메이저리그 선발투수 방어율의 통계적 예측

류현진 선수의 '13시즌 방어율 예측 및 분석

BUFS 부산외국어대학교 데이터경영학과
김선태, 송종명, 조현영

MAJOR LEAGUE BASEBALL

1. 주제선정

영화 머니볼은 야구에서 데이터를 이용하여 선수를 분석하는 '머니볼 이론' 을 다룬 영화다. 이 영화를 통해 데이터를 이용한 통계적 수치가 선수를 분석하는데 있어서 중요한 요소가 될 수 있다는 것을 확인 할 수 있다. 이러하듯 야구는 기록의 스포츠라고 불린다. 그만큼 야구에는 수치화 할 수 있는 데이터들이 많고 이 데이터를 바탕으로 한 다양한 통계적 분석방법들도 많이 활용되고 있다. 이에 우리는 메이저리그 선발투수의 방어율을 예측해 보자는 생각을 갖게 되었고, 이번 시즌 메이저리그에 진출하여 전반기가 끝난 현재 10승 3패로 순항중인 류현진 선수를 대상으로 '13시즌 방어율을 예측해 보고자 한다.

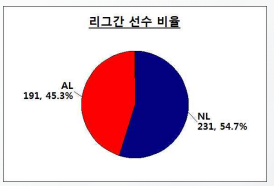
2. 수행방법

야구정보사이트 팬 그래프(www.fangraphs.com)를 통해 08-12시즌 동안 규정이닝(162이닝)을 소화한 메이저리그의 투수들의 자료를 수집 및 분석하였다.

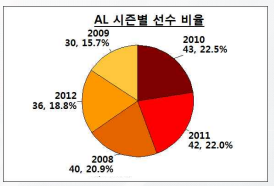
9개의 변수 ERA(방어율), WHIP(이닝 당 피출루율), AVG(피안타율), HR/9(9이닝당 홈런 수), BB/9(9이닝당 볼 넷 수), K/9(9이닝당 삼진 수), FIP(수비력을 제거한 순수 방어율), WAR(팀 내 기여도), HR/FB(플라이볼 중 홈런비율)를 선택하였고, ERA를 종속변수로 나머지 8개의 변수(WHIP, AVG, HR/9, BB/9, K/9, FIP, WAR, HR/FB)를 독립변수로 선정하고 ERA와 나머지 8가지 변수와의 상관분석을 통해 상관관계를 확인해 제거할 변수가 있다면 제거한다. 이어서 회귀분석을 통해서 방어율에 관한 회귀식을 도출하여 류현진 선수의 '13시즌 전반기 성적을 대입하여 '13시즌 전체 방어율을 예측한다.

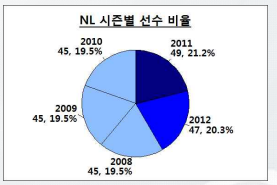
(*단, 류현진 선수의 후반기 성적은 전반기 성적과 같다는 가정하에 분석을 진행한다.)

3. 표본비율



전체 : '08-12년동안 규정이닝을 소화한 전체 투수는 중복을 포함하여(동일선수) 총 422명이고, 리그간 선수비율은 내셔널리그(NL)가 231명 아메리칸리그(AL)가 191명으로 전체 100% 중 각각 54.7%와 45.3%를 차지한다.



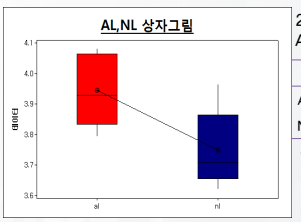


AL : 아메리칸리그의 시즌별 선수비율은 '08년부터 '12년까지 40명, 30명, 43명, 42명, 36명으로 전체 100% 중 각각 20.9%, 15.7%, 22.5%, 22%, 18.8% 를 차지한다.

NL : 내셔널리그의 시즌별 선수비율은 '08년부터 '12년까지 45명, 45명, 45명, 49명, 47명으로 전체 100% 중 각각 19.5%, 19.5%, 19.5%, 21.2%, 20.3% 를 차지한다.

4. 2표본 T 검정

현재 류현진 선수는 LA 다저스 소속으로 내셔널리그에 속해있다. 현재 데이터는 MLB리그 전체(AL,NL통합)를 대상으로 하였기 때문에 AL와 NL의 방어율 차이에 관한 T-TEST를 시행하여 두 리그간 방어율 차이가 없다면 전체데이터를 이용하여 앞으로의 분석을 진행하고, 차이가 있다면 AL 데이터를 배제하고 NL 데이터만을 이용하여 앞으로의 분석을 진행한다.



2-표본 T 검정 및 CI: AL, NL
AL 대 NL의 2-표본 T 검정

	N	평균	표준 편차	표준 오차
AL	5	3.945	0.119	0.053
NL	5	3.750	0.131	0.058

T-값 = 2.47 P-값 = 0.039 DF = 8
합동 표준 편차 0.1250를 사용.
(* 두 표본 등분산성을 가정한다.)

가설 H_0 : AL과 NL 간 방어율에 차이가 없다.
H_1 : AL과 NL 간 방어율에 차이가 있다.

T-Test 결과 AL과 NL의 방어율 추정치는 각각 3.945, 3.750이고 검정통계량(t-값)이 2.47로 유의수준 0.05가정 하에서 귀무가설(H_0)을 기각함을 확인 할 수 있다. 그러므로 AL과 NL간 방어율에 차이가 존재한다고 판단, 앞으로의 분석은 내셔널리그 투수들의 방어율을 바탕으로 진행한다.

5. 상관분석

	WHIP	AVG	HR/9	BB/9	K/9	FIP	WAR	HR/FB
AVG	0.772 0.000							
HR/9	0.253 0.000	0.346 0.000						
BB/9	0.613 0.000	−0.020 0.762	−0.045 0.498					
K/9	−0.349 0.000	−0.601 0.000	−0.154 0.019	0.167 0.011				
FIP	0.637 0.000	0.558 0.000	0.762 0.000	0.324 0.000	−0.578 0.000			
WAR	−0.646 0.000	−0.566 0.000	−0.650 0.000	−0.326 0.000	0.562 0.000	−0.900 0.000		
HR/FB	0.231 0.000	0.322 0.000	0.824 0.000	−0.032 0.625	−0.070 0.291	0.599 0.000	−0.555 0.000	
ERA	0.813 0.000	0.768 0.000	0.561 0.000	0.307 0.000	−0.369 0.000	0.731 0.000	−0.700 0.000	0.483 0.000

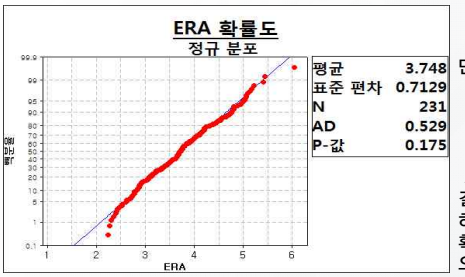
셀 내용 : Pearson 상관계수
P - 값

방어율에 대한 회귀식을 추정하기 전에 독립변수간 연관성이 강하면 다중공선성이 발생할 확률이 높기 때문에 이 문제가 발생하는 것을 방지하기 위하여 종속변수(ERA)와 독립변수(WHIP, AVG, HR/9, K/9, FIP, WAR, HR/FB)간 상관분석을 실시한다.

우선 임의로 상관계수가 0.8이상이면 강한 상관관계가 존재한다고 가정하고, 상관분석 결과를 확인하면 변수 HR/FB와 HR/9, 변수 FIP와 WAR 두 쌍의 조합이 강한 상관관계를 나타내는 것을 확인 할 수 있다.

그러므로 다중공선성의 문제가 예상되는 두 쌍의 변수 중 종속변수(ERA)와 상관관계가 낮은 변수인 HR/FB와 WAR를 제거한 후에, 다음 절차를 진행하기로 한다.

6. 회귀분석



정규성검정

회귀분석의 기본가정인 오차의 정규성을 만족하는지 확인하기 위해 정규성 검정 실시.

가설 H_0 : 정규성을 만족한다.
H_1 : 정규성을 만족하지 않는다

Anderson-Darling의 정규성 검정을 실시한 결과 P-값이 0.175로 유의수준 0.05가정 하에 귀무가설(H_0)을 기각 하지 못하는 것을 확인할 수 있다. 즉, 회귀분석의 기본가정인 오차의정규성을 만족한다고 할 수 있다.

단계적 회귀 분석: ERA 대 FIP, K/9, BB/9, HR/9, AVG, WHIP
6개의 예측 변수(N = 231)에서 반응변수 = ERA
입력 변수에 대한 α = 0.05, 제거 변수에 대한 α = 0.05

단계	1	2	3	4
상수	−1.6683	−1.8537	−2.1386	−0.7734
WHIP	4.24	3.74	4.43	9.17
T-값	21.14	23.20	22.21	5.64
P-값	0.000	0.000	0.000	0.000
HR/9		0.897	0.797	0.783
T-값		12.27	11.15	11.
BB/9			−0.178	−0.694
T-값			−5.35	−3.89
P-값			0.000	0.000
AVG				−24.0
T-값				−2.94
P-값				0.004
S	0.416	0.323	0.305	0.300
R-제곱	66.11	79.59	81.88	82.55
R-제곱(수정)	65.97	79.41	81.64	82.24
Mallows Cp	215.9	41.7	13.8	7.1

Stepwise

회귀식을 추정하는 방법으로 변수선택법의 방법중의 하나인 단계적 선택법 (Stepwise)을 이용했다. 분석결과 변수 WHIP,HR/9, BB/9, AVG가 유의한 변수임을 확인 할 수 있고, 재평가에서 제거된 변수는 없으므로 전진선택법 (Forward)을 이용해도 동일한 결과가 나오는 것을 확인 할 수 있다.

결정계수와 수정된 결정계수가 0.82 정도로 도출된 회귀식이 82%의 설명력을 가지는 것을 확인 할 수 있다.

아래에 WHIP, HR/9, BB/9, AVG로 구성된 ERA에 관한 회귀식이 도출되었다.

도출된 적합식

$$\widehat{ERA}$$

류현진의 전반기 성적은 WHIP = 1.25 HR/9 = 0.77 BB/9 = 3.01 AVG = 0.242이다. 위 성적을 도출된 적합식에 대입하면 후반기 방어율의 추정값이 3.39507나온다. 전반기의 등판 경기수는 18경기이고 방어율은 3.09이다. 그리고 후반기의 등판 경기수는 14경기로 예상되므로 14경기로 가정하고 추정된 후반기 방어율이 3.39507이므로 이를 바탕으로 '13 시즌 전체 방어율을 계산하면 (18/32)×3.09 + (14/32)×3.39507 = 3.223으로 예측할 수 있다.

7. 결론

각 분석의 결과

2표본 T-검정	AL와 NL간 방어율에 차이에 대한 T검정을 실시, 검정결과 두 리그간 방어율 차이가 존재한다는 결과 도출. 분석의 대상을 NL 투수들을 대상으로 선정
상관분석	상관계수가 0.8이상이면 강한 상관관계가 존재한다고 가정했을 때 변수 FR/FB와 HR/9, 변수 FIP와 WAR 두 쌍이 강한 상관관계를 가지는 것을 확인. 이 두 쌍의 변수 중 종속변수(ERA)와 상관관계가 낮은 변수 HR/FB, WAR를 제거한 후 다음 절차 진행
회귀분석	단계적선택법(Stepwise)을 이용하여 변수 K/9, FIP이 제거된 최종 회귀식을 도출.

최종결론

류현진 선수의 전반기 방어율은 3.09이고, 회귀분석을 통해 도출된 후반기 방어율은 3.39507이다.

이를 바탕으로 **'13 시즌 전체 방어율이 3.223**정도가 나올 것이라고 예측 할 수 있다.

예측된 3.223의 방어율은 내셔널리그 선발투수들의 '08 ~12시즌 방어율을 기준으로 했을 때 각 10위, 13위, 17위, 9위, 8위로 평균 12위 정도의 수준인 것을 확인 할 수 있다.

11	Madison Bumgarner	Giants	3.21
12	Tim Hudson	Braves	3.22
13	R.A. Dickey	Mets	3.28
14	Matt Garza	Cubs	3.32
15	Jeff Karstens	Pirates	3.38
16	Kyle Lohse	Cardinals	3.39
17	Chris Carpenter	Cardinals	3.45
18	Mat Latos	Padres	3.47
19	Daniel Hudson	Diamondbacks	3.49
20	Wandy Rodriguez	Astros	3.49

RYU