Factors that Affect the Salary for NBA Players

Rui Lu

1 Introduction

In this paper, I would like to find out what factors can affect the salary for NBA players. My research will be based on the performance data for all the NBA players during the 2019-2020 season. The dependent variable will be the salary of players, and the independent variables will include the age, experience, position, number of games, performance in shooting, and performance in assistance, etc.

Based on common sense, I would like to hypothesize that richer experience, more games played, more shoots, less turnovers and shooting guard can make higher salary. And this hypothesis will be tested in the following regression models.

2 Description of Data Set and Variables

2.1 Data Set

All the data for this analysis is collected from basketball-reference.com, whose data is provided by sportradar.com, the official stats partner of the NBA. I choose the data for the latest 2019-2020 season, and included all the 30 teams that have performed in this season. After manipulating with some missing values, I use the data for 521 players in my regression.

2.2 Variables

The dependent variable for this analysis is *Salary* for each player in dollar.

The independent variable for this analysis are as following:

Age – Numerical Variable - The player's age of February 1, 2019.

G – Numerical Variable - The number of games that the player has played

GS - Numerical Variable - The number of games that the player has stared

MP – Numerical Variable - The average minutes that the player has played per game

FG – Numerical Variable - The average field goals that the player has made per game

FGpct – Numerical Variable - The average field goals percentage (made/attempts) that the player has made per game

threeP – Numerical Variable - The average 3-point field goals that the player has

made per game

threePpct – Numerical Variable - The average 3-point field goals percentage (made/attempts) that the player has made per game

twoP – Numerical Variable - The average 2-point field goals that the player has made per game

twoPpct – Numerical Variable - The average 2-point field goals percentage (made/attempts) that the player has made per game

eFGpct – Numerical Variable - The effective field goal percentage (made/attempts). (This statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal)

FT - Numerical Variable - The average free throw that the player has made per game

FTpct - Numerical Variable - The average free throw percentage (made/attempts) that the player has made per game

ORB - Numerical Variable - The average offensive rebounds that the player has made per game

DRB - Numerical Variable - The average defensive rebounds that the player has made per game

AST - Numerical Variable - The average assists that the player has made per game

STL - Numerical Variable - The average steals that the player has made per game

BLK - Numerical Variable - The average blocks that the player has made per game

TOV - Numerical Variable - The average turnovers that the player has made per game

PFoul - Numerical Variable - The average personal fouls that the player has made per game

PTSperG - Numerical Variable - The average points that the player has made per game

PG - Binary Variable - Whether the position of the player is point guard - 1 for yes, 0 for no

PF - Binary Variable - Whether the position of the player is power forward - 1 for yes, 0 for no

SG - Binary Variable - Whether the position of the player is shooting guard - 1 for yes, 0 for no

SF - Binary Variable - Whether the position of the player is small forward - 1 for yes, 0 for no

C - Binary Variable - Whether the position of the player is center - 1 for yes, 0 for no

Exp - Numerical Variable - The years experience for the player in NBA/ABA (prior this season)

3 Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Age	521	26.031	4.147	19	23	29	43
G	521	40.474	22.598	1	20	61	74
GS	521	19.610	23.915	0	1	37	73
MP	521	20.299	8.915	1.000	13.200	27.900	37.500
FG	521	3.324	2.273	0.000	1.700	4.500	10.900
FGpct	519	0.447	0.106	0.000	0.403	0.488	0.833
threeP	521	1.002	0.875	0.000	0.300	1.500	4.400
threePpct	494	0.319	0.126	0.000	0.281	0.380	1.000
twoP	521	2.325	1.841	0.000	1.000	3.100	9.600
twoPpct	516	0.510	0.117	0.000	0.464	0.569	1.000
eFGpct	519	0.512	0.103	0.000	0.477	0.558	0.917
FT	521	1.459	1.414	0.000	0.500	1.900	10.200
FTpct	505	0.746	0.146	0.000	0.674	0.838	1.000
ORB	521	0.861	0.744	0.000	0.400	1.100	4.600
DRB	521	2.921	1.868	0.000	1.700	3.800	11.400
AST	521	1.992	1.787	0.000	0.800	2.500	10.200
STL	521	0.650	0.409	0	0.4	0.9	2
BLK	521	0.420	0.422	0	0.1	0.5	3
TOV	521	1.174	0.829	0.000	0.600	1.500	4.800
Pfoul	521	1.821	0.793	0.000	1.300	2.300	5.000
PTSperG	520	9.112	6.353	0.000	4.600	12.200	34.300
PG	521	0.154	0.361	0	0	0	1
PF	521	0.234	0.424	0	0	0	1
SG	521	0.244	0.430	0	0	0	1
SF	521	0.165	0.372	0	0	0	1
C	521	0.203	0.403	0	0	0	1
Exp	521	4.472	4.058	0	1	7	21
Salary	521	7,307,510.000	8,733,438.000	50,752	1,618,520	10,116,576	40,231,758

Table 1 Descriptive Statistics for Variables

This is the descriptive Statistics for all the variables, and it can give us an overview of my dataset. What worth mentioning is that the absolute value for *Salary* is extremely high comparing to the independent variables, so manipulation may needed towards *Salary* to enhance the interpretability of my model.

4 Initial Models

To build a model, I have first test the normality of my dependent variable.

```
        Shapiro-Wilk normality test

data:  d$Salary
W = 0.76142, p-value < 2.2e-16
```

Table 2 Shapiro-Wilk Normality Test to Salary

The null hypothesis is that the dependent variable *Salary* is normally distributed, but from this table, we can find that we should reject this null hypothesis, that is, Salary is not normally distributed, and the W statistic is expected to be closer to 1. As a result, I choose to log-transform the dependent variable Salary, aiming to make it normally distributed with a W statistic close to 1.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.601956   0.629969  21.591  < 2e-16 ***
Age         -0.005339   0.018467  -0.289 0.772644
G            0.009495   0.002608   3.641 0.000303 ***
GS          -0.003088   0.003026  -1.020 0.308133
MP           0.042447   0.017250   2.461 0.014235 *
FG           2.650643   1.329458   1.994 0.046772 *
FGpct        2.926287   2.475431   1.182 0.237769
threeP      -0.567558   1.164288  -0.487 0.626158
threePpct    0.336936   0.525998   0.641 0.522127
twoP        -1.418736   0.945090  -1.501 0.134005
twoPpct      0.293838   0.708456   0.415 0.678514
eFGpct      -3.423050   2.360130  -1.450 0.147644
FT           0.606677   0.555410   1.092 0.275276
FTpct       -0.138525   0.338238  -0.410 0.682329
ORB          0.100916   0.124574   0.810 0.418312
DRB          0.044607   0.052362   0.852 0.394721
AST          0.159938   0.056983   2.807 0.005219 **
STL          0.137697   0.157890   0.872 0.383608
BLK          0.280684   0.161140   1.742 0.082206 .
TOV          0.115039   0.148080   0.777 0.437640
Pfoul       -0.210492   0.098244  -2.143 0.032678 *
PTSperG     -0.625566   0.553930  -1.129 0.259355
PG          -0.927253   0.205613  -4.510 8.27e-06 ***
PF          -0.395195   0.152158  -2.597 0.009701 **
SG          -0.422607   0.172973  -2.443 0.014936 *
SF          -0.132450   0.178607  -0.742 0.458728
Exp          0.096878   0.019240   5.035 6.89e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9197 on 456 degrees of freedom
  (38 observations deleted due to missingness)
Multiple R-squared:  0.5497, Adjusted R-squared:  0.524
F-statistic: 21.41 on 26 and 456 DF,  p-value: < 2.2e-16
```

Table 3 Initial Model

This is my initial model, which is a multiple regression model with logged *Salary* as my dependent variable and all the independent variables that I have selected. From the result, we can see that the F-statistic is 21.41 and p-value is less than 2.2e-16, so we can reject the null hypothesis, and conclude that my independent variables as a whole can have a statistically significant influence to my dependent variable.

As for each variable separately, the statistically significant variables with 95% confidence level include *G* (the number of games), *PG* (point guard position), *Exp* (experience), *AST* (assists), *PF* (power forward position), *MP* (minutes played) , *FG* (field goal), *Pfoul* (personal fouls), *SG* (shooting guard position). Among them, number of games, experience, assists, minutes played and field goals are positively related with salary, while personal fouls and whether played as point guard, power forward, shooting guard can negatively affect the salary of the player.

However, there are still many problems existing in this model, and should be solved:

1. There might exist collinearity among the independent variables.
2. There might exist heteroscedasticity in my model.

After solving the above two problems, further improvement should be made to my model, such as the investigation of interaction.

5 Model Improvement

5.1 Collinearity

In the following table, I am checking the collinearity among my independent variables with VIF. For those variables that have VIFs over 10, they must have collinearity, and should be excluded in my final model.

From the result, we can find that variables with collinearity include *MP, FG, FGpct, threeP, twoP, eFGpct, FT, PTSperG*. This not only makes sense statistically, in common sense we can assume that there are close relationship between minutes

played and the number of games played; field goals and field goal percentage and other goals and their percentage as well.

Age	G	GS	MP	FG	FGpct
3.271716	1.835455	3.053657	12.354318	5128.708037	27.038244
threeP	threePpct	twoP	twoPpct	eFGpct	FT
583.533437	2.371518	1711.216126	3.258046	23.041539	360.009243
FTpct	ORB	DRB	AST	STL	BLK
1.321421	4.406259	5.092898	6.014238	2.321386	2.383176
TOV	Pfoul	PTSperG	PG	PF	SG
8.610972	3.030571	7001.058074	3.302385	2.339757	3.207500
SF	Exp				
2.567282	3.306859				

Table 4 VIF for Initial Model

To make an improvement, I will try to run another multiple regression model without *MP, FG, threeP, twoP, FT, FGpct* and *PTSperG*. (I kept *eFGpct*, because I think its collinearity can be moderated after I excluding *FGpct*)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.219833   0.614541  21.512  < 2e-16 ***
Age         -0.003142   0.018322  -0.171 0.863918
G            0.010693   0.002604   4.107 4.74e-05 ***
GS           0.001302   0.002870   0.454 0.650306
threePpct    0.487111   0.501252   0.972 0.331663
twoPpct      0.464687   0.675295   0.688 0.491719
eFGpct      -0.613482   0.965910  -0.635 0.525655
FTpct        0.172144   0.322177   0.534 0.593380
ORB          0.043882   0.106791   0.411 0.681324
DRB          0.078630   0.050692   1.551 0.121553
AST          0.188926   0.055725   3.390 0.000758 ***
STL          0.267824   0.152090   1.761 0.078905 .
BLK          0.314774   0.160818   1.957 0.050909 .
TOV          0.224165   0.121370   1.847 0.065390 .
Pfoul       -0.081669   0.088089  -0.927 0.354346
PG          -0.809349   0.205522  -3.938 9.48e-05 ***
PF          -0.278165   0.147600  -1.885 0.060112 .
SG          -0.245939   0.166966  -1.473 0.141435
SF           0.041553   0.168388   0.247 0.805194
Exp          0.099270   0.019319   5.139 4.09e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9354 on 463 degrees of freedom
  (38 observations deleted due to missingness)
Multiple R-squared:  0.5272, Adjusted R-squared:  0.5078
F-statistic: 27.17 on 19 and 463 DF,  p-value: < 2.2e-16
```

Table 5 Model 2: Modifying Collinearity

Age	G	GS	threePpct	twoPpct	eFGpct	FTpct	ORB
3.114106	1.768720	2.655299	2.082362	2.862239	3.731645	1.159239	3.130903
DRB	AST	STL	BLK	TOV	Pfoul	PG	PF
4.615272	5.561205	2.082704	2.295139	5.593256	2.355826	3.190267	2.128834
SG	SF	Exp					
2.889721	2.206388	3.223480					

Table 6 VIF for Model 2

From this result, we can find that none of the VIF for the independent variables have exceed 10, indicating the collinearity for the model has been moderated a lot by excluding those variables.

5.2 Heteroscedasticity

To test whether the model have violated the homoscedasticity assumption, I used the Breusch-Pagan test.

```
	studentized Breusch-Pagan test

data:  lm2
BP = 91.279, df = 19, p-value = 1.967e-11
```

Table 7 Breusch-Pagan test for Model 2

The result shows that it has a really low p-value, indicating this model has heteroscedasticity. To solve this problem, I try to further adjust the dependent variable with the Box-Cox method. Since the estimated lambda is 2, I implement:

var = (y ^ lambda - 1) / lambda

to *Salary.* As a result, the expected value for my model becomes:

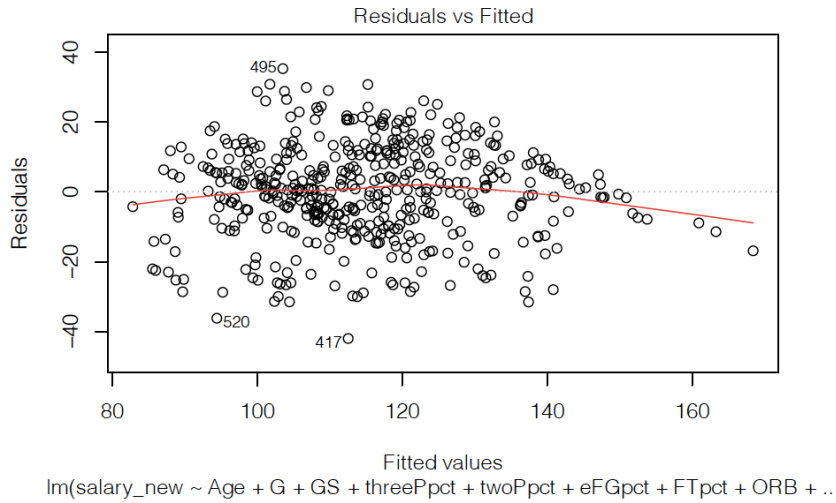
$$[ln(Salary)^2 - 1]/2$$

Next, I will run another multiple regression model based on this dependent value.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  85.66896    9.02314   9.494  < 2e-16 ***
Age           0.04340    0.26902   0.161 0.871899
G             0.13425    0.03823   3.512 0.000489 ***
GS            0.03327    0.04214   0.789 0.430253
threePpct     6.75553    7.35975   0.918 0.359148
twoPpct       5.08589    9.91518   0.513 0.608238
eFGpct       -9.24115   14.18221  -0.652 0.514982
FTpct         2.72760    4.73044   0.577 0.564487
ORB           0.68022    1.56799   0.434 0.664623
DRB           1.25644    0.74430   1.688 0.092066 .
AST           2.77960    0.81819   3.397 0.000739 ***
STL           3.82547    2.23310   1.713 0.087367 .
BLK           4.96971    2.36125   2.105 0.035856 *
TOV           3.65059    1.78204   2.049 0.041070 *
Pfoul        -1.46362    1.29339  -1.132 0.258379
PG          -11.43544    3.01762  -3.790 0.000171 ***
PF           -4.12261    2.16717  -1.902 0.057752 .
SG           -3.43546    2.45153  -1.401 0.161778
SF            0.85878    2.47239   0.347 0.728487
Exp           1.44048    0.28365   5.078 5.53e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.73 on 463 degrees of freedom
  (38 observations deleted due to missingness)
Multiple R-squared:  0.539,  Adjusted R-squared:  0.5201
F-statistic: 28.49 on 19 and 463 DF,  p-value: < 2.2e-16
```

Table 8 Model 3: Modifying Heteroscedasticity



Plot 1 Residuals vs Fitted for Model 3

With this transformation in my independent variable, we can find in the following plot that the residuals versus fitted has become a relative smooth and stable line around 0, indicating that heteroscedasticity has been successfully moderated.

5.3 Interaction

Interaction means that the association of one independent variable and my dependent variable varies according to levels of another independent variable. In my model, I may assume that the association of *eFGpct* and *Salary* varies according to levels of *twoPpct* and *threePpct*, because higher two-point percentage and three-point percentage may contribute to higher effective field goal.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               85.06952    7.69721  11.052  < 2e-16 ***
Age                        0.04473    0.26815   0.167 0.867589
G                          0.13557    0.03739   3.626 0.000320 ***
GS                         0.03460    0.04202   0.823 0.410745
FTpct                      3.17445    4.62295   0.687 0.492632
ORB                        0.37666    1.52652   0.247 0.805215
DRB                        1.30120    0.74071   1.757 0.079626 .
AST                        2.73730    0.81636   3.353 0.000865 ***
STL                        3.78119    2.22761   1.697 0.090286 .
BLK                        5.12587    2.35070   2.181 0.029715 *
TOV                        3.68915    1.77894   2.074 0.038649 *
Pfoul                     -1.42277    1.28504  -1.107 0.268787
PG                       -11.08015    2.95805  -3.746 0.000202 ***
PF                        -3.90969    2.13247  -1.833 0.067381 .
SG                        -3.16724    2.39288  -1.324 0.186283
SF                         1.14680    2.43761   0.470 0.638248
Exp                        1.45704    0.28336   5.142 4.01e-07 ***
threePpct:eFGpct:twoPpct  -1.04184   12.09513  -0.086 0.931395
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.72 on 465 degrees of freedom
  (38 observations deleted due to missingness)
Multiple R-squared:  0.5381, Adjusted R-squared:  0.5213
F-statistic: 31.87 on 17 and 465 DF,  p-value: < 2.2e-16
```

Table 9 Model 4: Modifying Interaction

From the result, we can find that the interaction between *threePpct*, *twoPpct* and *eFGpct* is not statistically significant, but the adjusted R-squared has raised from 0.5201 (previous lm3) to 0.5213, which implies this interaction have made some improvement to the model.

6 Final Models

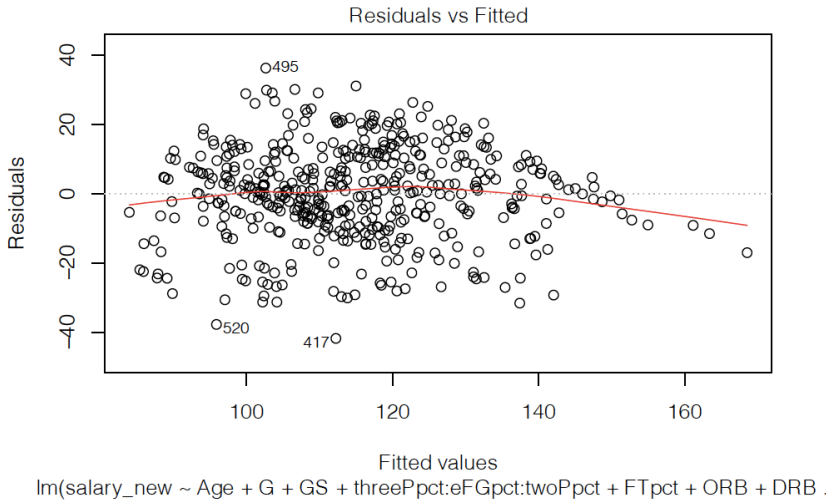
Table 9 is my final model. It's equation is:

$$[ln(Salary)^2 - 1]/2 = 85.070 + 0.048 * Age + 0.136 * G + 0.035 * GS + 3.174 * FTpct + 0.377 * ORB + 1.301 * DRB + 2.737 * AST + 3.781 * STL + 5.126 * BLK + 3.689 * TOV - 1.423 * Pfoul - 11.080 * PG - 3.910 * PF - 3.167 * SG + 1.147 * SF + 1.457 * Exp - 1.042 * threePpct * twoPpct * twoPpct$$

When comparing my initial model and my final model, we can find that the variables with collinearity have been excluded, the heteroscedasticity of the original model has been moderated and the interaction among independent variables haas been taken into consideration. (The following table and plot is used to check the collinearity and heteroscedasticity of my final model) When we evaluate the model as a whole, we can find that the F-statistic has been increased from 21.41 to 31.87, but the adjusted R-squared has been slightly decreased from 0.524 to 0.5213. However, I still believe that means the overall performance of my model has been improved, because when collinearity exits, higher adjusted R-squared doesn't necessarily mean that the model is better.

Age	G	GS
3.101862	1.696150	2.647657
FTpct	ORB	DRB
1.109929	2.974945	4.582300
AST	STL	BLK
5.550177	2.077665	2.280361
TOV	Pfoul	PG
5.587809	2.331339	3.073227
PF	SG	SF
2.066370	2.760016	2.150117
Exp	threePpct:eFGpct:twoPpct	
3.224966	1.087281	

Table 10 VIF for Final Model



Plot 2 Residuals vs Fitted for Final Model

Next, let me discuss what factors will affect the salary of a NBA player. For the factors with p-value less than 0.001, it includes the following variables:

G (the number of games played) - holding other factors constant, on average, every one more game the player played, his [(ln(Salary)^2 - 1) / 2] will increase by 0.136
AST (average assists per game) - holding other factors constant, on average, every one more assist the player has made, his [(ln(Salary)^2 - 1) / 2] will increase by 2.737
PG (position as point guard) - holding other factors constant, on average, if a player is in the position of point guard, his [(ln(Salary)^2 - 1) / 2] will decrease by 11.080
Exp (experiences year prior to season) - holding other factors constant, on average, every one year more experience, his [(ln(Salary)^2 - 1) / 2] will increase by 1.347

From this perspective, we can conclude that for NBA players who play in more games, make more assists, have more experiences, and do not play as point guard, on average, may be more likely to gain a higher pay.

7 Conclusion

The result of my model is partly consistent with my hypothesis, in that, richer experience, more games played can contribute to higher salary; but the new finding is that performance in shooting and turnovers doesn't matter much in the high or low of their salaries, rather, their performance in assists matters. The reason behind it may lies in basketball is a team sports, so the cooperation between team members may be more important than the personal performance. Additionally, position in shooting guard will not result in higher salary; rather, position in point guard will result in lower salary. This result is kind of counterintuitive, because when I calculate the mean salary for each position, the salary for point guards is indeed the highest, but here in my model, their income may be predicted to be lower if they are point guards. So, the reason behind this phenomena may need to be investigated.

However, there exists limitation in this research as well. One of the main limitations is that the salary for several star players may be extremely high while salary for new

players may be quite low (from the descriptive statistics we can see that the maximum salary is over 5 times of the mean while the minimum salary is less than 1% of the mean), so although I have used log-transformation to mitigate this influence, the existing of certain extreme values may reduce the representativeness of my conclusion. As a result, in further study, it might be better to exclude those extreme values.

As for further study, I think more research could be done to the panel data of player performance to see whether there are changes in the factors that affect the salary of players chronologically.