이 하고 있는지에 대해서 알아보려고 하며 변화 유무를 판단하는 근거로는 그래프와 상관계수를 사용하였다.

[1] (448-701) 경기도 용인시 죽전동 126번지, 단국대학교 응용통계학과, 교수. E-mail: jtlee@dankook.ac.kr

본 논문의 구성은 다음과 같다. 2절에서는 분석 자료에 대한 설명과 연구에서 사용한 타격 및 투수에 관한 통계량을 설명하였다. 3절에서는 연도와 타격 및 투수성적에 대한 상관계수와 선그래프를 작성하고 타격 성적들 간의 상관관계를 살펴보았다. 4절에서는 상관관계가 유의한 야구통계량에 대하여 ARIMA 예측모형을 제시하였으며 마지막으로 5절에서는 본 논문의 결론을 제시한다.

2. 연구방법

2.1. 데이터의 구성

본 연구에 사용된 데이터는 한국야구위원회 (KBO)에 기록되어 있는 1982년부터 2013년 사이에 있었던 32년간 데이터이다. 통계패키지 SPSS 21K 및 EViews7과 수집된 데이터를 이용하여 연도별로 야구기록의 평균을 구하고 그 결과를 분석에 사용하였다. 야구기록은 1982년부터 1985년까지는 6팀, 1986년부터 1990년까지는 7팀, 1991년부터 2012년까지는 8팀, 2013년은 9팀의 데이터를 이용하여 구하였다.

2.2. 타격 및 투수 성적 지표

본 연구에서 사용한 관심야구기록은 한국야구 위원회 기록대백과 (2009)에 요약된 야구 기록 중에서 팬들에게 친숙한 타격성적 기록 21개와 타격 성적과 겹치는 기록을 제외한 순수 투수성적 기록 6개를 관심의 대상으로 삼았다. 예를 들면 삼진의 개수는 타격 및 투수 성적으로 모두 간주되는 데, 본 연구에서는 타격 성적으로 분류하였다. 편의상 게임, 타수, 타석의 야구약어로 G, AB, PA를 각각 이용하고, 고려된 통계량을 구체적으로 나열하면 타격 성적으로 타율 (AVG), 득점 (R), 안타 (H), 1루타 (1B), 2루타 (2B), 3루타 (3B), 홈런 (HR), 타점 (RBI), 도루 (SB), 도루자 (CS), 희타 (SH), 희비 (SF), 4구 (BB), 고의4구 (IBB), 사구 (HBP), 삼진 (SO), 병살타 (GIDP), 실책 (E), 출루율 (OBP), 장타율 (SLG), 오피에스 (OPS), 순수 장타력 (ISO)이며, 투수 성적으로 방어율 (ERA), 완투 (CG), 완봉 (SHO), 이닝당 출루허용률 (WHIP), 폭투 (WP), 보크 (BK)와 같다. 타격 및 투수력의 평가를 위한 복합성과지표들은 매우 많지만 보편화되어있는 타격 성과지표들 중 OPS와 ISO, 투수 성과지표로는 WHIP을 본 연구에서 사용하였는데, 각 지표들에 대한 자세한 계산 방법은 Table 2.1에 정리되어 있다. OBP와 SLG의 합인 OPS, 순수 장타력 (isolated power; ISO)과 이닝당 출루허용률 (walks plus hits divided by innings pitched; WHIP)은 야구통계에 대한 이론적 접근의 정확성을 더하기 위하여 만든 세이버메트릭스 데이터 분석법에 이용되는 수치이다.

Table 2.1 Formulae for baseball statistics

Statistics	Statistics glossary
OBP	$OBP = [H + BB + HBP]/[AB + BB + HBP + SF]$
SLG	$SLG = [1B + 2(2B) + 3(3B) + 4(HR)]/AB$
OPS	$OPS = OBP + SLG$
ISO	$ISO = [(2B) + 2(3B) + 3(HR)]/AB$
WHIP	$WHIP = [H + BB]/IP$, $IP = innings\ pitched$

3. 상관분석

3.1. 타격 성적

Table 3.1 Correlation between batting statistics and year with p-value

AVG (.428)* (.014)	1B/AB (.069) (.707)	2B/AB (.592)** (.000)	3B/AB (-.659)** (.000)
HR/AB (.369)* (.038)	R/G (.441)* (.011)	RBI/G (.506)** (.003)	SB/G (-.293) (.104)
CS/G (-.808)** (.000)	E/G (-.863)** (.000)	BB/PA (.456)** (.009)	SO/PA (.881)** (.000)
GIDP/PA (.478)** (.006)	SH/PA (-.027) (.884)	SF/PA (.406)* (.021)	IBB/PA (.040) (.826)
HBP/PA (.597)** (.000)	OBP (.528)** (.002)	SLG (.397)* (.024)	OPS (.456)** (.009)
ISO (.344) (.054)			

$*p < 0.05,\ **p < 0.01$

프로야구 구단들의 타격성적을 기반으로 연도별 각 통계량들의 평균을 계산하였다. 그리고 생성된 32년 데이터를 이용하여 시간의 흐름과 각각의 통계량들이 서로 상관이 있는 지를 확인하기 위하여 피어슨 상관계수를 구하고 통계적 유의성을 살펴보았다. Table 3.1은 21개 야구 통계량과 연도의 상관계수와 p-값을 보여주는데, 유의수준 5%에서 유의하지 않은 통계량은 5개로 타수당 1루타 (1B/AB), 게임당 도루 (SB/G), 타석당 희타 (SH/PA), 타석당 고의4구 (IBB/PA), 순수 장타력 (ISO)이며, 유의수준 5%에서는 유의하나 유의수준 1%에서는 유의하지 않은 통계량은 5개로 타율 (AVG), 타수당 홈런 (HR/AB), 게임당 득점 (R/G), 타석당 희비 (SF/PA), 장타율 (SLG)이며, 유의수준 1%에서도 유의한 통계량은 모두 11개로 타수당 2루타 (2B/AB), 타수당 3루타 (3B/AB), 게임당 타점 (RBI/G), 게임당 도루자 (CS/G), 게임당 실책 (E/G), 타석당 4구 (BB/PA), 타석당 삼진 (SO/PA), 타석당 병살타 (GIDP/PA), 타석당 사구 (HBP/PA), 출루율 (OBP), 오피에스 (OPS)로 나타났다. 그리고 상관계수의 크기를 보아서 연도 대비 양의 선형관계가 큰 순서는 타석당 삼진 (SO/PA), 타석당 사구 (HBP/PA), 타수당 2루타 (2B/AB)의 순서 등이며, 연도 대비 음의 선형관계가 큰 순서는 게임당 실책 (E/G), 게임당 도루자 (CS/G), 타수당 3루타 (3B/AB) 등의 순서이다.

3.2. 타격 성적의 변화

Figure 3.1은 게임당 득점과 타점의 연도별 변화를 보여준다. 2개의 변수는 매우 밀접한 관계가 있으며 점진적으로 우상향 추세를 가진다고 할 수 있다. 득점과 타점의 피크는 모두 OPS가 1이 넘는 타자가 많았던 1999년이었다.

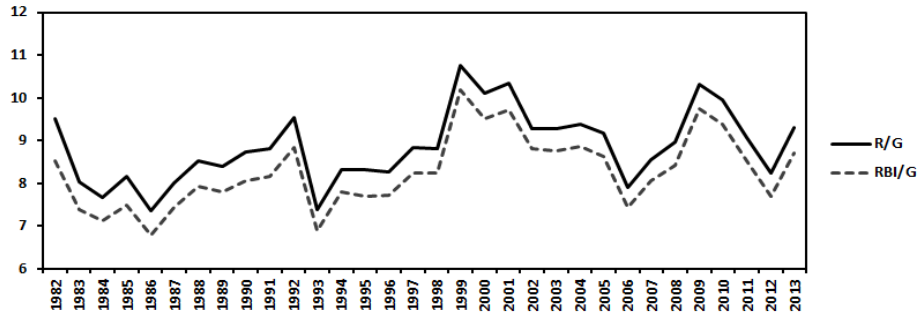

Figure 3.1 R and RBI per game, 1982-2013

Figure 3.2는 안타의 종류 중에서 유의수준 5%에서 연도 변화에 따른 유의성을 보인 타수당 2루타, 3루타, 홈런의 변화에 대한 선그래프이다.

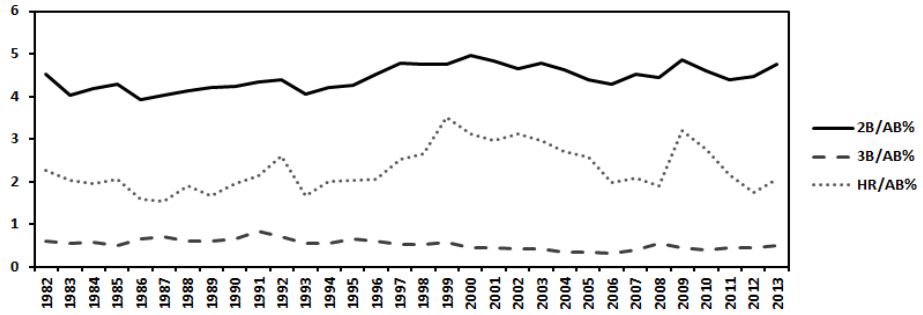

Figure 3.2 2B, 3B and HR per at bats, 1982-2013

타수당 2루타는 연도의 값이 증가할수록 증가 추세를 가지며, 타수당 3루타는 연도의 값이 증가할수록 감소 추세에 있다고 할 수 있다. 반면 타수당 홈런은 유명 홈런타자들이 많이 있었던 시기인 1999년 근방과 2009년 근방에서 피크가 나타나지만 전반적으로 뚜렷한 추세는 보이지 않는다. Figure 3.3는 출루율, 장타율 및 OPS에 대한 연도별 변화를 보여준다. 출루율은 변화가 크진 않지만 점진적으로 저점을 높여가며, 장타율과 OPS도 증가추세에 있지만 Figure 3.1과 Figure 3.2에서 알 수 있듯이 타점, 홈런이 많은 해가 역시 값이 큰 것을 확인할 수 있다.

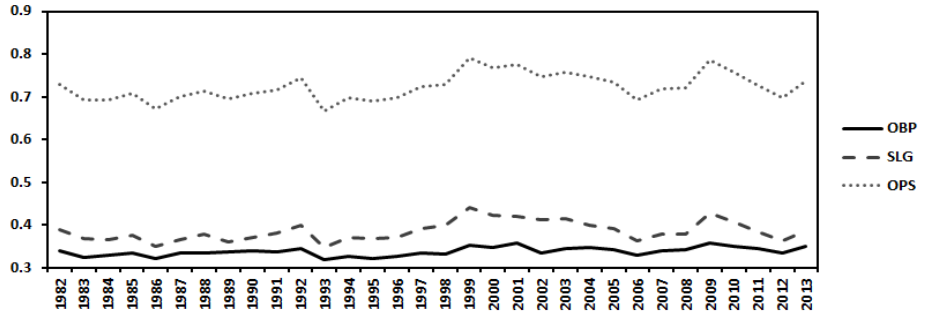

Figure 3.3 OBP, SLG and OPS, 1982-2013

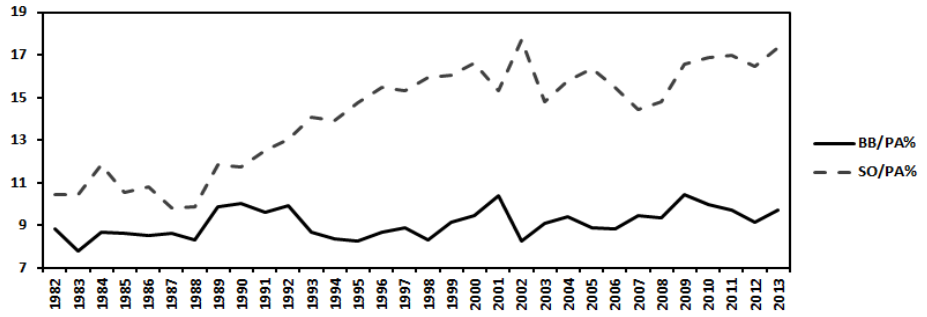

Figure 3.4 Walks and strike outs per plate appearance, 1982-2013

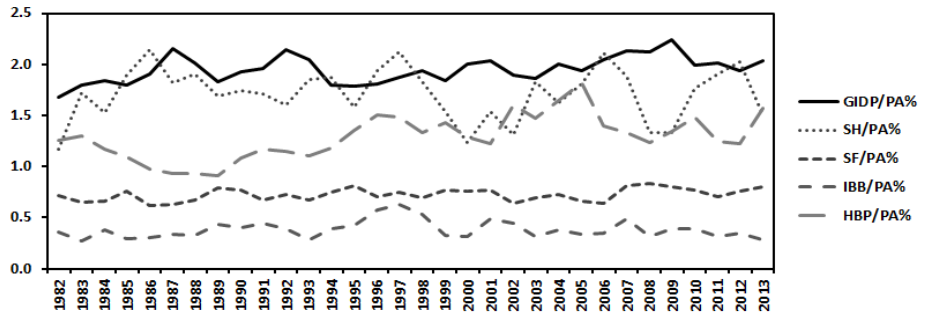

Figure 3.5 GIDP, SH, SF, IBB, and HBP per plate appearance, 1982-2013

Figure 3.4는 타석당 4구와 삼진 횟수에 대한 연도별 변화를 보여준다. 세월이 흐름에 따라 4구는 점진적으로 증가추세에 있으며 삼진은 Table 3.1에서도 알 수 있듯이 세월이 흐름에 따라 가장 뚜렷하게 증가 추세를 보이는 통계량이다. 삼진에 대한 결과는 여러 가지 해석이 가능하겠지만 타자들이 과거보다 현재에 이를수록 타석에서 좀 더 적극성을 보이는 영향으로 간주된다. Figure 3.5는 타석당 병살타, 희타, 희비, 고의사구와 사구 횟수에 대한 연도별 변화를 보여준다. 그래프에서 병살타, 희타, 사구, 희

비, 고의사구의 순서로 대략 빈번하게 발생되는 것을 알 수 있으며, 연도에 따라 점진적으로 값이 증가하는 것은 사구, 병살타, 희비의 횟수이다.

3.3. 투수 성적

Table 3.2은 6개의 투수 통계량과 연도의 상관계수와 p-값을 보여주는데, 유의수준 1%에서 모두 유의하게 나타났다. 상관계수 해석만으로는 연도에 따른 상관이 가장 큰 변수는 완투로 세월이 흐름에 따라 눈에 뜨게 격감하며 완봉, 보크도 감소한다. 반면 폭투는 눈에 뜨게 증가하며 방어율, WHIP도 증가추세라고 할 수 있다.

Table 3.2 Correlation between pitching statistics and year with p-value

ERA (.601)** (.000)	CG/G (-.924)** (.000)	SHO/G (-.575)** (.001)	WP/G (.899)**(.000)
BK/G (-.498)** (.004)	WHIP (.553)** (.001)		

$*p < 0.05$, $**p < 0.01$

3.4. 투수 성적의 변화

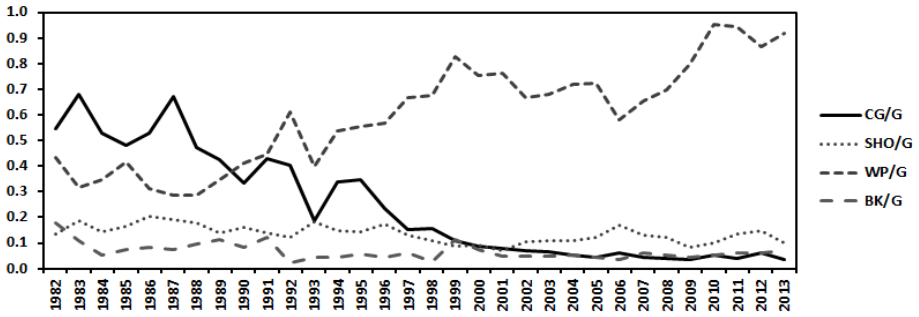

Figure 3.6 CG, SHO, WP and BK per game, 1982-2013

Figure 3.6은 게임당 완투, 완봉, 폭투, 보크 횟수의 연도별 변화를 보여준다. 프로야구 초창기엔 완투와 완봉의 수가 많았으며 또한 완투의 수가 완봉의 수보다 많았으나 2000년대에 들어서면서 완투와 완봉의 수가 줄었으며 2002년부터는 완투의 수보다 완봉의 수가 많아지는 현상이 생기고 있다. 한편 보크의 횟수는 세월이 흐를수록 투수들이 보크에 대한 대비를 철저히 하는 이유로 현저하게 줄었으며, 반면 폭투는 눈에 뜨게 증가했다. Figure 3.7은 방어율과 WHIP의 연도별 변화를 보여준다. 2개의 통계량 모두 값이 작은 것이 우수한 투수임을 설명하지만 세월이 흐를수록 점점 더 증가하는 것을 알 수 있다. 기량이 우수한 외국인 타자가 많아지고, 또 우수한 투수는 외국으로 진출하는 등 여러 이유가 있지만 우수한 투수들이 드물고 투수기술의 발달이 늦었던 것이 가장 큰 이유라고 간주된다.

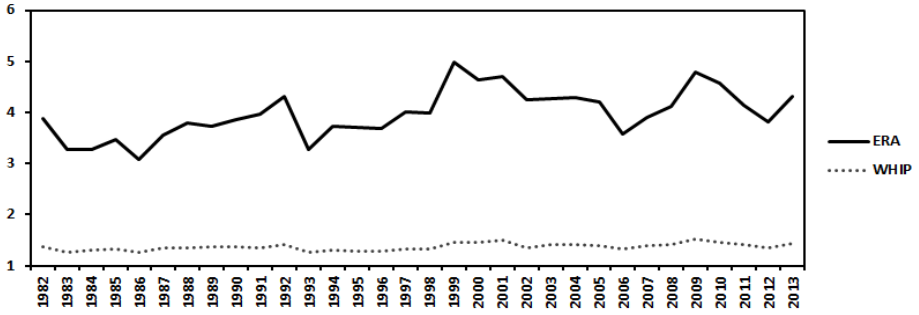

Figure 3.7 ERA and WHIP, 1982-2013

3.5. 타격 성적 간의 상관관계

Table 3.3 Correlation among statistics to determine the highest

Sign	Correlation coefficient ρ	Value
+	2B/AB with HR/AB	+0.815
	HBP/PA with SO/PA	+0.740
	2B/AB with SO/PA	+0.712
	HBP/PA with 2B/AB	+0.594
	HBP/PA with HR/AB	+0.575
-	3B/AB with HBP/PA	-0.620
	3B/AB with SO/PA	-0.618
	1B/AB with IBB/PA	-0.533
	3B/AB with 2B/AB	-0.473
	HR/AB with SH/PA	-0.460

*All ρ are significant at the 1% level.

타자가 타석에 들어서서 생기는 중요야구기록은 대략 11개로 1루타 (1B), 2루타 (2B), 3루타 (3B), 홈런 (HR), 희타 (SH), 희비 (SF), 4구 (BB), 고의4구 (IBB), 사구 (HBP), 삼진 (SO), 병살타 (GIDP)이다. 이와 같은 통계량들 사이에는 어떤 연관성이 있을까? Table 3.3은 통계량 사이의 상관관계가 큰 경우의 상위 5위까지 보여준다. 먼저 양의 상관관계가 가장 큰 1위부터 3위까지는 타수당 2루타와 타수당 홈런, 타석당 사구와 타석당 삼진, 타수당 2루타와 타석당 삼진이며, 음의 상관관계가 가장 큰 1위부터 3위까지는 타수당 3루타와 타석당 사구, 타수당 3루타와 타석당 삼진, 타수당 1루타와 타석당 고의4구이다. 나열된 변수의 조합들은 단지 두 변수간의 연관된 정도를 나타낼 뿐 인과관계를 설명하는 것은 아니다.

4. 시계열분석

4.1. 단위근 검정

Table 4.1 ADF unit root test results for baseball statistics with p-value

Statistics	Level			First differences		
	Test statistic		p-values	Test statistic		p-values
2B/AB	-3.177	(0)	0.1074	-6.677	(0)**	0.0000
3B/AB	-2.925	(0)	0.1688	-6.185	(0)**	0.0001
RBI/G	-3.699	(0)*	0.0375	-6.934	(0)**	0.0000
CS/G	3.062	(0)	0.1327	-7.806	(0)**	0.0000
E/G	-6.585	(0)**	0.0000	-7.403	(0)**	0.0000
BB/PA	-4.041	(0)*	0.0176	-7.831	(0)**	0.0000
SO/PA	-1.651	(1)	0.7477	-8.421	(0)**	0.0000
GIDP/PA	-3.687	(0)*	0.0384	-5.915	(0)***	0.0002
HBP/PA	-2.882	(0)	0.1815	-3.991	(0)*	0.0215
OBP	-4.600	(0)**	0.0047	-8.591	(0)**	0.0000
OPS	-3.620	(0)*	0.0443	-7.328	(0)**	0.0000
ERA	-3.664	(0)*	0.0404	-7.143	(0)**	0.0000
CG/G	0.113	(3)	0.9958	-6.171	(3)**	0.0001
SHO/G	-3.708	(0)*	0.0367	-7.224	(0)**	0.0000
WP/G	-3.405	(0)	0.0690	-7.193	(0)**	0.0000
BK/G	-5.621	(0)**	0.0004	-5.203	(0)**	0.0019
WHIP	-4.201	(0)*	0.0122	-8.053	(0)**	0.0000

$*p < 0.05,\ **p < 0.01$

일반적으로 시계열분석에서는 분석시계열들이 안정적인 (stationary) 것으로 가정하고 진행된다. 본 연구에서는 각 변수 시계열의 안정적 과정 여부를 결정하는 많은 검정 중에서 가장 보편화되어 있다고 할 수 있는 ADF (augmented Dickey Fuller) 단위근 (unit root) 검정을 실시하였다. 단위근 검정 결과에서 단위근이 존재한다는 귀무가설을 기각하지 못하면 그 시계열은 불안정적이며, 랜덤워크 과정을 따르는 것으로 간주한다. ADF 단위근 검정은 상수와 확정적 시간 추세의 포함여부에 따라 세 가지 형태의 자료생성과정을 가정하지만 본 연구에서는 상수항과 추세항이 모두 포함된 모형을 이용하였는데, 그 이유는 언급되어진 야구 통계량의 과다로 가장 제약이 약한 모형을 고려하여 안정적인 시계열로 확실히 분류할 수 있는 통계량만 취급하였다. Table 4.1은 유의수준 1%에서 상관계수가 유의한 야구통계량 17개에 대한 ADF 단위근 검정의 결과인데, 사용되어진 검정의 유의확률은 Mackinnon (1996)의 단측검정에 대한 값이며 단위근 검정에서 시계열의 시차가 분석결과에 중요한 영향을 미치기 때문에 분석에 앞서 EViews7의 디폴트 방법인 SIC (Schwarz information criterion) 통계량을 기준으로 시차를 결정하였으며 시차선정은 시차를 늘려가면서 SIC 통계량이 가장 적은 값을 선택하였다. 단위근 검정에 사용된 시차의 길이 (lag length)는 괄호 안에 표시되어있다. 검정의 결과, 원 시계열이 유의수준 5%에서 안정적인 시계열인 경우는 RBI/G, BB/PA, GIDP/PA, OPS, ERA, SHO/G, WHIP이며 유의수준 1%에서 안정적인 시계열인 경우는 E/G, OBP, BK/G와 같았다. 한편 고려된 변수들은 1차 차분하면 유의수준 5%에서는 모든 시계열이 안정적이며 유의수준 1%에서도 HBP/PA를 제외한 모든 변수가 안정적인 시계열로 나타났다.

4.2. 일변량 시계열모형

야구통계량 시계열에 적용될 수 있는 적절한 ARIMA 모형을 찾기 위하여 모형의 식별과 추정 및 검진과 같은 3단계에 걸친 분석을 실시하였다. 먼저 비정상 시계열은 1차 차분을 통해 정상 시계열로 변환하였으며, 모수를 절약하기 위해 AR(p) 모형을 우선적으로 추정하여 생성된 모형이 타당하면 그 모형을 채택하고 타당하지 않으면 MA(q) 계수들을 추가시켜 ARIMA(p,q) 모형을 구축하는 방법을 사용하였다. ARIMA(p,q)의 차수는 SPSS를 사용하여 시계열의 ACF와 PACF 도표, 정밀도의 측정도구인 평균제곱근오차 (root mean square error; RMSE), ARIMA 모형을 객관적으로 식별하기 위하여 정규화된 BIC (normalized Bayesian information criterion) 방법을 사용하였다.

Table 4.2 Results of fitting models for baseball statistics

Measure	Final Model	Model Fit statistics		Ljung-BoxSig.	Shapiro-WilkSig.
		RMSE	Normalized BIC		
BB/PA	ARIMA(1,1,0)	0.706	-0.475	0.736	0.939
CS/G	ARIMA(1,1,0)	0.114	-4.123	0.739	0.545
OBP	ARIMA(1,1,0)	0.010	-8.989	0.428	0.363
RBI/G	ARIMA(1,0,0)	0.718	-0.445	0.676	0.186
SO/PA	ARIMA(1,1,0)	0.977	+0.175	0.928	0.860
WHIP	ARIMA(1,1,0)	0.064	-5.263	0.276	0.509

본 연구에서는 원 시계열이 유의수준 5%에서 안정적인 야구통계량에는 4가지 모형 ARIMA(1,0,0), ARIMA(2,0,0), ARIMA(1,1,0), ARIMA(2,1,0)을 우선 고려하였으며, 1차 차분 시계열만이 안정적이면 ARIMA(1,1,0)와 ARIMA(2,1,0)을 고려하였다. 고려된 여러 가지 모형 중에서 모수의 추정이 유의하고, 적합도가 가장 높고 잔차분석에서 문제가 없는 적절한 모형을 최종 선택하였는데, 백색잡음 (white noise)의 정규성은 Shapiro-Wilk 검정, 독립성은 Ljung-Box의 Q-통계량을 이용하여 타당성 여부를 판단하였다. Table 4.2는 야구통계량들에 대한 시계열분석 결과를 보여주는 데, 언급되지 않은 통계량은 특별한 시계열 특성이 없는 것으로 나타났다. 제시된 모형은 모두 유의수준 5%에서 추정된 계

수가 유의하며 잔차분석을 통과한 모형들이다. 결과적으로 시계열적 특성을 갖는 통계량들은 대부분 가장 간단한 AR(1) 모형이면 충분하다는 결론을 내렸으며, AR(2)는 대부분 추정계수가 유의하지 않은 것으로 나타났다. Table 4.3은 AR 모형을 이용하여 추정된 여러 가지 야구통계량들에 대한 예측 모형식을 정리한 결과이며, Figure 4.1은 Table 4.3의 여러 가지 야구통계량에 대한 관측값과 예측값을 그림으로 나타낸 결과이다. 그림에서 실선이 관측값, 점선이 예측값인데 통계량에 따라 다소 이격이 있지만 전반적인 추세는 비교적 반영이 잘 되고 있다고 할 수 있다.

Table 4.3 Estimated model equations for baseball measures

Measure	Final Model	Model Equation
BB/PA	ARIMA(1,1,0)	$Z_t = 0.725 + 1.039Z_t - 1 - 0.039Z_t - 2$
CS/G	ARIMA(1,1,0)	$Z_t = 0.405 + 1.046Z_t - 1 - 0.046Z_t - 2$
OBP	ARIMA(1,1,0)	$Z_t = 0.803 + 1.008Z_t - 1 - 0.008Z_t - 2$
RBI/G	ARIMA(1,0,0)	$Z_t = 0.829 + 0.542Z_t - 1$
SO/PA	ARIMA(1,1,0)	$Z_t = 0.087 + 1.014Z_t - 1 - 0.014Z_t - 2$
WHIP	ARIMA(1,1,0)	$Z_t = 0.802 + 1.022Z_t - 1 - 0.022Z_t - 2$

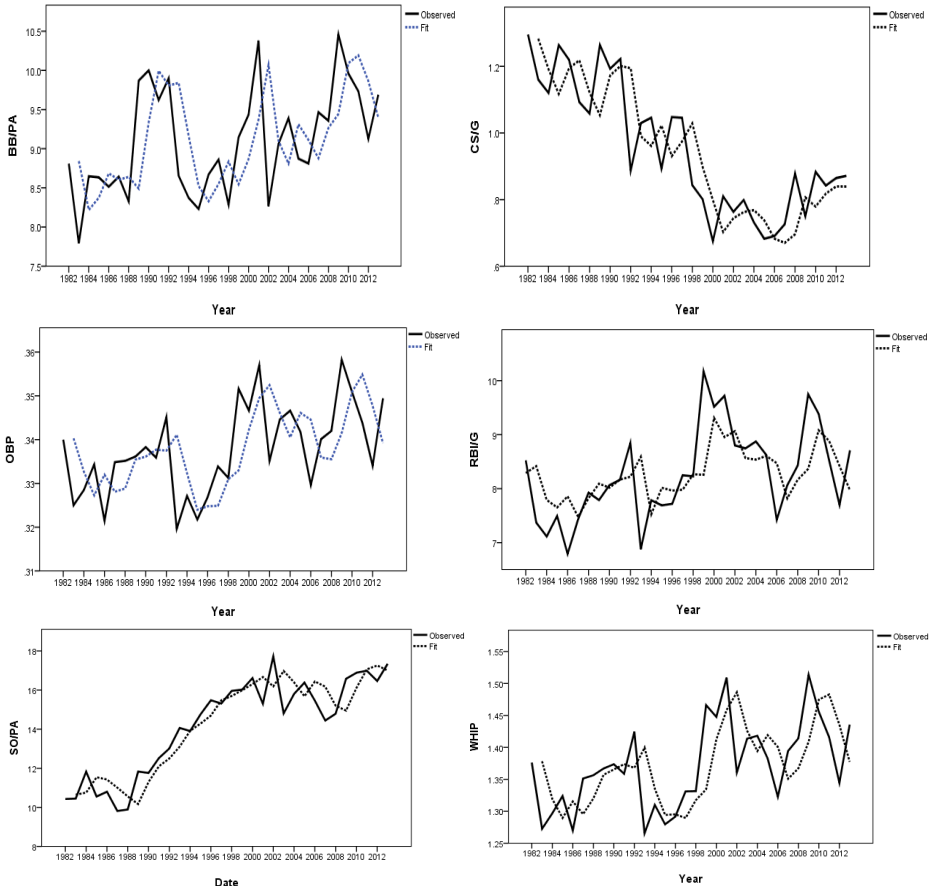

Figure 4.1 Time series plot for baseball measures with predicted values

5. 결론

야구의 기록은 야구를 더욱 풍성하고 흥미롭게 만든다. 이런 연유로 모든 야구 결과가 수치화되고 팬들은 선수들의 출중한 기록에 열광한다. 박찬호의 메이저리그 18승, 이승엽의 56호 홈런 등은 야구팬들이라면 누구나 절대 잊을 수 없는 기억들이다. 본 연구에서는 1982년 프로야구가 탄생한 이후 지속적인 성장을 거듭하던 한국프로야구에 대한 연도별 기록변화를 살펴보았다. 많은 기록들이 연도별로 변화를 보이는데, 정리하면 타격 성적에 대해서 도루의 횟수는 연도별로 큰 변화를 보이지 않으나 도루 실패의 횟수는 현저히 줄고 있다. 타격 내용에 대해서는 타율과 홈런 개수는 점진적 증가추세이지만 큰 폭의 상승이라고 하기에는 힘들며 1루타는 과거와 현재가 비슷한 수준이며 2루타 및 3루타는 눈에 띌만한 증가 및 감소추세에 있다. 득점과 타점도 점진적 증가추세이며 그밖에 고의4구 및 희타는 큰 변화가 없다. 가장 현저하게 변화를 보이는 타격 기록들은 삼진개수의 증가와 실책의 감소를 들 수 있다. 한편 투수 기록 중에서 눈에 띄게 연도별로 변화가 심한 것은 완투 및 완봉의 감소이다. 사실 미국 메이저리그와 일본프로야구를 보더라도 완투 및 완봉 수의 감소가 눈에 띄는데, 그 이유는 수준이 높은 외국인 타자 가세에 따른 타고투저 영향과 투수들의 분업화가 뚜렷해졌다는 것이 가장 큰 원인으로 간주된다. 방어율과 WHIP도 점점 높아져 가는데, 외국인 타자들의 등장, 프로야구 팀 수의 증가와 한국프로야구 대표 투수들이 연이어 해외로 빠져나간 영향으로 간주되며 반면 보크에 대한 투수들의 준비가 잘 된 탓으로 현저하게 보크 수는 줄었지만 스피드보다 코너워크를 주로 구사하는 투수들이 많아서 현저하게 폭투가 증가하는 것으로 간주된다.

References

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 (2009-2014). *2009-2014 official baseball guide*, Korea Baseball Organization, Seoul.

Korea Baseball Organization (2009). *The official baseball encyclopedia 2009 (1982-2008)*, Korea Baseball Organization, Seoul.

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

Lee, J. T. (2014b). Estimation of exponent value for Pythagorean method in Korean pro-baseball. *Journal of the Korean Data & Information Science Society*, **25**, 493-499.

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.

MacKinnon, J. G. (1996). Numerical distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics*, **11**, 601-618.

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
2015, **26**(1), 1–10

http://dx.doi.org/10.7465/jkdi.2015.26.1.1
한국데이터정보과학회지

Long term trends in the Korean professional baseball

Jang Taek Lee[1]

[1]Department of Applied Statistics, Dankook University

Received 16 August 2014, revised 2 October 2014, accepted 29 October 2014

Abstract

This paper offers some long term perspective on what has been happening to some baseball statistics for Korean professional baseball. The data used are league summaries by year over the period 1982-2013. For the baseball statistics, statistically significant positive correlations ($p < 0.01$) were found for doubles (2B), runs batted in (RBI), bases on balls (BB), strike outs (SO), grounded into double play (GIDP), hit by pitch (HBP), on base percentage (OBP), OPS, earned run average (ERA), wild pitches (WP) and walks plus hits divided by innings pitched (WHIP) increased with year. There was a statistically significant decreasing trend in the correlations for triples (3B), caught stealing (CS), errors (E), completed games (CG), shutouts (SHO) and balks (BK) with year (trend $p < 0.01$). The ARIMA model of Box-Jenkins is applied to find a model to forecast future baseball measures. Univariate time series results suggest that simple lag-1 models fit some baseball measures quite well. In conclusion, the single most important change in Korean professional baseball is the overall incidence of completed games (CG) downward. Also the decrease of strike outs (SO) is very remarkable.

Keywords: ARIMA model, completed games, correlations, Korean professional baseball, strike outs.

[1] Professor, Department of Applied Statistics, Dankook University, Yongin 448-701, Korea.
E-mail: jtlee@dankook.ac.kr