'잠실 구장 이적 효과'는 실재하는가?

박찬호와 김재환의 2026 성적 예측

야구 2조 : 김동혁, 김채연, 이지훈

목차

Introduction

분석 배경

2026 년 스토브리그, KBO 팬들의 이목은 두 명의 이적생에게 쏠렸다. 바로 두산 베어스로 이적한 박찬호와 SSG 랜더스로 둥지를 옮긴 김재환이다. 두 선수 모두 팀의 핵심 전력으로 기대받고 있지만, 팬들의 시선에는 기대와 우려가 공존하고 있다. 작년 9 위라는 충격적인 성적표를 받은 두산에게 박찬호는 절실한 보강 카드였으나, 과연 그가 '투수 친화적'인 잠실 구장에서도 기존의 퍼포먼스를 유지하며 주전 역할을 해낼 수 있을지는 미지수이다(박찬호의 잠실 원정 OPS 지표에 대한 우려는 여전함). 반면, "잠실이 너무 힘들다"며 탈(脫)잠실을 선택한 김재환의 경우, 타자 친화적인 SSG 랜더스필드에서 과연 유의미한 성적 향상 효과, 이른바 '탈잠실 효과'를 누릴 수 있을지가 초미의 관심사이다.

두산과 SSG 의 팬으로서, 그리고 야구 데이터를 사랑하는 분석가로서 본 칼럼은 단순한 예상을 넘어 이들의 이적이 가져올 변화를 보다 과학적인 방법으로 분석해보고자 한다.

분석 방법론 : 인과추론의 도입

우리는 데이터 분석을 진행할 때 흔히 "A 가 변하면 B 가 특정 방향으로 변화한다"라고 말한다. 하지만 이는 단순한 **상관관계(Correlation)**에 기반한 추측일 때가 많다.

- **상관관계 분석:** 단순히 데이터의 경향성을 관측함. "A 가 변할 때 B 도 변하더라"는 관측이지만, 이것이 반드시 A 때문에 B 가 변했다는 것을 보장하지 않는다. 역으로 B 의 변화가 A 에 영향을 주었을 수도 있고, 제 3 의 요인이 개입했을 수도 있다.
  머신러닝(ML)은 주로 이러한 상관성을 학습하여 미래를 예측하는 도구다.
- **인과관계 분석:** "A 가 변하면 B 는 **무조건** 변한다"는 과학적 인과성을 규명함. 이는 정책의 효과를 분석하거나 특정 처치(Treatment)의 효용을 판단할 때 사용되는 강력한 도구다.

제대로 된 인과분석을 위해서는 원인과 결과 외의 모든 '교란 변수(Confounding Variable)'를 통제해야 한다. 하지만 현실의 야구 경기에서 날씨, 팀 분위기, 상대 투수 등 모든 변수를 완벽히 통제하는 것은 불가능에 가깝다. 따라서 본 분석에서는 **준실험적 설계(Quasi-experimental Design)**를 통해 간접적으로 인과효과를 추정하고자 한다. 기존의 상관관계 분석과는 달리 정확한 통계량 산출이 어렵기에, 도메인 지식을 적극적으로 반영한 정성적 분석이 필수적으로 요구되는 영역이다.

분석 도구: 합성 통제법 (Synthetic Control)

본 칼럼에서 사용할 핵심 분석 도구는 **합성 통제법(Synthetic Control, 이하 SC)**이다. SC 는 **"만약 그 선수가 이적하지 않았다면 어땠을까?"**라는 반사실(Counterfactual)을 데이터로 구현하여, 실제 결과와 비교함으로써 순수한 인과효과(이적 효과)를 추정하는 방법론이다.

- **분석의 원리:** 분석 대상인 선수(처치 집단)와 유사한 특성을 가진 다른 선수들(통제 집단, 12~14 명 규모의 합성 풀)을 가중 조합하여, 가상의 **'합성된 선수(Synthetic Unit)'**를 생성한다.
- **적용:** 예를 들어, 실제 이적을 한 박해민 선수(처치 집단)의 성적과, 이적하지 않은 상황을 가정한 '합성 박해민'의 성적을 비교함으로써 구장 변화나 이적이 성적에 미친 순수 효과를 분리해내는 것이다.

이어지는 본론에서는 박해민, 서건창, 최주환, 오재일 등 주요 이적 사례를 처치 집단으로 설정하고, SC 방법론을 적용하여 구장 효과와 이적의 인과성을 면밀히 파헤쳐 보고자 한다. 이를 통해 박찬호와 김재환의 2026 시즌을 미리 가늠해볼 수 있는 유의미한 시사점을 도출하는 것이 목적이다.

Chapter 1. SC(합성통제법)

데이터 수집 및 전처리

Name	OPS	wRC+	WAR	K%	BABIP	IsoP
구자욱	0.863	121.2	3.3	17.8	0.355	0.171
손아섭	0.908	139.7	5.02	9.2	0.373	0.141
오지환	0.823	125.2	6.41	19.6	0.365	0.161
전준우	0.829	112.3	2.72	12.6	0.285	0.203
정수빈	0.764	113	2.94	10	0.325	0.098
황재균	0.882	129.5	5.36	16.3	0.347	0.2
홍창기	0.828	138.7	4.59	17.2	0.342	0.138
김혜성	0.744	100	2.88	17	0.336	0.114
박민우	0.877	131	5.45	9.1	0.363	0.13
김선빈	0.809	121.1	2.51	12.3	0.375	0.073
박성한	0.644	65.4	0.7	18.8	0.29	0.081
허경민	0.824	131.4	4.13	5.7	0.337	0.11
최지훈	0.644	65.4	0.7	15.4	0.307	0.068
강민호	0.836	106.8	2.66	13.7	0.291	0.2

박해민이 만일 이적하지 않았다면?을 알아보자.

위 엑셀 파일은 삼성에 잔류한 가상의 박해민을 합성하기 위한 재료들이다. 박해민과 비슷한 유형의 컨택 위주 타자들, 베테랑 선수들로 총 14명의 선수들의 데이터를 수집하였다. 2020~2023년도 OPS부터 IsoP까지 6개의 스탯에 대한 데이터가 담겨 있다. 이적 전 박해민의 성적을 완벽히 모방하기 위해 최적의 비율로 2020, 2021년도 데이터를 섞을 예정이며 모든 비율의 합은 1, 마이너스 비율은 없는 것을 조건으로 한다.

가상의 서건창 합성을 위해 역시 같은 재료를 사용한다.

Name	OPS	wRC+	WAR	K%	BABIP	IsoP
구자욱	0.771	104.1	2.54	16.7	0.296	0.177
김현수	0.807	130.5	3.15	8.7	0.316	0.133
오지환	0.717	107.9	5.26	20.7	0.311	0.126
전준우	0.84	126.4	3.05	11.7	0.309	0.18
최정	0.918	154.4	5.98	15.2	0.303	0.227
최형우	0.898	152.8	4.9	13.9	0.327	0.185
황재균	0.824	127	4.88	14	0.296	0.184
강민호	0.72	91.6	2.13	19.3	0.262	0.171
노시환	0.501	27.7	-1.36	37.5	0.305	0.074
한유섬	0.771	119.1	1.9	19.9	0.321	0.131
이대호	0.79	115.8	1.52	11.8	0.297	0.15
이정후	0.842	141.5	4.64	6.3	0.352	0.12

오재일과 최주환은 2019년~2022년 OPS형 히터, 거포형 타자들로 구성된 총 12명의 데이터를 이용하여 합성을 진행한다. 2019년~2020년 데이터를 섞어서 합성 진행할 예정이다.

Year	K%	BABIP	IsoP	WAR	OPS	wRC+
2019	14.1	0.275	0.089	2.17	0.646	76.5
2020	14.2	0.324	0.125	3.36	0.76	93.5
2021	15.1	0.34	0.086	3.45	0.76	109.6
2022	13.4	0.332	0.079	4.3	0.715	107
2023	13.3	0.325	0.074	3.46	0.707	103.4

박해민의 실제 성적

Year	K%	BABIP	IsoP	WAR	OPS	wRC+
2019	10.3	0.333	0.083	2.44	0.756	117
2020	9.7	0.299	0.109	2.87	0.776	117.3
2021	10.2	0.315	0.095	3.3	0.79	108
2022	17.8	0.27	0.082	-0.11	0.605	72.7
2023	11.1	0.222	0.082	-0.85	0.542	46.1

서건창의 실제 성적

서건창의 경우, 21년 시즌 중후반에 lg로 트레이드 되었기 때문에 키움의 성적을 베이스로 144경기 기준 환산하여 전처리하였다.

Year	K%	BABIP	IsoP	WAR	OPS	wRC+
2018	25.4	0.323	0.26	3.59	0.912	131.5
2019	18.7	0.331	0.202	4.18	0.864	148.2
2020	17.2	0.359	0.17	2.85	0.872	141.8
2021	21.9	0.319	0.227	2.81	0.878	132
2022	24.8	0.325	0.223	2.53	0.836	126.9

오재일의 실제 성적

Year	K%	BABIP	IsoP	WAR	OPS	wRC+
2018	15.1	0.355	0.249	5.9	0.979	150.5
2019	10.8	0.292	0.088	0.73	0.697	95.3
2020	11.5	0.32	0.167	4.66	0.839	127.9
2021	17.9	0.281	0.173	2.44	0.782	112.4
2022	16.5	0.229	0.151	-0.28	0.65	77.5

최주환의 실제 성적

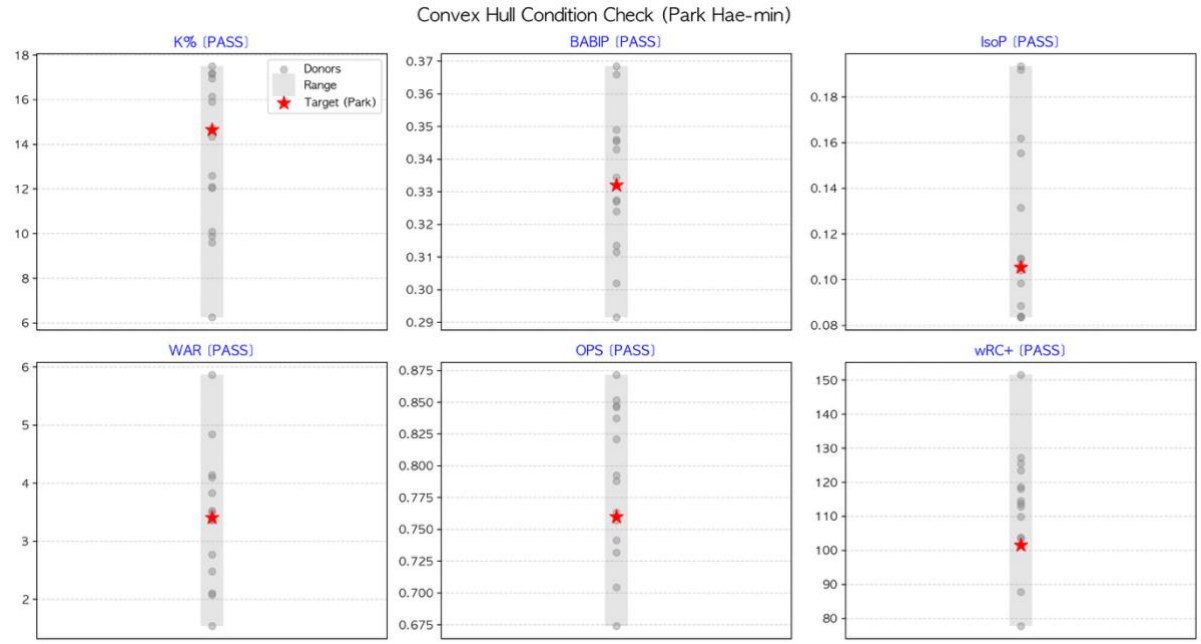
합성 선수들의 성적은 이적 전 실제 선수의 2년치 스탯과 동일해야하고 이적 후 2년동안 성적 차이가 발생하면 그만큼이 구장에 따른 인과효과라고 추정해볼 수 있다.

SC 분석 결과

본격적인 분석에 앞서 박해민, 서건창, 최주환, 오재일이 합성 풀에 있는 선수들로 합성이 가능한지 확인하기 위해서는 convex hull condition을 체크해야 한다.

- **Convex hull condition** : 박해민의 성적이 비교군 선수(합성 풀 선수)들의 범위 안에 있어야 적절한 가중치를 선택하여 합성을 진행할 수 있다.

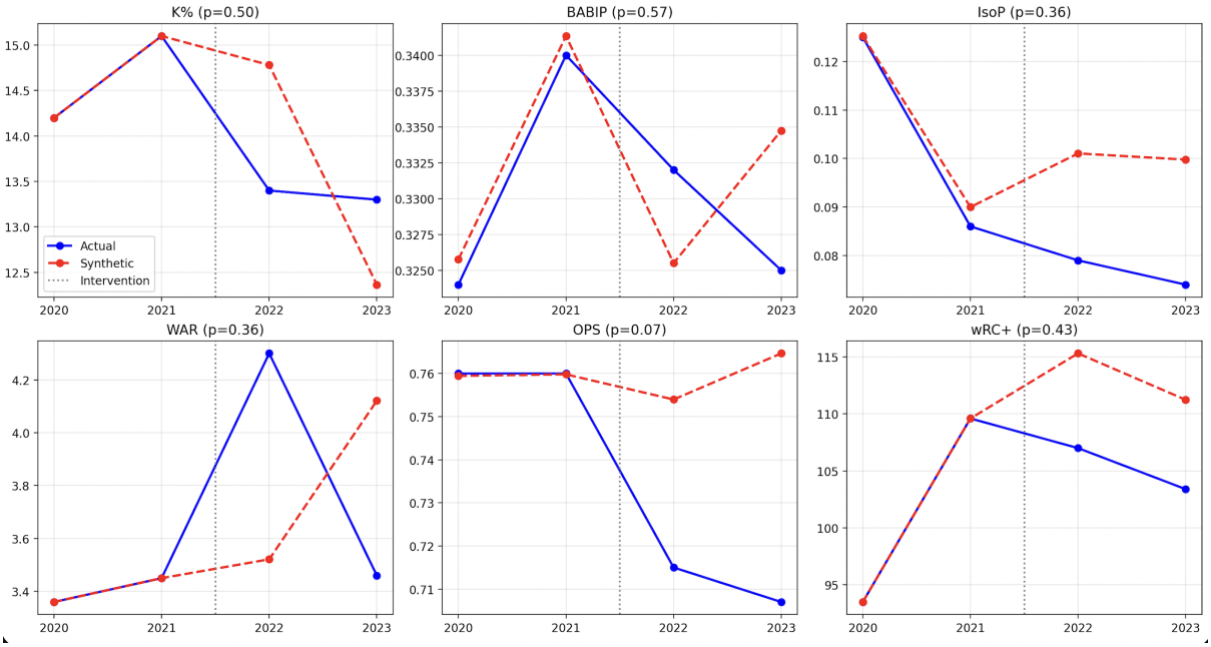
<박해민>



회색 점들은 합성 풀 선수들의 성적 분포, 빨간 별은 박해민의 성적을 의미하며 빨간 별이 회색 점들의 성적 범위 안에 존재하므로 합성이 가능하다. 현재 빨간 별이 회색 점들의 범위 사이에 위치하기 때문에 박해민은 SC가 가능하다.

본격적인 박해민의 SC 분석 후 결과는 아래와 같다.

Figure 1: Actual vs Synthetic Paths (Park Hae-min)



```
================================================
 >>> Final Summary Statistics (Park Hae-min) <<<
================================================
  Variable  P-value  RMSPE Ratio  Gap 2022  Gap 2023
0       K%    0.500          inf    -1.381     0.935
1    BABIP    0.571         5.30     0.007    -0.010
2     IsoP    0.357         8.50    -0.022    -0.026
3      WAR    0.357          inf     0.779    -0.661
4      OPS    0.071       119.61    -0.039    -0.058
5     wRC+    0.429          inf    -8.311    -7.838
================================================
```

박해민의 경우 이적 전(2020년, 2021년) 합성 박해민과 실제 박해민의 성적이 거의 일치하며 합성이 잘 이루어진 것을 확인할 수 있었다.

- RMSPE Ratio : (이적 후 오차) / (이적 전 오차). 이 값이 클수록 이적 후 값의 차이가 커짐 즉, 이적으로 인한 값의 변동이 크다는 것을 의미한다.
- Gap 2022 : 이적 직후 가상 박해민과 실제 박해민의 성적 차이다.

그 결과 BABIP, WAR은 증가, K%, wRC+, IsoP, OPS는 감소한 것을 볼 수 있다.

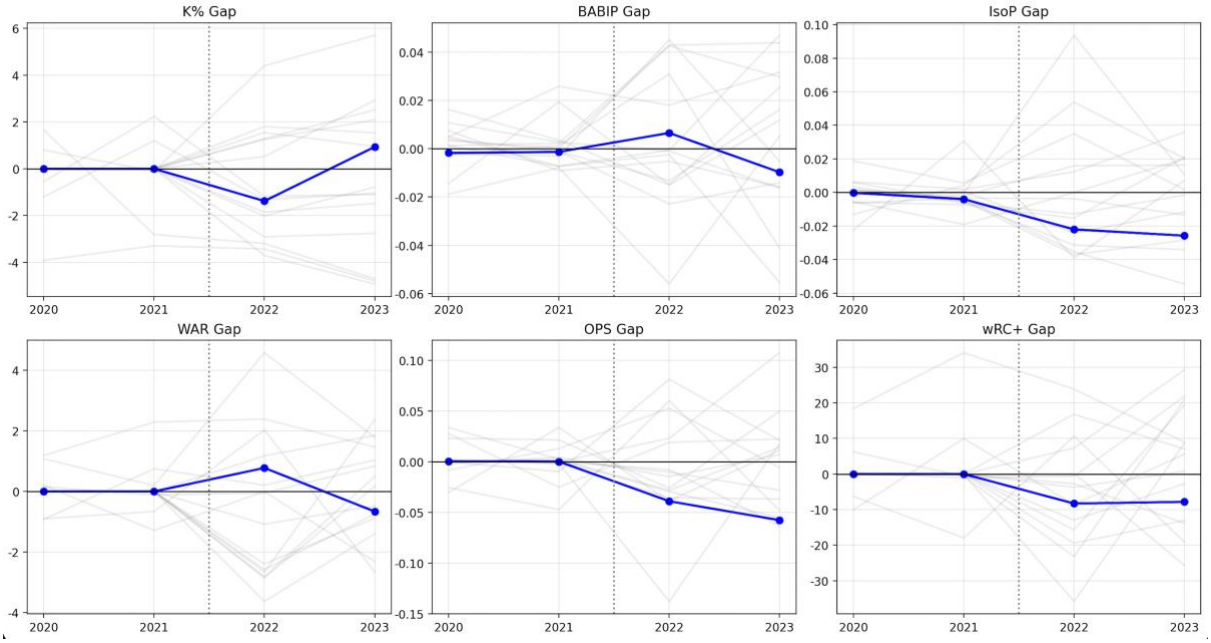
구장 크기가 증가함에 따라 BABIP 역시 증가하고 박해민 특유의 주루 툴과 수비 툴에서 잠실 구장에서 긍정적으로 작용해 증가한 것으로 해석해 볼 수 있다. 전체적인 타격 생산력은 떨어졌기 때문에 wRC+, IsoP, OPS는 감소한 것으로 확인된다. 그러나 이것만 가지고 구장의 변화에 따른 인과효과임을 주장하기에는 통계적인 근거가 부족하다. 이에 따라 추가적인 검정 통계량 체크 및 테스트 진행하였다. 이적의 효과가 통계적으로 우연인지 아닌지 검증하기 위해 합성 재료로 사용된 14명의 선수들에 대해 같은 합성통제법 분석을 진행하였다.

P-value = (박해민보다 더 변화가 심한 가짜 선수들의 수) / (전체 실험 참가자 수)

P-value값이 작을수록 박해민의 이적 후 성적 변화가 독보적이고 우연이 아님을 시사한다. 위의 경우, OPS가 0.07의 유의확률을 가지므로 확실한 인과효과임을 알 수 있다. 다만, kbo 특성상 꾸준히 많은 타석을 소화하는 선수가 적고 표본 자체가 적어 전체 실험 참가자 수가 적기 때문에 단순히 p-value의 수치만으로 인과효과의 통계적 유의성을 측정하는데에는 한계가 있음. 따라서 추가적인 플라시보 테스트를 진행하였다.

- 플라시보 테스트 : 도출된 인과효과가 단순한 우연인지 아닌지 확인하기 위해 실제로는 이적하지 않은 선수들에게도 동일한 분석을 적용하여 비교해보는 검증 절차이다.

Figure 2: Gap Analysis & Placebo Tests



파란색 선은 박해민, 나머지 회색 선은 control unit(14명의 재료 선수들)에 대해서도 동일한 합성통제법 적용했을 때의 결과다. 파란색 선이 회색선보다 양 혹은 음의 방향으로 많이 벗어나 있으면 통계적으로 유의하다고 할 수 있다.

p-value와 마찬가지로 OPS gap 부분에서 파란색 실선이 음의 방향에 대해 두 번째로 큰 변화를 보이고 있다.

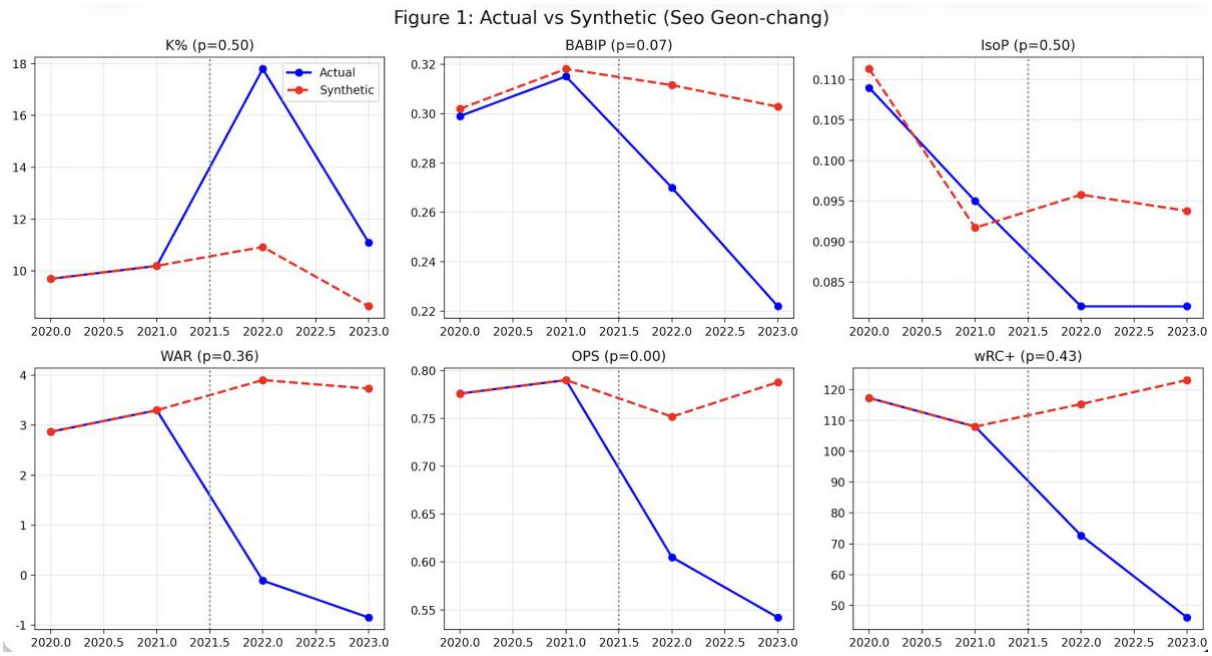
결론적으로 박해민의 경우, lg로 이적함에 따라 OPS 부분 4푼의 하락을 겪었고 이는 순수 구장에 따른 인과효과라고 볼 수 있다. 다만, 아쉽게도 OPS를 제외한 나머지 스탯에 대해서는 추가적인 인과효과를 측정할 수 없었다.

박해민 개별에 대한 인과효과임으로 박찬호에게 완전히 대입할 수는 없겠지만, 타격보다는 양호한 수비, 주루 능력을 통해 높은 war 스탯을 보이고 컨택과 출루에 집중하는 스타일이라는 점에서 두 사람은 유사하기 때문에 박찬호 역시 두산 이적에 따라 OPS의 하락(대략 3~5푼 정도)을 예상해볼 수 있다.

같은 방식으로 서건창, 오재일, 최주환에 대한 결과는 아래와 같다.

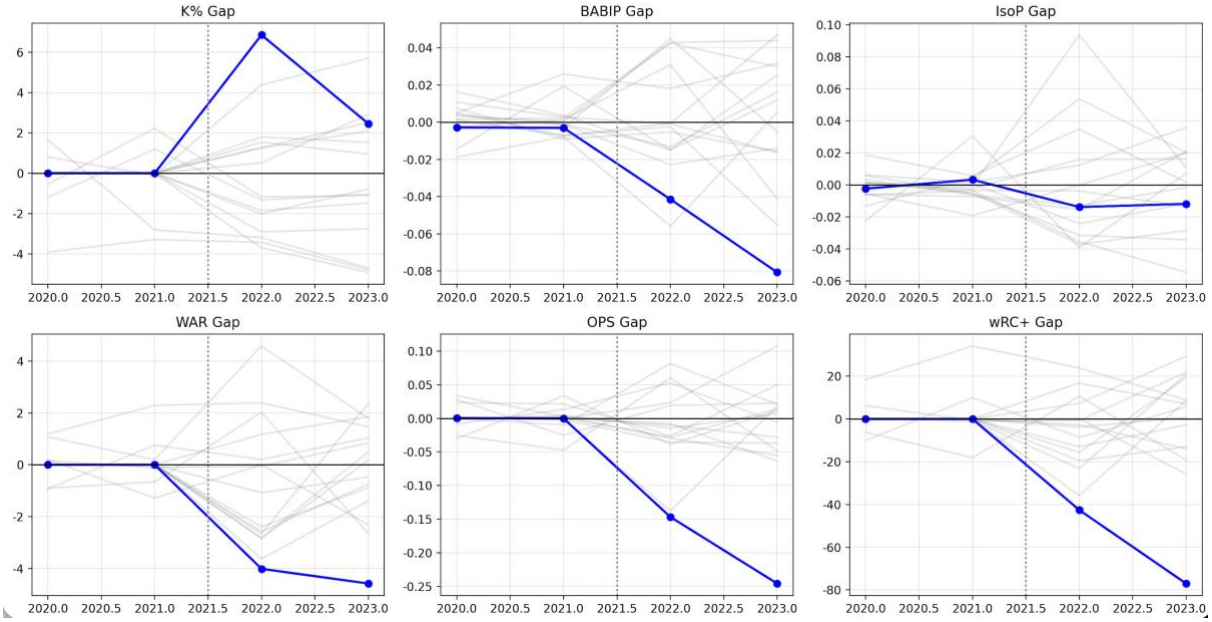
<서건창>

Convex hull condition 만족하므로 합성이 가능하다.



SC 결과 전체적인 스탯에서 뚜렷한 하락세를 관측할 수 있다. 이적에 따른 구장 변화(입잠실)가 전체적인 성적 하락에 영향을 끼친 것으로 확인된다.

Figure 2: Gap Analysis with Placebos (Seo Geon-chang)



IsoP를 제외하고 모든 스탯에 대해 파란색 실선이 회색 선들보다 양 or 음의 방향으로 더 치솟은 것을 확인할 수 있다. 이 중에서 박해민과 마찬가지로 OPS의 변화에 대해 강력한 통계적 근거를 찾을 수 있는데, 플라시보 테스트 결과를 보더라도 극단적으로 음의 방향으로 파란색 실선이 뻗어있고 p-value 역시 0에 수렴한다는 점에서 컨택형 타자들이 잠실로 오면 OPS가 하락하는 것은 인과적 효과임을 확인할 수 있었다. 다만, 서건창의 경우 21년부터 시작해서 심각한 부진에 빠진 시기라 이를 단순히 구장의 인과효과라고 일반화하기에는 무리가 있다고 판단하였다.(도메인 지식에 의거, 이상치로 간주하였음 → OPS에 대한 해석만 진행)

```
=== Seo Geon-chang Summary ===
  Variable  P-value  RMSPE Ratio  Gap 22  Gap 23
0       K%    0.500          inf   6.871   2.463
1    BABIP    0.071        21.47  -0.041  -0.081
2     IsoP    0.500         4.50  -0.014  -0.012
3      WAR    0.357          inf  -4.015  -4.583
4      OPS    0.000      1242.68  -0.147  -0.246
5     wRC+    0.429    904734.49 -42.613 -77.000

Process finished with exit code 0
```

<오재일>

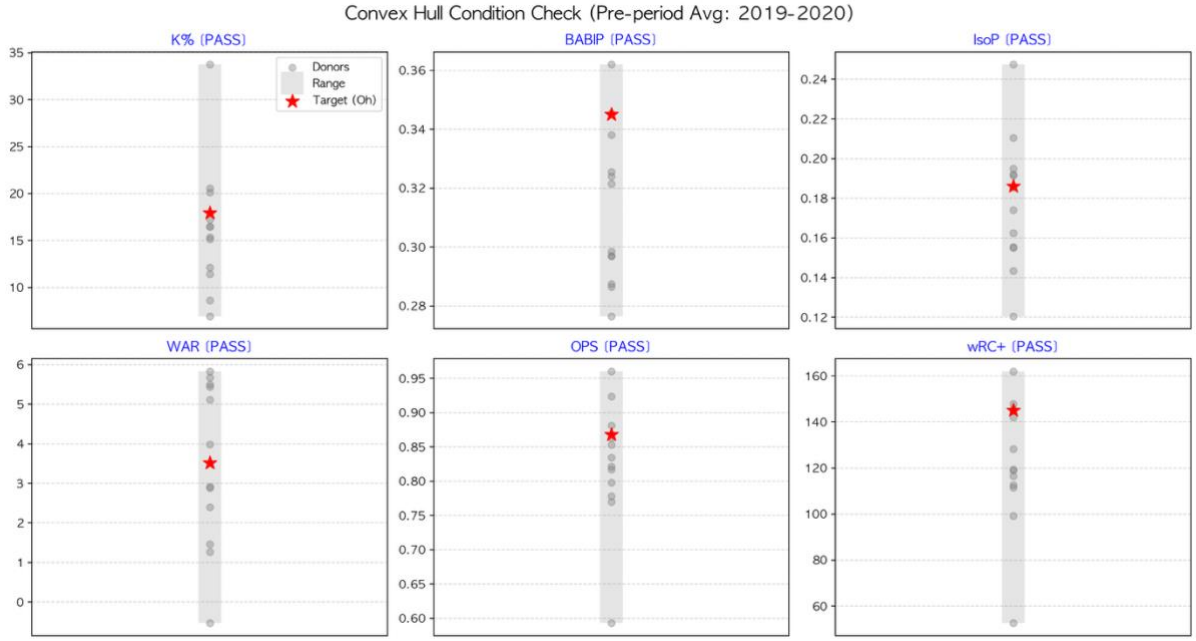


Figure 1: Actual vs Synthetic (Oh Jae-il)

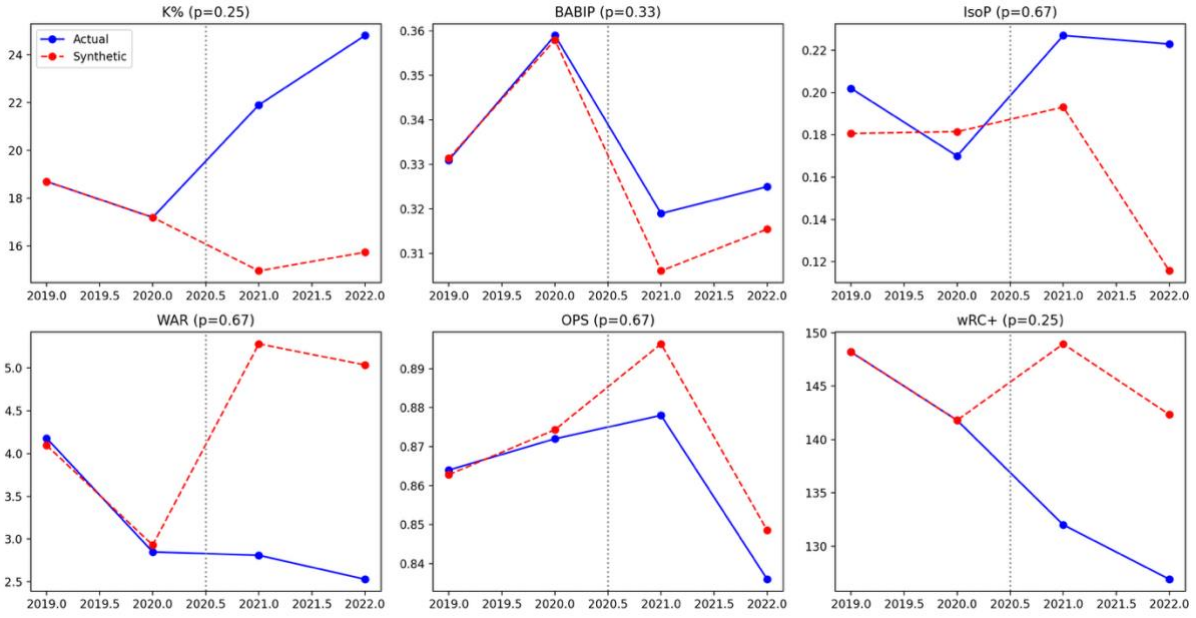
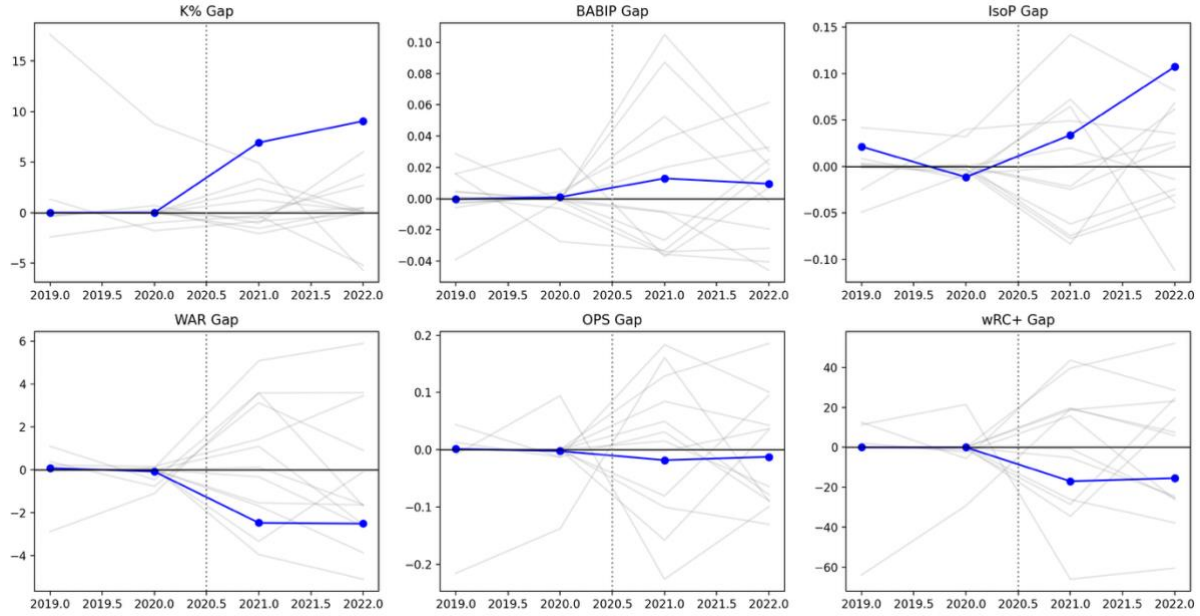


Figure 2: Gap Analysis (Oh Jae-il)

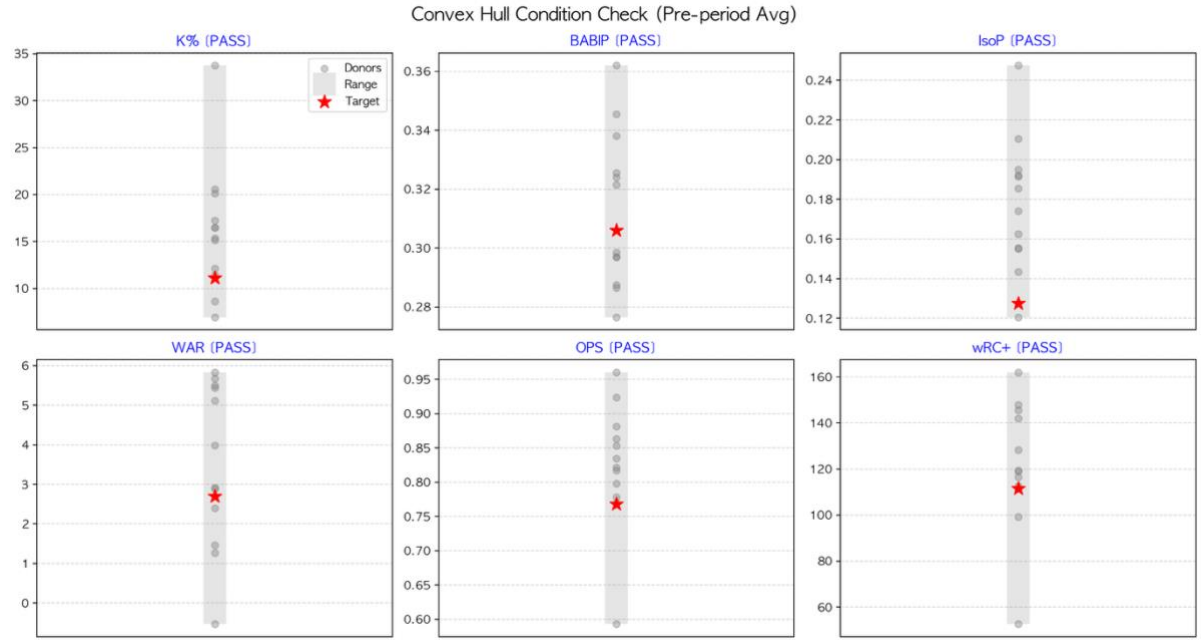



```
=== Oh Jae-il Summary ===
  Variable  P-value  RMSPE Ratio  Gap 21  Gap 22
0       K%    0.250          inf   6.934   9.053
1    BABIP    0.333        14.87   0.013   0.010
2     IsoP    0.667         4.64   0.034   0.107
3      WAR    0.667        30.67  -2.472  -2.506
4      OPS    0.667         8.47  -0.018  -0.013
5     wRC+    0.250          inf -16.959 -15.436
```

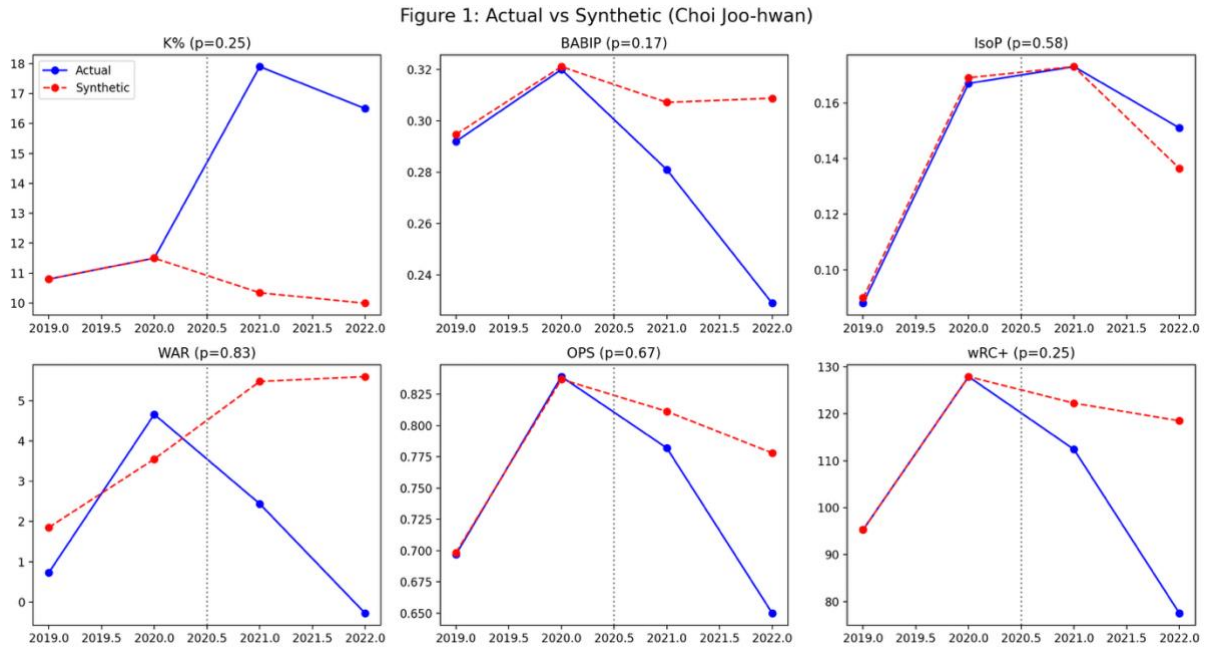
오재일과 최주환의 경우 control unit(합성에 사용된 재료)이 12명의 선수로 특히 더 적어서 p-value만으로 인과효과의 유의성을 검정하는 것에 한계가 있었고 플라시보 테스트 결과에 더 중점을 두어 해석을 진행하였다.

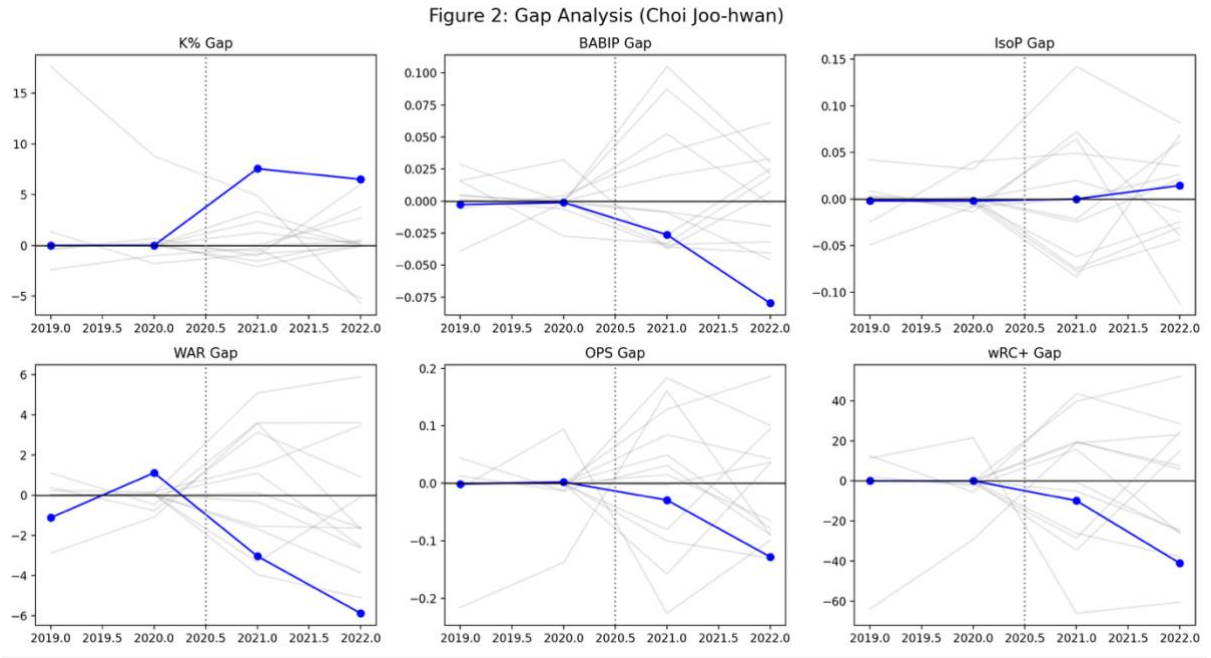
오재일은 순장타율은 증가하였지만, 그 외에 지표는 부정적인 방향으로 변화하였다. 즉, 구장 이적에 따라 홈런 등 장타의 개수는 증가했을지라도 구단의 승리 기여도 자체는 증가했다고 보기는 어려웠다. 스탯들 중에서 플라시보 테스트 결과, K%에서 유의한 인과효과를 관측할 수 있었고 탈잠실에 따라 오히려 삼진율이 약 7% 가량 증가한 것을 확인할 수 있었다.

<최주환>



Convex hull condition을 만족한다.






```
=== Choi Joo-hwan Summary ===
  Variable  P-value  RMSPE Ratio  Gap 21  Gap 22
0       K%    0.250          inf   7.557   6.509
1    BABIP    0.167        28.40  -0.026  -0.080
2     IsoP    0.583         5.18   0.000   0.015
3      WAR    0.833         4.19  -3.037  -5.876
4      OPS    0.667        51.59  -0.029  -0.128
5     wRC+    0.250          inf  -9.855 -41.005
```

최주환 역시 삼진율에서 유의미한 상승세를 관측할 수 있었다.(약 7.5%)

최종 결론

컨택형 : 잠실 구장으로 이적 시 OPS가 하락하는 인과관계 존재 (약 4푼)

거포형 : 탈잠실 시 K%가 증가하는 인과관계 존재 (약 7% 내외)

한계점 : 표본이 극히 적어서 그룹 별 처치효과(CATE) 추정이 불가하였다. 개인에 대한 인과효과를 토대로 박찬호와 김재환에게 적용하는 방법 밖에 없었다. 즉, SC를 적용한 박해민, 서건창, 오재일, 최주환에 대해 각각의 플레이 스타일을 공변량으로 하여 간접적으로 박찬호와 김재환에게 구장의 인과효과를 적용하는 것이기 때문에 절대적인 것은 아니라는 한계가 존재한다.

Chapter 2. GAM(일반화 가법 모형)

GAM이란?

단순 선형 회귀가 아닌 spline을 이용해 일정하지 않은 구불구불한 패턴의 데이터를 학습하고 각 feature에 따른 변화 요인을 따로 산출한 뒤 최종적으로 합산하여 미래값을 예측하는 모델링 기법이다. 선수의 미래 성적은 나이, 직전 년도의 성적, 기타 요소들에 의해 일정하지 않게 변화하기 때문에 이를 예측하기 위해 GAM 채택하였다.

데이터 수집

2016년~2025년 200타석을 넘게 소화한 kbo 타자들의 스탯티즈 '기본', '심화', '가치', '타구' 부분 데이터를 수집하였다.

Name	Year	Position	PA	G	oWAR	dWAR	PA	ePA	AB	R	H	2B	3B	HR	TB	RBI	SB
강민호	25	C	465	127	2.33	0.26	465	461	412	37	111	23	1	12	172	71	2
강백호	25	DH	369	95	1.74	-0.06	369	367	321	41	85	18	1	15	150	61	2
강승호	25	1B	400	115	0.65	0.9	400	396	360	51	85	19	3	8	134	37	14
고명준	25	1B	500	130	7.00E-02	0.72	500	500	471	46	131	20	1	17	204	64	2
고승민	25	2B	538	121	1.66	0.39	538	532	469	71	127	21	2	4	164	45	5
구본혁	25	3B	397	131	2.25	1.33	397	384	343	41	98	16	2	1	121	38	10
구자욱	25	LF	616	142	5.11	-0.63	616	611	529	106	169	43	2	19	273	96	4
권동진	25	SS	309	123	0.52	-0.11	309	303	271	34	61	12	3	1	82	25	3
권희동	25	LF	456	136	2.62	0.13	456	448	358	56	88	24	0	6	130	39	5
김건희	25	C	344	105	0.35	-0.05	344	337	322	24	78	20	2	3	111	25	2
김규성	25	2B	222	133	0.4	0.13	222	217	193	30	45	4	0	3	58	16	5
김기연	25	C	245	100	0.54	-0.24	245	241	219	19	54	9	0	2	69	24	1
김민석	25	LF	247	95	-0.8	-0.51	247	244	228	21	52	7	3	1	68	21	3
김민성	25	3B	254	96	0.75	-0.59	254	252	214	25	52	13	0	3	74	35	0
김민혁	25	LF	417	106	0.26	-0.42	417	412	380	52	109	12	2	0	125	35	11
김상수	25	2B	423	113	1.64	-0.46	423	417	355	42	90	14	0	5	119	47	3
김선빈	25	2B	308	84	2.74	-0.63	308	306	271	31	87	18	1	3	116	46	4
김성윤	25	RF	538	127	5.12	0.38	538	528	456	92	151	29	9	6	216	61	26
김영웅	25	3B	499	125	2.03	0.79	499	497	446	66	111	22	2	22	203	72	6
김인태	25	LF	225	106	0.79	0.02	225	225	183	17	39	10	1	3	60	25	0
김재환	25	DH	407	103	1.51	0.4	407	405	344	42	83	13	2	13	139	50	7
김주원	25	SS	624	144	6.27	0.06	624	620	539	98	156	26	8	15	243	65	44
김지찬	25	CF	373	90	1.41	-0.48	373	363	317	59	89	9	2	0	102	23	22
김태군	25	C	274	100	1.24	0.53	274	266	236	20	61	10	1	5	88	31	0
김태연	25	1B	340	120	7.00E-02	0.43	340	334	303	40	79	15	0	3	103	20	5
김태진	25	2B	304	94	0.06	-0.53	304	302	279	27	65	11	2	5	95	25	1
김현수	25	LF	552	140	3.61	-0.24	552	552	483	66	144	24	0	12	204	90	4
김형준	25	C	415	127	2.59	0.69	415	412	362	51	84	10	1	18	150	55	3
김호령	25	CF	381	105	2.57	0.24	381	373	332	46	94	26	3	6	144	39	12
김휘집	25	3B	500	142	2.78	0.09	500	498	429	64	107	18	2	17	180	56	10
나성범	25	RF	310	82	2.23	-0.32	310	309	261	30	70	16	0	10	116	36	0
나승엽	25	1B	392	105	0.66	-0.73	392	388	328	40	75	12	2	9	118	44	0

예시 : 2025년도 gam 데이터셋 – 이름 오름차순으로 정렬

(200타석 선정 이유 : 정규타석 70% 이상으로 수집 시 오버피팅의 문제 발생)

GAM 모델링

먼저 스탯 별로 단위가 다르기 때문에 StandardScaler를 이용해 스케일링 진행하였다.

그 후 모델 파이프라인을 구축하였고 그 과정은 아래와 같다.

a. 에이징커브를 계산하기 위해 선수별 만나이를 계산 후 age라는 리스트에 저장. 꾸준히 기록이 존재하는 선수들 혹은 주목할만한 선수들의 나이는 정확히 만 나이를 산출하였고 그 외 선수들은 데뷔 연차를 기준으로 나이 추산. (스탯티즈에 정확한 age 스탯이 없어서)

```python
birth_year_map = {
    '최형우': 1983, '김재환': 1988, '박찬호': 1995, '나성범': 1989, '황재균': 1987,
    '구자욱': 1993, '허경민': 1990, '박해민': 1990, '안치홍': 1990, '강민호': 1985,
    '박병호': 1986, '양의지': 1987, '최정': 1987, '김현수': 1988, '손아섭': 1988,
    '전준우': 1986, '이정후': 1998, '김혜성': 1999, '오지환': 1990, '김하성': 1995,
    '강백호': 1999, '박민우': 1993, '채은성': 1990, '오재일': 1986, '정수빈': 1990,
    '한동민': 1989, '한유섬': 1989, '이대호': 1982, '이용규': 1985, '김선빈': 1989,
    '노시환': 2000, '문동주': 2003, '김도영': 2003
}


def calculate_age(row):  1 usage
    if row['Name'] in birth_year_map: return row['Year'] - birth_year_map[row['Name']]
    if 'FirstYear' in row and pd.notnull(row['FirstYear']): return 24 + (row['Year'] - row['FirstYe
    return 27
```

b. 2025년 데이터셋을 test set, 그 전까지의 데이터는 train set으로 할당. 특히, 2025년 성적 예측을 위해 2024년, 2023년, 2022년 데이터셋에 각각 50%, 30%, 20%의 가중치를 부여(단순히 직전 연도의 데이터만 활용하는 것보다 최근 3개년에 차례대로 가중치를 부여하는 ETS 시계열 모델링 활용)

c. 3개년 가중 평균치를 한 칸 뒤로 밀어(shift 1) 실제 2025년 값과 매칭시킴

d. 추가적으로 직전 년도 성적을 기준으로 예측한 모델, 3개년 단순 평균치로 예측한 모델을 각각 shift 1해서 2025년 실제 성적과 매칭 시킨 후 각 모델 별 정확도를 비교해봄

```python
# ------------------------------------------------------------------------------
# [계산 로직] 1.가중평균(Model용), 2.작년성적(Base1), 3.단순평균(Base2)
# ------------------------------------------------------------------------------
def calc_3yr_weighted_avg(series):  1 usage
    vals = series.values
    n = len(vals)
    result = np.full(n, np.nan)
    w3, w2, w1 = [0.5, 0.3, 0.2], [0.6, 0.4], [1.0]

    for i in range(n):
        subset = vals[max(0, i - 2): i + 1][::-1]  # 최신순 정렬
        valid = subset[~np.isnan(subset)]
        if len(valid) == 0:
            continue
        elif len(valid) == 1:
            result[i] = valid[0]
        elif len(valid) == 2:
            result[i] = np.dot(valid, w2)
        else:
            result[i] = np.dot(valid[:3], w3)
    return pd.Series(result, index=series.index)
```

e. 성적에 가장 큰 영향을 미칠 것으로 예상되는 에이징 커브의 곡선 패턴 학습을 위해 SplineTransformer 사용. 나이에 따른 성적의 변화 패턴을 직선이 아닌 곡선의 형태로 표현하기 위해 사용.

Degree = 2로 2차 곡선, n_knots = 4로 4개의 관절 포인트를 설정함(야구 선수의 신인, 성장, 전성기, 노쇠화 패턴을 위해)

```python
for target in target_cols:
    cfg = feature_config.get(target, {'num': [f'WAVG_{target}', 'Age'], 'cat': []})

    # 파이프라인 구축
    num_pipe = Pipeline([
        ('imp', SimpleImputer(strategy='median')),
        ('scl', StandardScaler()),
        ('spl', SplineTransformer(n_knots=4, degree=2, include_bias=False))
    ])
    transformers = [('num', num_pipe, cfg['num'])]
    if cfg['cat']:
        cat_pipe = Pipeline(
            [('imp', SimpleImputer(strategy='most_frequent')), ('ohe', OneHotEncoder(handle_unknow
        transformers.append(('cat', cat_pipe, cfg['cat']))

    model = Pipeline([('pre', ColumnTransformer(transformers)), ('reg', Ridge(alpha=1.0))])
```

그 후 전처리된 데이터를 Ridge 모델에 입력시켜 오버피팅을 방지하였다.

모델에 사용된 feature는 아래와 같다.

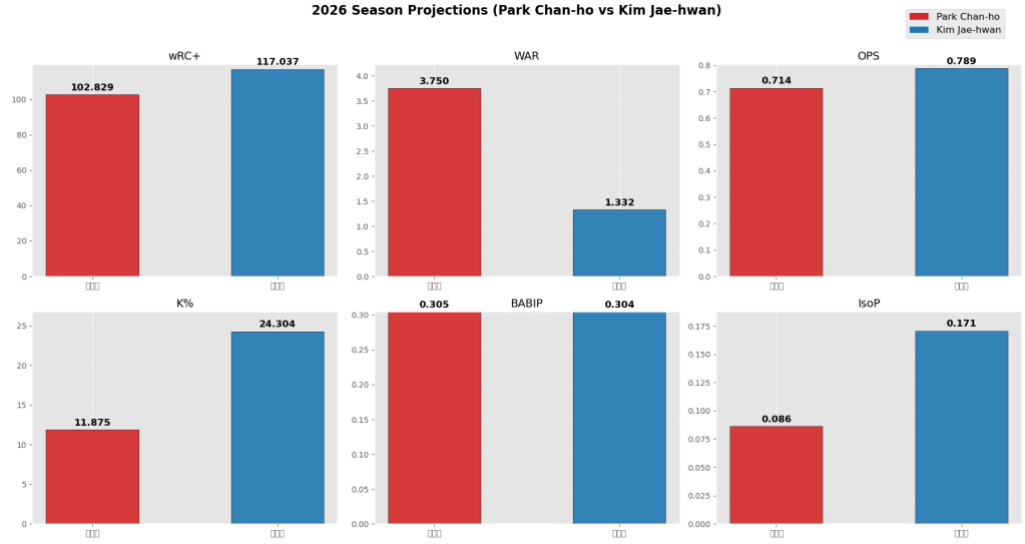
예측 대상 (Target)	주요 설명 변수 (Features, 지수가중적용)	선정 이유 (Why?)
wRC+ (타격 생산성)	**wRC+, BB/K, BABIP**	타격의 종합 지표(wRC+)에 선구안(BB/K)과 타구 질(BABIP) 추세를 반영하여 "운"을 보정하고 실력을 예측함.
WAR (승리 기여도)	**WAR, wRC+, 수비RAA**	WAR은 타격+수비의 종합임. 타격(wRC+)과 수비력(수비RAA)의 최근 추세를 합쳐 종합 기여도를 산출.
OPS (출루+장타)	**OPS, IsoP(순장타율)**	OPS의 구성 요소 중 변동성이 큰 장타력을 보완하기 위해 순수 파워 지표인 IsoP를 보조 지표로 활용.
K% (삼진율)	**K%, BB/K**	삼진은 타자의 컨택 능력과 직결됨. 볼넷/삼진 비율(BB/K)을 통해 타석 접근법(Approach)의 변화를 감지.
BABIP (인플레이 타율)	**BABIP, LD%(라인드라이브), ifFB%(내야뜬공)**	BABIP은 변동성이 크지만, 타구의 질(라인드라이브 비율)과 빗맞은 타구(내야뜬공) 비율을 알면 예측력이 상승함.
IsoP (순장타율)	**IsoP**	파워는 비교적 꾸준히 유지되거나 나이에 따라 완만하게 변하므로, 본인의 과거 파워 수치와 나이만으로도 예측력이 높음.

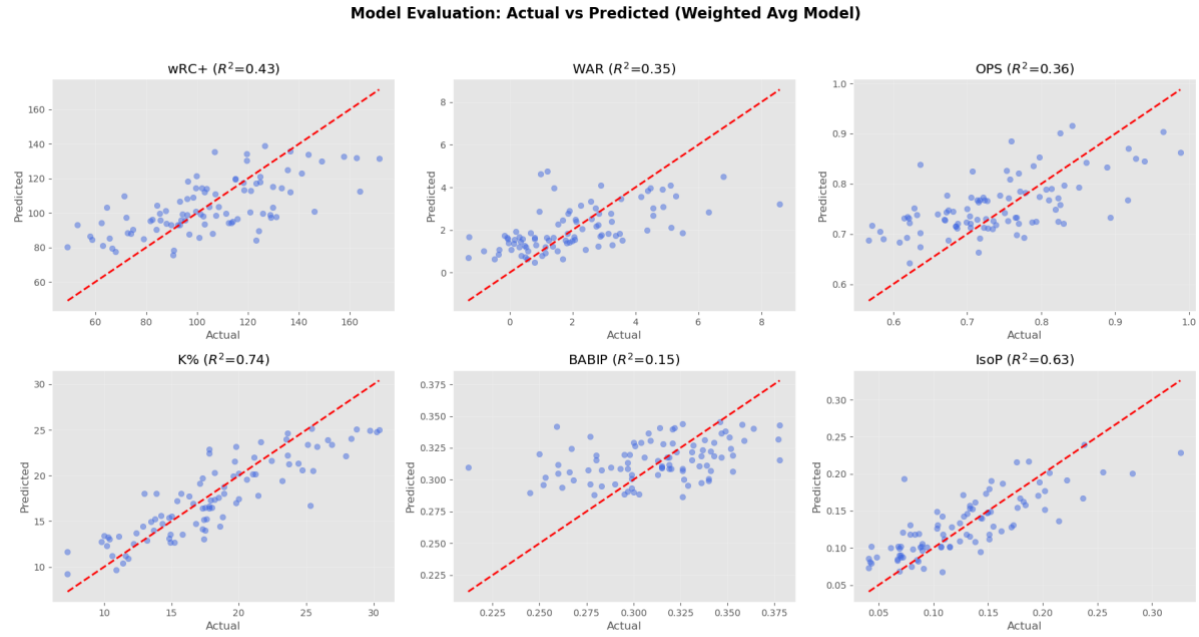
기본적으로 나이(age), 직전 년도 스탯은 주요 설명 변수로 포함시켰다.

모델링 결과 및 적합도

[최종 예측값]

Name	김재환	박찬호
Target		
BABIP	0.304344	0.304598
IsoP	0.170694	0.086285
K%	24.303808	11.875479
OPS	0.789071	0.713777
WAR	1.331716	3.749695
wRC+	117.036953	102.829020

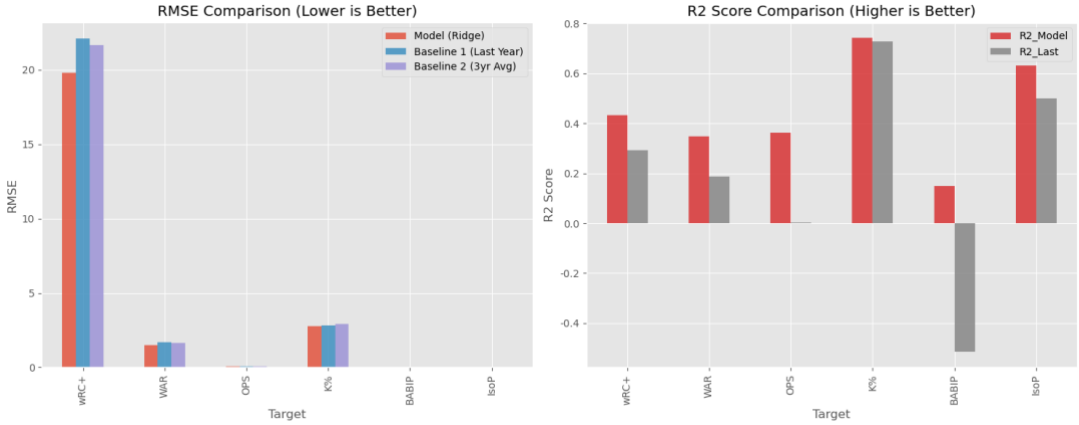




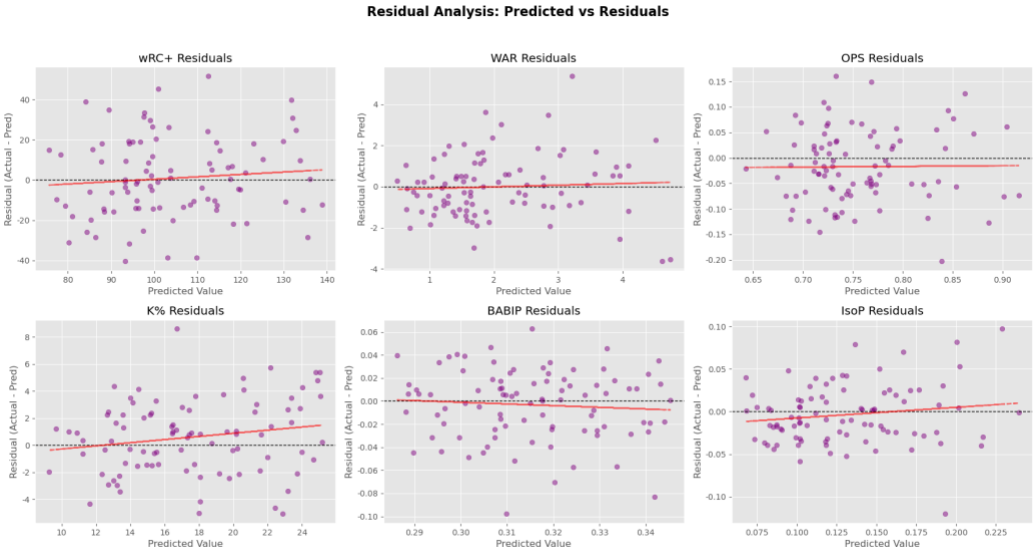
각 스탯별 모델의 R 스퀘어 값은 위와 같다. 야구의 경우, 팀 분위기, 운적인 요소, 날씨 등 다양한 측정 불가 요소의 영향을 받기 때문에 완벽한 학습과 예측은 쉽지 않고 대략

30~40%의 정확도를 목표로 모델링을 진행하였다.

BABIP의 경우 특히, 운적인 요소에 영향을 많이 받기 때문에 R^2값이 낮았다. 전체적으로 파란색 점들이 빨간색 점선과 비슷한 trend로 산포되어 있는 것으로 보아 학습이 나쁘지 않음을 확인할 수 있다.



3년 가중 평균(빨간색), 3년 단순 평균(보라색), 직전 연도(파란색)를 기준으로 25년 성적을 예측한 모델의 RMSE값과 R^2값은 위 그래프와 같다. RMSE는 작을수록 R^2는 높을수록 유의한데 3개년 가중 평균치 모델링이 가장 유의함을 확인할 수 있다. 우리가 사용한 3년 가중 평균 모델이 가장 효율적이다.



우리가 사용한 모델의 예측값에 따른 잔차 분석도이다. 전체적으로 보라색 점이 특정한 패턴 없이 퍼져있는 것을 확인할 수 있다. 빨간 실선의 기울기 역시 검은색 점선의 기울기와 거의 유사한 값을 가지고 있다. 다만, 몇몇 튀는 점(outlier)에 대한 전처리 작업이 진행되지 않아서 약간은 우상향 or 우하향하는 경향을 보이기도 하지만 박찬호와 김재환의 2026년 성적 예측에는 큰 영향이 없는 것으로 판단하였다.

결론

최종 26년도 성적 예측

```
[최종 예측값]
 Name            김재환          박찬호
Target
BABIP      0.304344    0.304598
IsoP       0.170694    0.086285
K%        24.303808   11.875479
OPS        0.789071    0.713777
WAR        1.331716    3.749695
wRC+     117.036953  102.829020
```

(GAM으로 예측한 26년도 성적)

GAM으로 예측한 26년도 성적에 SC를 통해 파악한 구장의 인과효과를 보정하여 최종 성적을 산출해보자.

<박찬호>

입잠실에서 통계적으로 유의한 인과관계가 관측된 것은 OPS뿐이다. 약 2~4푼 정도의 하락을 반영하면 박찬호의 2026 최종 OPS는 6할 후반대에서 7할 초반 정도로 계산된다(약 0.68).

IsoP와 wRC+, K%는 소폭 하락, WAR은 소폭 상승될 것으로 예상되지만 박해민의 SC에서 유의미한 구장의 인과효과를 측정하지 못했기 때문에 정확한 수치 계산보다는 방향성의 예상으로만 마무리하였다.

<김재환>

탈잠실에서 통계적으로 유의한 인과관계가 관측된 것은 K%뿐이다. 약 7% 증가하였으며 김재환의 2026 최종 K%는 20 후반에서 30 초반대로 계산됨(약 30% 정도)

IsoP는 최주환과 오재일의 SC 결과를 미루어 봤을 때, 소폭 상승될 것으로 예상되며 그 외 WAR, wRC+, OPS 자체에서는 소폭 하락 or 유의미한 증가가 기대되지는 않는다. 다만, 이 역시 유의미한 구장의 인과효과가 관측되지 않은 스탯들이기 때문에 정확한 수치 계산보다는 방향성에 초점을 맞추는 것으로 마무리하였다.

의의 및 한계점

- 파크팩터는 인과효과에 기반한 값이라기 보다는 단순 구장에 따른 성적의 비율로 산출해낸 값임(구장의 인과효과 추정이 별도로 필요한 부분). wRC+ 역시 인과효과에 따른 보정값은 아님.
- 탈잠실에 따른 성적 상승이 정말 효과가 있는가에 대한 근본론적인 접근이었다는 점에서

의의가 있음.

- 적지만 OPS 부분에서의 입잠실, K% 부분에서의 탈잠실 인과효과를 어느정도 유의미하게 측정해볼 수 있었음
- 다만, 표본 자체가 너무 적었다. 특히 합성 통제법을 위해서는 꾸준히 많은 시즌과 타석을 소화하며 처치 집단(박해민, 최주환, 오재일, 서건창)과 유사한 플레이 스타일을 가진 통제 집단을 확보해야했으나 너무 적었음.
- 이에 따라 k-means 등의 방식을 이용해 그룹화를 하고 그룹별 처치 효과(그룹별 구장 변화의 인과효과)를 측정하여 개개인 선수에게 적용하고 싶었으나 불가능하였음.
- 박해민, 최주환, 오재일 등 그나마 플레이 스타일 및 이적에 따른 구장 변화가 가장 유사한 선수들만 활용하였고 그들의 결과를 최대한 간접적으로 박찬호, 김재환에게 적용시키고자 하였음. 그래서 2026년도 인과효과를 바탕으로 한 성적 예측에는 한계가 존재할 수 밖에 없었다.