

다면량 통계학 및 실습
학기 말 과제 최종 발표

1조
통계학과 2015580020 엄호영
통계학과 2015580023 이동균
통계학과 2017580018 서용비

연구 주제 소개

NBA



NBA 선수들의 연봉에 영향 을 주는 요인이 무엇인가

사용한 데이터 : 2017-2018 시즌 NBA
선수들의 연봉과 각종 농구 수치 자료 이
용 (변수: 28개, 관측치: 485개)

변수 설명

변수명	유형	변수 설명	변수명	유형	변수 설명
Player	문자	선수 이름	TRB%	숫자	총 리바운드율
Salary	숫자	연봉	AST%	숫자	어시스트 비율
NBA_Country	문자	국가	STL%	숫자	스틸 비율
NBA_DraftNumber	숫자	드래프트 넘버	BLK%	숫자	블락 비율
Age	숫자	나이	TOV%	숫자	턴 오버율
Tm	문자	팀	USG%	숫자	공격 점유율
G	숫자	뛴 게임 수	OWS	숫자	공격적인 승리 공헌
MP	숫자	플레이 한 시간(단위:분)	DWS	숫자	수비적인 승리 공헌
PER	숫자	선수 효율성 지수	WS	숫자	승리 공헌
TS%	숫자	실제 슛 비율	WS/48	숫자	48분 간 승리 공헌
3PAr	숫자	3점 시도율	OBPM	숫자	수비적인 점수 마진
FTr	숫자	자유투 시도율	DBPM	숫자	공격적인 점수 마진
ORB%	숫자	공격 리바운드율	BPM	숫자	총 점수 마진
DRB%	숫자	수비 리바운드율	VORP	숫자	교체 선수 보다의 가치

데이터 전처리

- 경기 외적 요인은 고려하지 않기 위해 Player(선수 이름), NBA_Country(국가), NBA_DraftNumber(드래프트 넘버), Tm(팀) 변수는 제외
- G(뛴 게임 수) = 1인 경우는 해당 관측치에서 결측치가 발견되어 G=1인 117개의 데이터를 제외
- 변수들의 단위가 서로 다르므로 표준화 실시

데이터 분석방법 소개

1

설명변수의 차원을 줄이기 위해
주성분 분석 실시



주성분을 설명변수로 설정하여 선
형회귀분석 실시

2

설명변수의 차원을 줄이기 위해
인자 분석 실시

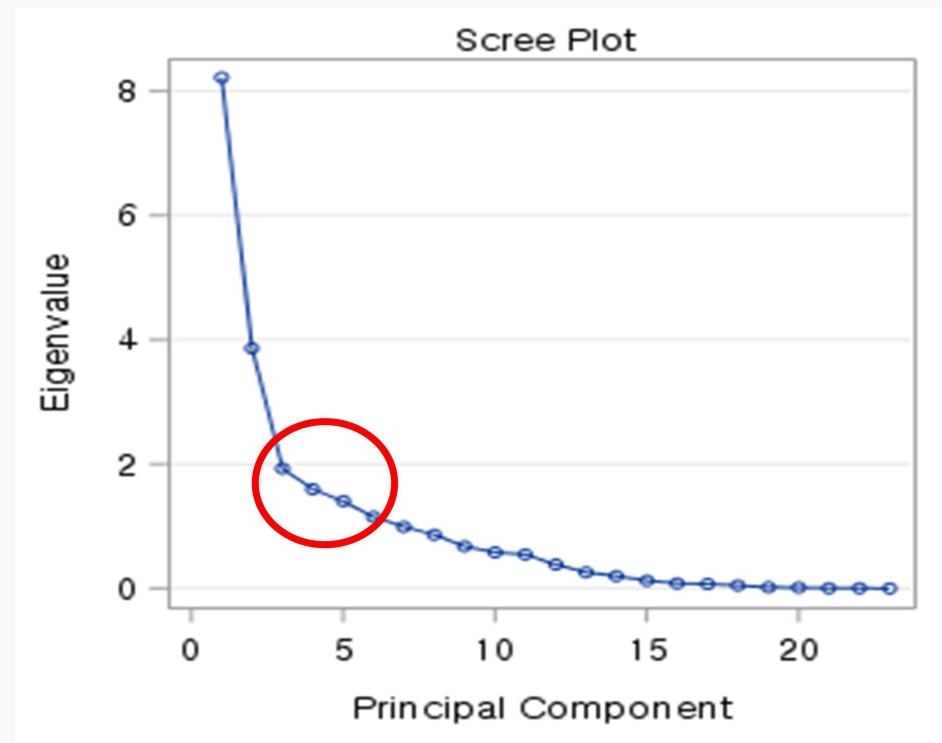


인자를 설명변수로 설정하여 선
형회귀분석 실시

데이터 분석

1-1. 주성분 분석

- ◆ scree plot



**Scree plot이 급격히 꺾이는 지
점을 보면 세 번째 ~ 네 번째 주
성분까지 선택하는 것이 적절해
보임.**

데이터 분석

1-1. 주성분 분석

◆ 주성분 분석 결과

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	8.22036185	4.35646332	0.3574	0.3574
2	3.86389853	1.93538198	0.1680	0.5254
3	1.92851655	0.33448668	0.0838	0.6093
4	1.59402987	0.19010748	0.0693	0.6786
5	1.40392239	0.25567715	0.0610	0.7396
6	1.14824524	0.15705254	0.0499	0.7895
7	0.99119270	0.12706434	0.0431	0.8326
8	0.86412836	0.18949178	0.0376	0.8702
9	0.67463658	0.09486149	0.0293	0.8995
10	0.57977509	0.03519619	0.0252	0.9247
11	0.54457890	0.16240478	0.0237	0.9484
12	0.38217412	0.12430958	0.0166	0.9650

- 하지만 네 번째 주성분의 누적 설명력은 약 68% 정도에 그침

✓ 누적 설명력을 고려해 다섯 번째 주성분까지 선택

데이터 분석

1-1. 주성분 분석

◆ 주성분 해석

	Prin1	Prin2	Prin3	Prin4	Prin5
Age	0.033523	-0.072227	-0.010363	-0.062804	0.139455
G	0.219178	-0.143153	0.250580	-0.238922	0.094387
MP	0.249828	-0.202256	0.243979	-0.157454	-0.095210
PER	0.302093	0.042154	-0.287779	0.095472	-0.056214
TS_	0.227435	0.004046	-0.383111	-0.088255	0.296657
_3PAr	-0.107500	-0.330755	0.026727	-0.193821	0.146938
FTr	0.060588	0.142635	-0.292548	0.095838	0.026262
ORB_	0.093975	0.410106	-0.080599	-0.070300	-0.089291
DRB_	0.141649	0.366320	0.120823	-0.068573	-0.161331
TRB_	0.139899	0.430631	0.045297	-0.078075	-0.146979
AST_	0.118264	-0.227317	-0.037996	0.528529	-0.167131
STL_	0.060519	-0.055987	0.239578	0.470319	0.274185
BLK_	0.109496	0.340962	0.022594	-0.004960	0.091880
TOV_	-0.014105	0.123844	0.066762	0.454402	0.044993
USG_	0.133644	-0.111969	-0.192178	0.225830	-0.553453
OWS	0.287279	-0.099612	0.032180	-0.116610	-0.188017
DWS	0.281722	-0.048988	0.308105	-0.067350	-0.080462
WS	0.311832	-0.088880	0.141191	-0.108042	-0.163622
WS_48	0.292007	0.037431	-0.288656	-0.058579	0.232623
OBPM	0.275513	-0.203015	-0.279033	-0.005669	0.135637
DBPM	0.180353	0.210488	0.373629	0.180403	0.347286
BPM	0.314009	-0.054404	-0.030884	0.088683	0.287906
VORP	0.287409	-0.070609	0.142010	0.055283	-0.170402

- 첫 번째 주성분: PER(선수 효율성 지수), WS(승리 공헌), BPM(총 점수 마진)
- 두 번째 주성분: ORB, DRB, TRB(공격/수비/총 리바운드율), BLK(블락 비율), 3PAr(3점 시도율)
- 세 번째 주성분: TS(실제 슛 비율), DWS(수비적인 승리 공헌), DBPM
- 네 번째 주성분: AST(어시스트 비율), STL(스틸 비율), TOV(턴 오버율)
- 다섯 번째 주성분: TS(실제 슛 비율), DBPM(수비적인 점수 마진), USG(공격 점유율)

데이터 분석

1-2. 회귀 분석

- ◆ 주성분 점수를 설명변수로 한 회귀분석 결과해석

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	6648327	261566	25.42	<.0001
Prin1	1	1435796	91326	15.72	<.0001
Prin2	1	-600086	133207	-4.50	<.0001
Prin3	1	495571	188551	2.63	0.0089
Prin4	1	-82580	207392	-0.40	0.6907
Prin5	1	-1165849	220988	-5.28	<.0001

- 추정된 회귀 방정식
 - $\text{Salary} = 6648327 + 1435796P1 - 600086P2 + 495571P3 - 82580P4 - 1165849P5$
- 네 번째 주성분은 연봉을 설명하는데 유의하지 않음
- 연봉에 많은 영향을 미치는 주성분은 첫 번째, 다섯 번째 주성분. 이 중 연봉에 가장 많은 영향을 미치는 주성분은 계수가 가장 큰 첫 번째 주성분.
- ✓ PER(선수 효율성 지수), WS(승리 공헌), BPM(총 접수 마진)이 클수록 연봉이 높음.

데이터 분석

2-1. 인자 분석

◆ 인자 분석 결과

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	8.22036185	4.35646332	0.3574	0.3574
2	3.86389853	1.93538198	0.1680	0.5254
3	1.92851655	0.33448668	0.0838	0.6093
4	1.59402987	0.19010748	0.0693	0.6786
5	1.40392239	0.25567715	0.0610	0.7396
6	1.14824524	0.15705254	0.0499	0.7895
7	0.99119270	0.12706434	0.0431	0.8326
8	0.86412836	0.18949178	0.0376	0.8702
9	0.67463658	0.09486149	0.0293	0.8995
10	0.57977509	0.03519619	0.0252	0.9247
11	0.54457890	0.16240478	0.0237	0.9484
12	0.38217412	0.12430958	0.0166	0.9650

- 배리맥스 회전(varimax rotation) 실시
- 주성분 방법(principal factor)을 사용 했기 때문에 인자가 설명하는 비율은 주 성분 분석에서의 결과와 동일
- 인자 개수로 5개 선택

데이터 분석

2-1. 인자 분석

◆ 베리맥스 회전 실시

Rotated Factor Pattern					
	Factor1	Factor2	Factor3	Factor4	Factor5
Age	0.10107	-0.13531	0.14228	-0.12006	-0.00328
G	0.79133	-0.07902	0.16356	-0.19766	-0.01990
MP	0.89616	-0.12167	0.12700	0.06336	-0.01121
PER	0.40118	0.32325	0.73066	0.36594	0.05025
TS_	0.18792	0.10687	0.68917	-0.04261	-0.06848
_3PAr	0.01428	-0.72615	-0.09253	-0.19297	-0.18968
FTr	-0.19584	0.28635	0.38367	0.14796	0.00169
ORB_	-0.05562	0.84952	0.14036	-0.02415	-0.09482
DRB_	0.22855	0.83836	-0.01337	-0.01559	-0.01090
TRB_	0.13390	0.94722	0.05852	-0.02211	-0.04860
AST_	0.22464	-0.29431	0.17133	0.61719	0.50453
STL_	0.13568	-0.10684	0.02210	-0.04406	0.76079
BLK_	0.03789	0.70039	0.16755	-0.17077	0.10682
TOV_	-0.20584	0.20142	-0.06633	0.14321	0.54187
USG_	0.24646	0.01293	0.12610	0.83562	-0.05237
OWS	0.77035	0.10465	0.33542	0.25438	-0.09201
DWS	0.89019	0.18628	0.11014	0.03189	0.14424
WS	0.88740	0.14659	0.28025	0.19211	-0.00901
WS_48	0.37711	0.24451	0.86393	0.00609	0.00732
OBPM	0.48968	-0.19045	0.80189	0.20039	0.00571
DBPM	0.41145	0.48232	0.08884	-0.38691	0.60785
BPM	0.60516	0.09579	0.68881	-0.03876	0.31839
VORP	0.77474	0.15604	0.23866	0.27649	0.15563

- 첫 번째 인자: G(뛴 게임 수), MP(플레이 한 시간), WS(승리 공헌), BPM(점수 마진), VORP(교체선수보다의 가치) → 게임 횟수, 승리에 대한 공헌을 설명해주는 인자
- 두 번째 인자: ORB, DRB, TRB(공격/수비/총 리바운드율), BLK(블락 비율), → 리바운드, 블락 능력에 대한 인자
- 세 번째 인자: PER(선수 효율성 지수), TS(실제 슛 비율), OBPM, BPM(공격/총 점수 마진). → 선수가 얼마나 공격을 잘하면서 효율적 인자 설명해주는 인자
- 네 번째 인자: AST(어시스트 비율), USG(공격 점유율). → 어시스트 와 공격 능력에 대한 인자
- 다섯 번째 인자: STL(스틸 비율), TOV(턴 오버율), DBPM(수비적인 점수 마진). → 선수 수비지표와 턴 오버에 대해 설명해주는 인자

데이터 분석

2-2. 회귀 분석

- ◆ 인자 점수를 설명변수로 한 회귀분석

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	6648327	266150	24.98	<.0001
Factor1	1	4032344	266432	15.13	<.0001
Factor2	1	397713	266432	1.49	0.1362
Factor3	1	1252231	266432	4.70	<.0001
Factor4	1	1651107	266432	6.20	<.0001
Factor5	1	135200	266432	0.51	0.6121

- 추정된 회귀방정식
 - $\text{Salary} = 6648327 + 4032344F1 + 397713F2 + 1252231F3 + 1651107F4 + 135200F5$
 - 두 번째, 다섯 번째 인자는 연봉을 설명하는 데 유의하지 않음.
 - 가장 많은 영향을 미치는 인자는 추정된 회귀계수가 가장 큰 첫 번째 인자.
- ✓ **플레이 시간과 승리 공헌도가 클수록 연봉이 높음**

분석 결과

- 높은 연봉을 받기 위해 선수에게 요구되는 능력은 다음과 같다.
- 주성분을 이용한 회귀분석 결과
 - PER(선수 효율성 지수), WS(승리 공헌), BPM(총 접수 마진)이 클수록 연봉이 높음.
- 인자를 이용한 회귀분석의 결과
 - 플레이 시간과 승리 공헌도가 높을수록 연봉이 높음.

분석의 타당성

- ◆ 기존의 23개 변수를 설명변수로 한 다중회귀분석 실시

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	6648327	232503	28.59	<.0001
Age	1	2118210	241952	8.75	<.0001
G	1	-3704646	602130	-6.15	<.0001
MP	1	4435330	902657	4.91	<.0001
PER	1	-2474409	1855829	-1.33	0.1831
TS_	1	-1311432	699837	-1.87	0.0616
_3PAr	1	-1242890	543186	-2.29	0.0226
FTr	1	-73584	287776	-0.26	0.7983
ORB_	1	-3375150	3883396	-0.87	0.3852
DRB_	1	-3603825	5952107	-0.61	0.5452
TRB_	1	6976101	8741708	0.80	0.4253
AST_	1	-620698	472486	-1.31	0.1896
STL_	1	-515990	405529	-1.27	0.2039
BLK_	1	375814	547612	0.69	0.4930
TOV_	1	393363	371789	1.06	0.2906
USG_	1	1157413	661575	1.75	0.0809
OWS	1	-6406167	8455313	-0.76	0.4491
DWS	1	-3606477	4631489	-0.78	0.4366
WS	1	10709777	11968965	0.89	0.3714
WS_48	1	-212843	1407925	-0.15	0.8799
OBPM	1	8885040	15444828	0.58	0.5654
DBPM	1	3588206	9771969	0.37	0.7136
BPM	1	-5645790	19002857	-0.30	0.7665
VORP	1	27842	821958	0.03	0.9730

- 대부분의 계수가 유의하지 않음.
- 이러한 결과가 나타난 이유로 다중공선성 (Multicollinearity) 문제를 생각해 볼 수 있음.

분석의 타당성

◆ 다중공선성 진단

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	6648327	232503	28.59	<.0001	0
Age	1	2118210	241952	8.75	<.0001	1.08064
G	1	-3704646	602130	-6.15	<.0001	6.69277
MP	1	4435330	902657	4.91	<.0001	15.04077
PER	1	-2474409	1855829	-1.33	0.1831	63.57711
TS_	1	-1311432	699837	-1.87	0.0616	9.04105
_3PAr	1	-1242890	543186	-2.29	0.0226	5.44655
FTr	1	-73584	287776	-0.26	0.7983	1.52874
ORB_	1	-3375150	3883396	-0.87	0.3852	278.38661
DRB_	1	-3603625	5952107	-0.61	0.5452	653.98275
TRB_	1	6976101	8741708	0.80	0.4253	1410.64434
AST_	1	-620698	472486	-1.31	0.1896	4.12099
STL_	1	-515990	405529	-1.27	0.2039	3.03578
BLK_	1	375814	547812	0.69	0.4930	5.53973
TOV_	1	393363	371789	1.06	0.2906	2.55164
USG_	1	1157413	661575	1.75	0.0809	8.07948
OWS	1	-6406167	8455313	-0.76	0.4491	1319.72764
DWS	1	-3606477	4631489	-0.78	0.4366	395.97362
WS	1	10709777	11968965	0.89	0.3714	2644.46649
WS_48	1	-212843	1407925	-0.15	0.8799	36.59181
OBPM	1	8885040	15444828	0.58	0.5654	4403.42959
DBPM	1	3588206	9771969	0.37	0.7136	1762.74321
BPM	1	-5645790	19002857	-0.30	0.7665	6665.96025
VORP	1	27842	821958	0.03	0.9730	12.47165

- 분산 팽창 지수(VIF) 값이 10을 넘으면 다중공선성이 있다고 판단.
 - 많은 변수들의 VIF값이 매우 큼. 따라서 여러 설명변수들이 이 다중공선성 문제를 발생시킨다고 판단.
- ✓ 주성분 분석과 인자 분석은 다중공선성 문제를 해결할 수 있는 방법이므로 Salary에 영향을 주는 변수를 찾는 과정에서 주성분 분석과 인자 분석은 타당하다고 볼 수 있음.

분석의 장점 및 한계점

- 장점

- 주성분/인자 분석은 설명 변수들 간에 다중공선성 문제를 해결.

- 한계점

- 주성분은 상관성이 높은 변수가 묶인 것이지 의미중심으로 묶인 것이 아니므로 인자와 다르게 뚜렷한 특성을 가지지 않음.

주성분은 인자와 비교해서 계수가 전체적으로 낮음.

	Prin1		Factor1
Age	0.033523	Age	0.10107
G	0.219178	G	0.79133
MP	0.249828	MP	0.89616
PER	0.302093	PER	0.40118
TS_	0.227435	TS_	0.18792
_3PAr	-.107500	_3PAr	0.01428
FTr	0.060588	FTr	-0.19584
ORB_	0.093975	ORB_	-0.05562
DRB_	0.141649	DRB_	0.22855
TRB_	0.139899	TRB_	0.13390
AST_	0.118264	AST_	0.22464
STL_	0.060519	STL_	0.13568
BLK_	0.109496	BLK_	0.03789
TOV_	-.014105	TOV_	-0.20584
USG_	0.133644	USG_	0.24646
OWS	0.287279	OWS	0.77035
DWS	0.281722	DWS	0.89019
WS	0.311832	WS	0.88740
WS_48	0.292007	WS_48	0.37711
OBPM	0.275513	OBPM	0.48968
DBPM	0.180353	DBPM	0.41145
BPM	0.314009	BPM	0.60516
VORP	0.287409	VORP	0.77474

제안사항

- 경기 외적인 요소가 선수의 연봉에 영향을 미칠 수 있음.
 - 샐러리 캡 제도: 프로스포츠 리그에 존재하는 팀 연봉 총액의 상한선으로 각 팀들은 샐러리 캡의 90%를 의무적으로 선수 연봉으로 지출해야 됨.
 - 각 팀들은 의무액수를 채우기 위해 주전 및 백업 선수들에게 능력에 비해 높은 연봉을 제시할 수 있다.
 - 루키 스케일 계약: 1라운드 드래프트로 입단한 모든 선수들은 4년동안 루키 스케일로 정해진 금액내에서만 계약이 가능.
 - 입단하지 얼마 안된 선수들은 좋은 능력에 비해 낮은 연봉을 받을 수 있다.
- 따라서 분석 전에 각 팀의 상황이나 선수의 루키 스케일 계약 여부를 확인한다면 더 좋은 분석을 실시할 수 있을 것.