Can NBA Player’s Salaries be Predicted from Statistics?

CALEB MILLER

ECON 381

Introduction

This presentation will be walking through my regression project

Will be asking “Can NBA Player’s salaries be predicted from statistics?”

Could be useful because NBA is an analytics driven league

Initial Variables: pts, pf, tov, blk, stl, ast, trb, orb, ftpercent, twopercent, threepercent, fgpercent, mp, lnage

- lnage to account for diminishing returns
- Semi-log form

Literature Review

Examined three studies asking the same question

Study A: R-squared of 0.674 F-statistic significant

- Free-throw percentage and games played most impactful

Study B: R-squared of 0.9728 F-statistic significant

- Assists and turnovers most impactful

Study C: R-squared of 0.613

- Field-goal percentage and points per game most impactful

Data Sources

Statistics for 2020-2021 season collected from basketballreference.com

- Reliable for up-to-date NBA statistics

Limited to 235 players who played 10+ minutes per game

- Range from minimum contract to superstar

Multicollinearity could be present due to orb, trb & two-point, three-point, and field-goal percentage

Summary Statistics

```
. estat sum

Estimation sample regress                    Number of obs =        232
```

Variable	Mean	Std. dev.	Min	Max
salaryMil	9.326825	9.554644	.358295	43.00636
pts	12.10474	6.401963	3.7	32
pf	1.97931	.5680554	.4	3.5
tov	1.409052	.8260137	.2	4.8
blk	.4788793	.3951355	0	2.7
stl	.8112069	.346103	.2	2.1
ast	2.7125	2.016183	.4	11.7
trb	4.575862	2.35446	.9	13.5
orb	.9965517	.7726304	.1	4.1
ftpercent	.7757759	.0992324	.444	.957
twopercent	.529944	.0697835	.38	.766
threepercent	.3393405	.1043944	0	1
fgpercent	.4720862	.0724792	.353	.763
mp	25.4625	6.459385	11	37.6
lnage	3.249145	.1599982	2.944439	3.610918

Skewed right – more low salaries
We can also see this in pts, tov, blk, ast, and orb

Skewed left – most players are good free-throw shooters

Somehow, a player shot 100% from three – Drew Eubanks – shot 2/2 on the season – extreme outlier
Second place: Joe Harris (47.5 percent)

The Model

Estimating signs of coefficients

- Positive: pts, blk, stl, ast, trb, orb, ftpercent, twopercent, threepercent, fgpercent, mp, lnage
- Negative: pf, tov

Estimating highest impact: pts, ast

Ideal sets of variables will account for:

- Offensive stats: pts, ast, trb, orb, ftpercent, twopercent, threepercent, fgpercent
- Defensive stats: blk, stl
- Usage/longevity stats: mp, lnage

Results: Correlation Matrix

I ran this to deal with multicollinearity

Correlation of 0.8 considered too high

	mp	fgperc~t	threep~t	twoper~t	ftperc~t	orb	trb	ast	stl	blk	tov	pf	pts	salary~l	lnage
mp	1.0000														
fgpercent	0.0252	1.0000													
threepercent	0.1249	-0.4023	1.0000												
twopercent	-0.0360	0.8440	-0.3266	1.0000											
ftpercent	0.2922	-0.3535	0.4754	-0.3625	1.0000										
orb	0.1607	0.7257	-0.4790	0.5612	-0.4169	1.0000									
trb	0.4870	0.5680	-0.2805	0.4393	-0.2188	0.8192	1.0000								
ast	0.6386	-0.0414	0.0952	-0.1355	0.2677	-0.0390	0.2533	1.0000							
stl	0.5746	-0.0966	0.0553	-0.1307	0.1349	-0.0026	0.1898	0.6704	1.0000						
blk	0.1362	0.5836	-0.3756	0.5178	-0.3644	0.6500	0.5966	-0.1314	0.0559	1.0000					
tov	0.7323	0.0964	0.0308	-0.0220	0.1492	0.1614	0.4669	0.8367	0.5257	0.0682	1.0000				
pf	0.4598	0.3087	-0.1565	0.2471	-0.1177	0.4414	0.5411	0.2120	0.2573	0.4600	0.4175	1.0000			
pts	0.8261	0.0983	0.1789	-0.0001	0.3658	0.1296	0.4364	0.6435	0.4273	0.0663	0.8144	0.3397	1.0000		
salaryMil	0.6052	0.1092	0.0097	0.0277	0.2236	0.1629	0.4183	0.5939	0.4625	0.1141	0.6003	0.2656	0.6566	1.0000	
lnage	0.0171	0.0031	-0.0083	0.0214	0.1278	0.0003	0.0250	0.0706	0.0550	0.0139	-0.0536	-0.0171	-0.0422	0.3537	1.0000

Orb correlated to trb 0.8192
Twopercent correlated to fgpercent 0.8440
Ast correlated to tov 0.8367
Pts correlated to mp 0.8261

Results: VIF Table

```
. estat vif
```

Variable	VIF	1/VIF
tov	8.92	0.112160
ast	6.51	0.153512
pts	6.47	0.154630
orb	6.40	0.156304
trb	5.94	0.168389
fgpercent	5.86	0.170776
mp	4.95	0.202014
twopercent	3.89	0.256782
stl	2.38	0.419643
blk	2.37	0.422804
ftpercent	1.99	0.502786
pf	1.83	0.544995
threepercent	1.59	0.630786
lnage	1.10	0.912229
Mean VIF	4.30	

VIF of 10+ extremely problematic VIF of 5+ moderately problematic

All VIFs under 10 so narrow examined variables to those with correlation issues

Orb (6.4), trb (5.94), twopercent (3.89), fgpercent (5.86), ast (6.51), tov (8.92)

Because of the high correlation and problematic VIFs, I look for redundant variables

Two-point, three-point, and field-goal percentage are redundant so fgpercent will be omitted from the regression

Assists and turnovers are redundant as well (high assists typically leads to more turnovers) so turnovers will be excluded

Offensive rebounds and total rebounds are redundant so offensive rebounds will be left out of this regression

Results: Regression Analysis

```
. regress salaryMil pts pf blk stl ast trb ftpercent twopercent threepercent mp lnage
```

Source	SS	df	MS
Model	13464.7545	11	1224.06859
Residual	7623.519	220	34.6523591
Total	21088.2735	231	91.2912274

Number of obs	=	232
F(11, 220)	=	35.32
Prob > F	=	0.0000
R-squared	=	0.6385
Adj R-squared	=	0.6204
Root MSE	=	5.8866

salaryMil	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
pts	.7690383	.1233057	6.24	0.000	.5260268	1.01205
pf	-.617709	.8820382	-0.70	0.484	-2.356035	1.120617
blk	.317883	1.483918	0.21	0.831	-2.606632	3.242398
stl	3.466186	1.684216	2.06	0.041	.146923	6.785449
ast	.7596816	.3302473	2.30	0.022	.1088285	1.410535
trb	.6243223	.2697608	2.31	0.022	.0926763	1.155968
ftpercent	-.1461917	5.212931	-0.03	0.978	-10.41986	10.12748
twopercent	-4.061483	7.020336	-0.58	0.563	-17.8972	9.774234
threepercent	-5.642433	4.454311	-1.27	0.207	-14.42101	3.136148
mp	-.0796474	.1294005	-0.62	0.539	-.3346706	.1753757
lnage	21.12414	2.506814	8.43	0.000	16.1837	26.06459
_cons	-69.06792	9.295706	-7.43	0.000	-87.38795	-50.74789

$salaryMil = -69.068 + 0.769(pts) - 0.618(pf) + 0.318(blk) + 3.466(stl) + 0.760(ast) + 0.624(trb) - 0.146(ftpercent) - 4.061(twopercent) - 5.642(threepercent) - 0.080(mp) + 21.124(lnage)$

Model passes the hypothesis test and is statistically significant

Adj. R-squared of 0.6204 meaning this model can account for 62.04 percent of variation in salaries

Five out of the eleven variables pass their t-tests and are significant
Unfortunately, pf, blk, ftpercent, twopercent, threepercent, and mp failed and are problematic

Most impactful statistics on salary (in order) are age, stl, twopercent, threepercent, and pts

The coefficients of two- and three-point percentage are misleading, however, since the results are percentages

Conclusion

To answer the question “Can NBA Player’s salaries be predicted from statistics”, I used a semi-log regression

This model was statistically significant at the 0.05 level

5/11 variables passed t-test showing significance

Five most impactful statistics from my regression (in order) are age, stl, twopercent, threepercent, and pts

There are commonalities between my findings and other literature

- The statistics that are most impactful varies from dataset to dataset

Further research is needed to conclude what the most impactful statistics on salary are

However, to answer the title of the project – yes – to an extent

- Not completely reliable