한국통계학회논문집
2012, 19권, 3호, 367–379

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.3.367

주성분회귀분석을 이용한 한국프로야구 순위

배재영[a], 이진목[a], 이제영[1,a]

[a]영남대학교 통계학과

요 약

야구경기에서 순위를 예측하는 것은 야구팬들에게 관심의 대상이 된다. 이러한 순위를 예측하기 위해서 2011년 한국프로야구 기록 자료를 바탕으로 산술평균방법, 가중평균방법, 주성분분석방법, 주성분회귀분석방법을 제시한다. 표준화를 통한 산술평균, 상관계수를 이용한 가중평균과 주성분 분석을 이용해서 순위를 예측하고, 최종모형으로 주성분회귀분석 모형이 선택되었다. 주성분 분석으로 축약된 변수를 이용해서 회귀분석을 실시하여, 투수부분, 타자부분, 투수와 타자부분의 순위예측 모형을 제안한다. 예측된 회귀모형을 통해서 2012년도 순위 예측이 가능하다.

주요용어: 산술평균, 가중평균, 주성분분석, 주성분회귀분석, 다중공선성.

1. 서론

한국프로야구는 일본이나 미국에 비해서 역사가 깊지 않지만, 최근 국제대회에서 좋은 성적을 내면서 세계적으로 한국야구에 대한 관심과 위상이 많이 높아졌다. 이런 한국프로야구가 출범 30년 동안 2011년에 관중 수가 600만을 돌파했다. 남녀노소 누구나 쉽게 즐길 수 있는 스포츠로 변화된 한국프로야구는 30년 동안 무수히 많은 명예로운 기록과 불명예스러운 기록이 수립되었다. 1982년도 OB팀의 첫 우승과 2011년 삼성 우승까지 수많은 경기를 치루면서, 한국프로야구는 많은 성장을 이뤘으며, 스타플레이어를 배출하면서 관중을 야구장으로 끌어 들이고 있다. 매년 프로야구가 개막을 하면, 전문가와 비전문가들은 항상 올해 우승팀을 예측한다. 이러한 예측이 관중들을 야구장으로 더욱 많이 찾아오게 한다. 팬들이 응원하는 팀이 올해 1등을 할 가능성이 높다는 평가가 나오면, 그 연고지의 팬들은 기대감을 가지고 야구장을 찾게 된다. 예를 들어 2011년도 시즌 종료 후 순위를 보면, 한화와 넥센은 하위권에 머물러 있었다. 하지만 시즌이 끝나고 선수보강을 많이 하여서 올해 2012년도 시즌에 상위권 도약이 기대되는 팀이다. 2012년도 시즌에는 한화와 넥센으로 많은 팬들이 집중을 하고, 관심을 가지게 될 것이다. 즉, 2012년도 시즌이 기다려지게 만드는 이유인 것이다.

스포츠에 관한 연구는 통계학자와 체육학자들 사이에서 활발히 연구되어 많은 자료들을 찾을 수 있다. 예를 들어 김응준 등 (2011)과 박철용과 이미숙 (2011)과 민대기 (2011)와 같이 많은 연구가 되었다. 체육학자들은 김응식 (2001)과 이장영과 강효민 (2001)과 같이 야구에 대한 연구가 많이 이루어졌다. 순위를 정하는데 있어서 승패를 이용을 하는데, 승패에 관한 연구는 김혁주와 이현정 (2011) 등의 연구가 이루어졌다. 하지만 투수부분과 타자부분에서 순위와 관련된 연구는 미비한 실정이다.

2011년도 한국프로야구 정규시즌의 순위는 삼성, 롯데, SK, KIA, 두산, LG, 한화, 넥센 순이었다. 본 논문에서는 2012년도 순위를 예측하기 위하여 한국야구위원회(www.koreabase ball.com)에 있는 총 37개의 변수를 이용한다. 투수부분에는 방어율, 승, 홀드, 세이브, 삼진, 피홈런 등 총 17개의 변수를 이용하고, 타자부분에는 타율, 타점, 득점, 안타, 2루타, 3루타, 홈런 등 20개의 변수를 이용한다. 이 변

[1] 교신저자: (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 교수. E-mail: jlee@yu.ac.kr

수들을 표준화하여서 산술평균을 구해서 순위를 예측하고, 상관계수를 통해서 가중평균순위를 예측한다. 하지만 산술평균은 유사한 능력이 있을 시, 동일한 가중치($1/n$)를 주는 문제가 있으며, 가중평균은 많은 변수의 상관계수가 복잡하여서, 상관계수의 크기만으로 분류하는 것은 쉽지 않다. 그래서 얻어진 변수들을 주성분분석을 통해 순위를 예측한다. 또한 주성분분석을 통해 얻어진 주성분변수를 독립변수로 선택을 하고, 주성분회귀분석을 실시하여서 순위를 예측한다. 그리고 정확도가 높은 주성분회귀분석을 최종모형으로 선택한다. 따라서 본 논문에서는 2011년도 각 팀이 낸 성적을 바탕으로 투수부분과 공격부분에 중요하게 영향을 미치는 요인들을 찾고, 그 요인들을 이용해서 순위를 예측하는 모형을 만들고자 한다. 타자부분에서는 타율순위를 이용을 하고, 투수부분에서는 방어율 순위, 정규시즌 순위는 승률을 이용해서 모형을 만든다.

2절에서는 주성분 분석에 대한 설명을 하고, 3절에서는 분석하기 위해 수집된 자료를 설명을 한 뒤, 4절에서 산술평균방법과 가중평균방법, 주성분분석방법, 주성분회귀분석방법으로 나누어서 분석을 한다. 5절에서는 결론을 맺는다.

2. 주성분회귀분석

2011년 한국프로야구 순위를 예측하기 위해서 한국야구위원회에서 가져온 데이터인 투수부분 17개, 타자부분 20개로 총 37개의 변수를 이용해서 산술평균방법과 가중평균방법, 주성분 분석방법을 이용하였다. 총 37개 변수를 표준화하여 산술평균을 구한 뒤, 각 8팀의 순위를 측정하였다. 하지만 모든 변수들이 동일한 가중치($1/n$)로 반영이 되었으므로, 투수부분에 방어율($X1$)과 실점($X16$), 자책점수($X17$) 등과 타자부분의 타율($X18$)과 1루타($X22$), 출루율($X37$)과 같은 유사한 능력을 측정하는 경우 이 부분의 능력이 높은 팀이 높은 점수를 받을 것이다. 이러한 단점을 보안하기 위해서 모든 변수의 상관계수를 이용해서 구한 가중평균으로 순위를 측정하였다. 상관계수가 높은 것끼리 그룹으로 묶은 후, 각 다른 가중치를 부여함으로 써 순위를 예측할 수 있다. 또한 37개의 변수를 모두 사용하여 다중회귀분석을 하는 경우 설명변수들 사이의 높은 상관관계에 의해 다중공선성(multicollinearty) 문제를 야기시킬 수 있다 (권세혁, 2008). 이러한 다중공선성 문제를 해결하기 위해서, 본 논문에서는 주요 분석방법으로 주성분분석을 통해 주성분변수를 얻어 이를 설명변수로 이용함으로써 다중공선성 문제를 해결하였다 (오경주 등, 2012).

그리고 주성분의 개수를 선택할 때, 상관계수행렬을 이용할 시 일반적으로 고유치 값이 1이상인 주성분과 총 변동의 설명력이 80%이상인 주성분 변수를 선택할 수 있다. 성분 부하 값이 크다는 것은 그에 대응하는 원 변수의 영향이 크다는 것을 의미하므로 성분 부하 값이 큰 변수를 파악하여 주성분의 이름을 부여하면 된다. 주성분 이름을 부여한 뒤, 주성분 점수를 구하게 되는데, 다음의 식을 이용하게 된다.

$$y_{rj} = \underline{e}_j' \frac{\left(\underline{x}_r - \underline{\mu}\right)}{\underline{\sigma}} \tag{2.1}$$

y_{rj}는 r번째 개체의 j번째 주성분 점수를 뜻하며 $\underline{e}_j'$는 j번째 주성분의 고유벡터를 뜻하며, $\underline{x}_r$는 r번째 개체의 측정치 벡터를 나타낸다. $\underline{\mu}$는 모평균벡터이고, $\underline{\sigma}$는 모표준편차벡터를 나타낸다. SAS에서 OUT 옵션을 이용하면 주성분 점수 데이터를 얻을 수 있다. 주성분 분석을 통해 다중공선성 문제를 해결 할 수 있다. 주성분 변수가 회귀분석의 설명변수 측정치가 되는 회귀모형은 아래와 같다 (성웅현, 1998).

$$y_i = \beta_0 + \beta_1 \mathrm{Prin}_1 + \beta_2 \mathrm{Prin}_2 + \cdots + \beta_p \mathrm{Prin}_p + \epsilon_i, \quad i = 1, \ldots, k. \tag{2.2}$$

표 1: 2011년 한국프로야구 변수설명

투수 변수(X1~X17)		타자 변수(X18~X37)	
방어율(X1)	한 게임 동안 준 점수의 평균율	타율(X18)	한 게임 동안 친 안타의 평균율
승(X2)	이긴 경기 경우	타석(X19)	타자가 타석에 선 횟수
홀드(X3)	중간에서 점수를 지키는 경우	타수(X20)	희생타, 4구를 제외한 타석
세이브(X4)	2점차이내 점수 차를 지키는 경우	득점(X21)	안타시 홈을 밟는 경우
총이닝(X5)	투수들이 총 던진 이닝	1루타(X22)	타격시 1루까지 가는 경우
상대타자(X6)	투수가 상대한 타자수	2루타(X23)	타격시 2루까지 가는 경우
투구수(X7)	투수들이 던진 공의 수	3루타(X24)	타격시 3루까지 가는 경우
피안타(X8)	투수가 맞은 안타의 수	홈런(X25)	타격시 홈까지 가는 경우
피홈런(X9)	투수가 맞은 홈런의 수	총루타(X26)	베이스를 밟은 수
희생타(X10)	타자아웃시키고, 실점하는 경우	타점(X27)	안타시 주자가 홈을 밟는 경우
4사구(X11)	타자를 걸어서 1루로 보내는 경우	도루(X28)	주자가 다음 베이스로 가는 것
고의사구(X12)	작전상 볼넷을 주는 경우	도루실패(X29)	도루시 아웃되는 경우
탈삼진(X13)	스트라이크로 아웃 잡는 경우	희생타(X30)	자신은 아웃, 점수를 내는 경우
폭투(X14)	투수 실책	4사구(X31)	타자가 볼넷으로 걸어가는 경우
보크(X15)	투수 반칙	고의사구(X32)	일부러 볼넷을 주는 경우
실점(X16)	그 팀이 준 모든 점수	삼진(X33)	스트라이크 3번으로 아웃되는 경우
자책점수(X17)	투수에 의해 준 점수	병살(X34)	한번 타격으로 2아웃 되는 경우
		실책(X35)	수비시 실수로 주자를 보내는 경우
		장타율(X36)	타격시 멀리 가는 평균율
		출루율(X37)	타자가 베이스로 나가는 평균율

본 논문에서 k는 팀의 수가 되기 때문에 8이 된다. $\text{Prin}_1, \text{Prin}_2, \ldots, \text{Prin}_p$는 주성분 변수가 되고, $\beta_0, \beta_1, \ldots, \beta_p$는 회귀계수 추정치이며, ϵ_i는 평균벡터가 0, 공분산행렬이 $\text{cov}(\epsilon) = \sigma^2 I$인 확률오차벡터이다. 식 (2.2)에서 추정된 y값을 통해서 각 팀의 순위를 예측한다. 투수부분의 방어율과 타자부분의 타율, 정규시즌 순위의 승률 값을 추정을 통해서, 각 팀의 추정 값을 찾을 수 있다. 이 추정 값을 통해서 각 부분의 순위예측을 할 수 있다. 이 주성분회귀분석을 통해서 3절에서 2011년도 순위를 예측하였다.

3. 데이터 소개

본 연구는 2011년 한국프로야구의 순위예측에 관한 연구를 하기 위하여 2011년 4월 2일부터 2011년 10월 19일까지 한국프로야구의 각 8팀별 133경기를 종합한 기록 데이터를 바탕으로 순위 예측을 한다. 데이터는 한국야구위원회 홈페이지(www.koreabase ball.com; KBO) 기록실에 게시되어있는 데이터를 텍스트파일로 코딩 후 분석 하였다. 변수는 투수관련 변수 17개와 타자 관련 변수 20개 총 37개의 변수를 이용하여 순위예측을 하였다. 표 1은 37개의 변수를 소개한 표이다. 하지만 투수부분에 방어율($X1$)과 실점($X16$), 자책점수($X17$) 등과 타자부분의 타율($X18$)과 1루타($X22$), 출루율($X37$) 등은 유사한 능력을 측정한 항목이다. 즉, 변수들 값이 다중공선성이 존재하게 된다. 그러므로 주성분 분석을 실시하여 변수축약을 한 뒤, 주성분회귀분석을 실시한다. 분석에 사용된 프로그램은 EXCEL2010 과 통계패키지 SAS 9.2를 사용하였다.

4. 분석을 통한 순위예측

투수부분과 타자부분순위를 예측하기 위해서, 4.1절에서는 변수들의 측정단위가 많이 차이 나기 때문에, 변수의 표준화 값을 이용하여 산술평균값을 구한 뒤 순위를 예측한다. 4.2절에서는 상관계수를 이용하여 그룹을 나눈 뒤 가중평균 값을 이용하여 순위를 예측한다. 4.3절에서는 주성분 분석을 통

표 2: 산술평균에 의한 순위와 2011년 순위의 비교

최종순위	투수부분 산술평균점수		KBO 투수순위	최종순위	타자부분 산술평균점수		KBO 타자순위
1	삼성	0.9379	삼성(3.35)	1	롯데	0.8106	롯데(0.288)
2	SK	0.4675	SK(3.59)	2	KIA	0.3672	두산(0.271)
3	KIA	0.4281	KIA(4.10)	3	두산	0.3421	KIA(0.269)
4	롯데	−0.2216	LG(4.15)	4	삼성	0.2075	LG(0.266)
5	LG	−0.2429	롯데(4.20)	5	SK	0.0405	SK(0.263)
6	넥센	−0.2987	두산(4.26)	6	LG	−0.1901	삼성(0.259)
7	두산	−0.4023	넥센(4.36)	7	한화	−0.6171	한화(0.255)
8	한화	−0.6679	한화(5.11)	8	넥센	−1.0207	넥센(0.245)

해 변수를 축약한 뒤 순위를 예측한다. 4.4절에서는 주성분 변수를 이용해서 주성분회귀분석을 이용해 순위를 예측한다.

4.1. 산술 평균방법을 이용한 순위측정결과

투수부문, 타자부문의 측정단위가 서로 다르고, 값의 크기도 차이가 크기 때문에 변수를 표준화 하여 분석 하였다. 표준화한 변수의 값을 산술평균(AVGTeam)으로 계산하면,

$$\mathrm{AVG}_{Team} = \frac{(Z_1 + Z_2 + Z_3 + \cdots + Z_n)}{n}, \qquad Z_i = \frac{(X_i - \mu_i)}{\sigma_i}, \quad i = 1, \ldots, n. \tag{4.1}$$

식 (4.1)과 같다. n이 투수부분에서는 17이 되고, 타자부분에서는 20이 된다. 이 값을 계산한 뒤 순위를 측정하였다. 평균값이 계산될 때 모든 측정 항목들이 동일한 가중치가 반영되므로, 유사한 능력을 측정하는 변수가 여러 개 있다면 이 분야 값이 높은 야구팀의 산술평균값이 높게 나오고, 변수의 개수가 많아지면 측정하는데 시간과 비용이 많이 든다. 경기의 승패에 영향을 많이 미치는 변수와 그렇지 않은 변수의 가중치가 같다는 문제점을 안고 있다. 표 2는 투수부문과 타자부문의 산술평균(AVGTeam) 계산한 뒤 2011년 투수순위(방어율), 타자순위(타율)와 비교한 결과표이다.

표 2에서 산술평균에 의한 투수부분 산술점수를 보면 삼성, SK, KIA, 롯데, LG, 넥센, 두산, 한화 순으로 순위가 측정이 되었지만, 실제 투수순위에서는 삼성, SK, KIA, LG, 롯데, 두산, 넥센, 한화 순서인 것을 확인 할 수 있었다. 롯데, LG, 넥센, 두산의 4팀의 순위가 바뀐 것을 확인 하였다. 산술평균에 의한 타자부분 산술점수를 보면 롯데, KIA, 두산, 삼성, SK, LG, 한화, 넥센 순으로 순위가 측정이 되었지만, 실제 순위에서는 롯데, 두산, KIA, LG, SK, 삼성, 한화, 넥센 순으로 두산, KIA, LG, 삼성 4개의 팀에서 차이를 보이고 있다. 앞에서 언급한 산술평균의 문제점 중 유사한 능력을 측정하는 변수가 여러 개 있다면 이 분야 점수가 높은 야구팀의 산술평균점수가 높게 나오는 문제점을 가지고 있다. 가령, 삼성의 경우 방어율($X1$), 실점($X16$), 자책점수($X17$)가 다른 팀보다 월등히 좋기 때문에 1위를 하는데 많은 영향을 주었으며, 롯데 역시 타율($X18$)과 안타부분($X22$~$X25$)에서 점수가 높기 때문에 1위에 많은 영향을 주었다. 이러한 문제점을 보완하기 위하여 가중평균을 이용한 분석을 4.2절에서 언급하기로 한다.

4.2. 가중 평균방법을 이용한 순위측정결과

투수부문과 타자부문의 변수에 가중치를 부여할 때, 주관적으로 가중치를 부여하는 방법보다 객관적인 방법으로 상관계수가 높은 변수를 그룹화 하여 가중치를 설정하였다. 산술평균에서의 문제점이었던, 방어율($X1$), 실점($X16$), 자책점수($X17$)와 같은 유사한 능력을 지닌 항목을 상관계수를 통해서 그룹화 하였다. 표 3과 표 4는 투수부분 변수 17개와 타자부분 변수 20개의 상관계수 일부분을 나타낸 표이다.

표 3: 투수부분 상관계수 일부분

	$X1$(방어율)	$X2$(승)	$X3$(홀드)	$X4$(세이브)	$X5$(이닝)	$X6$(상대타자)
$X12$(고의사구)	−0.1641	0.15523	0.0176	0.2386	−0.2788	−0.1642
(p-value)	(0.6979)	(0.7136)	(0.9670)	(0.5693)	(0.5037)	(0.6979)
$X13$(삼진)	0.0148	0.1155	−0.1110	0.0499	−0.1747	0.2824
(p-value)	(0.9723)	(0.7935)	(0.7935)	(0.9066)	(0.6790)	(0.4979)
$X14$(폭투)	0.8320	0.6466	0.4130	0.6001	0.3697	0.6595
(p-value)	(0.0104)	(0.0832)	(0.3091)	(0.1157)	(0.3674)	(0.0752)
$X15$(보크)	−0.3688	−0.4111	0.2351	−0.2275	−0.2945	−0.3693
(p-value)	(0.3686)	(0.3116)	(0.5752)	(0.5879)	(0.4790)	(0.3680)
$X16$(실점)	**0.9849**	0.6762	0.6802	0.7809	0.4235	0.8161
(p-value)	(< **0.0001**)	(0.0656)	(0.0634)	(0.0222)	(0.2958)	(0.0135)
$X17$(자책점)	**0.9977**	0.6992	0.6991	0.8085	0.4926	0.7719
(p-value)	(< **0.0001**)	(0.0536)	(0.0537)	(0.0151)	(0.2150)	(0.0248)

표 4: 타자부분 상관계수 일부분

	$X18$(타율)	$X19$(타석)	$X20$(타수)	$X21$(득점)	$X22$(1루타)	$X23$(2루타)
$X18$(타율)	1	0.9161	0.8334	0.9201	0.9842	0.7627
(p-value)		(0.0014)	(0.0102)	(0.0012)	(< .0001)	(0.0278)
$X19$(타석)	**0.9161**	1	0.8608	0.8668	0.9292	0.5970
(p-value)	**(0.0014)**		(0.0061)	(0.0053)	(0.0008)	(0.1182)
$X20$(타수)	0.8334	0.8608	1	0.7297	0.9174	0.6140
(p-value)	(0.0102)	(0.0061)		(0.0399)	(0.0013)	(0.1054)
$X21$(득점)	**0.9201**	0.8668	0.7297	1	0.8948	0.7181
(p-value)	**(0.0012)**	(0.0053)	(0.0399)		(0.0027)	(0.0448)
$X22$(1루타)	**0.9842**	0.9292	0.9174	0.8948	1	0.7490
(p-value)	(< **0.0001**)	(0.0008)	(0.0013)	(0.0027)		(0.0325)
$X23$(2루타)	0.7627	0.5970	0.6140	0.7181	0.7490	1
(p-value)	(0.0278)	(0.1182)	(0.1054)	(0.0448)	(0.0325)	

표 5: 상관계수에 의한 그룹화

Group	투수부분변수	Group	타자부분 변수
선발과 불펜의 호흡력	$X1$, $X2$, $X4$, $X6$, $X7$ $X14$, $X16$, $X17$	공격능력	$X18$, $X19$, $X20$, $X21$, $X22$, $X23$, $X25$, $X26$, $X27$, $X30$, $X33$, $X34$, $X36$, $X37$
불펜력	$X3$, $X15$	타자의 체력	$X24$, $X31$, $X35$
투수의 체력	$X5$	기동력	$X28$, $X32$
투수의 악영향	$X8$, $X9$, $X10$, $X11$, $X12$	실책	$X29$
제구력	$X13$		

표 3을 보면, $X1$(방어율), $X16$(실점), $X17$(자책점) 등은 상관계수가 매우 높은 것을 확인 할 수 있다. 표 4에서는 $X18$(타율), $X19$(타석), $X21$(득점), $X22$(1루타) 등이 상관계수가 높아서 같은 그룹으로 만들 수 있다. 상관계수가 높다는 것은 유사한 능력을 가지고 있는 것이기 때문에 같은 그룹이 된다. 즉 투수부문의 17개의 변수를 5개의 그룹으로 분류하여 새로운 그룹 이름을 붙였으며, 타자부문의 20개의 변수를 4개의 그룹으로 분류하여 그룹이름을 붙였다. 표 5는 변수를 그룹화 하여 나타낸 표이다.

표 5의 결과를 보면 한국프로야구위원회에서는 투수 17개 변수, 타자 20개 변수를 측정하였는데, 실제는 투수 5개, 타자 4개의 그룹능력을 측정한 결과와 동일하다. 투수에서 Group 1은 8개, Group 4

표 6: 가중평균을 이용한 순위와 실제 투수, 타자 순위 비교

최종순위	투수부분 가중평균점수		KBO 투수순위	최종순위	타자부분 가중평균점수		KBO 타자순위
1	SK	0.7685	삼성(3.35)	1	삼성	0.1506	롯데(0.288)
2	삼성	0.6781	SK(3.59)	2	KIA	0.0964	두산(0.271)
3	LG	−0.1067	KIA(4.10)	3	두산	0.0725	KIA(0.269)
4	KIA	−0.2125	LG(4.15)	4	롯데	0.0572	LG(0.266)
5	두산	−0.2487	롯데(4.20)	5	SK	−0.0386	SK(0.263)
6	한화	−0.2879	두산(4.26)	6	LG	−0.0444	삼성(0.259)
7	넥센	−0.2940	넥센(4.36)	7	넥센	−0.1119	한화(0.255)
8	롯데	−0.2968	한화(5.11)	8	한화	−0.1819	넥센(0.245)

표 7: 투수부분과 타자부분, 투수와 타자부분 고유치와 누적설명력

투수부분			타자부분			투수와 타자부분		
	고유치	누적설명력		고유치	누적설명력		고유치	누적설명력
1	**7.7244**	**0.4544**	1	**11.5269**	**0.5763**	1	**14.7737**	**0.3993**
2	**3.3713**	**0.6527**	2	**3.1278**	**0.7327**	2	**7.8843**	**0.6124**
3	**2.0348**	**0.7724**	3	**2.2342**	**0.8444**	3	**5.0233**	**0.7481**
4	**1.8062**	**0.8786**	4	**1.2009**	**0.9045**	4	**4.0463**	**0.8575**
5	**1.1216**	**0.9446**	5	0.9588	0.9524	5	**2.6144**	**0.9282**
6	0.6643	0.9837	6	0.6182	0.9833	6	1.4589	0.9676
7	0.2773	1.0000	7	0.3332	1.0000	7	1.1991	1.0000

5개이고, 타자에서 Group1은 14개의 항목이 측정되었다. 상관계수를 이용하여 변수를 그룹화 하고 가중평균을 구하려면 이 문제를 해결할 수 있다. 가중평균의 가중치는 아래와 같이 둘 수 있다.

$$\begin{aligned}\mathrm{AVG}_{pitcher} &= [(X1+X2+X4+X6+X7+X14+X16+X17)/8+(X8+X9+X10+X11+X12)/5\\ &\quad +(X3+X15)/2+X13+X5]/5, \qquad (4.2)\\ \mathrm{AVG}_{batter} &= [(X18+X19+X20+X21+X22+X23+X25+X26+X27+X30+X33+X34\\ &\quad +X36+X37)/14+(X24+X31+X35)/3+(X28+X32)/2+X29]/4. \qquad (4.3)\end{aligned}$$

식 (4.2)와 식 (4.3)의 식을 이용하여 가중평균을 구할 수 있다. 표 6에 투수부문과 타자부문의 가중평균점수값과 2011년 투수순위(방어율), 타자순위(타율)와 비교한 결과표이다.

표 6의 가중평균에 의한 투수부분 가중점수를 보면, SK, 삼성, LG, KIA 순으로 순위가 측정이 되었지만, 실제 투수순위는 삼성, SK, KIA, LG로 차이가 나는 것을 볼 수 있었다. 가중평균에 의한 타자부분 가중점수를 보면 삼성, KIA, 두산, 롯데 순으로 순위가 측정이 되었지만, 실제 순위에서는 롯데, 두산, KIA, LG 순으로 일치하는 것이 전혀 없었다. 이것은, 37개인 측정변수가 너무 많아 변수들 간의 다중공선성 때문이라 여겨지며, 상관계수의 크기만으로 변수들을 분류하는 것은 쉽지 않을 것이다. 그래서 변수를 정량적으로 축약하는 주성분분석을 4.3절에서 활용하였다.

4.3. 주성분 분석방법을 통한 순위측정결과

원 자료의 모든 변수 37개를 이용해서 상관계수의 크기를 분류하는 것은 쉽지 않다. 이러한 문제를 해결하기 위해서 주성분 분석을 통해 변수를 축약하였다. 타자부분의 20개 변수들을 주성분 분석을 통하여 차수를 줄일 수 있었다. 원 자료에 있는 투수부분의 17개 변수와 주성분 분석을 통해서 나온 고유치와 누적 설명력을 이용해서, 그에 맞는 합당한 변수들로 축약을 할 수 있는데, 일반적으로 고유치 1이상이고 누적 설명력이 80%이상인 주성분을 선택한다. 표 7은 투수부분과 타자부분, 투수와 타자부분

표 8: 주성분 분석에 의해 얻어진 투수부분 고유벡터

	투수부분($Prin_1$)	투수실투($Prin_2$)	제구력($Prin_3$)	불펜력($Prin_4$)	투수체력($Prin_5$)
$X1$	**0.351614**	−0.096471	−0.015415	0.074337	−0.032792
$X2$	0.28118	0.25438	0.193821	0.047349	−0.157250
$X3$	0.207549	−0.243685	−0.159334	**0.460850**	0.147512
$X4$	0.285786	0.002358	−0.037697	**0.393732**	0.040178
$X5$	0.18124	0.007377	−0.251716	0.27108	**−0.626137**
$X6$	0.282122	−0.027362	0.286143	−0.247083	0.107928
$X7$	0.260311	0.30153	−0.019672	−0.287795	0.071832
$X8$	0.186802	**−0.34988**	0.368066	−0.102943	−0.008069
$X9$	0.184334	0.319309	−0.20417	−0.212443	**0.35058**
$X10$	0.005926	**0.382252**	0.371149	0.188865	0.027725
$X11$	0.201164	**0.43557**	−0.071303	0.036058	−0.090487
$X12$	−0.024423	0.348501	0.031743	0.322233	**0.580998**
$X13$	−0.002083	−0.120169	**0.646109**	0.057472	−0.142801
$X14$	**0.324843**	0.075663	−0.146059	−0.15506	0.174131
$X15$	−0.185029	−0.212479	0.159883	**0.4225**	0.143329
$X16$	**0.344522**	−0.135099	0.066100	0.038947	0.029486
$X17$	**0.352025**	−0.0982	0.006194	0.059666	0.018723

표 9: 주성분 분석에 의해 얻어진 타자부분 고유벡터

	공격력($Prin_1$)	선구안($Prin_2$)	기동력($Prin_3$)	타자실책($Prin_4$)
$X18$	**0.28745**	0.021784	−0.135005	0.050695
$X19$	**0.272632**	0.018474	0.048103	0.189205
$X20$	**0.245879**	−0.245201	−0.054388	0.068185
$X21$	**0.280882**	0.032164	0.046130	0.166105
$X22$	**0.285521**	−0.066571	0.11253	0.048167
$X23$	**0.232856**	−0.044352	0.062675	−0.243234
$X24$	0.162279	**−0.331004**	0.060843	−0.076931
$X25$	**0.237107**	0.221949	0.141352	0.191907
$X26$	**0.288284**	−0.02826	0.108469	0.031687
$X27$	**0.276851**	0.091288	0.03285	0.199605
$X28$	0.0965	−0.018268	**0.611924**	0.065543
$X29$	0.088045	0.165935	0.207543	**−0.710797**
$X30$	−0.187129	**0.290247**	−0.281723	0.256577
$X31$	0.069407	**0.396354**	0.382819	0.016194
$X32$	0.076585	−0.342543	0.438536	0.281941
$X33$	0.212279	**0.295072**	−0.124306	−0.160641
$X34$	−0.254803	0.068575	0.036688	**0.285031**
$X35$	−0.062546	**0.478220**	0.232824	0.144822
$X36$	**0.282907**	0.063471	−0.113182	0.023534
$X37$	**0.262569**	0.229418	0.041017	0.025048

원 자료를 이용해 주성분 분석을 한 결과, 선택한 주성분 변수의 고유치와 누적설명력을 나타낸 표이다.

표 7의 투수부분에서는 5개의 변수로 축약이 되었으며, 타자부분에서는 4개의 변수로 축약되었고, 투수와 타자 부분에서는 5개의 변수로 축약하였다. 투수부분에서는 5개의 주성분 변수가 94%의 누적설명력을 가지고 있으며, 타자부분의 4개의 주성분 변수가 90%의 누적설명력을 가지고 있고, 투수와 타자 부분에서는 5개의 주성분 변수가 93%의 누적설명력을 가지고 있다. 선택된 주성분 변수에 의해서 고유벡터를 표 8, 표 9, 표 10에 나타냈다. 이 고유벡터를 바탕으로 주성분 변수로 변수 축약을 할 수 있었다.

표 10: 주성분 분석에 의해 얻어진 투수와 타자부분 고유벡터

	공격력(Prin_1)	투수력(Prin_2)	제구력(Prin_3)	투수실투(Prin_4)	투수체력(Prin_5)
$X1$	0.130231	**0.295506**	−0.07007	−0.07159	0.076469
$X2$	0.192532	0.15716	0.110795	0.126071	0.195003
$X3$	−0.0276	**0.278596**	−0.25401	0.004208	0.124185
$X4$	0.060071	**0.301664**	−0.09232	0.169261	0.082124
$X5$	−0.10267	−0.09959	0.215562	−0.07163	**−0.28207**
$X6$	0.146591	0.203809	0.172584	−0.16393	0.011656
$X7$	**0.216463**	0.10318	0.100342	0.075038	−0.15657
$X8$	−0.01687	**0.238315**	0.137802	−0.31649	0.127253
$X9$	−0.000077	**0.216844**	−0.11273	−0.25529	−0.32281
$X10$	0.001286	0.021315	0.263274	**0.330533**	0.127607
$X11$	**0.205676**	0.050411	0.029452	0.290466	0.01286
$X12$	−0.04285	0.04484	0.062638	**0.428212**	−0.17882
$X13$	−0.08666	0.096423	**0.335994**	−0.07696	0.258234
$X14$	0.165417	**0.226728**	−0.04131	−0.01427	−0.17867
$X15$	−0.18815	−0.0374	−0.07698	−0.01588	**0.279346**
$X16$	0.115811	**0.30449**	−0.0243	−0.11075	0.066599
$X17$	0.126952	**0.300836**	−0.05458	−0.07664	0.057745
$X18$	**0.231011**	−0.14494	0.028232	−0.07848	0.076238
$X19$	**0.233198**	−0.09127	0.00143	0.036817	0.085502
$X20$	0.195849	−0.1367	−0.18917	−0.02266	0.078665
$X21$	**0.24183**	−0.08528	0.033579	0.06394	0.095372
$X22$	**0.228546**	−0.14894	−0.04365	−0.0649	0.074576
$X23$	**0.20892**	0.018111	−0.06826	−0.24999	0.052382
$X24$	0.115728	−0.09813	−0.27351	−0.06009	0.064676
$X25$	**0.206279**	−0.06173	0.20032	−0.04201	0.157614
$X26$	**0.236724**	−0.11575	−0.01694	−0.0949	0.098924
$X27$	**0.240408**	−0.07287	0.0791	0.050274	0.12146
$X28$	0.139306	0.181617	−0.10239	**0.275556**	−0.11446
$X29$	0.118658	0.157154	0.037447	−0.15854	−0.41496
$X30$	−0.16301	0.0643	0.267464	−0.12553	**0.239169**
$X31$	0.089034	0.072982	0.258594	0.214593	−0.20239
$X32$	0.095221	0.103584	−0.30291	0.2445	0.151499
$X33$	0.169003	−0.12838	0.216455	−0.10587	−0.15262
$X34$	−0.19238	0.192323	0.062356	0.04259	**0.210445**
$X35$	−0.0087	0.209356	0.322299	0.045242	0.023431
$X36$	**0.235642**	−0.09623	0.054328	−0.10913	0.09858
$X37$	**0.223646**	−0.09345	0.167867	0.021756	−0.04559

각각의 주성분 내에서 고유벡터 값을 큰 변수들끼리 묶은 후 이를 이용해서 주성분에 이름을 부여할 수 있다. 표 8의 투수부분에서 제 1주성분(Prin_1)의 계수의 크기에 의하면 $X1$(방어율), $X14$(폭투), $X16$(실점), $X17$(자책점) 변수의 부하 값이 크므로 제일 주성분은 투수력이라 할 수 있다. 제 2주성분(Prin_2)에서는 $X10$(희생타), $X11$(4사구), $X12$(고의사구)의 부하 값이 크므로 투수실투라 할 수 있다. 제 3주성분(Prin_3)은 $X13$(삼진)의 부하 값이 크므로 제구력 변수로 이름 지으면 된다. 제 4주성분(Prin_4)은 $X3$(세이브), $X4$(홀드), $X15$(보크)의 부하 값이 크므로 불펜력이라 이름 지을 수 있다. 표 9의 타자부분에서 제 1주성분(Prin_1)의 계수의 크기를 보면, $X18$(타율), $X19$(타석), $X20$(타수), $X21$(득점), $X22$(1루타), $X23$(2루타), $X25$(홈런), $X26$(총루타), $X27$(타점), $X36$(장타율), $X37$(출루율) 변수의 부하 값이 크므로 제 1주성분은 공격력이라 할 수 있다. 제 2주성분(Prin_2)에서는 $X24$(3루타), $X30$(희생타), $X31$(4사구), $X33$(삼진)의 부하값이 크므로 선구안이라 할 수 있다. 제 3주성분(Prin_3)은 $X28$(도

표 11: 주성분 점수에 의한 2011년 투수부분 순위

순위	팀	투수력	투수실투	제구력	불펜력	투수체력	SUM	실순위
1	삼성	4.4015	1.7173	−0.5293	1.9562	0.6889	8.2347	삼성
2	KIA	1.1387	0.8845	1.5672	−2.2499	1.3663	2.7069	SK
3	SK	2.1454	−2.4933	2.0912	0.4831	−1.2377	0.9887	KIA
4	한화	−4.8970	1.8356	1.3784	1.2241	−0.1236	−0.5825	LG
5	롯데	−0.1805	1.4181	−0.9235	−0.6264	−0.8942	−1.2065	롯데
6	LG	0.0998	−0.3693	−1.4317	−0.9377	−0.4807	−3.1197	두산
7	두산	−0.6095	−0.1447	−1.1624	−0.5329	−0.8215	−3.2711	넥센
8	넥센	−2.0983	−2.8482	−0.9897	0.6836	1.5024	−3.7505	한화

표 12: 주성분 점수에 의한 2011년 타자부분 순위

순위	팀	공격력	선구안	기동력	타자실책	SUM	실순위
1	롯데	6.1394	−1.7971	−1.1895	0.3788	3.5317	롯데
2	삼성	−0.0026	−0.3931	0.6332	0.6332	3.5295	두산
3	KIA	1.2709	3.4253	−1.3582	−1.3582	3.2231	KIA
4	두산	2.1222	−0.0286	−0.2776	−0.2776	2.7651	LG
5	SK	−1.1760	1.1121	0.9904	0.9904	−0.0943	SK
6	한화	−3.3342	0.6635	1.6477	1.6477	−1.8792	삼성
7	LG	−0.0832	−0.919	−0.6787	−0.6787	−2.4731	한화
8	넥센	−4.9366	−2.0628	−1.3356	−1.3356	−8.6028	넥센

루)의 부하 값이 크므로 기동력으로 이름 지으면 된다. 제 4주성분(Prin_4)은 $X29$(도루실패), $X34$(병살) 부하 값이 크므로 실책이라 이름 지을 수 있다. 표 10의 투수와 타자 부분에서 제 1주성분(Prin_1)의 계수의 크기를 보면, $X18$(타율), $X19$(타석), $X21$(득점), $X22$(1루타), $X23$(2루타), $X25$(홈런), $X26$(총루타), $X27$(타점), $X36$(장타율), $X37$(출루율) 변수의 부하값이 크므로 공격력이라 이름 지을 수 있으며, 제 2주성분(Prin_2)은 $X1$(방어율), $X4$(세이브), $X16$(실점), $X17$(자책점)의 부하값이 크므로 투수력이라 할 수 있다. 제 3주성분(Prin_3)은 $X13$(탈삼진)의 부하값이 크므로 제구력이라 할 수 있다.

이런 각 주성분의 이름을 정한 뒤, 주성분 점수를 구할 수 있다. 아래의 식의 각 부분은 얻어진 주성분의 점수 산출식이다.

$$\text{Pitcher}_{score} = 0.3516Z_1 + 0.2812Z_2 + \cdots + 0.3520Z_{17}, \tag{4.4}$$

$$\text{Batter}_{score} = 0.2875_{18} + 0.2726Z_{19} + \cdots + 0.2626Z_{37}, \tag{4.5}$$

$$\text{Total}_{score} = 0.0.1302Z_1 + 0.0.1925Z_2 + \cdots + 0.2236Z_{37}. \tag{4.6}$$

위 식에서 Z_i는 각 변수를 표준화한 값이며, 투수부분 17개 변수, 타자부분 20개의 변수, 총 37개의 변수를 표준화하여 주성분 점수를 구하였다. 위의 식에서 주성분 점수에 의한 투수 주성분 점수와 타자 주성분 점수, 투수와 타자의 주성분 점수의 순위를 각각 표 11, 표 12, 표 13에 나타냈다.

표 11의 주성분 분석에 의한 투수부분 순위에서는 삼성이 타 구단에 비해 월등히 높은 것을 확인 할 수 있었다. 올해 삼성이 우승을 하는데 있어서 투수부분에서 높은 점수가 원동력이 되었다. 표 12의 주성분 분석에 의한 타자부분 순위에서는 롯데가 1위를 하였는데, 롯데, 삼성, KIA가 공격력 부분에 선두권을 유지하고 있었다. 롯데가 투수력에서는 많이 부족하지만 타자부분에서 높은 점수를 얻어서 정규시즌 2위를 할 수 있었다. 표 13는 투수와 타자변수를 모두 주성분 분석을 하였을 때 2, 3, 4위 순위가 바뀌어 있는 것을 확인 할 수 있었다. 축약된 변수를 이용해서 회귀분석을 통해서 다시 순위를 측정한 것을 4.4절에서 다루도록 하겠다.

표 13: 주성분 점수에 의한 2011년 투수와 타자부분 순위

순위	팀	공격력	투수력	제구력	투수실투	투수체력	SUM	실순위
1	삼성	2.297	4.592	−1.220	3.320	0.014	9.004	삼성
2	SK	−0.692	2.860	1.336	−2.417	2.831	3.889	롯데
3	KIA	1.979	0.697	4.007	−0.971	−2.18	3.529	SK
4	롯데	5.569	−3.577	−1.522	−0.416	1.235	1.289	KIA
5	두산	1.729	−1.468	−0.740	0.070	−0.897	−1.308	두산
6	한화	−4.954	−3.186	2.068	2.740	1.047	−2.284	한화
7	LG	0.007	−0.443	−1.264	−0.910	−1.018	−3.628	LG
8	넥센	−5.934	0.554	−2.664	−1.416	−1.030	−6.564	넥센

표 14: 투수부분 적합회귀모형을 이용한 시즌 순위 추정

순위	팀	추정값(방어율)	2011년 투수순위(방어율)
1	삼성	3.3487	삼성(3.35)
2	SK	3.5766	SK(3.59)
3	KIA	4.0858	KIA(4.10)
4	LG	4.1311	LG(4.15)
5	두산	4.2520	롯데(4.20)
6	롯데	4.2543	두산(4.26)
7	넥센	4.3826	넥센(4.36)
8	한화	5.0885	한화(5.11)

4.4. 주성분분석을 이용한 회귀분석모형 결과

4.3절에서 축약된 변수를 이용해서, 회귀분석을 하여 각 부분에 대한 적합된 회귀모형을 만들고, y값을 추정하여 순위를 예측하였다.

4.4.1. 투수부문 주성분변수의 회귀분석

주성분 분석을 통해 투수부문에서 17개의 변수를 5개의 주성분변수로 축약하였고, 주성분 점수를 구하였다. 그리고 주성분 점수와 각 팀의 방어율을 이용하여 회귀분석을 하고자 한다. 투수부분 적합 회귀모형식은 식 (4.7)과 같다.

$$y = 4.14 + (-0.18489) \times \text{투수력} + (-0.03909) \times \text{불펜력} + 0.01724 \times \text{투수체력}. \tag{4.7}$$

식 (4.7)의 적합모형은 유의확률이 0.0003으로 모형이 매우 유의하며, 99.76%의 설명력을 가지고 있었다. 이 모형으로 충분히 투수순위를 설명할 수 있었다. 위의 회귀모형에 각 주성분점수를 대입 결과 표 14의 결과를 얻었다.

표 14는 적합회귀모형을 이용해서 구한 각 팀의 방어율 추정값을 나타내었다. 회귀모형을 통해서 나온 추정값과 2011년도 시즌 방어율 성적을 보면, 롯데와 두산의 순위만 차이가 있고, 다른 순위는 차이가 없는 것을 확인하였다. 투수부분 적합 회귀모형이 투수부분 순위를 예측하는데, 산술평균이나 가중평균 보다 잘 예측 하였다. 투수부분 적합 회귀모형이 투수부분 순위를 예측하는데 적합한 모형으로 선택되었다.

4.4.2. 타자부문 주성분변수의 회귀분석

주성분 분석을 통해 타자부문에서 20개의 변수를 4개의 주성분변수로 축약하였고, 주성분 점수를 구하였다. 그리고 주성분 점수와 각 팀의 타율을 이용하여 회귀분석을 실시하였다. 타자부분 적합 회

표 15: 타자부분 적합회귀모형을 이용한 시즌 순위 추정

순위	팀	추정값(타율)	2011년 타자순위(타율)
1	롯데	0.2891	롯데(0.288)
2	두산	0.2704	두산(0.271)
3	KIA	0.2684	KIA(0.269)
4	LG	0.2651	LG(0.266)
5	SK	0.2626	SK(0.263)
6	삼성	0.2592	삼성(0.259)
7	한화	0.2548	한화(0.255)
8	넥센	0.2461	넥센(0.245)

표 16: 투수와 타자부분 적합회귀모형을 이용한 시즌 순위 추정

순위	팀	추정값(승률)	2011년 팀순위(승률)
1	삼성	0.615649	삼성(0.612)
2	롯데	0.55632	롯데(0.563)
3	SK	0.544155	SK(0.546)
4	두산	0.494881	KIA(0.526)
5	KIA	0.491447	두산(0.466)
6	LG	0.470246	LG(0.450)
7	한화	0.437854	한화(0.450)
8	넥센	0.391449	넥센(0.389)

귀모형식은 식 (4.8)과 같다.

$$y = 0.2645 + 0.00364 \times \text{공격력} + (-0.00171) \times \text{기동력} + 0.00064125 \times \text{타자실책}. \quad (4.8)$$

식 (4.8)의 적합모형은 유의확률이 0.0001로 모형이 매우 유의하며, 99.62%의 설명력을 가지고 있었다. 이 모형으로 충분히 타자순위를 설명할 수 있었다. 위의 회귀모형에 각 주성분점수를 대입 결과 표 15의 결과를 얻을 수 있었다.

표 15는 적합회귀모형을 이용해서 구한 각 팀의 타율 추정값을 나타내었다. 회귀모형을 통해서 나온 추정값 순위와 2011년도 시즌 타율 성적을 보면, 정확하게 일치하는 것을 확인 할 수 있다. 타자부분 적합 회귀모형이 타자부분 순위를 예측하는데, 산술평균이나 가중평균 보다 매우 잘 예측하였다. 즉 타자부분 적합 회귀모형이 타자부분 순위를 예측하는데, 매우 우수한 모형으로 선택되었다

4.4.3. 투수와 타자부문 주성분변수의 회귀분석

주성분 분석을 통해 투수부문 변수와 타자부문 변수 37개를 5개의 주성분변수로 축약하였고, 주성분 점수를 구하였다. 그리고 주성분 점수와 각 팀의 승률을 이용하여 회귀분석을 실시하였다. 투수와 타자부분 적합 회귀모형식은 식 (4.9)와 같다.

$$y = 0.5 + 0.0143 \times \text{공격력} + 0.011 \times \text{투수력} + 0.0096 \times \text{투수실투} + 0.0162 \times \text{투수체력}. \quad (4.9)$$

식 (4.9)의 적합모형은 유의확률이 0.0348로 모형이 매우 유의하며, 93.02%의 설명력을 가지고 있었다. 이 모형으로 충분히 투수와 타자순위를 설명할 수 있었다. 위의 회귀모형에 각 주성분점수를 대입 결과 표 16의 결과를 얻을 수 있었다.

표 16은 적합회귀모형으로 구한 각 팀의 승률 추정값을 나타내었다. 회귀모형을 통해 나온 추정값과 2011년도 시즌 승률 보면 KIA와 두산의 순위가 바뀌었고, 다른 팀 순위는 완벽히 일치하는 것을 확

인 하였다. 투수와 타자부분 적합 회귀모형이 정규시즌 순위를 예측하는데, 산술평균이나 가중평균 보다 잘 예측하였다. 투수와 타자부분 적합 회귀모형이 정규시즌 순위를 예측하는데 우수한 모형으로 선택되었다. 즉, 주성분 분석에 의한 회귀분석은 산술평균이나 가중평균에 의한 순위예측 보다 잘 예측한 것을 확인 할 수 있었고, 각 부분의 적합된 회귀모형을 통해 한국프로야구 순위를 예측 할 수 있었다.

5. 결론 및 토의

본 논문은 순위에 영향을 미치는 변수 37개를 사용하여 산술평균, 가중평균, 주성분 분석, 주성분 분석에 의한 회귀분석을 하였다. 첫 번째로 변수의 측정단위가 다르기 때문에 변수를 표준화 하여 산술평균을 구하고 순위 비교하였다. 두 번째로 37개변수의 상관관계를 이용하여 9개의 그룹으로 나눈 뒤, 가중평균을 계산하여 순위를 비교하였다. 앞의 두 방법의 경우 유사한 항목이 많아서 가중치 문제와 많은 변수에 의한 다중공선성 문제에 의해서 세 번째로 주성분 분석을 실시하였다. 주성분 분석을 통해 37개의 변수를 5개의 주성분변수를 축약하고, 주성분 점수를 계산하여 순위를 비교하였다. 마지막으로 주성분 분석을 통해 축약된 변수와 승률을 이용하여 적합회귀모형을 만들고, 값을 추정하여 2011년도 한국프로야구 정규시즌순위를 비교하였다.

우리는 산술평균과 가중평균, 주성분점수, 주성분변수를 이용한 회귀모형을 적용하였을 때, 가장 정확도가 높고 효율적인 분석 방법으로 주성분변수를 이용한 회귀모형을 최종 모형으로 선택하였다. 승률에 영향을 미치는 37개의 변수를 주성분분석을 통해 5개의 주성분변수로 축약하고, 회귀분석을 하여 모형을 만들었다.

본 논문의 결과와 같이 올해 2012년 시즌 순위를 예측 하는데 이 모형을 사용하면 가능 할 것이다. 아직 2012년도에 대한 데이터가 없기 때문에, 각 팀에서는 이 모형에 쓰인 주성분 변수 4개에 대한 목표를 정하고, 적절한 훈련을 통해서 2012년도 팀 순위를 예측할 수 있을 것이다. 그리고 각 팀에서는 이 모형을 바탕으로 팀의 부족한 부분을 찾아내어 적은비용으로 효과적인 훈련을 실시하여 팀의 승률을 올릴 수 있을 것이다. 마지막으로 승률을 올리기 위해서는 여러 방법으로 올릴 수 있지만, 가능하면 적은비용으로 승률을 올리는 것이 팀에 이득이 될 것이다.

참고 문헌

권세혁 (2008). <다변량 데이터 분석과 활용>, 자유아카데미, 서울.

김응식 (2001). 한국프로야구 선수의 경기력과 연봉과의 관계, <한국스포츠사회학회지>, **14**, 15–24.

김응준, 김종규, 이남주, 이미숙 (2011). 스포츠영재들의 자아존중감 문항적합도, <한국데이터정보과학회지>, **22**, 487–494.

김혁주, 이현정 (2011). 새로운 승률 계산 방식이 2009년과 2010년의 한국프로야구에 미친 영향 및 보완할 점, <응용통계연구>, **24**, 169–175.

민대기 (2011). 2010 미국프로골프협회 자료를 활용한 경로분석을 통한 경기력의 평균타수에 미치는 영향력 비교, <한국데이터정보과학회지>, **22**, 65–71.

박철용, 이미숙 (2011). 스포츠영재성 검사 항목과 코스타스 점수간의 연관성 분석, <한국데이터정보과학회지>, **22**, 57–64.

성웅현 (1998). <응용다변량분석>, 탐진출판사, 서울.

오경주, 안재준, 심경식 (2012). 성분 분석과 로지스틱 회귀분석을 이용한 다국 통화 포트폴리오 전략, <한국데이터정보과학회지>, **23**, 151–159.

이장영, 강효민 (2001). 한국프로야구 투수의 경기수행과 연봉책정의 관계, <한국스포츠사회학회지>, **14**, 115–125.

한국야구위원회 공식 홈페이지. http://www.koreabaseball.com.

2012년 2월 20일 접수; 2012년 3월 19일 수정; 2012년 3월 22일 채택

Predicting Korea Pro-Baseball Rankings by Principal Component Regression Analysis

Jae-Young Bae[a], Jin-Mok Lee[a], Jea-Young Lee[1,a]

[a]Department of Statistics, Yeungnam University

Abstract

In baseball rankings, prediction has been a subject of interest for baseball fans. To predict these rankings, (based on 2011 data from Korea Professional Baseball records) the arithmetic mean method, the weighted average method, principal component analysis, and principal component regression analysis is presented. By standardizing the arithmetic average, the correlation coefficient using the weighted average method, using principal components analysis to predict rankings, the final model was selected as a principal component regression model. By practicing regression analysis with a reduced variable by principal component analysis, we propose a rank predictability model of a pitcher part, a batter part and a pitcher batter part. We can estimate a 2011 rank of pro-baseball by a predicted regression model. By principal component regression analysis, the pitcher part, the other part, the pitcher and the batter part of the ranking prediction model is proposed. The regression model predicts the rankings for 2012.

Keywords: Arithmetic average, weighted average, principal component regression analysis, multicollinearity.

[1] Corresponding author: Professor, Department of Statistics, Yeungnam University, Gyungsan 712-749, Korea.
Email: jlee@yu.ac.kr