다변량 해석

요인분석을 활용한 새로운 타격지표 제안

2017110525 송민철
2017110504 신준식
2018110476 배정민
2019110464 정유정

Ⅰ. 서론

1. 데이터 분석 배경 및 목적

Major League Baseball(이하 MLB)의 각 팀은 선수의 능력을 항목별로 나누고 이에 점수를 매기는 방식을 사용하여 선수를 평가한다. 이때, 점수를 매기는 방식은 평균 50, 표준편차 10을 가지는 정규 분포를 기준으로 하는 20-80 Scale로 평가한다. 일반적으로 스카우팅 리포트에서 Hitting은 타율, Power는 홈런의 개수, Speed는 홈에서 1루까지 걸리는 시간을 기준으로 평가한다. 하지만, 선수가 가진 Hitting 능력이 좋아 안타를 많이 칠 경우, Power가 같아도 더 많은 홈런을 칠 수 있다. 이런 방식으로 타자의 여러 능력이 하나의 지표에 복합적으로 영향을 미치기 때문에 각 능력에 대한 정확한 지표가 될 수 없다는 점이 문제가 된다. 그렇기에, 여러 타격 지표들을 통해서 타자의 능력들을 성공적으로 분리해 낼 수 있다면, 선수에 대한 정확한 평가가 가능해질 것이다. 따라서, 타격 지표에 영향을 미칠 것으로 생각되는 능력들을 여러 지표에서 요인 분석을 통해서 분리해내고, 이 요인들을 타자의 타격 능력에 대한 종합적인 지표로 볼 수 있는 xwoba에 대한 설명력을 기반으로 기존의 측정 방식과 비교하여 합리성을 확인하는 것을 목표로 한다. 또한, 합리성이 검증된다면, 요인 분석을 통해 도출한 요인들을 이용해서 선수에게 실제 적용하는 단계까지 진행할 것이다.

Ⅱ. 본론

1. 데이터 설명

분석에 사용된 데이터는 총 3가지로 stat0525는 2022년 5월 26일 기준, 2015년부터 2022년까지 100타석 이상의 타자들의 타격 스탯캐스트 관련 기록, FanGraphs_wrc와 FanGraphs_value는 2022년 5월 20일 기준, 2015년부터 2022년까지 100타석 이상의 타자들의 타격 관련 기록이다. 각각의 데이터에서 분석에 사용되는 주요 변수들과 데이터 구조의 세부적인 특징은 다음과 같다.

가. stat0525

Baseball Savant의 데이터로 스탯캐스트 자료들을 다수 포함하고 있는 데이터이다. 3,229개의

관측치와 21개의 변수로 이루어진 데이터이며, 각 변수의 이름과 설명은 다음과 같다.

변수명	변수 설명 (단위)	변수 유형
last_name	last name	character
first_name	first name	character
player_id	선수별 구분을 위한 ID	numeric
year	연도	numeric
b_total_pa	타석 수	numeric
b_total_pitches	상대 투수 투구 수 총합	numeric
xba	기대 타율	numeric
xwoba	기대 가중 출루율	numeric
xobp	기대 출루율	numeric
xiso	기대 순수 장타율	numeric
exit_velocity_avg	출구 속도 (miles for hour)	numeric
launch_angle_avg	발사 각도 (°)	numeric
sweet_spot_percent	스윗 스팟에 맞출 확률 (%)	numeric
barrel_batted_rate	배럴 타구 확률 (%)	numeric
z_swing_percent	스트라이크 존 안에 들어오는 공에 스윙할 확률 (%)	numeric
oz_swing_percent	스트라이크 존 밖에 나가는 공에 스윙할 확률 (%)	numeric
oz_contact_percent	스트라이크 존 밖에 나가는 공을 컨택할 확률 (%)	numeric
iz_swing_percent	스트라이크 존 안에 들어오는 공을 컨택할 확률 (%)	numeric
n_bolts	최고 스프린트 스피드가 30ft/sec 이상이었던 횟수	numeric
hp_to_1b	홈에서 1루까지 걸리는 시간 (초)	numeric
sprint_speed	최고 스프린트 스피드 (feet for second)	numeric

[표 2-1-1-1] 데이터 설명

나. FanGraphs_wrc, FanGraphs_value

1) FanGraphs_wrc

Fangraph의 데이터로 자료들을 기본적인 타격 자료와 세이버매트릭스 자료를 포함하고 있는 데이터이다. 3,202개의 관측치와 24개의 변수로 이루어진 데이터이며, 주요 변수의 이름과 설명은 다음과 같다.

변수명	변수 설명 (단위)	변수 유형
Season	연도	numeric
Name	Full name	character
HR	홈런 개수	numeric
K%	삼진 확률 (%)	numeric
BABIP	인플레이 타율	numeric

[표 2-1-2-1] FanGraphs_wrc 데이터 설명

2) FanGraphs_value

Fangraph의 데이터로 자료들을 기본적인 타격 자료와 세이버매트릭스 자료를 포함하고 있는 데이터이다. 3,202개의 관측치와 15개의 변수로 이루어진 데이터이며, 주요 변수의 이름과 설명은 다음과 같다.

변수명	변수 설명 (단위)	변수 유형
Season	연도	numeric
Name	full name	character
Batting	평균에 비해서 타격으로 얻은 점수	numeric
Base running	평균에 비해서 주루로 얻은 점수	numeric

[표 2 - 1 - 2 - 2] FanGraphs_wrc 데이터 설명

2. 전처리

원 자료 <stats0525.csv>는 3229행 21열로 이루어진 데이터로 구조는 다음과 같다.

last_name	first_name	player_id	year	b_total_pa	b_total_pitches	xba	xwoba	xobp	xiso	exit_velocity_avg	launch_angle_avg	sweet_spot_percent
Cabrera	Miguel	408234	2022	147	540	0.254	0.309	0.308	0.158	90.4	5.8	37.5
Cruz Jr.	Nelson	443558	2022	170	700	0.273	0.356	0.343	0.222	90.2	8.4	33.3
Peralta	David	444482	2022	149	579	0.249	0.342	0.341	0.219	90.4	19.1	31.6
Escobar	Alcides	444876	2022	109	433	0.201	0.230	0.252	0.059	81.0	15.1	35.1
Blackmon	Charlie	453568	2022	158	580	0.263	0.327	0.321	0.175	84.9	11.7	31.1
Maldonado	Martin	455117	2022	114	461	0.172	0.259	0.259	0.140	90.7	18.7	28.4
Cain	Lorenzo	456715	2022	107	430	0.224	0.256	0.289	0.050	88.5	-0.3	20.8

barrel_batted_rate	z_swing_percent	oz_swing_percent	oz_contact_percent	iz_contact_percent	n_bolts	hp_to_1b	sprint_speed
8.3	72.8	28.4	55.0	73.8	NA	5.14	23.4
10.8	73.1	29.8	57.8	78.3	NA	4.85	25.0
10.2	67.5	29.7	45.5	85.3	NA	4.48	27.2
1.4	70.7	36.1	61.3	82.4	1	4.38	28.2
6.6	70.2	26.2	67.6	87.1	NA	4.36	27.8
6.0	63.8	27.0	39.3	81.1	NA	NA	20.6
0.0	69.2	31.0	53.7	80.4	NA	4.46	28.0

[그림 2 - 2 - 1] 원천 데이터

결측치는 n_bolts 와 hp_to_1b에서 각각 2289개, 63개가 존재한다. n_bolts의 결측치는 30feet/sec가 선수들이 내기 어려운 기록임을 감안하여 0으로 대체하였다. hp_to_1b의 결측치는 n_bolts, hp_to_1b, sprint_speed 등 주루 능력과 관련된 변수들에 knn 알고리즘을 적용해 대체하였다.

추가로 <FanGraphs_value.csv>의 Batting, Base.Running 변수와 <FanGraphs_wrc.csv>의 HR, K., BABIP 변수를 사용하였다. 각 변수들에 결측치는 존재하지 않는다. 이 5개의 변수는 <stats0525.csv>의 last_name, first_name 변수로 name 변수를 생성하고 형식을 통일한 뒤, year 변수명을 season으로 변경하여 원 자료와 합하였다.

percent값을 나타내는 sweet_spot_percent, barrel_batted_rate, z_swing_percent, oz_swing_percent, oz_contact_percent, iz_contact_percent 변수들에 대해서는100으로 나누어 백분율 값을 확률 값으로 변환하였다.

다음으로 원 변수를 이용하여 분석에 사용될 5개의 새로운 변수를 생성하였다. 첫 번째, 투구수(b_total_pitch)를 타석에 들어서는 횟수(b_total_pa)로 나누어 “타석 당 투구 수”를 나타내는 ball_per_pa 변수를 생성하였다. 두 번째, 기대출루율(xobp)에서 기대타율(xba)를 빼서 기대순수출루율를 나타내는 xobp_iso 변수를 생성하였다. 세 번째, 존 밖 스윙을 나타내는 변수 oz_swing_percent은 선구안 측면에서 높을수록 부정적인 의미를 갖는다고 판단하여, 1-oz_swing_percent로 변환하였다. 네 번째, 홈에서 1루까지 도달시간이 가장 느린 선수 대비

얼마나 빠른 지를 대략적으로 나타내 주는 변수인 1.1*max - hp_to_1b를 생성하였다. 이 변수는 클수록 긍정적인 의미를 갖는다. 다섯 번째, 삼진 비율을 나타내는 변수 K 역시 높을수록 부정적인 의미를 갖는다고 판단하여 확률 값으로 변환한 뒤, 1-K로 대체하였다.

단축 시즌이었던 2020년 현재 진행중인 2022년의 데이터는 분석에서 제외하였다. 또한, MLB의 로스터는 일반적으로 12명의 타자와 13명의 투수로 구성한다. 12명의 타자 중 충분한 타석 수를 보장받을 수 있는 타자는 한 팀에 10명~11명이다. 6년간의 기록이라는 점을 고려해 봤을 때, 적합한 obs의 개수는 1,800~1,980명이다. 이에 부합하는 타석 수의 기준은 250타석이기에 250타석 이상의 기록을 가진 타자들을 분석에 사용하기로 하였다.

정리하자면, 본 분석에 사용할 변수는 다음과 같다.

```
> colnames(aa0518)
 [1] "xba"                 "xobp_iso"            "xiso"                "exit_velocity_avg"
 [5] "launch_angle_avg"    "sweet_spot_percent"  "barrel_batted_rate"  "z_swing_percent"
 [9] "1 - oz_swing_percent" "oz_contact_percent"  "iz_contact_percent"  "n_bolts"
[13] "1.1*max - hp_to_1b"  "sprint_speed"        "HR"                  "1-K"
[17] "BABIP"               "ball_per_pa"         "Name"                "Season"
[21] "xwoba"               "Batting"             "Base.Running"
```

[그림 2 - 2 - 2] 분석에 사용할 변수명

현재 전처리 결과는 다음과 같으며, 1863행 23열의 데이터에 대해 추가 분석을 진행한다.

xba	xobp_iso	xiso	exit_velocity_avg	launch_angle_avg	sweet_spot_percent	barrel_batted_rate	z_swing_percent	1 - oz_swing_percent	oz_contact_percent
0.259	0.070	0.212	89.3	11.2	0.327	0.104	0.641	0.745	0.488
0.214	0.110	0.169	90.4	10.2	0.305	0.097	0.679	0.773	0.371
0.263	0.071	0.153	88.8	11.9	0.340	0.031	0.682	0.776	0.599
0.231	0.065	0.126	89.1	10.7	0.285	0.046	0.707	0.764	0.576
0.248	0.114	0.181	86.7	11.7	0.300	0.075	0.616	0.797	0.593
0.262	0.120	0.204	89.7	13.2	0.342	0.088	0.651	0.817	0.650

iz_contact_percent	n_bolts	1.1*max - hp_to_1b	sprint_speed	HR	1-K	BABIP	ball_per_pa	Name	Season	xwoba	Batting	Base.Running
0.788	1	1.265	28.9	19	0.748	0.328	3.866505	AaronAltherr	2017	0.341	10.4	-1.9
0.775	1	1.285	29.0	8	0.681	0.247	4.052632	AaronAltherr	2018	0.317	-9.0	-1.2
0.848	5	1.455	28.4	11	0.831	0.285	3.871795	AaronHicks	2015	0.329	-1.9	0.4
0.807	5	1.495	28.5	8	0.812	0.248	3.825485	AaronHicks	2016	0.288	-15.7	-0.4
0.779	3	1.505	28.5	15	0.814	0.290	4.108033	AaronHicks	2017	0.348	12.3	2.6
0.795	7	1.505	28.1	27	0.809	0.264	4.280551	AaronHicks	2018	0.372	19.8	7.0

[그림 2 - 2 - 3] 전처리 후 데이터

3. 요인분석

가. 요인분석의 필요성

요인분석은 변수들 간의 상호 연관성을 분석하여, 변수들 간에 공통적으로 작용하고 있는 핵심적인 내재요인으로 변수를 축약하는 기법이다. 이를 통해 과다한 정보로 인한 다중공선성 문제를 해결할 수 있고, 정보를 보다 효과적으로 전달할 수 있다.

wRC는 타자의 득점 생산력을 나타내는 변수로, 현존하는 타격 스탯 중 가장 정확하다. 우리는 wRC를 계산하는데 직접적으로 사용되는 기대가중출루율(xwoba)을 18개의 변수(1열~18열 변수)를 사용하여 예측하고자 한다. 이때 18개의 변수를 모두 사용하는 것은 다중공선성 문제를 야기할 수 있기에 요인분석을 통해 선수를 평가하기 위한 핵심 내재요인을 뽑아내려고 한다.

나. Uniqueness

18개의 변수에 대해 요인분석을 진행하였다. 참고로 H0: 7개의 Factor가 충분한가? 에 대한 가설검정을 진행하였을 때, p-value=1.83e-133으로 아주 작은 값이 나왔으나, 이보다 적은 수의 Factor에 대해서는 p-value=0이였다는 점과 실제 자료임을 감안하여 7개의 Factor를 사용하는 것이 적절하다고 판단하였다. 또한, 야구 데이터인만큼 Factor들간에 상관성이 존재할 수밖에 없다고 판단하여 이를 허용하는 promax 요인 회전을 적용하였다. Factor Score는 Bartlett방식을 통해 구하였다. Uniqueness는 다음과 같다.

```
Uniquenesses:
               xba            xobp_iso                xiso    exit_velocity_avg     launch_angle_avg
             0.036               0.204               0.011               0.298               0.326
  sweet_spot_percent   barrel_batted_rate      z_swing_percent 1 - oz_swing_percent   oz_contact_percent
             0.135               0.058               0.074               0.219               0.270
  iz_contact_percent            n_bolts   1.1*max - hp_to_1b        sprint_speed                  HR
             0.191               0.726               0.141               0.045               0.289
               1-K               BABIP         ball_per_pa
             0.005               0.429               0.272
```

[그림 2-3-2-1] uniqueness

7개의 요인은 Uniqueness<0.3인 변수들의 변동성을 잘 설명하고 있다고 판단하였다. 따라서 launch_angle_avg, n_bolts, BABIP 변수를 제외한 변수들에 대해 이후 분석을 진행하였다.

다. Loadings기준으로 Factor별 해석

```
Loadings:
                      Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
xba                    0.410   0.619                   0.544   0.205
xobp_iso                                       0.935                   0.145
xiso                   0.965
exit_velocity_avg      0.774  -0.129  -0.140           0.273  -0.192
launch_angle_avg       0.267                          -0.643   0.389
sweet_spot_percent                                     0.130   0.946
barrel_batted_rate     0.874  -0.266
z_swing_percent                                                0.910
1 - oz_swing_percent  -0.114   0.160           0.915
oz_contact_percent    -0.101   0.793                                  -0.106
iz_contact_percent    -0.131   0.777                                  -0.233
n_bolts                                0.478
1.1*max - hp_to_1b                     0.920
sprint_speed                           0.959
HR                     0.919   0.168   0.107          -0.166
1-K                            1.024           0.133                   0.120
BABIP                 -0.110  -0.188   0.131           0.696   0.274
ball_per_pa                   -0.307           0.488           0.134  -0.308

                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
SS loadings       3.459   2.958   2.060   1.986   1.340   1.249   1.037
Proportion Var    0.192   0.164   0.114   0.110   0.074   0.069   0.058
Cumulative Var    0.192   0.356   0.471   0.581   0.656   0.725   0.783
```

[그림 2-3-3-1] Loading

Factor1는 xiso, HR, barrel_batted_rate, exit_velocity_avg 등과 밀접하게 관련되어 있다. 이 변수들은 장타율, 홈런, 좋은 타구를 만들어내는 능력과 관련된 변수로, 이를 종합하여 판단하였을 때 Factor1은 선수의 POWER를 대표하는 요인이다.

Factor2는 1-K, oz_contact_percent, iz_contact_percent, xba 등과 밀접하게 관련되어 있다. 이 변수들은 삼진 비율, 컨택 능력, 타율과 관련된 변수로, 컨택 능력이 높을수록 삼진 비율은 낮아지고 컨택퍼센트와 타율은 높아진다는 특징이 있다. 따라서 Factor2는 선수의 CONTACT를 대표하는 요인이다.

Factor3는 1.1*max - hp_to_1b, sprint_speed 등과 밀접하게 관련되어 있다. 이 변수들은 선수의 스피드와 변수로, Factor3는 선수의 SPEED를 대표하는 요인이다.

Factor4는 xobp_iso, 1 - oz_swing_percent 등과 밀접하게 관련되어 있다. 이 변수들은

순수출루율과 스윙(컨택 능력)과 관련된 변수로, 순수출루율은 볼넷을 많이 얻을수록 높은 값을 갖는다. 즉, 순수 출루율은 좋은 공을 가려내는 능력과 안 좋은 공에 대한 참을성을 평가하는 지표가 된다. 이러한 선구안을 바탕으로 좋은 공에 대한 스윙(컨택) 또한 이뤄낼 수 있다. 따라서 Factor4는 선수의 PATIENT + EYE를 대표하는 요인이다.

Factor5는 xba와 관련된 변수로, BATTING SKILL을 대표하는 요인이다.

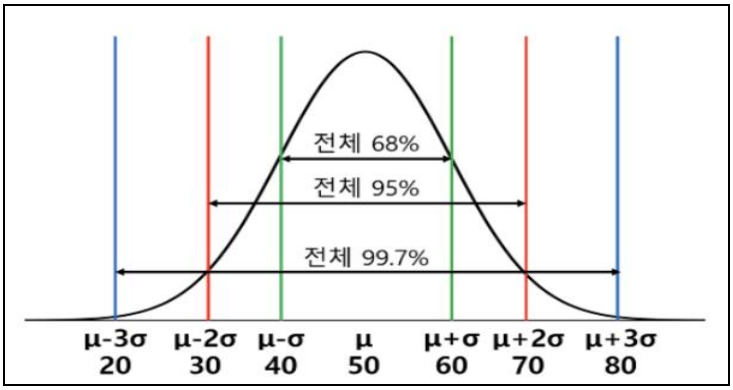
Factor6은 sweet_spot_percent와 관련된 변수로, 정타를 치는 능력을 의미한다. sweet spot은 배트에서 운동에너지를 가장 잘 전달할 수 있는 부분으로, 이는 2루 이상의 안타(장타)를 만들어낼 수 있는 능력을 의미하기도 한다. 따라서 Factor6는 BATTING SKILL FOR EXTRA BASE HIT을 대표하는 요인이다.

Factor7은 z_swing_percent와 관련된 변수로, 이는 타자가 타석에서 투수와 얼마나 적극적으로 승부를 하는지에 대해 나타내며, 더불어 선구안에 대한 능력도 요구된다. 따라서 Factor6는 AGGRESSIVE + EYE를 대표하는 요인이다.

본 조는 Loading값을 바탕으로 각 데이터에 대한 Factor Score값을 구하고, 이에 대한 추가 분석을 진행하였다.

라. 20-80 Scale 변환

20-80 Scale은 선수 스카우팅 리포트에서 사용하는 측정기준이다. nothing을 의미하는 0과 완벽함을 의미하는 100을 기준으로, 0-20구간과 80-100구간을 제외한 20-80 범위를 평가지표를 사용하는 것이다. 이때 양 끝 구간을 제외하는 이유는 정규분포를 기준으로 상위 0.1%, 하위 0.1%를 제외한 전체 99.8%가 20-80 구간 안에 위치하기 때문이다. 20-80 Scale은 전체 99.8%에 속하는 인원을 5점 단위로 나누어 14단계로 구분한다. 즉, 평균 50, 표준편차 10의 분포를 사용하여 선수를 평가한다. 우리는 위에서 구한 요인 별 대표값의 Scale을 통일하기 위해 요인 별 평균, 표준편차 값을 통해 각 변수를 스케일링 한 뒤, 20-80 Scale로 재 변환하였다.



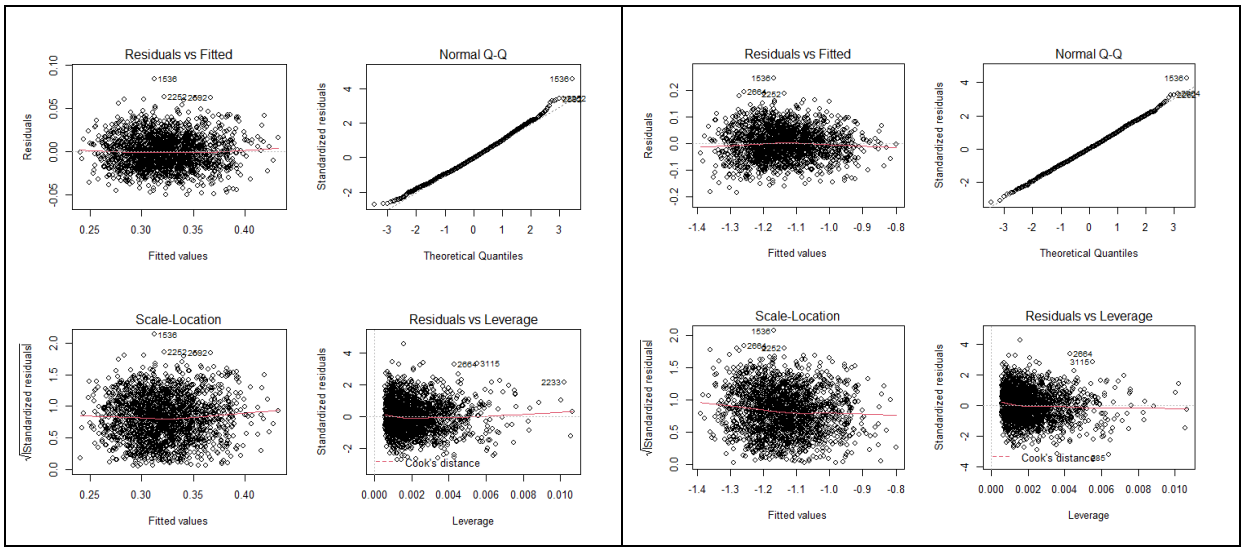
[그림 2 - 3 - 4 - 1] 20-80 Scale

4. 회귀

일반적으로 스카우팅 리포트에서는 기존의 단순한 타격지표들을 통해 타자의 능력을 평가해왔다. 이 변수들로 타격의 대표적인 종합지표인 xwOBA 를 목적변수로 하는 회귀모형을 적용시켜보고, 본 프로젝트에서 분리해낸 7 개의 요인들로 xwOBA 에 회귀를 적용시킨 모형의 결과를 비교함으로써 검증한다.

가. 가정과 진단

회귀분석의 가정인 선형성, 정규성, 등분산성, 독립성을 확인하고 회귀모델을 수정한다.



[그림 2-4-1-1] 설명변수 : 기존 지표 - 변환 전

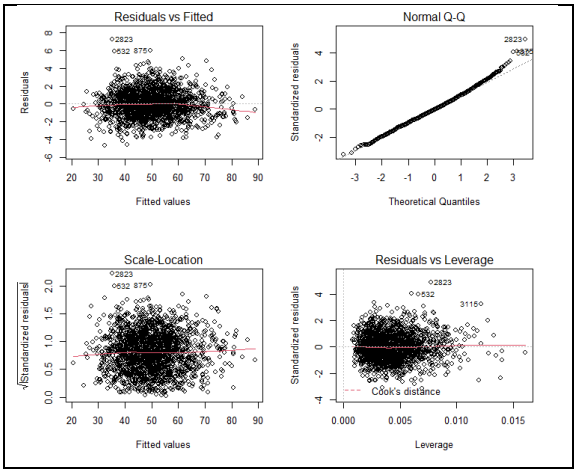
[그림 2-4-1-2] 설명변수 : 기존 지표 - 변환 후

[그림 2 - 4 - 1 - 1] 그래프에서 선형성과 잔차의 등분산성을 확인할 수 있다. Q-Q plot 보면 우측 상단의 값이 벗어나는 것으로 보아 정규성이 위배됨을 알 수 있다. 따라서 car 패키지의 powerTransform() 함수를 활용해 Box-Cox 변환을 수행했다. P-value 가 0 에 가까우므로 H0 : λ=1 을 기각하고 추정된 최적 λ 값은 -0.054 로 0 에 근사하기 때문에 *log*변환을 진행했다. [그림 2 - 4 - 1 - 2] 를 통해 정규성까지 만족함을 확인했다.

	xba	`1.1*max-hp_to_1b`	HR
VIF	1.103343	1.025369	1.118951

[표 2-4-1-1] 기존지표 설명변수들 간의 VIF

[표 2-4-1-1] 를 보면 VIF값이 모두 10보다 작기 때문에 다중공선성의 문제가 없으며 설명 변수 간의 독립성을 가정할 수 있다.



[그림 2-4-1-3] 설명변수 : 요인

[그림 2 - 4 - 1 - 3] 그래프에서 선형성과 잔차의 등분산성, 정규성까지 확보되었다고 판단하였다. 그리고 요인분석을 통해 얻은 요인들로 설명변수가 구성되어 있기 때문에 설명변수 간의 다중공선성 문제는 없으며, 독립성을 만족한다.

나. 모형 설명력 비교

```
Call:
lm(formula = logoba ~ xba + `1.1*max - hp_to_1b` + HR, data = aa0530)

Residuals:
      Min        1Q    Median        3Q       Max
-0.184449 -0.039375 -0.001107  0.038618  0.245018

Coefficients:
                        Estimate Std. Error  t value Pr(>|t|)
(Intercept)          -1.7661820  0.0149697 -117.984   <2e-16 ***
xba                   2.4159417  0.0549664   43.953   <2e-16 ***
`1.1*max - hp_to_1b` -0.0566985  0.0062891   -9.015   <2e-16 ***
HR                    0.0056006  0.0001414   39.597   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05745 on 1859 degrees of freedom
Multiple R-squared:  0.7351,    Adjusted R-squared:  0.7347
F-statistic:  1720 on 3 and 1859 DF,  p-value: < 2.2e-16
```

[그림 2 - 4 - 2 - 1] 기존 지표와 로그 xwOBA 에 대한 회귀모형

$$\log(xwOBA) = -1.7661820 + 2.4159417 \times (xBA) - 0.0566985 \times (1.1 * max - \text{hp\_to\_1b}) + 0.056006 \times (\text{HR})$$

기존 지표들을 설명변수로 하는 회귀모형식은 위와 같으며, Adjusted_R-squred 값은 0.7347 이다.

```
Call:
lm(formula = xwoba ~ power + `patient+eye` + speed + contact +
    `batting skill` + `aggresive+eye` + `batting skill for extra base hit`,
    data = matrix0520_23)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6722 -0.9961 -0.0442  0.9045  7.2620

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -48.729340   0.530498 -91.856   <2e-16 ***
power                             0.774581   0.004230 183.131   <2e-16 ***
`patient+eye`                     0.338546   0.004346  77.897   <2e-16 ***
speed                             0.003201   0.003629   0.882    0.378
contact                           0.359088   0.003942  91.098   <2e-16 ***
`batting skill`                   0.291897   0.003641  80.163   <2e-16 ***
`aggresive+eye`                   0.072521   0.004100  17.689   <2e-16 ***
`batting skill for extra base hit`   0.134751   0.003782  35.629   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.48 on 1855 degrees of freedom
Multiple R-squared:  0.9782,    Adjusted R-squared:  0.9781
F-statistic: 1.189e+04 on 7 and 1855 DF,  p-value: < 2.2e-16
```

[그림 2 - 4 - 2 - 2] 요인변수들과 xwOBA 에 대한 회귀모형 - 변수 제거 전

```
Call:
lm(formula = xwoba ~ power + `patient+eye` + contact + `batting skill` +
    `aggresive+eye` + `batting skill for extra base hit`, data = matrix0520_23)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6394 -0.9957 -0.0505  0.9151  7.3468

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -48.556983   0.493168  -98.46   <2e-16 ***
power                             0.773705   0.004111  188.21   <2e-16 ***
`patient+eye`                     0.338482   0.004345   77.90   <2e-16 ***
contact                           0.358579   0.003899   91.97   <2e-16 ***
`batting skill`                   0.292659   0.003537   82.74   <2e-16 ***
`aggresive+eye`                   0.072731   0.004093   17.77   <2e-16 ***
`batting skill for extra base hit`  0.134983   0.003773   35.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.479 on 1856 degrees of freedom
Multiple R-squared:  0.9782,    Adjusted R-squared:  0.9781
F-statistic: 1.387e+04 on 6 and 1856 DF,  p-value: < 2.2e-16
```

[그림 2 – 4 – 2 – 2] 요인변수들과 xwOBA에 대한 회귀모형 – 변수 제거 후

$$\log(xwOBA) = -\ 48.7729340\ + 0.774581 \times (POWER) + 0.338546 \times (PATIENT + EYE) + 0.359088 \times (CONTACT) + 0.291897 \times (BATTING\ SKILL) + 0.072521 \times (AGGRESIVE + EYE) + 0.134751 \times (BATTING\ SKILL\ FOR\ EXTRA\ BASE\ HIT)$$

뽑아낸 요인들을 설명변수로 하는 회귀모형식에서 SPEED 요인은 유의하지 않기 때문에 제외했으며 회귀모형식은 위와 같다. Adjusted_R-squred 값은 0.9781 이며 뽑아낸 요인들로 만든 회귀 모형의 설명력이 기존 지표로 만든 회귀 모형보다 설명력이 더 높기 때문에 보다 합리적인 모형이라고 볼 수 있다. SPEED 요인은 목적변수와의 설명력이 떨어져 제외되었지만, 이 때 요인들은 세부적이고 독립적인 각각의 타격 능력 및 성향을 의미하므로 범용성 또한 갖추고 있다.

5. 활용방안

위와 같은 절차를 거쳐 합리성이 확인된 요인들로 타자의 능력을 평가하는 구체적인 종합 지표를 만들고 적용시킬 수 있다. 연산과정은 먼저 Loading을 보고 요인을 해석했던 방법을 그대로 활용하여 각 요인에 영향을 미치는 주요 변수만을 뽑는다. 특정 요인에 대해 뽑아낸 변수를 표준화하고 Score값을 곱해주는 행렬 연산을 시행한 뒤, 20-80스케일로 변환하는 공식을 만들었다. 행렬연산 과정에서 $\mu_2 \times D_2^{-1}$은 고정된 값을 갖게 되고 모든 원소가 매우 작아 0으로 처리한다.

계산식 : $50 + 10 \times \{(\mathrm{DN} - \mu_1) \times \mathrm{D}_1^{-1} \times S \times D_2^{-1} - \mu_2 \times D_2^{-1}\} = R$
DN : 새로운 obs
D_1^{-1} : $diag\left(\text{분석 데이터의 변수 별 표준편차}\right)$
D_2^{-1}: $diag\left(\text{factor별 표준편차}\right)$
μ_1 :분석 데이터의 변수 별 평균 , μ_2 : 요인별 평균
S : 변수별 표준화한 분석 데이터에 곱하면 요인 분석의 Scores 값이 나오는 행렬
R : 요인마다 $20-80$Scale로 변환된 행렬

[표 2 – 5 – 1] 새로운 변수 계산을 위한 행렬식

연산과정을 거쳐 요인 별 공식을 도출하면 다음과 같다.

$$\text{F1 : POWER} = \{(\text{xiso} - 0.17) \times 13.70 + (exit\ velocity\ avg - 88.56) \times 0.01 + (barrel\ batted\ rated - 0.07) \times 3.67 + (HR - 16.19) \times 0\} \times 10 + 50$$

$$\text{F2 : CONTACT} = \{(\text{xba} - 0.25) \times 2.55 + (\text{oz contact percent} - 0.59) \times 0.36 + (\text{iz contact percent} - 0.83) \times 0.94 + ((1 - \text{k}) - 0.79) \times 14.52\} \times 10 + 50$$

$$\text{F3 : SPEED} = \{(\text{n}_{\text{bolts}} - 4.96) \times 0.00 + \{(1.1 \times 5.15 - hpto1b) - 1.23\} \times 1.14 + (sprint\ speed - 27.03) \times 0.52\} \times 10 + 50$$

$$\text{F4 : PATIENT + EYE} = \{(XOBP\ ISO - 0.07) \times 19.19 + \{(1 - oz\ swing\ percent) - 0.72\} \times 7.08\} \times 10 + 50$$

$$\text{F5 : BATTING SKILL} = \{(\text{xba} - 0.25) \times 45.55\} \times 10 + 50$$

$$\text{F6 : BATTING SKILL FOR EXTRA BASE HIT} = \{(\text{launch angle avg} - 12.43) \times 0.05 + (\text{sweet spot percent} - 0.33) \times 20.06\} \times 10 + 50$$

$$\text{F7 : AGGRESIVE + EYE} = \{(\text{z swing percent} - 0.67) \times 16.31\} \times 10 + 50$$

특정 선수의 18개의 변수들을 입력하면 보다 직관적인 요인 별 점수가 반영된다. 이 점수는 20-80 스케일을 따르기 때문에 직관적으로 이해하기도 편하며 선수들의 장단점과 특징을 파악하기에 유용하다. 이렇게 파악한 정보를 통해 효율적인 선수 영입과 선수단 관리 계획에 활용할 수 있을 것이다.

Ⅲ. 결론

본 프로젝트는 야구에서 타자의 특정 능력을 평가하는 데에 사용되는 지표들이 평가하고자 하는 능력을 정확하게 나타내지 못하고 있다는 생각에서 시작됐다. 기존에 사용되는 여러가지 타격지표들을 가져와 요인분석을 실행했고, 그 결과로 나온 요인들이 타자의 어떤 능력을 나타내는지를 해석한 뒤 새로운 지표를 만들면 기존의 방식보다 정확하게 타자의 능력을 평가할 수 있다고 생각했다. 그리고 이렇게 추출된 요인들이 실제로 타자의 능력을 잘 설명한다고 볼 수 있는지 확인하기 위해 타자의 종합적인 타격 능력을 가장 잘 나타낸다고 평가받는 지표인 xwOBA를 공통 목적변수로 두고, 기존에 사용되던 타격 지표와 요인들을 각각 설명변수로 하는 두 회귀모형을 만들어서 비교해 보았다. 그 결과로 두 회귀모형의 설명력을 통해 추출된 요인들이 충분히 타자의 능력을 잘 설명한다는 걸 확인할 수 있었다. 최종적으로는 합리성이 확인된 이 요인들을 이용해서 타자의 특정 능력을 직관적으로 나타내는 새로운 지표를 만들어 활용하는 방안을 제시하였다.

요약하자면, 스카우팅 리포트에서 타율이나 장타율을 통해 단순하게 평가하는 방식 대신에, 추출한 요인들을 통해 더 구체적인 지표로 접근하는 방식을 제안했다. 이를 통해 여러 선수들의 특정 능력에 대한 비교와 평가를 더 객관적으로 할 수 있게 되었다는데 의의가 있다. 한편, 이번 프로젝트에서는 타자가 타석에서 보여줄 수 있는 능력 위주의 특정 타격지표만을 사용하였고 타격 관련 능력만을 수치화 하는데 그쳤지만, 더 많은 지표를 추가 사용한다면 타격 이외의 야수의 수비 능력이나 투수의 투구 능력을 포함한 선수의 전체적인 능력을 세분화해서 수치화 할 수 있을 것이라 생각한다. 향후 추가적인 검증 과정을 연구한다면 요인분석을 통해 만든 야구 지표의 범용성은 더욱 커질 것이라 기대한다.

IV. 참고문헌

1. FANGRAPHS (https://www.fangraphs.com/)

2. BASEBALL SAVANT (https://baseballsavant.mlb.com/)

V. 부록

```
> colSums(is.na(aa0518))
        last_name        first_name         player_id              year       b_total_pa
                0                 0                 0                 0                0
  b_total_pitches               xba              xwoba              xobp             xiso
                0                 0                 0                 0                0
 exit_velocity_avg  launch_angle_avg sweet_spot_percent barrel_batted_rate  z_swing_percent
                0                 0                 0                 0                0
 oz_swing_percent oz_contact_percent iz_contact_percent           n_bolts        hp_to_1b
                0                 0                 0              2289               63
     sprint_speed
                0
```

[Figure 1] <stats0525.csv> 데이터 결측치 확인

Season	Name	Batting	Base.Running
2018	Mookie Betts	62.4	6.9
2018	Mike Trout	64.2	5.0
2015	Bryce Harper	74.1	3.6
2015	Mike Trout	55.7	3.0
2017	Aaron Judge	61.7	0.0
2015	Josh Donaldson	44.7	4.0
2016	Mike Trout	57.3	9.6

[Figure 2] <FanGraphs_value.csv> (이하 행 생략)

Season	Name	HR	K.	BABIP
2018	Mookie Betts	32	14.8%	0.368
2018	Mike Trout	39	20.4%	0.346
2015	Bryce Harper	42	20.0%	0.369
2015	Mike Trout	41	23.2%	0.344
2017	Aaron Judge	52	30.7%	0.357
2015	Josh Donaldson	41	18.7%	0.314
2016	Mike Trout	29	20.1%	0.371

[Figure 3] <FanGraphs_wrc.csv> (이하 행 생략)

```
> factanal(aa0518[,1:18],7,rotation="promax",scores="Bartlett")

Call:
factanal(x = aa0518[, 1:18], factors = 7, scores = "Bartlett",     rotation = "promax")

Uniquenesses:
                xba             xobp_iso                 xiso    exit_velocity_avg     launch_angle_avg
              0.036                0.204                0.011                0.298                0.326
  sweet_spot_percent   barrel_batted_rate      z_swing_percent 1 - oz_swing_percent   oz_contact_percent
              0.135                0.058                0.074                0.219                0.270
  iz_contact_percent             n_bolts   1.1*max - hp_to_1b         sprint_speed                   HR
              0.191                0.726                0.141                0.045                0.289
              1-K                BABIP         ball_per_pa
              0.005                0.429                0.272


Loadings:
                     Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
xba                   0.410   0.619                   0.544   0.205
xobp_iso                                      0.935                   0.145
xiso                  0.965
exit_velocity_avg     0.774  -0.129  -0.140           0.273  -0.192
launch_angle_avg      0.267                          -0.643   0.389
sweet_spot_percent                                    0.130   0.946
barrel_batted_rate    0.874  -0.266
z_swing_percent                                                       0.910
1 - oz_swing_percent -0.114   0.160           0.915
oz_contact_percent   -0.101   0.793                                  -0.106
iz_contact_percent   -0.131   0.777                                  -0.233
n_bolts                                0.478
1.1*max - hp_to_1b                     0.920
sprint_speed                           0.959
HR                    0.919   0.168   0.107          -0.166
1-K                           1.024           0.133                   0.120
BABIP                -0.110  -0.188   0.131           0.696   0.274
ball_per_pa                  -0.307           0.488           0.134  -0.308

                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
SS loadings       3.459   2.958   2.060   1.986   1.340   1.249   1.037
Proportion Var    0.192   0.164   0.114   0.110   0.074   0.069   0.058
Cumulative Var    0.192   0.356   0.471   0.581   0.656   0.725   0.783
Factor Correlations:
        Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
Factor1  1.0000  0.2833  0.0451  0.3118  0.1973  0.0213  0.1254
Factor2  0.2833  1.0000 -0.1891  0.2872 -0.0433  0.4162  0.1803
Factor3  0.0451 -0.1891  1.0000 -0.0974 -0.2219 -0.0611  0.0599
Factor4  0.3118  0.2872 -0.0974  1.0000  0.1558  0.1826 -0.4544
Factor5  0.1973 -0.0433 -0.2219  0.1558  1.0000  0.0208 -0.0357
Factor6  0.0213  0.4162 -0.0611  0.1826  0.0208  1.0000  0.0386
Factor7  0.1254  0.1803  0.0599 -0.4544 -0.0357  0.0386  1.0000

Test of the hypothesis that 7 factors are sufficient.
The chi square statistic is 782.79 on 48 degrees of freedom.
The p-value is 1.83e-133
```

[Figure 4] 요인분석 시행 결과

power	contact	speed	patient+eye	batting skill	batting skill for extra base hit	aggresive+eye
0.803849088	-0.62605458	0.947006989	0.237481910	0.7582603221	-0.192387330	-0.457041678
0.042903821	-2.07554996	0.911963463	1.442859979	-0.0981247943	-1.063308641	0.173054541
-0.323825427	0.61486736	1.019801754	0.417835582	0.2473680278	0.180545246	0.247411774
-0.695459667	0.06408109	1.040613657	0.265722879	-0.5367259380	-1.450345698	0.764620220
0.228503648	0.18057865	1.097677154	1.784071974	-0.3934318558	-0.700349031	-0.584691045
0.670619519	0.15310378	0.908124980	2.097427220	-0.1802318644	0.295430288	-0.154169071
0.049641565	-1.43394922	0.394439975	1.354293102	-1.1161909021	-1.723650631	-0.750772166

[Figure 5] Factor Score 값 (이하 행 생략)

[R코드]

```
##### 0518
rm(list=ls())
setwd("C:/Users/vz050/OneDrive/바탕 화면/다변량/data/4")
##install.packages("DMwR")
##install.packages("GPArotation")
library(corrplot)
library(DMwR2)
library(tm)
library(GPArotation)
library(caret)
library(dplyr)
library(car)
aa0518<-read.csv("stats0525.csv", encoding = 'UTF-8')


summary(aa0518)
str(aa0518)
aa05188<-aa0518[,-22]
### 0528/ 1열 열이름 변경
colnames(aa0518)[1]<-"last_name"
####
aa0518[is.na(aa0518$hp_to_1b),]
aa0518[is.na(aa0518$n_bolts),"n_bolts"]<-0
aa0518<-aa0518[,c(1:7,9:21,8)]
aa0518[,18:20]<-knnImputation(aa0518[,18:20])
str(aa0518)
```

```
######0520 추가
aa0518_value<-read.csv("FanGraphs_value.csv")[,c(1,2,4,5)]
str(aa0518_value)
aa0518_wrc<-read.csv("FanGraphs_wrc.csv")[,c(1,2,6,11,13)]
str(aa0518_wrc)

### 0527/ name
aa0518_y<-merge(aa0518_value,aa0518_wrc,by=c("Season","Name"))
aa0518_y$Name<-gsub("\\.","",aa0518_y$Name)
aa0518$Name<-gsub("III","",aa0518$Name)
aa0518_y$Name<-gsub("II","",aa0518_y$Name)
aa0518_y$Name<-gsub(" ","",aa0518_y$Name)
aa0518$Name<-paste(gsub(" ","",aa0518[,2]),aa0518[,1])
aa0518$Name<-gsub("\\.","",aa0518$Name)
aa0518$Name<-gsub("III","",aa0518$Name)
aa0518$Name<-gsub("II","",aa0518$Name)
aa0518$Name<-gsub(" ","",aa0518$Name)

###
colnames(aa0518)[4]<-"Season"
str(aa0518_y)


aa0518<-merge(aa0518,aa0518_y,by=c("Name","Season"))
######
aa0518$ball_per_pa<-aa0518$b_total_pitches/aa0518$b_total_pa
aa0518$xobp<-aa0518$xobp-aa0518$xba
colnames(aa0518)[9]<-"xobp_iso"
aa0518[,13:18]<-aa0518[,13:18]/100


###0520/ name(first,last), id, pa, total pitch 3:7
###0527 / %in%
aa0518<-aa0518[(aa0518$Season %in% c(2015,2016,2017,2018,2019,2021))
                &(aa0518$b_total_pa>=250),-c(3:7)]
str(aa0518)
summary(aa0518)
head(aa0518)
savedata<-aa0518
#####
```

```
aa0518<-aa0518[,c(3:16,20:23,1:2,17:19)]
### oz_swing 9 1-v
aa0518$oz_swing_percent<-1-aa0518$oz_swing_percent
colnames(aa0518)[9]<-paste("1 -",colnames(aa0518)[9])
### hp_to_1b 13 1.1 * max - v
aa0518$hp_to_1b<-max(aa0518$hp_to_1b)*1.1 - aa0518$hp_to_1b
### 0528/ 1-k로 변경
colnames(aa0518)[c(13,16)]<-c("1.1*max - hp_to_1b","1-K")
aa0518[,16]<-1-(as.numeric(gsub("%","",aa0518[,16]))/100)
### factanal / 7
### 기존 방식의 요인들 공선성이 높게 나와서 임의로 회전방식 추가

faa<-factanal(aa0518[,1:18],7,rotation="promax",scores="Bartlett")

matrix0520_222<-as.data.frame(cbind(faa$scores,aa0518[,19:23]))
colnames(matrix0520_222)[1:7]<-
c("power","contact","speed","patient+eye","batting skill",
                                    "batting skill for extra base
hit","aggresive+eye")
model0520_222<-lm(xwoba~power+`patient+eye`+speed+contact+`batting skill`+
                        `aggresive+eye`+`batting skill for extra base
hit`,data=matrix0520_222[,c(1:7,10)])
summary(model0520_222)

par(mfrow=c(2,2))
plot(model0520_222)
summary(model0520_222)
#정규성검정
par(mfrow=c(1,1))
hist(rstandard(model0520_222))
shapiro.test(rstandard(model0520_222)) #정규성 위배

###################
#스케일링 코드
matrix0520_23<-matrix0520_222
meansd<-cbind(apply(matrix0520_23[,c(1:7,10)],2,mean),
apply(matrix0520_23[,c(1:7,10)],2,sd))



for(i in c(1:7)){
```

```
  matrix0520_23[,i]<-50 + 10*(matrix0520_23[,i]-meansd[i,1])/meansd[i,2]
}
matrix0520_23[,10]<-50 + 10*(matrix0520_23[,10]-meansd[8,1])/meansd[8,2]

summary(matrix0520_23)
str(matrix0520_23)
########################
#스케일링한 모형 : 요인(설명변수)
model0520_23<-lm(xwoba~power+`patient+eye`+speed+contact+`batting skill`+
                   `aggresive+eye`+`batting skill for extra base
hit` ,data=matrix0520_23)
summary(matrix0520_23)
summary(model0520_23) #SPEED는 유의하지 않음
#plot : 회귀 가정확인
par(mfrow=c(2,2))
plot(model0520_23)
##정규성 확인
shapiro.test(rstandard(model0520_23)) #H0기각 정규성 만족안됨.

##정규성가정 위배 : lamda가 0.5에 근접
powerTransform(matrix0520_23$xwoba) #0.4265044 ->sqrt(x)변환
summary(powerTransform(matrix0520_23$xwoba))

par(mfrow=c(1,1))

str(matrix0520_23)
##등분산성
spreadLevelPlot(model0520_23) #0.9417023이므로 H0기각 -> 등분산성 만족

#다중공선성 : 문제없음
corrplot( cor(matrix0520_23[,c( 1:7)]) ,method="number")
vif(model0520_23)
########################
###SPEED는 유의하지 않음->SPEED빼고 회귀 모형 다시
model0520_24<-lm(xwoba~power+`patient+eye`+contact+`batting skill`+
                   `aggresive+eye`+`batting skill for extra base
hit` ,data=matrix0520_23)
summary(model0520_24)

########################################
```

```
##sqrt(x)변환
#df변환 -> 별로임
matrix0520_23_plus=matrix0520_23%>%
  mutate(sqrt_xwoba=sqrt(xwoba))
#모델변환
model0520_23_plus<-
lm(sqrt_xwoba~power+`patient+eye`+speed+contact+`batting skill`+
                    `aggresive+eye`+`batting skill for extra base
hit` ,data=matrix0520_23_plus)
shapiro.test(rstandard(model0520_23_plus))
powerTransform(matrix0520_23_plus$sqrt_xwoba)
#plot
par(mfrow=c(2,2))
plot(model0520_23_plus)
############################
##log(x)변환 ->별로임
#df변환
matrix0520_23_plus=matrix0520_23%>%
  mutate(log_xwoba=log(xwoba))
#모델변환
model0520_23_plus2<-
lm(log_xwoba~power+`patient+eye`+speed+contact+`batting skill`+
                    `aggresive+eye`+`batting skill for extra base
hit` ,data=matrix0520_23_plus)
shapiro.test(rstandard(model0520_23_plus2))
powerTransform(matrix0520_23_plus$sqrt_xwoba)
#plot
par(mfrow=c(2,2))
plot(model0520_23_plus2)

###변환이 별로 효과가 없으며,
###결과적으로 요인들을 설명변수로 할 때는 그래프 상에서 정규성 보이니까 변환x



############################################################
##설명변수 : 기존 변수
aa0521<-aa0518[,c(1,13,15,19:21)]
summary(aa0521)
aa0522<-aa0521
#xba,`1.1*max - hp_to_1b`,HR,xwoba
```

```
model_aa0522<-lm(aa0522$xwoba~.,data=aa0522[,c(1:3,6)])
summary(model_aa0522)

a<-lm(xwoba~xba+`1.1*max - hp_to_1b`+HR,data=aa0522)
summary(a)
par(mfrow=c(2,2))
plot(a)


##등분산성
spreadLevelPlot(a) #0.8082565이므로 H0기각 -> 등분산성 만족

##정규성 검정
shapiro.test(aa0522$xwoba) #h0기각 ->정규성 위배

powerTransform(aa0522$xwoba) #-0.05376256 ->log변환
summary(powerTransform(aa0522$xwoba))
##
corrplot( cor(matrix0520_23[,c( 1:7)]),method="number")
vif(a)

###기존 지표들은 정규성 위배됐으므로 변환 필요
#기존지표 log변환
aa0530<-aa0522
aa0530$logoba<-log(aa0530$xwoba)
b<-lm(logoba~xba+`1.1*max - hp_to_1b`+HR,data=aa0530)
par(mfrow=c(2,2))
plot(b)
#정규성 확보됐는지 확인
shapiro.test(aa0530$logoba) #H0채택 : 정규성 확보

summary(b)

##############################################
####6/1 추가코드 : 행렬연산
meansd2<-cbind(apply(aa0518[,1:18],2,mean), apply(aa0518[,1:18],2,sd))
d1<-faa$loadings[1:18,1:7]
p1<-diag(faa$uniquenesses)
p2<-solve(p1)
p2 %*% d1 %*% solve(t(d1) %*% p2 %*% d1)
```

```
#연산
d2_inv<-solve(diag(meansd2[,2])) #18*18
mat_s<-p2 %*% d1 %*% solve(t(d1) %*% p2 %*% d1) #18*7
d1_inv<-solve(diag(meansd[,2][1:7])) #7*7
mu_2<-meansd2[,1] #1*18
mu_1<-meansd[,1][1:7] #1*7
mat_t<-d2_inv %*% mat_s %*% d1_inv #18*7
```