제목 : 2021 KBO 타자&투수 데이터 분석

과 목 명 :	다변량통계학
담당교수님 :	김규성 교수님
학 과 :	통계학과
학 번 :	2017580002
성 명 :	권휘성
제 출 일 :	20211128

차례

1. 서론

1.1 연구목적

야구와 관련된 데이터들을 많이 발견할 수 있고 관련된 연구들도 많이 발견할 수 있다. 영화 '머니볼'에서도 알 수 있듯이 예전부터 그에 관한 연구는 계속되어왔다. 그리고 이것을 '세이버메트릭스'라고 불러오고 있다. 최근에는 야구 데이터를 향한 관심이 점점 많아지면서 각 구단마다 전력 분석팀에서 데이터 분석원들을 전문적으로 채용하여 야구 데이터들을 실제 야구 게임에 활용하는 '데이터 야구' 시대로 접어들고 있다. 비슷한 예로 전력 분석원이던 사람이 한 구단의 감독으로 일하고 있는 사례도 있다. 이렇듯 수비와 타격을 불문하고 가능한 모든 부분에서 데이터를 만들어내고 이를 이용해 경기에서 효과를 볼 수 있도록 현대 야구는 노력하고 연구하고 있다. 그리고 모든 구단이 활용한다는 점에서 한 경기 승리하는 데 효과가 있다는 것을 짐작해 볼 수 있다. 야구 데이터가 점점 고차원적으로 변함에 따라 사용되는 변수 또한, 상당히 많아지고 있다. 이러한 변수들에 대해 파악해보고 변수들 간의 밀접한 연관이 있다고 보아 관계에 대해 더욱 파악해보고 각 야구 변수들의 요인들을 분석해보고 싶어 2021 KBO 투구와 타격 데이터를 가지고 연구를 시작해 본다.

1.2 문헌연구

KBReport에서 2020년 8월 16일에 쓰인 '[썩빡꾸의 세이버메트릭스] 타자의 클러치(clutch) 능력은 연속성을 가질까?' 라는 칼럼에서는 자신의 평소 성적 대비 중요도가 높은 상황에서 얼마나 더 잘했는지를 나타내는 클러치 스코어에 대해 연구한 것을 말해주고 있다. 도루와 타율과 같은 다양한 변수들과 연관 지으면서 분석을 한 결과 클러치 스코어를 만들어내는 능력이 통계적으로 확인되지 않는다고 말하고 있다. 이렇듯 야구에서 많은 변수들을 엮어서 설명하는 것은 가능하며 변수들간에는 당연히 특정 관계가 존재할 것이고 많은 변수들을 그보다 더 낮은 차원으로 표현이 가능할 것이라고 생각하였다.

1.3 데이터 설명

먼저 2021년 시즌 KBO 타자 데이터를 KBReport에서 크롤링하여 받아왔다. WAR(대체선수 대비 승리 기여도)를 기준으로 상위 100명의 선수들의 데이터들을 이용하였다.

player	team	game	PA	AB	Hit	HR	R	RBI	BB	SO	SB	BABIP	AVG
홍창기	LG	144	651	524	172	4	103	52	109	95	23	0.393	0.328
이정후	Hero	123	544	464	167	7	78	84	62	37	10	0.373	0.36
강백호	KT	142	628	516	179	16	76	102	104	85	10	0.385	0.347
최정	SSG	134	555	436	121	35	92	100	84	102	8	0.277	0.278
양의지	NC	141	570	480	156	30	81	111	69	60	2	0.315	0.325
알테어	NC	143	565	492	134	32	83	84	57	156	20	0.331	0.272
전준우	롯데	144	619	552	192	7	88	92	53	71	6	0.384	0.348
구자욱	삼성	139	610	543	166	22	107	88	48	98	27	0.331	0.306
정은원	한화	139	608	495	140	6	85	39	105	105	19	0.346	0.283
추신수	SSG	137	580	461	122	21	84	69	103	123	25	0.315	0.265

OBP	SLG	OPS	wOBA	WAR
0.456	0.408	0.864	0.412	7.1
0.438	0.522	0.96	0.429	7.0E
0.451	0.521	0.972	0.434	6.97
0.41	0.562	0.972	0.425	6.8E
0.414	0.581	0.995	0.432	6.4E
0.358	0.514	0.872	0.384	4.8E
0.405	0.469	0.874	0.395	4.8E
0.361	0.519	0.88	0.382	4.4E
0.407	0.384	0.791	0.376	4.41
0.409	0.451	0.86	0.396	4.3

변수들에 대해 설명을 해보겠다.

player : 선수명

team : 소속팀명

game : 경기 수

PA : 타석

AB: 타수

Hit : 안타

HR : 홈런

R : 득점

RBI : 타점

BB : 볼넷&사구

SO : 삼진

SB : 도루

BABIP : 인플레이 타구 타율

AVG : 타율

OBP : 출루율

SLG : 장타율

OPS : 출루율 + 장타율

wOBA : 타석당 득점 생산력

WAR : 대체선수 대비 승리 기여도

다음으로 2021년 시즌 KBO 투수 데이터를 KBReport에서 크롤링하여 받아왔다. WAR(대체선수 대비 승리 기여도)를 기준으로 상위 100명의 선수들의 데이터들을 이용하였다.

player	team	win	lose	sv	hld	bsv	game	start	inning	soper9	bbper9	hrper9	BABIP	LOB	ERA	RA9WAR	FIP	kFIP	WAR
미란다	두산	14	5	0	0	0	28	28	173.2	11.66	3.26	0.57	0.305	81.7	2.33	8.46	2.67	2.14	8.49
루친스키	NC	15	10	0	0	0	30	30	178.2	8.92	2.77	0.6	0.298	71.6	3.17	6.61	3.32	3.05	6.37
뷰캐넌	삼성	16	5	0	0	0	30	30	177	8.24	3	0.66	0.308	75.5	3.1	6.27	3.46	3.24	5.74
고영표	KT	11	6	0	1	0	26	25	166.2	7.02	1.46	0.49	0.286	73.5	2.92	6.63	3.2	3.04	5.51
폰트	SSG	8	5	0	0	0	25	25	145.2	9.7	2.78	0.74	0.271	66.8	3.46	4.03	3.28	2.93	5.43
데스파이	KT	13	10	0	0	0	33	33	188.2	7.87	3.72	0.48	0.301	72.8	3.39	6.11	3.62	3.51	5.24
요키시	Hero	16	9	0	0	0	31	31	181.1	6.5	2.28	0.6	0.295	73.5	2.93	6.59	3.54	3.45	5.12
수아레즈	LG	10	2	0	0	0	23	22	115.1	9.83	3.2	0.31	0.294	79	2.18	5.38	2.77	2.42	4.82
킹험	한화	10	8	0	0	0	25	25	144	8.25	2.56	0.69	0.267	71.1	3.19	4.96	3.43	3.2	4.79
스트레일	롯데	10	12	0	0	0	31	31	165.2	8.96	3.64	0.65	0.328	68.9	4.07	4.26	3.62	3.41	4.78

변수들에 대해 설명을 해보겠다.

player : 선수명

team : 소속팀명

win : 승리

lose : 패배

sv : 세이브

hld : 홀드

bsv : 블론세이브

game : 출전 경기

start : 선발 경기

inning : 이닝

SOper9 : 9이닝 당 삼진

bbper9 : 9이닝 당 볼넷

hrper9 : 9이닝 당 홈런

BABIP : 인플레이 피타율

LOB : 잔루율

ERA : 평균자책점

RA9WAR : RA9-WAR(실점 기반 war)

FIP : 수비 무관 평균자책점
kFIP : 한국 기준 수비 무관 평균자책점
WAR : 대체선수 대비 승리 기여도

1.4 분석 방법

모든 분석 과정은 SAS 프로그램을 기반으로 진행할 예정이다. 타자 데이터와 투수 데이터 모두 다음과 같은 분석을 진행할 것이다. 먼저 변수들의 평균 표준편차와 같은 기초 통계량들을 살펴봄으로써 각 변수들의 특성과 관계들에 대해 파악할 것이다. 특성들이 파악되면 갖고 있는 데이터의 차원이 작은 편이 아니기에 갖고 있는 차원보다 낮은 차원으로 설명이 가능할지 주성분 분석을 시행해볼 것이다.

1.5 결과활용 및 기대효과

데이터들 변수 간 관계를 파악한다면 기록들로 각 선수들의 특성을 파악할 수 있고 종합적인 능력까지 고려가 가능하다. 이를 통해 각 구단에서 작전을 세우는 데 반영할 수 있고 선수들을 파악해 연봉이나 평가들에 반영이 가능하다. 일단 변수들 간 상관있는 결과가 예상된다. 예를 들어, 타자 데이터에서 홈런 변수와 득점 변수, 안타 변수들 간에는 강한 상관관계가 있을 것으로 추측된다. 따라서 이 세 가지 변수들을 하나의 변수로 표현할 수 있을 것이라고 본다. 즉, 여러 변수들이 있지만 이 변수들을 더 적은 개수의 변수들로 표현할 수 있다고 예상한다. 투수 데이터의 경우에도 마찬가지이다. 평균자책점이 낮으면 9이닝 당 피홈런이 낮고 볼넷이 낮을 것이라고 자연스럽게 예측이 가능하다. 따라서 타자 데이터, 투수 데이터 모두 차원의 저주를 해결해주는 유의미한 주성분들이 나올 것이라고 예상한다. 그리고 이러한 분석을 실제 선수들 평가에 활용될 수 있기를 기대한다.

2. 본론

2.1 분석방법 소개

모든 분석 과정은 SAS 프로그램을 통해 진행할 것이다. 먼저 타자 데이터, 투수 데이터 각각 평균과 상관계수와 같은 기초 통계량 값들을 파악해볼 것이다. 그렇게 변수들의 통계량을 파악한 후 princom 프로시져를 통해 주성분 분석을 해볼 것이다. 현재 갖고 있는 데이터의 차원이 크기 때문에 차원의 저주를 피할 수 있도록 현재 데이터의 차원보다 낮은 차원으로 설명을 해볼 것이다. 이 과정은 타자 데이터, 투수 데이터 각각에 대해 진행할 것이다.

2.2 데이터 분석 및 결과 설명

먼저 타자 데이터와 투수 데이터를 각각 hitter와 pitcher에 저장하였다.

```
data hitter;
input no player$ team$ game PA AB Hit HR R RBI BB SO SB BABIP AVG OBP SLG OPS wOBA WAR @@;
cards;
1 홍창기 LG 144 651 524 172 4 103 52 109 95 23 0.393 0.328 0.456 0.408 0.864 0.412 7.10
2 이정후 Hero 123 544 464 167 7 78 84 62 37 10 0.373 0.360 0.438 0.522 0.960 0.429 7.06
3 강백호 KT 142 628 516 179 16 76 102 104 85 10 0.385 0.347 0.451 0.521 0.972 0.434 6.97
4 최정 SSG 134 555 436 121 35 92 100 84 102 8 0.277 0.278 0.410 0.562 0.972 0.425 6.83
5 양의지 NC 141 570 480 156 30 81 111 69 60 2 0.315 0.325 0.414 0.581 0.995 0.432 6.48
6 알테어 NC 143 565 492 134 32 83 84 57 156 20 0.331 0.272 0.358 0.514 0.872 0.384 4.89
7 전준우 롯데 144 619 552 192 7 88 92 53 71 6 0.384 0.348 0.405 0.469 0.874 0.395 4.86
8 구자욱 삼성 139 610 543 166 22 107 88 48 98 27 0.331 0.306 0.361 0.519 0.880 0.382 4.48
9 정은원 한화 139 608 495 140 6 85 39 105 105 19 0.346 0.283 0.407 0.384 0.791 0.376 4.41
10 추신수 SSG 137 580 461 122 21 84 69 103 123 25 0.315 0.265 0.409 0.451 0.860 0.396 4.30
11 김재환* 두산 137 566 475 130 27 86 102 81 127 2 0.316 0.274 0.382 0.501 0.883 0.393 4.26
```

```
data pitcher;
input no player$ team$ win lose sv hld bsv game start inning soper9 bbper9 hrper9 BABIP LOB ERA RA9WAR FIP kFIP WAR;
cards;
1 미란다 두산 14 5 0 0 0 28 28 173.2 11.66 3.26 0.57 0.305 81.7 2.33 8.46 2.67 2.14 8.49
2 루친스키 NC 15 10 0 0 0 30 30 178.2 8.92 2.77 0.60 0.298 71.6 3.17 6.61 3.32 3.05 6.37
3 뷰캐넌 삼성 16 5 0 0 0 30 30 177.0 8.24 3.00 0.66 0.308 75.5 3.10 6.27 3.46 3.24 5.74
4 고영표 KT 11 6 0 1 0 26 25 166.2 7.02 1.46 0.49 0.286 73.5 2.92 6.63 3.20 3.04 5.51
5 폰트 SSG 8 5 0 0 0 25 25 145.2 9.70 2.78 0.74 0.271 66.8 3.46 4.03 3.28 2.93 5.43
6 데스파이네 KT 13 10 0 0 0 33 33 188.2 7.87 3.72 0.48 0.301 72.8 3.39 6.11 3.62 3.51 5.24
7 요키시 Hero 16 9 0 0 0 31 31 181.1 6.50 2.28 0.60 0.295 73.5 2.93 6.59 3.54 3.45 5.12
8 수아레즈 LG 10 2 0 0 0 23 22 115.1 9.83 3.20 0.31 0.294 79.0 2.18 5.38 2.77 2.42 4.82
9 킹험 한화 10 8 0 0 0 25 25 144.0 8.25 2.56 0.69 0.267 71.1 3.19 4.96 3.43 3.20 4.79
10 스트레일리 롯데 10 12 0 0 0 31 31 165.2 8.96 3.64 0.65 0.328 68.9 4.07 4.26 3.62 3.41 4.78
```

위 과정은 데이터를 저장하는 과정의 일부이고 모든 데이터의 행은 총 100개이다. 또한 선수 이름에 *표시가 있는 것은 마약이나 음주운전 같은 불미스러운 범죄를 범한 선수들에게 표시한다.

```
proc means data=hitter;
run;
```

먼저 타자 데이터에 관하여 통계량들을 살펴보았다.

SAS 시스템

MEANS 프로시저

변수	N	평균	표준편차	최솟값	최댓값
no	100	50.5000000	29.0114920	1.0000000	100.0000000
game	100	106.4000000	32.4824739	17.0000000	144.0000000
PA	100	384.4300000	174.4981407	26.0000000	668.0000000
AB	100	330.0500000	151.8093817	22.0000000	589.0000000
Hit	100	91.7700000	47.6041877	8.0000000	192.0000000
HR	100	9.0600000	8.7071548	0	35.0000000
R	100	48.9400000	26.8572056	2.0000000	107.0000000
RBI	100	48.6300000	28.2684301	1.0000000	111.0000000
BB	100	42.5600000	23.4552314	2.0000000	109.0000000
SO	100	63.2000000	33.4564393	5.0000000	156.0000000
SB	100	6.7000000	9.2206400	0	46.0000000
BABIP	100	0.3161400	0.0416132	0.2220000	0.4710000
AVG	100	0.2722400	0.0326636	0.1800000	0.3640000
OBP	100	0.3632800	0.0355431	0.3010000	0.4560000
SLG	100	0.4052100	0.0677959	0.2400000	0.6350000
OPS	100	0.7684900	0.0880566	0.6170000	1.0910000
wOBA	100	0.3535400	0.0338498	0.2960000	0.4770000
WAR	100	2.0692000	1.6905350	0.3200000	7.1000000

결과는 위와 같고 몇 개의 변수들을 살펴보도록 하겠다. 우선 HR(홈런)의 경우에는 평균이 약 9, 표준편차가 약 8.7임을 알 수 있다. AVG(타율)의 경우에는 평균이 약 0.272임을 알 수 있다. 또한, R(득점)과 RBI는 평균이 48-49 수준으로 비슷한 모습을 보이고 표준편차도 각각 약 26.9, 28.2로 비슷한 값을 보여 비슷한 분포를 따를 것이라고 생각한다. 또한 BB(사사구)와 SO(삼진)을 보았을 때 삼진이 사사구의 약 150%라는 것을 알 수 있고 사사구를 얻는 것보다 삼진을 당하는 일이 더 빈번했다고 단순히 생각할 수 있고 WAR(대체 선수 대비 승리기여도) 상위 100명 중에 홈런 타자가 많기에 당연히 삼진이 많은 선수들이 많이 있을 것이라고도 볼 수 있다.

다음으로는 투수 데이터에 대해 살펴보겠다.

```
proc means data=pitcher;
run;
```

SAS 시스템

MEANS 프로시저

변수	N	평균	표준편차	최솟값	최댓값
no	100	50.5000000	29.0114920	1.0000000	100.0000000
win	100	5.1000000	4.2533409	0	16.0000000
lose	100	4.3400000	3.2759300	0	12.0000000
sv	100	2.6500000	8.2147625	0	44.0000000
hld	100	3.9600000	6.3164368	0	26.0000000
bsv	100	0.9800000	1.8311557	0	8.0000000
game	100	34.3100000	17.4081689	5.0000000	70.0000000
start	100	11.0400000	11.6280052	0	33.0000000
inning	100	82.9090000	49.2622155	12.1000000	188.2000000
soper9	100	7.7674000	1.6302436	3.7700000	13.5000000
bbper9	100	3.6530000	1.1432768	1.4600000	7.2100000
hrper9	100	0.6458000	0.3241280	0	1.6700000
BABIP	100	0.3013400	0.0435909	0.2280000	0.5220000
LOB	100	71.9440000	7.1807447	53.6000000	85.1000000
ERA	100	3.8306000	1.2875651	1.9300000	7.3300000
RA9WAR	100	2.3678000	1.9579187	-1.2300000	8.4600000
FIP	100	3.8574000	0.6631371	2.4200000	5.3300000
kFIP	100	3.7532000	0.7818999	2.1400000	5.3900000
WAR	100	1.8494000	1.7214275	0.3100000	8.4900000

우선 투수 기록의 가장 기본적인 era(평균자책점)을 보게 되면 평균이 약 3.83이고 표준편차가 약 1.28임을 알 수 있다. 그 다음 가장 먼저 들어온 변수들은 soper9(9이닝당 탈삼진), bbper9(9이닝당 사사구)였다. 앞에서 타자 데이터를 분석할 때와 비교해볼 수 있었다. 타자 데이터를 볼 때 당한 삼진이 얻어낸 볼넷보다 1.5배 많았다는 점을 고려해보면 투수 또한, 9이닝당 평균 탈삼진이 9이닝당 평균 사사구보다 두 배가량 많다는 것을 무리없이 받아들일 수 있다고 생각한다. 또한 inning(이닝수) 변수를 보면 평균은 약 82.9, 표준편차는 49.26정도이다. 표준편차가 다른 변수들에 비하면 상당히 높은 값을 보인다. 그 이유를 추측해보면 야구는 투수의 보직에 따라 이닝 수가 다른데 war이 높은 100명의 투수들을 뽑아왔기에 구원 투수, 선발 투수가 섞여 있어서 이닝 수의 표준편차가 다른 변수들에 비해 눈에 띄게 높기 때문이라고 볼 수 있다. 비슷한 이유로 구원 투수가 대부분 기록하는 기록인 홀드와 세이브 또한 크게 유의한 점을 발견할 수 없었다.

다음으로는 변수들의 상관관계를 파악해보겠다. 먼저 타자 데이터를 보도록 하겠다.

	no	game	PA	AB	Hit	HR	R	RBI	BB	SO	SB	BABIP	AVG	OBP	SLG	OPS	wOBA	WAR
no	1.00000	-0.67771 <.0001	-0.79696 <.0001	-0.77463 <.0001	-0.82185 <.0001	-0.55462 <.0001	-0.81359 <.0001	-0.76682 <.0001	-0.80285 <.0001	-0.56240 <.0001	-0.37428 0.0001	-0.21680 0.0303	-0.57438 <.0001	-0.53934 <.0001	-0.59645 <.0001	-0.67691 <.0001	-0.65925 <.0001	-0.91801 <.0001
game	-0.67771 <.0001	1.00000	0.90065 <.0001	0.89976 <.0001	0.85896 <.0001	0.49073 <.0001	0.84334 <.0001	0.74409 <.0001	0.73380 <.0001	0.65699 <.0001	0.43280 <.0001	-0.01997 0.8436	0.27230 0.0061	0.05481 0.5881	0.20515 0.0406	0.18007 0.0730	0.11769 0.2435	0.61183 <.0001
PA	-0.79696 <.0001	0.90065 <.0001	1.00000	0.99654 <.0001	0.97245 <.0001	0.55981 <.0001	0.93574 <.0001	0.85114 <.0001	0.84937 <.0001	0.71394 <.0001	0.46319 <.0001	-0.00021 0.9984	0.38745 <.0001	0.17387 0.0836	0.34407 0.0005	0.33508 0.0007	0.27203 0.0062	0.73324 <.0001
AB	-0.77463 <.0001	0.89976 <.0001	0.99654 <.0001	1.00000	0.97605 <.0001	0.55388 <.0001	0.93013 <.0001	0.85014 <.0001	0.80597 <.0001	0.70733 <.0001	0.45876 <.0001	0.00130 0.9898	0.39085 <.0001	0.13054 0.1955	0.33643 0.0006	0.31172 0.0016	0.23863 0.0168	0.70249 <.0001
Hit	-0.82185 <.0001	0.85896 <.0001	0.97245 <.0001	0.97605 <.0001	1.00000	0.49532 <.0001	0.92969 <.0001	0.83562 <.0001	0.79489 <.0001	0.61397 <.0001	0.47203 <.0001	0.13948 0.1663	0.55130 <.0001	0.26585 0.0075	0.39339 <.0001	0.41018 <.0001	0.35315 0.0003	0.78100 <.0001
HR	-0.55462 <.0001	0.49073 <.0001	0.55981 <.0001	0.55388 <.0001	0.49532 <.0001	1.00000	0.55312 <.0001	0.82040 <.0001	0.48315 <.0001	0.69990 <.0001	-0.05173 0.6092	-0.22073 0.0273	0.07272 0.4721	0.03552 0.7257	0.71996 <.0001	0.56864 <.0001	0.42407 <.0001	0.54079 <.0001
R	-0.81359 <.0001	0.84334 <.0001	0.93574 <.0001	0.93013 <.0001	0.92969 <.0001	0.55312 <.0001	1.00000	0.78903 <.0001	0.80882 <.0001	0.69558 <.0001	0.58986 <.0001	0.10599 0.2939	0.43995 <.0001	0.26936 0.0067	0.41190 <.0001	0.42585 <.0001	0.37727 0.0001	0.77959 <.0001
RBI	-0.76682 <.0001	0.74409 <.0001	0.85114 <.0001	0.85014 <.0001	0.83562 <.0001	0.82040 <.0001	0.78903 <.0001	1.00000	0.70766 <.0001	0.70859 <.0001	0.13943 0.1665	-0.04346 0.6676	0.36656 0.0002	0.16611 0.0986	0.63513 <.0001	0.55604 <.0001	0.43367 <.0001	0.73305 <.0001
BB	-0.80285 <.0001	0.73380 <.0001	0.84937 <.0001	0.80597 <.0001	0.79489 <.0001	0.48315 <.0001	0.80882 <.0001	0.70766 <.0001	1.00000	0.62112 <.0001	0.39965 <.0001	0.04352 0.6673	0.34421 0.0005	0.44627 <.0001	0.34471 0.0004	0.44553 <.0001	0.45803 <.0001	0.79643 <.0001
SO	-0.56240 <.0001	0.65699 <.0001	0.71394 <.0001	0.70733 <.0001	0.61397 <.0001	0.69990 <.0001	0.69558 <.0001	0.70859 <.0001	0.62112 <.0001	1.00000	0.31270 0.0015	0.00279 0.9780	0.00882 0.9306	-0.04611 0.6487	0.36780 0.0002	0.26456 0.0078	0.17288 0.0854	0.50221 <.0001
SB	-0.37428 0.0001	0.43280 <.0001	0.46319 <.0001	0.45876 <.0001	0.47203 <.0001	-0.05173 0.6092	0.58986 <.0001	0.13943 0.1665	0.39965 <.0001	0.31270 0.0015	1.00000	0.25573 0.0102	0.28200 0.0045	0.18925 0.0593	-0.04456 0.6598	0.04208 0.6776	0.08030 0.4271	0.35601 0.0003
BABIP	-0.21680 0.0303	-0.01997 0.8436	-0.00021 0.9984	0.00130 0.9898	0.13948 0.1663	-0.22073 0.0273	0.10599 0.2939	-0.04346 0.6676	0.04352 0.6673	0.00279 0.9780	0.25573 0.0102	1.00000	0.69996 <.0001	0.56029 <.0001	0.14657 0.1456	0.33901 0.0006	0.42701 <.0001	0.27155 0.0063
AVG	-0.57438 <.0001	0.27230 0.0061	0.38745 <.0001	0.39085 <.0001	0.55130 <.0001	0.07272 0.4721	0.43995 <.0001	0.36656 0.0002	0.34421 0.0005	0.00882 0.9306	0.28200 0.0045	0.69996 <.0001	1.00000	0.66209 <.0001	0.47239 <.0001	0.63094 <.0001	0.64282 <.0001	0.60875 <.0001
OBP	-0.53934 <.0001	0.05481 0.5881	0.17387 0.0836	0.13054 0.1955	0.26585 0.0075	0.03552 0.7257	0.26936 0.0067	0.16611 0.0986	0.44627 <.0001	-0.04611 0.6487	0.18925 0.0593	0.56029 <.0001	0.66209 <.0001	1.00000	0.39307 <.0001	0.70627 <.0001	0.86153 <.0001	0.59066 <.0001
SLG	-0.59645 <.0001	0.20515 0.0406	0.34407 0.0005	0.33643 0.0006	0.39339 <.0001	0.71996 <.0001	0.41190 <.0001	0.63513 <.0001	0.34471 0.0004	0.36780 0.0002	-0.04456 0.6598	0.14657 0.1456	0.47239 <.0001	0.39307 <.0001	1.00000	0.92857 <.0001	0.79881 <.0001	0.61762 <.0001
OPS	-0.67691 <.0001	0.18007 0.0730	0.33508 0.0007	0.31172 0.0016	0.41018 <.0001	0.56864 <.0001	0.42585 <.0001	0.55604 <.0001	0.44553 <.0001	0.26456 0.0078	0.04208 0.6776	0.33901 0.0006	0.63094 <.0001	0.70627 <.0001	0.92857 <.0001	1.00000	0.96277 <.0001	0.71393 <.0001
wOBA	-0.65925 <.0001	0.11769 0.2435	0.27203 0.0062	0.23863 0.0168	0.35315 0.0003	0.42407 <.0001	0.37727 0.0001	0.43367 <.0001	0.45803 <.0001	0.17288 0.0854	0.08030 0.4271	0.42701 <.0001	0.64282 <.0001	0.86153 <.0001	0.79881 <.0001	0.96277 <.0001	1.00000	0.70368 <.0001
WAR	-0.91801 <.0001	0.61183 <.0001	0.73324 <.0001	0.70249 <.0001	0.78100 <.0001	0.54079 <.0001	0.77959 <.0001	0.73305 <.0001	0.79643 <.0001	0.50221 <.0001	0.35601 0.0003	0.27155 0.0063	0.60875 <.0001	0.59066 <.0001	0.61762 <.0001	0.71393 <.0001	0.70368 <.0001	1.00000

위의 표는 타자 데이터의 상관관계를 보여준다. 아무래도 hit(안타), R(득점), RBI(타점), HR(홈런)과 같은 변수들은 서로가 연관되어 있기에 상관계수가 높을 수밖에 없다. WAR 지표 같은 경우에도 잘하는 선수가 높은 값을 가지므로 좋은 기록인 Hit, R, BB, AVG와 같은 변수들과 상관계수가 높은 것을 확인할 수 있었다. AB(타석수)와도 상관관계가 높은 것으로 보아 타석수가 많을수록 아무래도 기회를 많이 받는다는 뜻이기에 가능한 것으로 보인다. AB와의 상관계수는 누적 기록 변수들 같은 경우에는 무조건 높게 나오는 경향을 보이지만 비율 기록 변수들과는 그렇게 상관관계가 있다고 보기는 어렵다. 또한, 누적 기록 변수들(홈런, 타점 등등)은 서로 상관관계가 있어 보인다. 이렇듯 타자 데이터의 상관관계를 보았을 때 대부분의 기록들은 서로 높은 상관관계를 보이고 변수들끼리 연관이 있다고 충분히 생각이 가능하다.

다음으로는 투수 데이터의 상관분석이다.

```
proc corr data=pitcher;
run;
```

피어슨 상관 계수, N = 100
H0: Rho=0 가정하에서 Prob > |r|

	no	win	lose	sv	hld	bsv	game	start	inning	soper9	bbper9	hrper9	BABIP	LOB	ERA	RA9WAR	FIP	kFIP	WAR
no	1.00000	-0.76284 <.0001	-0.58455 <.0001	-0.01753 0.8626	0.29722 0.0027	0.10990 0.2764	0.13679 0.1747	-0.76042 <.0001	-0.83842 <.0001	-0.19122 0.0567	0.39934 <.0001	0.02268 0.8228	0.13791 0.1712	-0.28681 0.0038	0.36115 0.0002	-0.84558 <.0001	0.31965 0.0012	0.34550 0.0004	-0.88966 <.0001
win	-0.76284 <.0001	1.00000	0.57168 <.0001	-0.16406 0.1029	-0.24123 0.0156	-0.21632 0.0306	-0.11051 0.2737	0.79133 <.0001	0.87277 <.0001	0.03633 0.7197	-0.24684 0.0133	0.15527 0.1229	-0.13644 0.1759	0.20626 0.0395	-0.18975 0.0586	0.84113 <.0001	-0.02304 0.8200	-0.04450 0.6602	0.79993 <.0001
lose	-0.58455 <.0001	0.57168 <.0001	1.00000	-0.12090 0.2308	-0.31322 0.0015	-0.20092 0.0450	-0.17669 0.0787	0.75325 <.0001	0.76345 <.0001	0.04036 0.6901	0.05485 0.5878	0.24127 0.0156	-0.00294 0.9768	-0.16119 0.1091	0.18650 0.0632	0.47479 <.0001	0.23187 0.0203	0.20716 0.0386	0.55561 <.0001
sv	-0.01753 0.8626	-0.16406 0.1029	-0.12090 0.2308	1.00000	-0.07386 0.4652	0.60186 <.0001	0.46618 <.0001	-0.30821 0.0018	-0.15507 0.1234	0.12295 0.2230	-0.03584 0.7233	-0.07476 0.4598	-0.11662 0.2479	0.38774 <.0001	-0.31846 0.0012	0.05870 0.5619	-0.21704 0.0301	-0.22835 0.0223	-0.08461 0.4026
hld	0.29722 0.0027	-0.24123 0.0156	-0.31322 0.0015	-0.07386 0.4652	1.00000	0.45492 <.0001	0.69221 <.0001	-0.55999 <.0001	-0.38229 <.0001	0.02653 0.7933	0.00863 0.9321	-0.07434 0.4623	-0.14622 0.1466	0.12751 0.2061	-0.15280 0.1291	-0.26894 0.0068	-0.08083 0.4240	-0.07755 0.4431	-0.35625 0.0003
bsv	0.10990 0.2764	-0.21632 0.0306	-0.20092 0.0450	0.60186 <.0001	0.45492 <.0001	1.00000	0.69700 <.0001	-0.50281 <.0001	-0.28312 0.0043	0.09851 0.3295	0.02517 0.8037	-0.18411 0.0667	-0.06724 0.5063	0.35574 0.0003	-0.31240 0.0016	-0.09485 0.3479	-0.24127 0.0156	-0.23926 0.0165	-0.23928 0.0165
game	0.13679 0.1747	-0.11051 0.2737	-0.17669 0.0787	0.46618 <.0001	0.69221 <.0001	0.69700 <.0001	1.00000	-0.49827 <.0001	-0.18087 0.0717	0.05551 0.5833	-0.00279 0.9780	-0.09647 0.3397	-0.21161 0.0346	0.33140 0.0008	-0.33068 0.0008	-0.03481 0.7310	-0.17919 0.0744	-0.18074 0.0719	-0.20349 0.0423
start	-0.76042 <.0001	0.79133 <.0001	0.75325 <.0001	-0.30821 0.0018	-0.55999 <.0001	-0.50281 <.0001	-0.49827 <.0001	1.00000	0.92888 <.0001	-0.02670 0.7921	-0.12412 0.2185	0.20140 0.0445	-0.02368 0.8151	-0.05431 0.5915	0.05977 0.5547	0.71333 <.0001	0.16088 0.1098	0.14641 0.1461	0.79205 <.0001
inning	-0.83842 <.0001	0.87277 <.0001	0.76345 <.0001	-0.15507 0.1234	-0.38229 <.0001	-0.28312 0.0043	-0.18087 0.0717	0.92888 <.0001	1.00000	0.00534 0.9579	-0.20260 0.0432	0.17944 0.0740	-0.11777 0.2432	0.11232 0.2659	-0.11309 0.2626	0.84559 <.0001	0.05570 0.5820	0.03696 0.7150	0.86240 <.0001
soper9	-0.19122 0.0567	0.03633 0.7197	0.04036 0.6901	0.12295 0.2230	0.02653 0.7933	0.09851 0.3295	0.05551 0.5833	-0.02670 0.7921	0.00534 0.9579	1.00000	0.23934 0.0165	0.10909 0.2800	0.16392 0.1032	0.08061 0.4253	0.00777 0.9389	0.08683 0.3903	-0.34306 0.0005	-0.47002 <.0001	0.22828 0.0224
bbper9	0.39934 <.0001	-0.24684 0.0133	0.05485 0.5878	-0.03584 0.7233	0.00863 0.9321	0.02517 0.8037	-0.00279 0.9780	-0.12412 0.2185	-0.20260 0.0432	0.23934 0.0165	1.00000	0.06950 0.4920	0.15658 0.1198	-0.31608 0.0014	0.49098 <.0001	-0.36664 0.0002	0.48319 <.0001	0.46048 <.0001	-0.33761 0.0006
hrper9	0.02268 0.8228	0.15527 0.1229	0.24127 0.0156	-0.07476 0.4598	-0.07434 0.4623	-0.18411 0.0667	-0.09647 0.3397	0.20140 0.0445	0.17944 0.0740	0.10909 0.2800	0.06950 0.4920	1.00000	-0.06335 0.5312	-0.09949 0.3247	0.39303 <.0001	-0.04579 0.6510	0.68223 <.0001	0.57956 <.0001	-0.01559 0.8777
BABIP	0.13791 0.1712	-0.13644 0.1759	-0.00294 0.9768	-0.11662 0.2479	-0.14622 0.1466	-0.06724 0.5063	-0.21161 0.0346	-0.02368 0.8151	-0.11777 0.2432	0.16392 0.1032	0.15658 0.1198	-0.06335 0.5312	1.00000	-0.29812 0.0026	0.52899 <.0001	-0.25148 0.0116	-0.05995 0.5535	-0.07248 0.4736	-0.06476 0.5221
LOB	-0.28681 0.0038	0.20626 0.0395	-0.16119 0.1091	0.38774 <.0001	0.12751 0.2061	0.35574 0.0003	0.33140 0.0008	-0.05431 0.5915	0.11232 0.2659	0.08061 0.4253	-0.31608 0.0014	-0.09949 0.3247	-0.29812 0.0026	1.00000	-0.82581 <.0001	0.46159 <.0001	-0.33009 0.0008	-0.33577 0.0006	0.18929 0.0593
ERA	0.36115 0.0002	-0.18975 0.0586	0.18650 0.0632	-0.31846 0.0012	-0.15280 0.1291	-0.31240 0.0016	-0.33068 0.0008	0.05977 0.5547	-0.11309 0.2626	0.00777 0.9389	0.49098 <.0001	0.39303 <.0001	0.52899 <.0001	-0.82581 <.0001	1.00000	-0.50261 <.0001	0.58417 <.0001	0.55519 <.0001	-0.27576 0.0055

RA9WAR	-0.84558 <.0001	0.84113 <.0001	0.47479 <.0001	0.05870 0.5619	-0.26894 0.0068	-0.09485 0.3479	-0.03481 0.7310	0.71333 <.0001	0.84559 <.0001	0.08683 0.3903	-0.36664 0.0002	-0.04579 0.6510	-0.25148 0.0116	0.46159 <.0001	-0.50261 <.0001	1.00000	-0.27523 0.0056	-0.28632 0.0039	0.89400 <.0001
FIP	0.31965 0.0012	-0.02304 0.8200	0.23187 0.0203	-0.21704 0.0301	-0.08083 0.4240	-0.24127 0.0156	-0.17919 0.0744	0.16088 0.1098	0.05570 0.5820	-0.34306 0.0005	0.48319 <.0001	0.68223 <.0001	-0.05995 0.5535	-0.33009 0.0008	0.58417 <.0001	-0.27523 0.0056	1.00000	0.98781 <.0001	-0.30846 0.0018
kFIP	0.34550 0.0004	-0.04450 0.6602	0.20716 0.0386	-0.22835 0.0223	-0.07755 0.4431	-0.23926 0.0165	-0.18074 0.0719	0.14641 0.1461	0.03696 0.7150	-0.47002 <.0001	0.46048 <.0001	0.57956 <.0001	-0.07248 0.4736	-0.33577 0.0006	0.55519 <.0001	-0.28632 0.0039	0.98781 <.0001	1.00000	-0.33886 0.0006
WAR	-0.88966 <.0001	0.79993 <.0001	0.55561 <.0001	-0.08461 0.4026	-0.35625 0.0003	-0.23928 0.0165	-0.20349 0.0423	0.79205 <.0001	0.86240 <.0001	0.22828 0.0224	-0.33761 0.0006	-0.01559 0.8777	-0.06476 0.5221	0.18929 0.0593	-0.27576 0.0055	0.89400 <.0001	-0.30846 0.0018	-0.33886 0.0006	1.00000

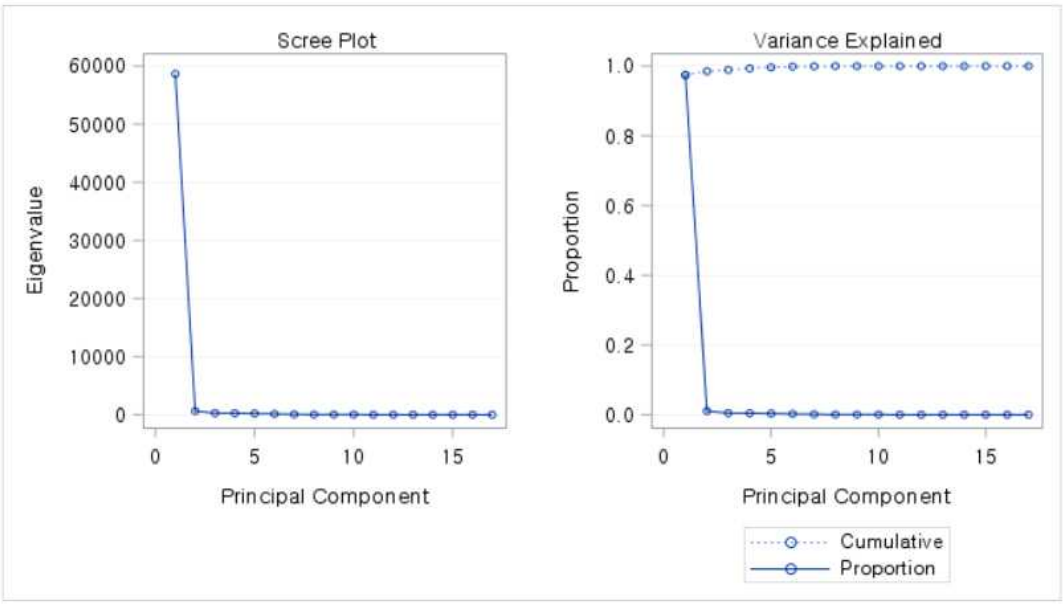
먼저 war과 다른 변수들 간의 상관관계를 보도록 하겠다. 승리와 패배와 같은 기록과는 상관관계가 크지만 홀드, 세이브와 같이 구원 투수들과 관련된 기록들과는 상관관계가 그렇게 크다고 판단되지 않는다. 이러한 점에서 상대적으로 승리나 패배의 기록이 많은 선발투수들이 war이 더 높을 가능성이 구원 투수들보다 높다고 생각하였다. 또한, 소화 이닝이 많을수록 war이 더 높을 것이라고 생각할 수 있다. 9이닝당 삼진, 9이닝당 사사구와 같은 기록들이 승패나 war과 상관이 있을 것이라고 생각하였지만 의외로 큰 상관이 없었다. 게다가 9이닝당 삼진은 평균자책점과도 무관하다고 나온다. 9이닝당 사사구와 홈런은 많을수록 평균자책점이 커지는 것과 상관이 있다는 걸 확인할 수 있었다. FIP(수비 무관 평균자책점)은 ERA(평균자책점)과 상관관계가 존재할 수밖에 없지만 0.58정도로 예상보다 큰 상관관계는 안 보였으며 수비와 관련된 부분이 평균자책점에 큰 영향을 준다고 판단하였다. 이렇듯 투수 데이터의 상관관계를 보았을 때 대부분의 기록들은 서로 높은 상관관계를 보이고 변수들끼리 연관이 있다고 생각하였다. 다음으로는 주성분 분석을 시행해보았다. 먼저 타자 데이터들에 대하여 진행하였고 결과는 아래와 같다.

```
proc princomp data=hitter out=out1 covariance;
var game--war;
run;
```

Total Variance	60170.564355

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	58628.8091	57998.3576	0.9744	0.9744
2	630.4515	360.5748	0.0105	0.9849
3	269.8768	20.6903	0.0045	0.9893
4	249.1865	65.5855	0.0041	0.9935
5	183.6010	61.5590	0.0031	0.9965
6	122.0420	75.2590	0.0020	0.9986
7	46.7830	23.2158	0.0008	0.9993
8	23.5672	14.4705	0.0004	0.9997
9	9.0967	2.1521	0.0002	0.9999
10	6.9446	6.7425	0.0001	1.0000
11	0.2020	0.1991	0.0000	1.0000
12	0.0029	0.0022	0.0000	1.0000
13	0.0007	0.0004	0.0000	1.0000
14	0.0003	0.0002	0.0000	1.0000
15	0.0001	0.0001	0.0000	1.0000
16	0.0000	0.0000	0.0000	1.0000
17	0.0000		0.0000	1.0000

고유치들의 결과는 위와 같은데 분산을 80% 이상 설명하는 주성분을 고르고자 하면 하나의 주성분만으로 충분히 설명이 가능하다. scree plot도 그려서 주성분이 몇 개가 필요한지 판단해보겠다.



주성분의 축이 1개 이후로 완만해지기 때문에 여기서도 1개의 주성분으로 충분히 설명이 가능하다고 판단하였고 하나의 주성분만으로 설명을 해보겠다. 고유벡터의 결과는 다음과 같다.

	Eigenvectors									
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
game	0.121150	0.022938	0.270852	-.444717	0.799991	0.238451	0.124684	0.010351	0.018293	0.028764
PA	0.720224	-.018164	0.227873	0.251920	0.047486	-.183714	-.118035	0.068022	-.047683	-.552565
AB	0.626058	-.105700	-.250343	-.358467	-.173480	-.140068	-.076491	-.061526	0.040010	0.580527
Hit	0.191749	-.252888	-.197925	0.096925	-.155635	0.469394	0.737549	-.038252	0.196560	-.118428
HR	0.020305	0.184329	-.205407	0.117707	0.130396	0.082809	-.293339	-.103488	0.885299	-.047831
R	0.103925	0.026069	0.143655	0.067582	-.163057	0.714610	-.436912	-.430777	-.220094	0.033153
RBI	0.099911	0.241341	-.648154	0.359234	0.355708	0.191472	-.099133	0.327270	-.311507	0.073868
BB	0.080756	0.064253	0.451299	0.639703	0.154243	-.045925	0.171871	-.028167	0.071134	0.561055
SO	0.099154	0.908151	0.105659	-.170931	-.222793	0.036936	0.268116	-.025100	-.030544	-.023374
SB	0.017586	-.047417	0.274032	-.112719	-.244670	0.329660	-.174198	0.827770	0.136716	0.082849
BABIP	0.000001	-.000149	0.000180	0.000129	-.000720	0.001821	0.003610	0.000379	-.001025	0.000846
AVG	0.000053	-.000536	-.000364	0.000503	-.000303	0.001266	0.001834	0.000259	0.000654	-.000060
OBP	0.000024	-.000295	0.000382	0.001426	-.000092	0.001224	0.001421	-.000071	0.000673	-.000979
SLG	0.000098	0.000690	-.001885	0.001841	0.000155	0.002258	0.000009	-.000355	0.005281	-.000562
OPS	0.000121	0.000395	-.001503	0.003267	0.000063	0.003482	0.001430	-.000426	0.005953	-.001541
wOBA	0.000037	0.000051	-.000267	0.001397	-.000038	0.001368	0.000796	-.000253	0.001998	-.000931
WAR	0.005076	0.000473	0.001249	0.047296	0.005549	0.055490	0.046178	-.002469	0.056296	-.108469

첫 번째 고유벡터만 살펴보면 PA(타수), AB(타석)의 값들이 각각 0.72, 0.62 수준이고 다른 변수들의 값은 크게 높지 않은 편이다. 따라서 주성분 1은 타석과 타수

에 의해서 충분히 설명된다고 볼 수 있다. 각 선수의 주성분 점수를 보겠다.

```
proc print data=out1;
var player prin1;
run;
```

OBS	player	Prin1
1	홍창기	348.036
2	이정후	220.791
3	강백호	328.371
4	최정	215.503
5	양의지	255.783
6	알테어	262.128
7	전준우	341.640
8	구자욱	328.452
9	정은원	289.625
10	추신수	249.405
11	김재환*	251.463
12	박건우	201.099
13	최재훈	93.539
14	한유섬	188.403
15	강민호	112.182
16	김혜성	356.038
17	노시환	97.712

87	이성곤	-250.131
88	이흥련	-325.572
89	문보경	-79.701
90	김민수	-410.114
91	서건창	-163.786
92	이형종	-144.897
93	김준태	-351.816
94	알몬테	-181.706
95	최용제	-362.928
96	김주형	-443.299
97	김찬형	-496.895
98	강승호	-54.762
99	안재석	-215.087
100	이지영	-167.523

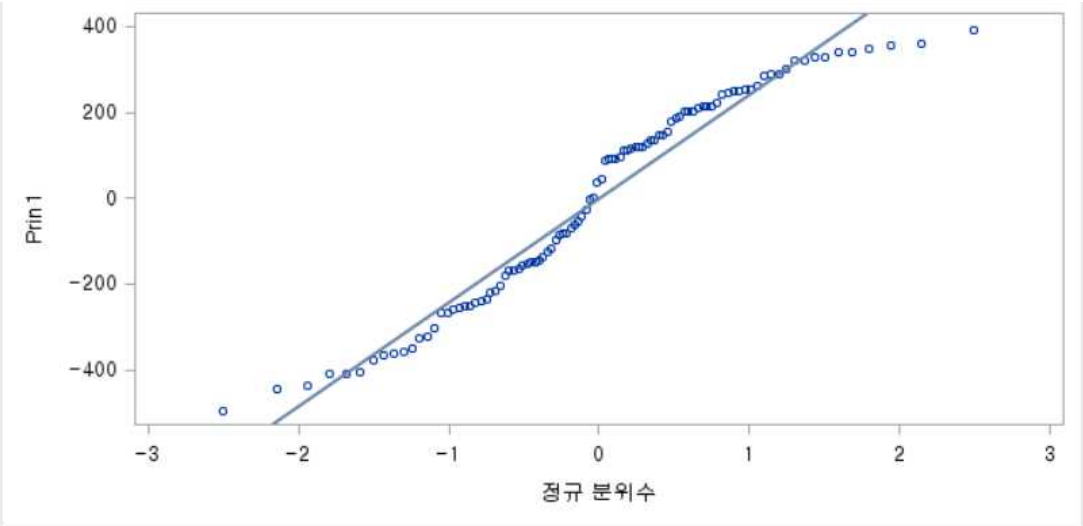
상위권 일부 선수와 하위권 일부 선수의 값들을 보게 되면 하위권 선수들의 값이 음수를 보이며 상위 선수일수록 주성분 점수가 높고 하위 선수일수록 주성분 점수가 낮다고 판단할 수 있다. war이 타석과 타수의 영향을 많이 받기에 주성분 1로도 이러한 차이가 발생하였다고 볼 수 있다.

일변량 기초통계 분석을 통해 이상치를 탐색해보았다.

```
proc univariate data = out1 plot;
var Prin1;
run;
```

위치모수 검정: Mu0=0				
검정	통계량		p 값	
스튜던트의 t	t	0	Pr > |t|	1.0000
부호	M	2	Pr >= |M|	0.7644
부호 순위	S	14	Pr >= |S|	0.9619

mu가 0이라는 가설검정을 무리 없이 받아들일 수 있다.



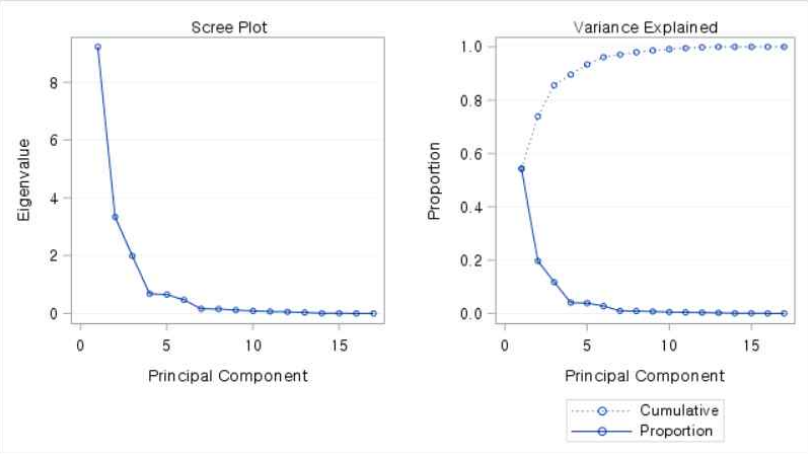
또한 정규 분위수 그림을 그려보았을 때 나름 직선 위를 잘 따라간다고 판단하였고 큰 이상치도 없다고 판단하였다.

주성분 분석에 공분산 행렬이 아닌 상관계수행렬을 사용해보았다.

```
proc princomp data=hitter out=out1;
var game--war;
run;
```

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	9.23026892	5.89373227	0.5430	0.5430
2	3.33653665	1.34837602	0.1963	0.7392
3	1.98816063	1.31045944	0.1170	0.8562
4	0.67770120	0.02937651	0.0399	0.8960
5	0.64832469	0.18356168	0.0381	0.9342
6	0.46476301	0.30971573	0.0273	0.9615
7	0.15504728	0.00422602	0.0091	0.9706
8	0.15082126	0.03731761	0.0089	0.9795
9	0.11350365	0.02743734	0.0067	0.9862
10	0.08606631	0.02414092	0.0051	0.9912
11	0.06192539	0.01129618	0.0036	0.9949
12	0.05062921	0.01987620	0.0030	0.9979
13	0.03075300	0.02678998	0.0018	0.9997
14	0.00396302	0.00265607	0.0002	0.9999
15	0.00130695	0.00107812	0.0001	1.0000
16	0.00022883	0.00022883	0.0000	1.0000
17	0.00000000		0.0000	1.0000

총 분산의 80% 이상을 설명하고자 한다면 적어도 3개의 주성분이 필요할 것이다. scree plot을 그려봤을 때도 3개의 주성분이면 충분히 설명이 가능하다고 판단할 수 있다.



	Eigenvectors									
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
game	0.260423	-.251712	0.111200	0.033792	-.156669	0.071264	0.714444	-.394849	-.363673	0.048152
PA	0.299763	-.198182	0.085049	-.026446	-.142639	-.000819	-.060785	0.220267	-.032489	-.048973
AB	0.294870	-.209438	0.086974	0.045685	-.167006	-.044376	-.042919	0.258694	-.006176	-.124178
Hit	0.303762	-.119523	0.139345	0.070135	-.241910	-.097683	-.125931	0.146787	0.039741	-.221875
HR	0.222950	-.071323	-.467695	0.125234	0.158158	0.006212	0.228764	-.148440	0.528335	0.310154
R	0.303619	-.129670	0.120386	-.014346	0.082924	-.145236	0.028558	0.033558	0.097443	-.409118
RBI	0.292817	-.097974	-.212790	0.142366	-.188898	0.018727	-.028153	0.063654	0.345162	0.141038
BB	0.281562	-.072488	0.074049	-.454264	0.006275	0.316645	-.232861	0.091938	-.271781	0.590330
SO	0.229390	-.219460	-.135190	0.242164	0.496067	0.448767	-.146243	0.174880	-.125073	-.137960
SB	0.138856	-.088966	0.458966	-.059538	0.615099	-.506360	0.048733	-.018877	0.124568	0.233744
BABIP	0.066174	0.333999	0.370708	0.551862	0.132853	0.425683	0.024705	-.101345	0.000804	0.033974
AVG	0.189433	0.314880	0.263062	0.320443	-.358045	-.188951	-.095804	0.062480	0.075177	0.357680
OBP	0.147848	0.411636	0.167394	-.416391	0.013932	0.244196	0.295383	0.177452	0.291586	-.136964
SLG	0.210224	0.247982	-.376498	0.192369	0.103456	-.321921	-.079880	0.020084	-.449517	0.033200
OPS	0.221531	0.357077	-.222304	-.019965	0.085275	-.149284	0.057728	0.087090	-.228393	-.029723
wOBA	0.203221	0.405095	-.118559	-.188828	0.110615	0.005736	0.166319	0.125902	-.017358	-.188326
WAR	0.295971	0.114599	0.025043	-.181038	-.020118	0.037122	-.449258	-.755645	0.084078	-.199275

세 개의 주성분이 적절하다고 판단하였기에 세 개의 주성분만을 보면 첫 번째 주성분은 모든 변수의 값들을 골고루 반영하고 있고 두 번째 주성분은 AVG, OBP, OPS, wOBA 같은 비율형 기록을 주로 반영하고 있다. 세 번째 주성분은 SB, BABIP 변수를 주로 설명하고 있다.

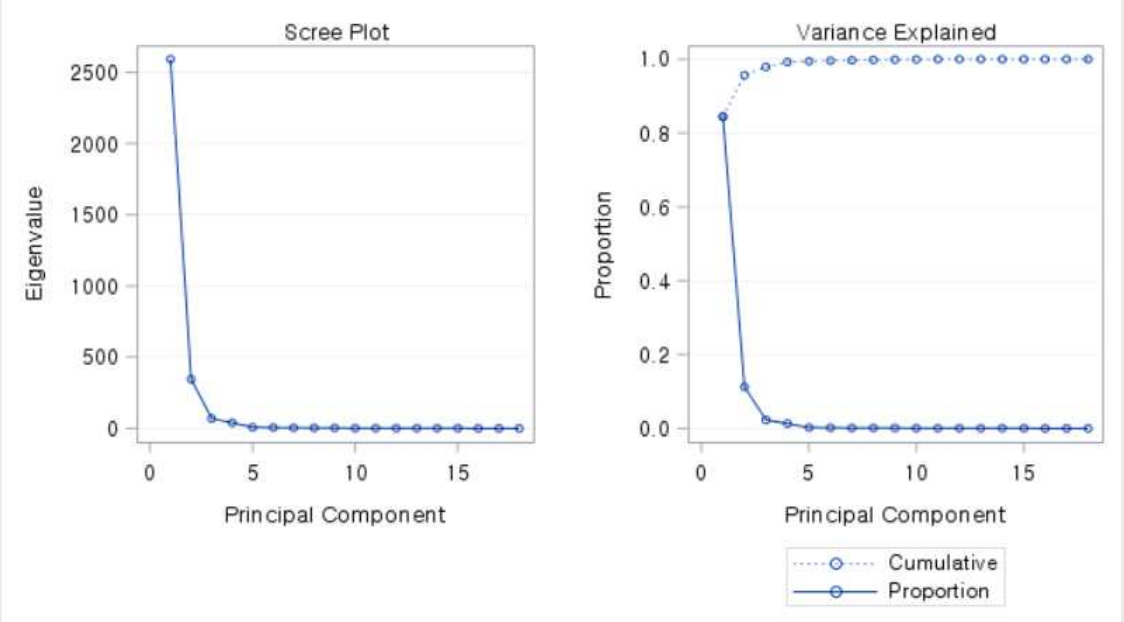
다음으로 투수 데이터를 가지고 주성분 분석을 해보았다.

```
proc princomp data=pitcher out=out2 covariance;
var win--war;
run;
```

Total Variance	3069.7167113

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2592.17883	2247.46219	0.8444	0.8444
2	344.71665	275.33613	0.1123	0.9567
3	69.38051	30.65232	0.0226	0.9793
4	38.72820	31.54302	0.0126	0.9919
5	7.18518	2.12353	0.0023	0.9943
6	5.06165	1.66980	0.0016	0.9959
7	3.39185	0.26044	0.0011	0.9970
8	3.13141	0.91563	0.0010	0.9981
9	2.21578	0.73533	0.0007	0.9988
10	1.48046	0.34595	0.0005	0.9993
11	1.13450	0.57023	0.0004	0.9996
12	0.56427	0.31469	0.0002	0.9998
13	0.24958	0.07850	0.0001	0.9999
14	0.17108	0.05238	0.0001	1.0000
15	0.11870	0.11118	0.0000	1.0000
16	0.00753	0.00702	0.0000	1.0000
17	0.00051	0.00050	0.0000	1.0000
18	0.00001		0.0000	1.0000

총 분산의 80% 이상을 설명하기 위해서는 최소한 하나의 주성분이 필요하다. scree plot도 살펴보겠다.



scree plot을 그려봤을 때도 하나의 주성분으로도 충분할 것이라고 말할 수 있다.

	Eigenvectors									
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
win	0.072729	0.022298	-.013642	0.107276	0.048599	-.595118	0.686913	0.167296	-.332883	0.018121
lose	0.049275	-.003728	-.035059	-.150961	0.257167	0.570274	0.136259	0.372256	-.563321	-.280875
sv	-.028666	0.222534	0.766726	-.401028	0.398599	-.109615	-.009525	-.119201	0.053040	0.028334
hld	-.051191	0.202637	-.404706	0.264065	0.807136	-.088926	-.120644	-.167397	0.076125	0.061462
bsv	-.011359	0.066617	0.050308	-.012511	0.132081	0.007536	-.061504	0.060474	-.247248	0.203944
game	-.077965	0.906954	-.188298	-.133085	-.241215	0.090832	0.151786	0.030485	0.118756	-.103856
start	0.215851	-.184853	-.003204	-.037356	0.169403	0.278786	0.552136	-.086780	0.568099	-.328598
inning	0.966409	0.127888	-.018599	-.016379	-.018662	-.036937	-.170365	-.027252	-.061949	0.100620
soper9	0.000071	0.006426	0.019689	0.000735	0.107400	-.108555	-.143221	0.809470	0.355199	0.119925
bbper9	-.004451	-.005190	-.019301	-.056195	0.005865	0.159273	0.162877	0.210390	0.164139	0.516838
hrper9	0.001163	-.001143	-.002630	-.005228	0.012652	0.021771	0.030812	0.006305	0.000160	0.116506
BABIP	-.000091	-.000614	-.000365	-.001391	-.000475	-.000286	-.000430	0.005552	0.000507	0.003329
LOB	0.013288	0.172727	0.447009	0.829121	-.050942	0.237860	0.062424	0.046285	-.008003	0.051731
ERA	-.002390	-.029850	-.061057	-.125472	0.026586	0.058556	0.105010	0.055291	0.017012	0.353388
RA9WAR	0.032223	0.021079	0.054715	0.084361	0.013487	-.199442	-.058990	0.039696	-.011780	-.182530
FIP	0.000856	-.007351	-.017923	-.024473	0.008918	0.118458	0.130432	-.109754	-.013550	0.314008
kFIP	0.000726	-.008978	-.021972	-.028064	0.002500	0.143132	0.156760	-.175950	-.034356	0.342341
WAR	0.029170	0.001197	0.021628	0.027214	0.034325	-.203417	-.162042	0.170270	0.080481	-.271950

고유벡터 결과를 통해 첫 번째 주성분을 보게 되면 첫 번째 주성분은 inning만으로도 충분히 설명이 가능하다는 것을 알 수 있다. 다음은 주성분 점수를 살펴보도록 하겠다.

```
proc print data=out2;
var player prin1;
run;
```

OBS	player	Prin1
1	미란다	92.903
2	루친스키	98.076
3	뷰캐넌	96.765
4	고영표	85.179
5	폰트	64.565
6	데스파이	107.971
7	요키시	101.031
8	수아레즈	35.169
9	킹험	63.768
10	스트레일	85.222

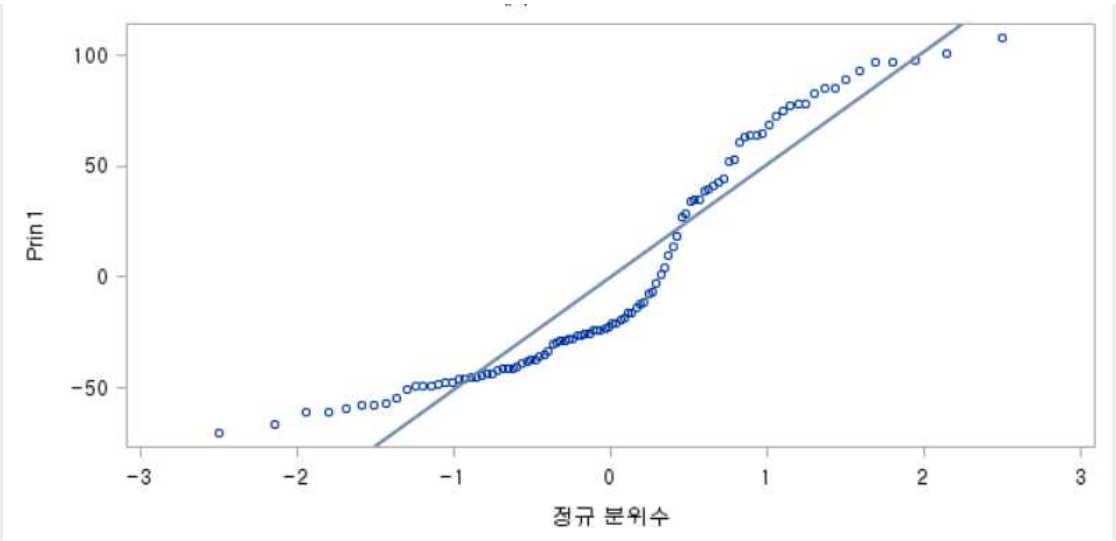
91	최성훈	-44.862
92	이재희	-59.079
93	류진욱	-42.195
94	이승현(5	-45.856
95	박종기	-23.630
96	이준영	-50.323
97	엄상백	-28.190
98	최영환	-41.397
99	김진성	-48.033
100	김기탁	-70.473

상위권 일부 선수와 하위권 일부 선수의 값들을 보게 되면 하위권 선수들의 값이 음수를 보이며 상위 선수일수록 주성분 점수가 높고 하위 선수일수록 주성분 점수가 낮다고 판단할 수 있다. war이 이닝 수의 영향을 많이 받기에 주성분 1로도 이러한 차이가 발생하였다고 볼 수 있다. 일변량 기초통계 분석을 통해 이상치를 탐색해보았다.

```
proc univariate data = out2 plot;
var Prin1;
run;
```

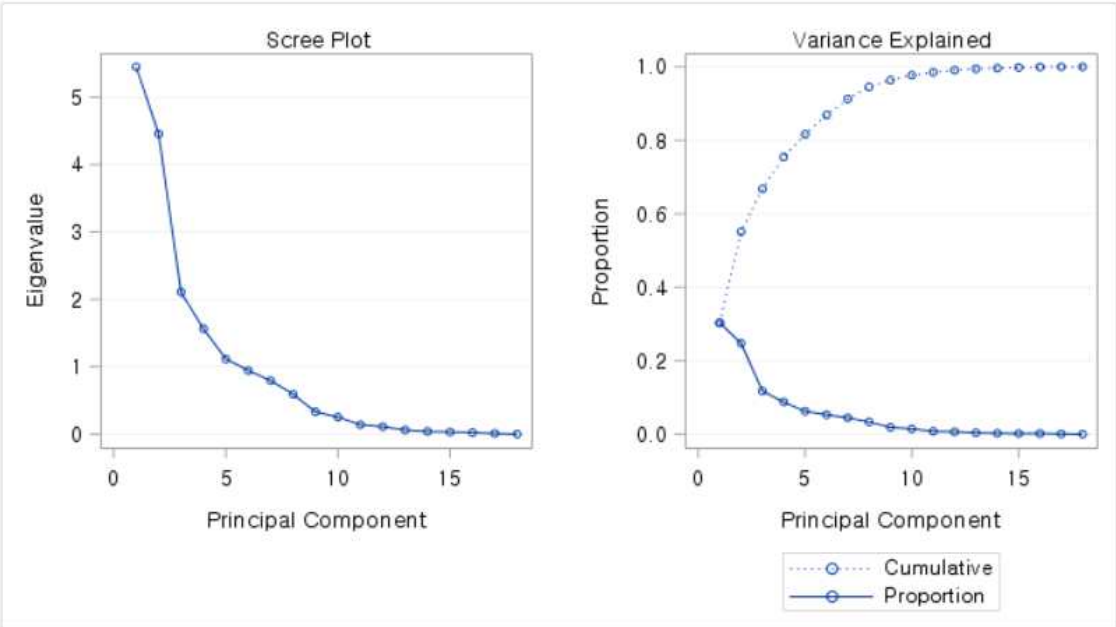
위치모수 검정: Mu0=0				
검정	통계량		p 값	
스튜던트의 t	t	0	Pr > |t|	1.0000
부호	M	-12	Pr >= |M|	0.0210
부호 순위	S	-55	Pr >= |S|	0.8511

mu가 0이라는 귀무가설을 쉽게 받아들일 수 있다.



하지만 정규 분위수 그림을 보게 되면 직선위에 있다기 보다는 특정한 패턴을 보이며 움직이므로 정규성 가정을 만족한다고는 보기가 어렵다. 이상치는 보이지 않는다고 판단된다. 주성분 분석에 공분산 행렬이 아닌 상관계수행렬을 사용해보았다.

```
proc princomp data=pitcher out=out2;
var win--war;
run;
```



scree plot을 보게 되면 적어도 다섯 개의 주성분이 필요할 것으로 보인다. 고유치를 통해 최종 주성분의 수를 정해보도록 하겠다.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	5.45166199	0.99502608	0.3029	0.3029
2	4.45663591	2.35047136	0.2476	0.5505
3	2.10616455	0.54328711	0.1170	0.6675
4	1.56287745	0.44903705	0.0868	0.7543
5	1.11384040	0.17132747	0.0619	0.8162
6	0.94251293	0.14936960	0.0524	0.8685
7	0.79314333	0.20341446	0.0441	0.9126
8	0.58972887	0.26193622	0.0328	0.9454
9	0.32779265	0.07708688	0.0182	0.9636
10	0.25070578	0.11264080	0.0139	0.9775
11	0.13806498	0.03036536	0.0077	0.9852
12	0.10769962	0.04851399	0.0060	0.9912
13	0.05918563	0.02023157	0.0033	0.9944
14	0.03895406	0.00779883	0.0022	0.9966
15	0.03115523	0.00910267	0.0017	0.9983
16	0.02205256	0.01424533	0.0012	0.9996
17	0.00780723	0.00779039	0.0004	1.0000
18	0.00001684		0.0000	1.0000

총 분산의 80% 이상을 설명하려 한다면 적어도 5개의 주성분이 필요하며 이 주성분들을 고유벡터들을 통해 파악해보겠다.

	Eigenvectors									
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
win	0.374792	-.056682	0.127566	0.057534	0.167663	0.024341	0.047942	0.163867	0.366728	-.194636
lose	0.304996	0.116706	0.127571	0.250466	0.071269	0.275625	-.047126	-.271088	-.681394	0.232047
sv	-.095495	-.226274	0.193030	0.263553	-.600353	0.198396	0.121616	-.219842	0.225880	0.239451
hld	-.221177	-.130603	0.225786	0.044700	0.667448	-.052000	0.044748	0.065374	-.064552	0.053727
bsv	-.187779	-.251803	0.251442	0.294868	-.045560	0.300961	0.094833	0.021424	-.180398	-.766461
game	-.168337	-.245386	0.358477	0.232840	0.270939	0.187805	0.074771	-.070559	0.249801	0.443468
start	0.409594	0.104866	-.003620	0.007169	-.012271	0.068293	-.052117	0.028040	-.079111	-.052994
inning	0.408472	-.004661	0.121000	0.085981	0.077236	0.127294	0.003685	0.007531	0.026917	0.055680
soper9	0.017418	-.101293	-.200971	0.588090	0.030718	-.563286	-.121296	-.096416	-.069348	-.030857
bbper9	-.119183	0.235555	0.071522	0.392326	-.056135	0.020764	-.676816	0.277138	0.121502	0.030168
hrper9	0.065641	0.239159	0.332977	0.146468	-.075012	-.505004	0.492353	-.096293	-.015060	-.043112
BABIP	-.044498	0.131040	-.385387	0.314058	0.010711	0.278802	0.454555	0.579791	-.053130	0.155852
LOB	0.046855	-.339361	0.230837	-.081550	-.215375	-.246277	0.007545	0.576460	-.317114	0.149789
ERA	-.057884	0.412639	-.122784	0.249349	0.083967	0.129212	0.183839	-.105724	0.202951	-.052469
RA9WAR	0.366586	-.203274	0.097595	0.018284	-.001283	0.011322	-.053363	0.131325	0.209412	-.017161
FIP	-.008518	0.387679	0.379687	-.025702	-.074962	-.044122	-.009937	0.139753	0.019354	-.022776
kFIP	-.017528	0.382368	0.374163	-.106128	-.071887	0.068137	-.049753	0.163685	0.026558	-.015107
WAR	0.386196	-.130002	-.076691	0.113821	0.072367	-.010499	0.016575	-.046785	0.186525	-.061400

첫 번째 주성분의 경우 start, inning, war이 대부분 설명한다. 두 번째 주성분의 경우는 era, fip, kfip가 대부분 설명하고 있다. 세 번째 주성분의 경우에는 game, fip, kfip가 대부분 설명하고 있다. 그리고 네 번째 주성분의 경우에는 soper9이 대부분 설명하고 있다. 마지막으로 다섯 번째 주성분의 경우에는 hld가 대부분 설명하고 있다. 투수 데이터와 타자 데이터 모두 상관계수 행렬보다 공분산 행렬을 사용하였을 때 더 적은 변수로 분산의 80% 이상을 설명할 수 있기에 공분산 행렬을 이용할 것이다.

2.3 분석의 타당성 설명

타자 데이터와 투수 데이터의 통계량을 보고 데이터들의 분포에 대해 살펴보았고 상관분석을 해보았다. 상관분석의 결과, 타자 데이터와 투수 데이터 모두 상관관계가 있는 변수들이 존재하였고 변수들 간 연관성이 존재한다고 판단하였다. 그리고 주성분 분석을 통해 주성분들을 도출해낸 결과 각각의 선수들의 많은 변수들을 선형결합하여 하나의 주성분 점수로 도출해낼 수 있었다.

3.결론

3.1 분석 결과 요약

먼저 타자 데이터에 대해 말해보겠다. R과 RBI의 경우에는 동일한 분포를 띠따를 것이라고 생각하였고 SO가 BB의 1.5배라는 점을 발견하였다. 상관관계의 경우에는 타석의 수가 중요하게 작용하여 타석의 수가 많을수록 유리한 누적형 기록인 Hit, R, BB 등은 타석의 수와 상관계수가 높았고 누적형 기록끼리도 상관계수가 높은 걸 볼 수 있었다. 하지만 비율 기록 변수들은 타석과는 그렇게 상관이 있지는 않은 걸 발견할 수 있었다. 공분산 행렬로 주성분 분석을 하였을 시에는 하나의 주성분으로 전체 분산의 80% 이상을 설명할 수 있다는 걸 알았다. 그리고 그 주성분은 타수와 타석이 큰 비중을 설명하였고 타수와 타석을 주로 설명하는 주성분 하나로도 데이터를 80% 이상 설명할 수 있다고 말할 수 있다.

투수 데이터의 경우에는 soper9가 bbper9의 두 배 가량 된다는 점을 발견했고 투수들의 보직에 따라 이닝의 편차가 크기 때문에 이닝의 표준편차가 50에 가까운 걸 확인할 수 있었다. 투수 데이터의 상관분석은 보직들이 섞여 있어서 타자만큼 유의한 정보를 얻어낼 순 없었다. 그렇지만 era와 fip가 많이 유사하지만 수비 부문도 평균자책점에 큰 영향을 미친다는 점과 이닝의 수와 war이 상관정도가 크기에 선발투수가 war이 더 높을 수밖에 없을 것이라고 생각하였다. 공분산 행렬로 주성분 분석을 하였을 시에는 하나의 주성분으로 전체 분산의 80% 이상을 설명할 수 있다는 걸 알았다. 그리고 그 주성분은 이닝이 큰 비중을 설명하였고 이닝을 주로 설명하는 주성분 하나로도 데이터를 80% 이상 설명할 수 있다고 말할 수 있다.

3.2 분석의 장점 및 한계점 설명

분석의 장점으로는 변수가 많은 데이터였지만 자세히 들여다보면 서로 연관이 많은 데이터였다는 점을 파악할 수 있고 이를 토대로 점수를 많이 내는 방식이나 점수를 덜 주는 방식에 대해 힌트를 줄 수 있다고 생각하였다. 또한, 변수가 많은 데이터를 훨씬 낮은 차원의 데이터로 쉽게 선수의 특성을 알아볼 수 있다는 점이 장점이라고 본다. 하지만 한계점으로는 시즌별로 전체적인 기록의 형태가 다르게 나오기에 여러 시즌의 데이터를 통해 변수들의 특성을 분석하는 게 낫다고 본다. 물론 너무 예전의 데이터는 현재 트렌드에 맞지 않기에 적당한 시점을 정해 그 이후 시즌의 데이터를 이용해야 한다. 이 외에도 투수 데이터의 경우 보직에 따라 특히 이닝과 같은 변수의 특성이 너무 다르기에 보직을 나누어 분석을 진행한다면 더욱 유의미한 결과가 관측될 것이라고 생각한다.

참고문헌

http://www.kbreport.com/main (데이터 출처)

http://www.kbreport.com/statDic/detail?seq=6763&contentsType=a301