

경영통계학

한국 프로야구 팀별 타자의 기록분석과 홈런과 삼진의 상관관계

경영정보학과
2019011033 김재용



CONTENTS

01

주제 선정 동기

02


데이터 수집 및 설명

03

데이터 분석

04

결론 및 느낀점



1. 주제 선정 동기



평소 야구에 대한 관심이 있었다. 응원하는 팀의 야구 경기를 지켜보다가 팀의 홈런 타자가 득점권 타석일 때 자주 삼진을 당하는 모습을 보았다. 다른 팀의 타자들의 특성에 대한 궁금증도 생겨서 이 연구를 진행하였다.

데이터 수집 방법 및 출처



KBO 기록실에서 데이터 수집

<https://www.koreabaseball.com/Record/Player/Hitter/Basic/Basic1.aspx>



수집 대상

2022년 정규 시즌 KBO 타자 규정타석(446타석) 충족 선수



데이터 범위

52명의 선수의 이름, 팀명, 타석, 홈런, 삼진의 개수를 파악해 데이터를 정리했다.

변수와 가설 설정

가설 설정: 홈런이 많을 수록 삼진이 많을 것이다.

독립 변수: 홈런(HR)

종속 변수: 삼진(SO)

=> 홈런과 삼진의 상관분석 및 회귀분석 진행

데이터 정리 및 요약

- 기술통계법으로 표현

홈런(HR)		삼진(SO)	
평균	12.69230769	평균	85.65384615
표준 오차	1.123094485	표준 오차	3.671553505
중앙값	11.5	중앙값	84
최빈값	4	최빈값	100
표준 편차	8.098749508	표준 편차	26.47594885
분산	65.58974359	분산	700.9758673
첨도	-0.408516946	첨도	-0.514382525
왜도	0.555157155	왜도	0.168566198
범위	34	범위	105
최소값	1	최소값	32
최대값	35	최대값	137
합	660	합	4454
관측수	52	관측수	52
신뢰 수준(95.0%)	2.254706261	신뢰 수준(95.0)	7.370951228

52명의 평균 홈런은 약 13개
이고 삼진은 약 86개 이다.

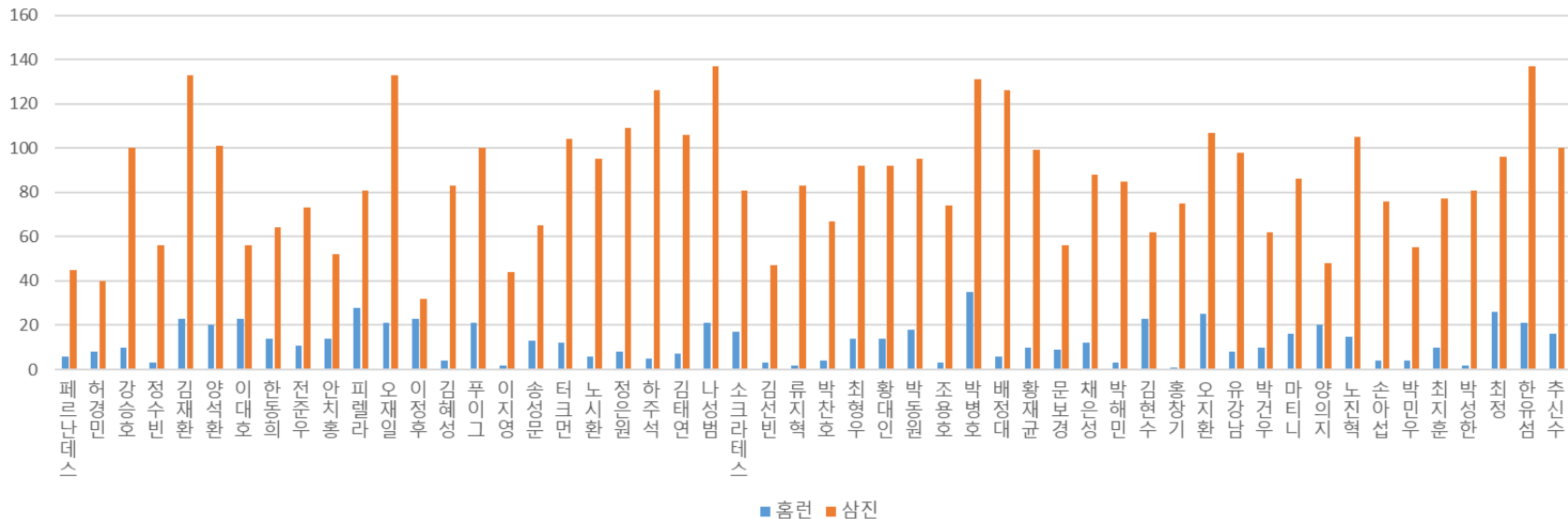
홈런의 최대값은 35개 최소값
은 1개

삼진의 최대값은 137개 최소
값은 32개이다.

데이터 정리 및 요약

52명의 선수의 홈런과 삼진의 개수를 막대 그래프로 요약

선수별 홈런과 삼진 개수



상관분석

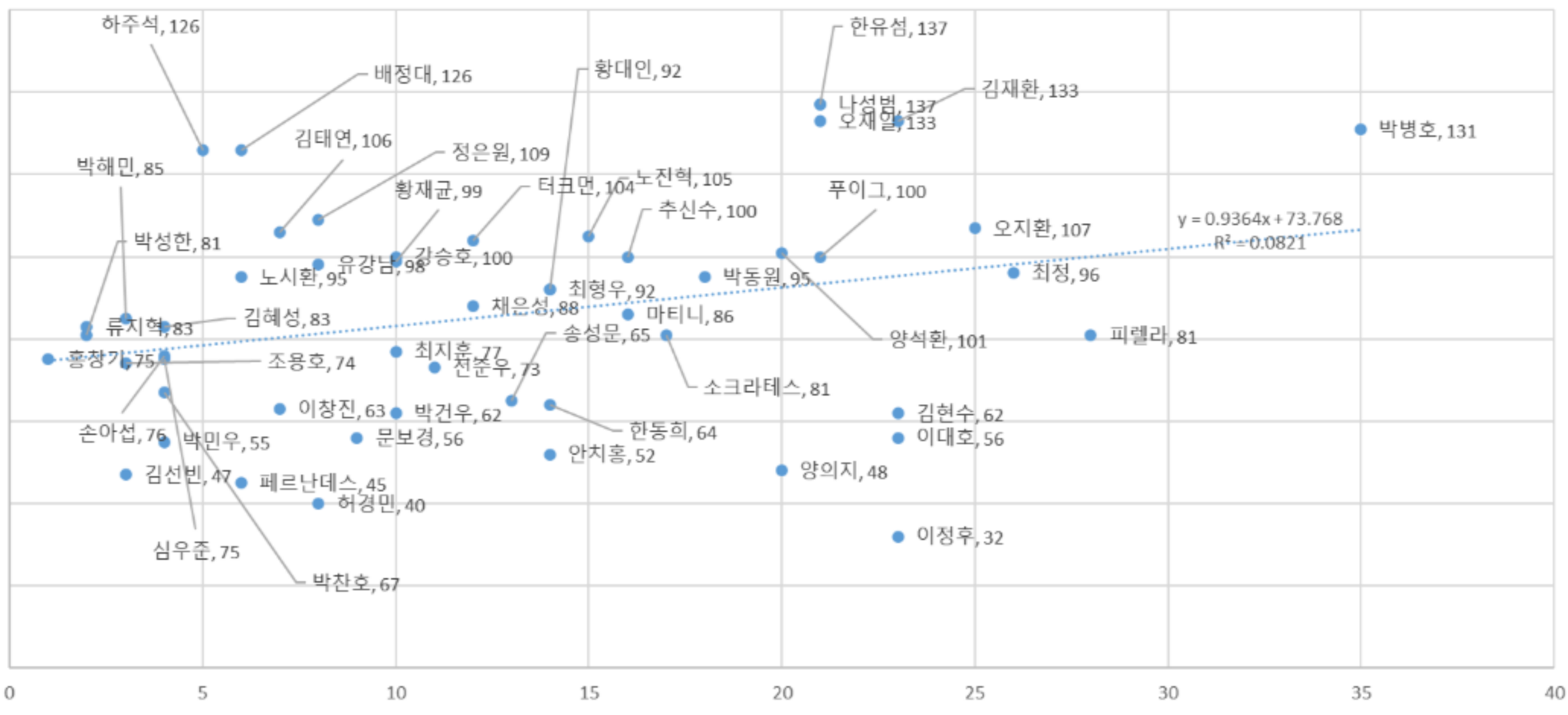
상관계수와 산점도와 추세선을 바탕으로 분석

	홈런(HR)	삼진(SO)
홈런(HR)	1	
삼진(SO)	0.286448183	1

상관계수

상관계수가 약 0.28이므로 상관관계가 약하다고 볼 수 있다.

홈런과 삼진의 상관관계



회귀분석

H0 (귀무가설) : 홈런과 삼진 사이에는 관계가 있다.

H1 (대립가설) : 홈런과 삼진 사이에는 관계가 없다.

요약 출력									
회귀분석 통계량									
다중 상관계수	0.286448183								
결정계수	0.082052561								
조정된 결정계수	0.063693613								
표준 오차	25.61890282								
관측수	52								
분산 분석									
	자유도	제곱합	제곱 평균	F 비	유의한 F				
회귀	1	2933.360134	2933.360134	4.46935	0.03952				
잔차	50	32816.4091	656.3281819						
계	51	35749.76923							
	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%	
Y 절편	73.76827025	6.65053615	11.09207868	4.4E-15	60.4103	87.1263	60.4103	87.1263	
홈런(HR)	0.936439314	0.442952821	2.114083644	0.03952	0.04674	1.82614	0.04674	1.82614	

회귀분석을 통한 데이터를 분석해볼 결과 여기서 홈런의 p-의 값이 0.03952므로 0.05보다 작다. 그래서 처음에 홈런이 삼진에 대해 통계적으로 유의한 영향을 가진다는 것으로 생각했다.

하지만...

결정계수가 0.082로 이 지표만을 참고했을 땐 설명력이 낮다고 볼 수 있다. 결정계수가 1에 가까워야 모델이 데이터를 잘 설명한다고 볼 수 있다. 또한 상관계수가 0.28으로 상관관계가 약하다고 볼 수 있다.

따라서 H0의 가설을 기각했다.

데이터 추가 수집



이대로 가설이 기각되는 것이 아쉬워
규정타석을 충족한 선수가 52명이라 좀 더 추가적인 데
이터를 수집하기로 했다.
이번에는 200타석을 이상 출전한 선수 107명을 수집
한 데이터를 바탕으로
홈런과 삼진의 관계에 관한 통계분석을 했다.

표본 재추출 후 상관분석 및 회귀분석

	홈런(HR)	삼진(SO)
홈런(HR)	1	
삼진(SO)	0.510317	1

요약 출력		
회귀분석 통계량		
다중 상관계수	0.510317386	
결정계수	0.260423834	
조정된 결정계수	0.253380251	
표준 오차	23.02801498	
관측수	107	

상관계수가 0.510317로 51명의 데이터보다는 상관 관계가 높은 것을 알 수 있다.
또한 결정계수가 0.260423으로 이전 결정계수에 비해 높음을 알 수 있다.

표본 재추출 후 상관분석 및 회귀분석

요약 출력									
회귀분석 통계량									
다중 상관계수	0.510317								
결정계수	0.260424								
조정된 결정계수	0.25338								
표준 오차	23.02801								
관측수	107								
분산 분석									
	자유도	제곱합	제곱 평균	F 비	유의한 F				
회귀	1	19606.5	19606.5	36.97321	1.96E-08				
잔차	105	55680.39	530.2895						
계	106	75286.9							
	계수	표준 오차	t 통계량	p-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%	
Y 절편	57.3297	3.388096	16.92092	9.15E-32	50.61174	64.04767	50.61174	64.04767	
홈런(HR)	1.800349	0.296083	6.08056	1.96E-08	1.213271	2.387426	1.213271	2.387426	

홈런의 p-값도 0.05보다 작기 때문에 처음에는 유의미한 결과라고 판단했다.

하지만 여전히 결정계수의 값이 작기 때문에 유의미한 관계를 나타낸다고 보기는 어렵다.

따라서 홈런이 많을 수록 삼진이 많다는 결론은 기각이다.

결론

- 타자의 홈런과 삼진과의 상관관계는 낮음.
- 독립변수와 종속변수의 재설정 필요.
- 타자의 홈런이 많을 수록 삼진이 많을 것이라는 나의 생각은 기분탓이었음.

한계 및 보완점



1. 데이터 수집의 한계

야구 기록 데이터를 입력하기 위해서는 엑셀 파일이 별도로 존재하지 않아서, 데이터를 직접 입력해야 하는 불편함이 있었다.



2. 독립변수와 종속변수 재설정 필요

처음에는 홈런과 삼진 사이의 상관관계를 설정하였으나, 통계 분석 결과를 분석한 결과로써, 장타율과 삼진 사이의 상관관계가 더욱 유의미하다는 생각이 들었다.



3. 통계분석 노력 필요

엑셀의 기능을 활용하여 처음 데이터 분석을 진행했기 때문에, 주제 선정부터 데이터 수집, 그리고 분석 결과 도출까지의 과정에서 상당한 시간이 소요되었다. 이 프로젝트를 토대로 엑셀을 더욱 활용하여 다양한 데이터 분석을 수행하고자 생각하였습니다.



THANK YOU

경영정보학과
2019011033 김재용

