Research on the productivity performance of NBA players and its influence factor

Xuanwu Wang[1,*], Juntao Shen[2], Jiaye Hou[2]

[1]Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China
[2]Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China
**Corresponding author: q030018077@mail.uic.edu.cn*

Abstract: *In this paper, we employ common statistical regression models to investigate the factors influencing NBA player performance. In addition, the accuracy of each model was evaluated accordingly.*

Keywords: *Regression, Statistics, p-value, homoscedasticity*

1. Introduction

1.1 Background

The National Basketball Association (NBA) is the most commercial and the highest competitive level of professional basketball league in the world, and some scholars have studied the NBA from different perspectives. Professor Li believes that the different training guiding ideology leads to significant differences in the training contents and methods, which is a main factor causing the obvious gap between the competitive level of China and the United States. Throughout the existing research results, the research on the theory and method of NBA player training, talent training, and selection system has been come up with, but there are few studies on the influencing factors of the productivity (performance) data of NBA players.

On this basis, this project focuses on the 2021-2022 NBA playoffs of 465 players, implements analysis, and constructs a regression model. In this project, we want to find out the main influence factors affecting the NBA players' productivity (performance), so that it can provide certain theoretical support and decision-making basis for Chinese basketball players and coaches to carry out competition and training work[1].

1.2 Experimental Method

In this paper, we first selected the recent quantifiable data of NBA players on and off the court and carried out a simple data mining. Then, based on OLS method, regression analysis was conducted on players' annual salary and its possible influencing factors, and reasonable inference and verification were made based on realistic information. Finally, the effectiveness and rigor of the experiment were evaluated.

1.3 Notation

Notations that we use in the model are shown in the following table 1:

Table 1: Notations in the model

Variables	Explanation	Variables	Explanation
Salary	Annual Salary in Million	*GP*	Games played
Height	Height in Meters of Players	*PPG*	Pointes Per Game
Weight	Weight in Kilograms of Players	*Pos*	Position
Experience	NBA Experience in Seasons of Players	*MPG*	Minutes Per Game
Country	Group of Players	*FT%*	Free Throw Percentage
2PA	Average 2-Point Field Goals Attempted	*TOPG*	Turnovers Per Game
3PA	Average 3-Point Field Goals Attempted	*RPG*	Rebound Per Game
2P%	2-Point Field Goal Percentage	*APG*	Average Per Game
3P%	3-Point Field Goal Percentage	*SPG*	Steal Per Game
TS%	True Shooting Percentage	*BPG*	Block Per Game
EFG%	Effective Field Goal Percentage		

1.4 Data

Some of the cross-sectional data for the experiments are shown above (465 obs. of 23 variables)

● The units of Salary are per thousand dollars

● The symbols carrying "%" are decimal values from 0 to 1.

Dummy variables (with values of 0 or 1):

● Pos: whether the striker is a striker, with a value of 1 for the striker.

● Country: whether the player is a US athlete, 1 for US athletes.

2. Model Assumptions

Let me assume that the data used for the experiment are real and reflect the abstract field reality as objectively as possible. Our model satisfies the MLR requirements, and the estimates obtained in the model are the best linear unbiased estimators (BLUEs).

The models we set up in this project are all linear in parameters, and by checking, they are meaningful in reality, then the models are satisfied to the MLR.1. And we randomly selected 465 players as observations, then clearly, the models are satisfied to the MLR.2. Besides, after picking up the appropriate factors as the regressors, we can make sure that none of the explanatory variables is constant, and there is no perfect collinearity among these explanatory variables, which is satisfied to the MLR.3. What's more, by revising the models, the error u has no longer included any content related to the explanatory variables, then the models are satisfied to MLR.4. Therefore, under assumptions MLR.1 through MLR.4, we know that the OLS estimators we got are unbiased. Then, we also checked the homoskedasticity to ensure that our models are satisfied with MLR.5.

3. Model building and analysis

In order to reflect as objectively as possible, the competitive level of a professional basketball player over a certain period of time, we selected the annual $Salary$ of the player, a more convincing criterion in the industry and sporting events, as the dependent variable to construct an economic model for quantitative analysis[2-3].

Since the indicators of in-game data for evaluating players are complicated and do not produce large correlations between the players' personal identification information and physiological attributes, we split the observed variables and related studies into two parts:

① The influence of in-game indicators on Points per game (PPG)

② The influence of the players' own information and in-game composite indicators on their annual $Salary$

4. In-field indicator model (PPG) section

First, we set the scoring average as the dependent variable and then look for independent variables related to it, which can affect the outcome. Based on the experience of watching the game and combined with the existing data collected, we try to set the POS, MPG, FT, $2PA$, $2P$, $3PA$, $3P$, RPG, APG, SPG, BPG, and $TOPG$ to become the explanatory variables of the initial regression.

In fact, preliminary regression results are not entirely accurate since there are too many control variables and it is easy to involve irrelevant variables, resulting in heteroscedasticity of regression. Therefore, this regression does not have much reference value for the influencing factors of players' average points per game. However, it can help us initially exclude variables that have **little connection** with the average score per game. The regression results are as follows:

Table 2: Statistics data of the preliminary regression

	Coefficient	*SE*	*t Statistics*	*p-value*
Intercept	1.251	0.373	3.357	0.001
POS	0.274	0.222	1.240	0.215
MPG	0.093	0.030	3.081	0.002
FT%	0.659	0.416	1.582	0.114
2PA	0.057	0.004	14.622	6.337
2P%	1.875	0.558	3.361	0.001
3PA	0.046	0.005	8.831	2.301
3P%	2.284	0.592	3.858	0.001
RPG	0.010	0.079	0.132	0.895
APG	0.281	0.101	2.792	0.005
SPG	-0.695	0.340	1.793	0.074
BPG	0.763	0.300	2.548	0.011
TOPG	1.563	0.246	6.356	0.001

According to the *t* statistic and p-value in the above table 2, we can first exclude the variable of *RPG* in the next regression. Because **the *t* statistic in the initial regression of this variable is 0.13198**, it is statistically insignificant even at the level of α=0.005. At the same time, according to the information searched, the definition of a rebound is the behavior of a player regaining control of a live ball after a shot attempt, which has little influence on the player's scores.

By the way, by using the alternative **White test** with the form of $\hat{u}^2$=$\delta_0 + \delta_1\hat{y}$+$\delta_2\hat{y}^2 + error$ and the hypothesis of H_0: $\delta_1 = \delta_2 = 0$, we can conclude again that the former regression does not give out a proper linear function, only giving indications related to the selection of the explanatory variables because **the F-value is 28.4465** and we can't reject the null hypothesis lastly, which means it must exist heteroskedasticity among the variables. However, it does not matter since the regression equation is not used to show the linear relationships.

Combining the results of the initial regression, with scoring average as the explained variable, *POS*, *MPG*, *2PA*, *2P%*, *3PA*, *3P%*, *APG*, *BPG,* and *TOPG* are predictor variables for the second regression. The regression results are as follows:

$$\widehat{PPG} = -1.45 + 0.3POS + 0.068MPG + 0.0058twoPA + 1.77twoP + 0.046threePA$$

$$(0.356) \quad (0.220) \quad (0.025) \qquad (0.004) \qquad\qquad (0.550) \qquad (0.005)$$

$$+2.123threeP + 0.240APG + 0.812BPG + 1.596TOPG$$

$$(0.588) \qquad (0.096) \qquad (0.280) \qquad (0.242)$$

$$R^2 = 0.904, \bar{R}^2 = 0.903, n = 465$$

Frankly, this is not a good result by all accounts, but it is not the model we ultimately want, so we may as well skip it for now.

Among the model, the ***t* statistic of *2PA* is the largest, which is 15.14504**. According to the rules of the basketball game, there are three ways for players to score, two-pointers inside the restricted area, three-pointers outside the restricted area and free throws (one-pointer). Naturally, players have obviously more opportunities to shoot in the penalty area than the latter two. So, apparently, the number of two-pointers made is the most important indicator of a player's scoring ability[4-5].

Besides, the variable of *3PA* also has a statistically significant ***t*-value of 9.4**, with the reason that three-pointers are the most points that can be obtained in every single hit, and their hits also greatly affect the scoring efficiency of players. What's more, it is worth noting that the ***t* statistic of the number of** *TOPG* **is 6.59**, which cannot be ignored, and **the coefficient of** *TOPG* **is 1.596**, with a positive sign, which seems counterintuitive.

However, according to Yao (2008), turnovers per game also to some extent measure the number of scoring attempts by players on the field. And in the process of trying to break through and shoot by players, mistakes often occur. Therefore, more turnovers means that a player has more chances to shoot and score, and the more points he could score.

Fortunately, **except for the two variable** ***POS*** **and** ***APG*****, the *t* statistic values of the other variables are all above 3**, and in the case of degrees of freedom=467-9-1=457, those variables are all significant at the level of α=0.005.

Referring to the statistical method of the article "Influencing Factors and Regression Analysis of NBA Players' Total Career Score in Playoffs—Based on the Panel Data of NBA Playoffs from 1948 to 2017" and the results of exploratory factor analysis in the article, we use *TS%*, *3PA*, *RPG*, *SPG*, *TOPG* and *BPG* as regressors for regression. The regression results are as follows *Table 3*:

Table 3: Statistics data of the revised regression

	Coefficients	SE	*t* Stat	*p*-value
Intercept	2.34579835	0.445896	-5.26086	2.21E-07
TS%	3.74462577	0.767775	4.877246	1.49E-06
3PA	0.076447	0.004776	16.0049	4.09E-46
RPG	0.47774696	0.079961	5.974739	4.63E-09
SPG	0.86142927	0.352837	2.441435	0.015007
TOPG	3.64700785	0.219349	16.62647	6.55E-49
BPG	0.59236083	0.357102	1.6588	0.097841

Among them, **the *t* statistics of** *SPG* **and** *BPG* **are 2.44 and 1.66**, respectively, which belong to a comparative level. **Considering that they do not directly contribute to the score, this result is also very reasonable**. The other results are almost consistent with the regression results that we obtained above.

Lastly, in order **to prove that there is no heteroskedasticity in the above regressions**, it is supposed to give a test, which uses the methodology of White test as well as Breusch-Pagan test for heteroskedasticity. **With the reasonable *t* statistic values of 4.88 and 5.97**, the form of the new regression has to be $\hat{u}^2$=$\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 {x_1}^2 + \delta_4 {x_2}^2 + \delta_5 x_1 x_2 + error$, where x_1 and x_2 are supposed to be *RPG* and *TS%*, respectively. In addition, the hypothesis is that δ_1=δ_2=δ_3=δ_4=δ_5=0. The regression results are as below *Table 4*:

Table 4: Statistics data of the new regression

	Coefficients	*STD*	*t Stat*
Intercept	-2.21924428	12.414735	-0.17876
RPG	4.657388855	6.8250452	0.682397
TS%	-13.4335308	42.74812	-0.31425
RPG^2	-0.84716017	0.3309613	-2.5597
$(TS\%)^2$	5.965091701	36.326688	0.164207
$RPG{\cdot}TS$%	18.20082591	10.707321	1.699849

More importantly, **the F-value of the test is 2.33748**, which is not large enough to reject the former hypothesis. Therefore, it is certain that the regression is reasonable, at least in terms of the two variables, which are *RPG* and *TS*, without the phenomenon of heteroskedasticity.

However, **the normal White test below remains some flaws because the regression does not consider enough possible forms of explanatory variables**, so the results of *F*-value may be not informative enough. Nevertheless, if adding all the control variables and then conducting the White test, the calculation may be so complex because of the including of interaction terms as well as the square and higher power terms.

Therefore, as following, an alternative of White test will be conducted, with the form of $\hat{u}^2$=$\delta_0 +$ $\delta_1\hat{y}$+$\delta_2\hat{y}^2 + error$ and the hypothesis of H_0: $\delta_1 = \delta_2 = 0$. And the regression information is as below *Table 5*:

Table 5: Statistics data of the new regression

	Coefficients	*STD*	*t Statistics*
Intercept	1.833	1.434	1.278
y	3.113	0.301	10.349
y^2	0.243	0.013	18.247

Similarly, the F-value of the regression is large enough, which is 3.41, whose corresponding p-value is very far from 0, with degrees of freedom=2, 462 (465-3). Then, we are not supposed to reject the null hypothesis, which means there exists almost no heteroskedasticity in the regression model.

In addition, with the R-square of 0.068, LM $= nR_{\hat{u}}{}^2 \sim \chi 2_2$, which is 29.172, which verifies that it does not exist heteroskedasticity of the regression because we can certainly reject the null hypothesis entirely in the significance level of α=0.1.

In conclusion, we figure out what influence players' wages and scoring average.

In terms of points, *TS, 3PA, RPG, SPG TOPG, BPG, 2PA* as well as *2P* are important influential factors. Nevertheless, *TOPG* as well as *3PA* are two vital regressors as for the regressions with the dependent variable of PPG.

5. General model (Salary) section

Since the volatility of the dependent variable (salary) is relatively large, we **take a logarithmic approach ($\boldsymbol{log(Salary)}$) to average out the effect of the magnitude of the fluctuations as its value grows**, and thus reduce the heteroskedasticity.

$$\log(\widehat{Salary}) = -0.612 + 1.189Height - 0.007Weight - 0.032Age + 0.128Experience$$

$$(1.359) \quad (0.736) \qquad (0.005) \qquad (0.022) \qquad (0.024)$$

$$-0.085Country + 0.08PPG + 0.052APG$$

$$(0.089) \qquad (0.008) \qquad (0.029)$$

$$R^2 = 0.557, \bar{R}^2 = 0.549, n = 465$$

In this regression model we observed that:

- **Age vs. Experience**

Generally speaking, people may think that professional athletes' athletic performance will gradually decline with age, and the negative coefficient of Age in the model is also in line with this. Unexpectedly, the effect of Age is minimal compared to the regression of good performance in field $Experience$, **with the absolute value of t-statistic only 1.487**, which is obviously not significant.

However, this result is also acceptable. On the one hand the metric for increasing base salary in the NBA's salary system is the player's length of service. On the other hand, using playing age ($Experience$) to measure maturity and growth in skill level is indeed more convincing than Age alone.

- **Country**

For the dummy variable of group, we tried to find out whether the NBA system is conducive to the development of American national players. However, the result is contrary to our conjecture like, **the negative coefficient indicates that foreign players tend to be more competitive.** (Although this observation is insignificant with an absolute value of 0.949 for the t-statistic)

To put it another way, this may be reasonable. It is not that foreign players receive certain subsidies to play in the NBA, but rather that foreigners who are eligible to play in NBA games inherently cross a higher competitive threshold.

- **Height & Weight**

Neither of these two factors performed better in the regression, which is a side-effect of the fact that a small physical difference between players at the NBA level does not have much of an impact. However, we are willing to retain $Height$ as an explanatory variable for subsequent studies.

- ***PPG***

- Fortunately, **the most significant item in this model is *PPG***, which coincides with our previous suspicions. In fact, the higher scoring players tend to get more attention from fans and sponsors, thus bringing revenue to the club, which will be reflected in the individual's salary.

After deleting the variables that we do not want to discuss anymore, we obtain:

$$\begin{gathered}\log(\widehat{Salary}) \;= -0.892 + 0.658 Height + 0.091 Experience - 0.08 Country \\ (1.067) \quad (0.521) \qquad (0.009) \qquad\qquad (0.089) \\ +0.081 PPG + 0.057 APG \\ (0.008) \qquad (0.029) \\ R^2 = 0.552, \bar{R}^2 = 0.547, n = 465\end{gathered}$$

Then, we want to make a new attempt:

Combining the dummy variable of group with in-field location, while relaxing the assumption that "in-field location raises are the same for group everywhere" and considering their joint effect.

$$\begin{gathered}\log(\widehat{Salary}) \;= -0.657 + 0.514 Height + 0.09 Experience + 0.08 PPG + 0.059 APG \\ (1.154) \quad (0.568) \qquad (0.009) \qquad\qquad (0.008) \quad (0.029) \\ -0.016 AmerFord - 0.062 AmernonFord + 0.109 nonAmerFord \\ (0.124) \qquad\qquad (0.131) \qquad\qquad (0.149) \\ R^2 = 0.553, \bar{R}^2 = 0.546, n = 465\end{gathered}$$

In this regression model we observed that:

It is easy to see that several of the added dummy variables do not perform well in the model, and their F-statistics are nearly zero, which is clearly **not jointly significant**. Since their respective regressions performed equally poorly, **we eventually decided to drop their discussion and obtained the following model:**

$$\begin{gathered}log(\widehat{Salary}) \;= -1.266 + 0.813 Height + 0.09 Experience + 0.08 PPG + 0.059 APG \\ (0.998) \quad (0.497) \qquad (0.009) \qquad\qquad (0.008) \quad (0.029) \\ R^2 = 0.551, \bar{R}^2 = 0.547, n = 465\end{gathered}$$

Where the confidence interval for $Height$ is $0.813 \mp 1.96 \times 0.487$.

The situation reflected of residual plot **is largely acceptable** to us.

6. Conclusions and model evaluation

6.1 Our Results

- In terms of on-court stats, the most important factors affecting scoring are 3-pointers per game (*3PA*) and turnovers per game (*TOPG*).

- In terms of measuring a player's value, points per game (*PPG*) ranked first, followed by on-field experience (*Experience*) and assists per game (*APG*), while other off-field factors were not important.

- Considering **lagged dependent variable** to help reduce the impact of unquantifiable explanatory variables is **difficult** to do due to the mechanics of NBA salary contracts.

6.2 Model Assessments

Advantages:

- We used the ordinary least squares (OLS) regression method with Gauss-Markov theorem to find out the proper regression models based on the collected data, which can help reduce the error, and then we got satisfactory results.

● Various modifications were made to the variables in the model to reduce the heteroskedasticity of the model.

Disadvantages:

● Some relevant variables that cannot be omitted are not statistically significant enough.

● There is correlation between variables in the model, which has bad effect on the regression model.

● The fitting effect is not good enough in the model that we obtained.

References

[1] Cheng Qiyun, Sun Caixin, Zhang Xiaoxing, et al. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic [J]. Transactions of China Electrotechnical Society, 2004, 19(10): 53-58.

[2] Meng Q, Jia L. Research on total factor productivity of Shanghai services and its influence factors[C]// 2011 IET International Conference on Smart and Sustainable City. SILC, Shanghai University, China;, 2011.

[3] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability [J]. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[4] Ma Kunlong. Short term distributed load forecasting method based on big data [D]. Changsha: Hunan University, 2014.

[5] Shi Biao, Li Yu Xia, Yu Xhua, Yan Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model [J]. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.