Proceedings of CONF-CDS 2025 Symposium: Application of Machine Learning in Engineering
DOI: 10.54254/2755-2721/2025.AST26387

Application of Linear Regression in Basketball Analytics: Illustrated by Two Case Studies

Hao Deng

School of Physics, Mathematics and Statistics, Newcastle University, Newcastle Upon Tyne, United Kingdom
c3043227@newcastle.ac.uk

Abstract. As sports have evolved over the last hundred years, scientific approaches have proven essential for studying athletic phenomena. Linear Regression, a method reveals relationship between dependent variable and independent variables, is favored by statisticians by its prediction power and computational efficiency. This article explores the application of linear regression as a scientific method in basketball analytics by separately examining key factors influencing NBA players' skill levels and salaries. Among independent variables such as duration of playtime, steals per game, efficiency rating and professional age in correspondence of dependent variables, the skill levels and salaries of players are selected based on data provided by American Basketball League. To investigate this further, the author constructed regression models with R program to assess significant factors affecting two dependent variables respectively. Moreover, visualization charts and plots were made to ensure the validity of the regression models. Lastly, results indicated that determinants like duration of playtime, professional age showed low p-value and VIF (variance inflation factors) implying significance and no risk of multicollinearity.

Keywords: Multiple Linear Regression, NBA, Salaries of Players, Skill Level of Players

1. Introduction

The origin of basketball can be traced to 134 years ago, when it was invented by James Naismith as a recreational indoor sport. With the refinement of the rules, basketball has become a symbol of competitive sports. Nowadays, people's enthusiasm for basketball has swept across the world. NBA, the national basketball association, is the forefront where it gathers top basketball athletes from all over the world. According to data from ESPN [1], the 2024 NBA final captivated 3,922,000 viewers watching the game via YouTube TV proving the popularity of this sport. Correspondingly, the huge profit and potential of the league drive sponsors to study not only how to improve players' performance but also what are the decisive factors behind players' salaries.

So far, researchers have leveraged multiple scientific approach to analyze phenomena in sports. In the work by Papageorgiou et al., the authors mainly focus on selecting optimal method for performance prediction of basketball players [2]. Eventually, the researchers encouraged random forest due to its high applicability to performance indicators specifically for context. Whereas the other project researched on recognition of basketball strategies by means of standard machine

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.AST26387

learning approach [3]. The researchers used three classification models: Quadratic SVM, K-NN Classifier and Decision Trees. This machine learning technique was able to achieve up to 69% accuracy in the end. However, this paper adopts linear regression that is a well-established estimation method renowned for its high precision and user-friendliness. Scientists employed linear regression to investigate market value of football players in forward position [4]. This paper offers insights demonstrating the potential applicability of linear regression across various sports, for example basketball.

The research topic is the application of linear regression in basketball, as illustrated by two examples: evaluating basketball players' skill level and finding significant factors affecting basketball player's salary. This paper will use R programming language to generate scatter plots of different independent variable presenting their mutual independence. Subsequently, the relative statistical information produced by R will indicate key determinants. Finally, diagnostic plots including Normal Q-Q plots will be analyzed to validate the regression model's assumptions, supporting the robustness of the conclusions.

2. Methodology

2.1. Simple linear regression model

The linear regression model is used to investigate the influence of factors towards a variable to be analyzed or to be predicted. Technically speaking, the linear regression model demonstrates to what extent explanatory/independent variable affecting response/dependent variable. The research by Anandhi and Nathiya vividly outlines the advantages of linear regression [5]. First, it performs optimally when the underlying dataset exhibits linear separability. Secondly, it measures important factors that tends to influence dependent variable. Thirdly, it is easy to understand and implement, making it accessible for both beginners and experts. The coefficients directly indicate the relationship between predictors and the target variable. Finally, it produces transparent predictions, while its accuracy scales with dataset quality and size.

Linear regression is a widely used method in fields like medicine, where it helps improve the accuracy of surgical time predictions. For example, models may use surgeon-controlled time as an explanatory variable alongside other factors [6]. Beyond healthcare, it also aids business analytics. Yue Shi's enhanced linear regression model, for instance, was designed to optimize precision in e-commerce sales forecasting [7].

Suppose that the observed values of the response variables $z_1, z_2, z_3, \cdots, z_n$ are realizations of random variables $Z_1, Z_2, Z_3, \cdots Z_n$. An appropriate simple linear regression model is

$$Z_i = \delta_0 + \delta_1 g_i + \epsilon_i \tag{1}$$

for $i = 1 \cdots n$ (given that a dataset of n realizations is provided) where δ_0 and δ_1 are called regression coefficients and Z_i is the response variable also g_i is the explanatory variable. Moreover, to ensure the feasibility of this statistical model, certain probabilistic assumptions are indispensable. Firstly, ϵ_i, the error term, is regarded as a normally distributed random variable. Secondly, ϵ_i has expectation whose value is 0 and a variance of value of σ^2. Thirdly, each ϵ_i is independent of each other, considering the existence of multiple error terms. In other words, it could be written in a formal way as $\varepsilon_i \sim N\left(0, \sigma^2\right)$ for $i = 1,2,3, \cdots, n$. In addition, the way of

DOI: 10.54254/2755-2721/2025.AST26387

computing the maximum likelihood estimator which will be explained later in the next section of regression coefficients (δ_0 and δ_1) is:

$$\widehat{\delta_0} = \bar{z} - \widehat{\delta_1}\bar{g}, \widehat{\delta_1} = \frac{\sum g_i z_i - n\bar{g}\bar{z}}{\sum g_i^2 - n\bar{g}^2}. \tag{2}$$

2.2. Multiple linear regression model

The multiple linear regression model is an advanced form of the simple linear regression model by introducing two or more explanatory variables into the model. This is due to the fact that, in practice, certain phenomena are influenced by multiple factors rather than single factor (eg. For predicting a person's weight, Height is an important explanatory variable, yet it is likely that Age will also help to explain the variation in weights). Therefore, the simple linear regression model can be expanded to contain p predictor variables $G_1, G_2, G_3 \cdots G_p$. This gives the multiple linear regression model:

$$Z_i = \delta_0 + \delta_1 g_{i1} + \delta_2 g_{i2} + \ldots + \delta_p g_{ip} + \epsilon_i \tag{3}$$

where it happens the same to the error terms in multiple linear regression model that ϵ_i are independent, for $i = 1,2,3,\cdots,n$. However, it is significant to highlight the word "Linear" which stands for the linearity between regression coefficients δ_i for $i = 0,1,2,\cdots,p$. For example, if $Z_i = \delta_0 + \delta_1 g_{i1} + \delta_2 g_{i2}^2 + \epsilon_i$, it is a well-defined multiple linear regression model. By contrast, if $Z_i = \delta_0 + e^{\delta_1 g_{i1} + \delta_2 g_{i2}} + \epsilon_i$, it is not a multiple linear regression model as the linearity is violated in between δ_1 and δ_2 .

Suppose there are n observations, the multiple linear regression model then becomes:

$$Z_1 = \delta_0 + \delta_1 g_{11} + \delta_2 g_{12} + \ldots + \delta_p g_{1p} + \epsilon_1$$

$$Z_2 = \delta_0 + \delta_1 g_{21} + \delta_2 g_{22} + \ldots + \delta_p g_{2p} + \epsilon_2$$

$$\vdots$$

$$Z_n = \delta_0 + \delta_1 g_{n1} + \delta_2 g_{n2} + \ldots + \delta_p g_{np} + \epsilon_n$$

Most of time, the value of n is way larger than the value of p . This model can also be expressed in an alternative form

DOI: 10.54254/2755-2721/2025.AST26387

$$\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix} = \begin{bmatrix} 1 & g_{11} & g_{12} & \cdots & g_{1p} \\ 1 & g_{21} & x_{22} & \cdots & g_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & g_{n1} & x_{n2} & \cdots & g_{np} \end{bmatrix} \begin{bmatrix} \delta_0 \\ \delta_1 \\ \vdots \\ \delta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \tag{4}$$

Alternatively, it may be more compactly rewritten as

$$Z = G\delta + \varepsilon \tag{5}$$

where $\boldsymbol{Z}$ is called response vector, $\boldsymbol{G}$ is noted as design matrix, $\boldsymbol{\delta}$ is the parameter vector of partial regression coefficients, $\boldsymbol{\varepsilon}$ is the error vector.

To find relationship between the response variable and explanatory variables, the major part is the estimation of regression coefficients. Hence, a proper multiple linear regression model of estimation is needed:

$$\widehat{Z} = \widehat{\delta_0} + \widehat{\delta_1} G_1 + \widehat{\delta_2} G_2 + \ldots + \widehat{\delta_k} G_k. \tag{6}$$

Among various estimation methods, the maximum likelihood estimation prevails for its consistency and asymptoticity meaning that it gives reliable estimation for large amount of sample size meanwhile the estimator follows normal distribution. Despite the computational challenge of the MLE method, it is still widely favored as a way of estimation in statistics due to its outstanding estimation precision. The maximum likelihood estimator of $\boldsymbol{\delta}$ is calculated as:

$$\delta = \left(\boldsymbol{G}^T \boldsymbol{G}\right)^{-1} \boldsymbol{G}^T Z \tag{7}$$

3. Application

3.1. Evaluating basketball player skill levels

In the work by Wu, multiple linear regression model has been introduced as a way of studying the skill level of basketball players [8]. The study analyzes data from a sample of 300 NBA basketball players, consisting of the top 10 performers in each game across the 82-game regular season (excluding playoffs) of the 2022-2023 season. The dataset was sourced from Sina's website [9].

To leverage multiple linear regression model, selection of factors/explanatory variables is crucial. The author picks these 11 factors affecting response variable Z (average score of each player) based on data from sina.com, see Table 1.

DOI: 10.54254/2755-2721/2025.AST26387

Table 1. Statistical description of the 11 explanatory variables

Index	Explanatory Variable	Representation	Index	Explanatory Variable	Representation
1	G_1	Height	7	G_7	Turnovers Per Game
2	G_2	Weight	8	G_8	Steals Per Game
3	G_3	Minutes Played Per Game	9	G_9	Blocks Per Game
4	G_4	Field Goal Percentage	10	G_{10}	Personal Fouls Per Game
5	G_5	Rebounds Per Game	11	G_{11}	Assist Loss Ratio
6	G_6	Assist Per Game			

The article describes using R program to process the data, the commands pairs() and boxplot() are used to create a scatter plot matrix (see Figure 1) which shows the relationship between each explanatory variable, but most importantly, the independence between each variable.

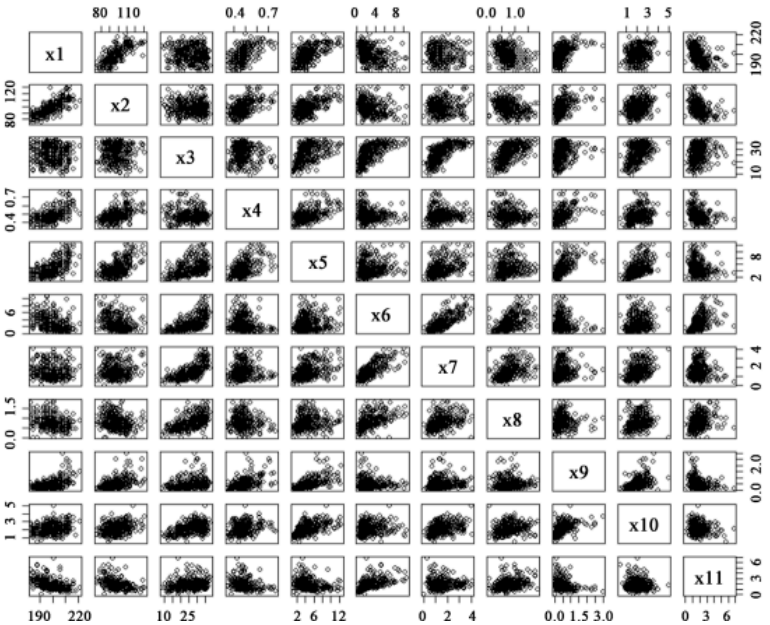

Figure 1. Scatter plot matrix between different independent variables [8]

After implementing necessary commands such as: lm() and summary(), R provides the statistical information about the analyzed target as shown in Table 2. In this table, the p-value column represents the value for the hypothesis test on each regression coefficient, testing whether it significantly differs from zero. The lower the p-value is, the stronger the evidence it provides to

DOI: 10.54254/2755-2721/2025.AST26387

reject null-hypothesis: $\delta_i = 0$ and to accept alternative hypothesis: $\delta_i \neq 0$, meaning that the corresponding explanatory variable is significant to response variable (average score of the player). In this case, the significance level, the threshold at which the null hypothesis is rejected if the p-value falls below it, is set at 0.05. It is then clear to see that, the variables $G_3, G_4, G_6, G_7, G_{10}$ are significant factors affecting average score of each player, while others contribute minimally. However, G_7 (turnovers per game) is removed due to the little research significance it holds. Since the errors a player makes do not influence their personal points scored in the game. Thus, it is meaningless to calculate its VIF (variance inflation factor) value which is being noted as NA instead.

Table 2. Descriptive statistical information (case 1)

Variable	Estimate	T value	P-value	Variance Inflation Factor
(Intercept)	-6.905750	-0.99g	0.31987	NA
Minutes Played per Game (G_3)	0.601878	13.812	< 2e-16	2.1392
Field Goal Percentage (G_4)	7.325942	2.467	0.01421	1.1145
Assist per Game (G_6)	-0.676724	-2.607	0.00961	1.7356
Turnovers per Game (G_7)	4.976979	8.355	2.80e-15	NA
Personal Fouls per Game (G_{10})	-1.943013	-5.889	1.08e-08	1.4392
R-squared value: 0.8108	Adjusted R-squared value: 0.8035	NA	NA	NA
F-ratio value: 112.2 on 11 and 288 Degree of Freedom	F-ratio p-value: < 2.2e-16	NA	NA	NA

Based on the analysis above, duration of played time(G_3), shooting percentage(G_4), assist per game(G_6), personal fouls per game (G_{10}) are the four major factors with influencing the average score of each player(Z). Furthermore, there is no such worry of multicollinearity due to their low VIF values. Eventually, a specific multiple linear regression is obtained as

$$Z = -6.905750 + 0.601878G_3 + 7.325942G_4 - 0.676724 - 1.943013G_{10} + \epsilon. \quad (8)$$

3.2. Exploring salary determinants for NBA players

Following a similar approach, Zhao investigated salary influencers among NBA players through multiple linear regression modeling [10]. Based on data sourced from the basketball-reference website [11], the author of that paper selects 14 determinants as explanatory variables, see Table 3.

DOI: 10.54254/2755-2721/2025.AST26387

Table 3. Statistical description of the 14 explanatory variables

Index	Variable	Representation	Index	Variable	Representation
1	G_1	height	8	G_8	box plus/minus
2	G_2	weight	9	G_9	value over replacement player
3	G_3	age	10	G_{10}	team winning percentage
4	G_4	professional age of player	11	G_{11}	continent of player
5	G_5	playing time per game of player	12	G_{12}	position
6	G_6	efficiency rating of player	13	G_{13}	market value of the team
7	G_7	win share	14	G_{14}	all-star team experience

Scatterplots which are shown in Figure 2 of each explanatory variable between salary are obtained via R program. It can be seen that variables like: G_1, G_2, G_3, G_4, G_8 do not possess obvious linear correlation with the salary of basketball player. Yet, the plots manifest clear linear correlation between each variable such as G_5, G_6, G_7, G_9 and the salary of basketball player.

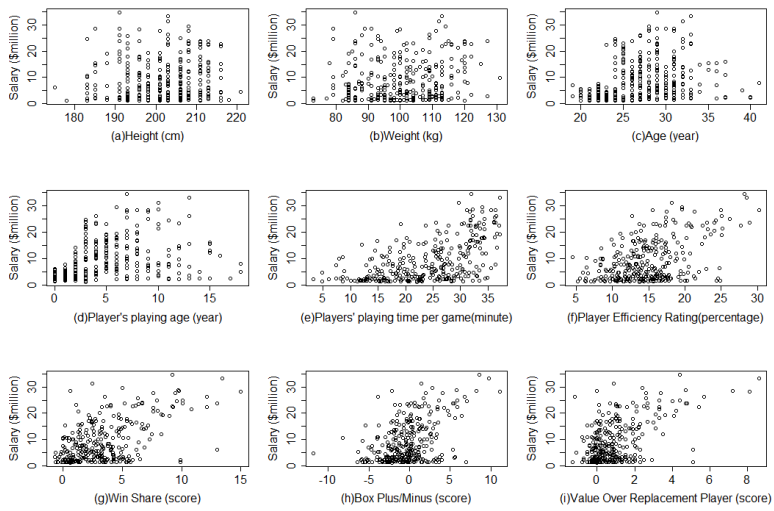

Figure 2. Scatterplots of explanatory variables and salaries [10]

The Table 4 shows the statistical information for certain regression coefficients that only pass the significance test whose significance value is set to be 0.05 as well. This is due to their noticeably

DOI: 10.54254/2755-2721/2025.AST26387

low p-value, which serve as compelling evidence of their statistical importance. Furthermore, the fact that all VIF values less than 5 symbolizes no multicollinearity between the independent variables. Meanwhile R-squared value suggests that approximately 58.39% of the variation of salaries is explained by the model. Consequently, the regression model achieves its final optimized form as:

$$Z = -10.079 + 0.0761G_1 + 0.414G_4 + 0.323G_5 + 0.707G_9 + 8.850G_{14} + \epsilon \tag{9}$$

Table 4. Data of multiple linear regression model

Variable	Estimate	T value	P-value	VIF
Intercept	-10.079	-3.298	0.00109	NA
Weight of Player (G_1)	0.076	2.885	0.00421	1.07
Professional age of Player (G_4)	0.414	5.351	1.76e-07	1.14
Playing Time per Game of Player (G_5)	0.323	6.628	1.61e-10	1.66
Value Over Replacement Player (G_9)	0.707	2.517	0.01237	1.84
All-Star Team Experience (G_{14})	8.850	8.277	4.43e-15	1.61
R-squared value: 0.5908	Adjusted R-squared value: 0.5839	NA	NA	NA
F-ratio value: 85.18 on 5 and 295 Degree of Freedom	F-ratio p-value: < 2.2e-1	NA	NA	NA

Lastly, the Residual vs fitted value plot and Q-Q plot fully justifies the assumptions:

(i). the relationship between $E[Z]$ (the expectation of response variable Z) and g is linear;

(ii). ϵ_i are normally distributed;

(iii). ϵ_i are independent and therefore uncorrelated;

(iv). each error term, ϵ_i , has zero mean;

(v). each error term, ϵ_i , has variance σ^2 .

Since in residual vs fitted value plot as shown in Figure 3, there is no clear pattern in the residuals and variance does not appear to change in a systematic way. Additionally, the majority of points lie in between the section from -2 to 2. Thus, there is no clear evidence of a departure from assumptions of constant variance and a straight-line relationship. In addition, studentized residuals lie fairly close to line of equality in Q-Q plot referring no clear evidence of departure from an assumption of normal errors. Overall, it indicates the validity of multiple linear regression model.

DOI: 10.54254/2755-2721/2025.AST26387

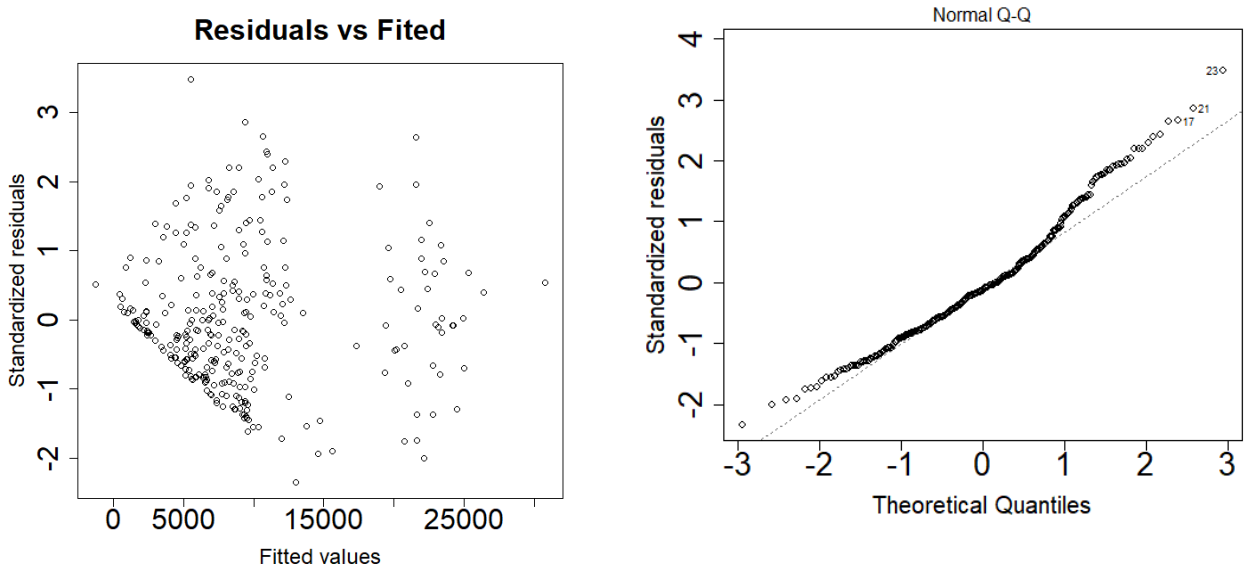

Figure 3. Residuals vs Fitted value plot & Normal Q-Q plot [11]

4. Conclusion

This paper applies multiple linear regression to identify factors influencing average score and salaries of each NBA player. Based on the findings, it is evident that Field Goal Percentage and Minutes Played serve as crucial factors of average score of players. Even the factor as Turnovers Per Game has low p-value, it has been extracted out due to its minimal research value. Whereas indicators related to physical condition as Height and Weight show negligible impact. On the other hand, Professional Age and Popularity of players demonstrate direct influence towards players' salaries. This suggests that a player's value heavily depends on his on-court performance and charismatic influences. Most importantly, the issue of multicollinearity and departure of regression assumption was avoided which can be fully explained by low VIF value and Residual fitted & Normal Q-Q plots. Given that the accuracy of the model had not reached to its optimal level owing to a lack of sufficient data, however the multiple linear regression model still successfully managed to reveal key factors in a straightforward way.

Changes can be made for the improvement of regression model. To begin with, the model can be enhanced by increasing sample size instead of picking only top 10 players for each game which leads to data insufficiency for every position. As factor like Injury Resilience which is not included also affects the skill level and salaries of players, thus the model can introduce more independent factors to optimize accuracy. Future research, such as improving player skill levels and increasing salaries, could be conducted to enhance the team's win rate and create a better-paid environment for players.

References

[1] ESPN. (n.d.). NBA Countdown on ABC during the 2024 NBA Finals presented by YouTubeTV delivers largest viewership since 2019. ESPN Press Room. https: //espnpressroom.com/us/

[2] Papageorgiou, G., Sarlis, V., & Tjortjis, C. (2024). Evaluating the effectiveness of machine learning models for performance forecasting in basketball: A comparative study. Knowledge and Information Systems, 66, 4333–4375.

DOI: 10.54254/2755-2721/2025.AST26387

[3] Tian, C., De-Silva, V., Caine, M., & Swanson, S. (2019). Use of machine learning to automate the identification of basketball strategies using whole team player tracking data. Applied Sciences, 10(1), 24.
[4] Koloğlu, Y., Birinci, H., Kanalmaz, S. I., & Özyılmaz, B. (2018). A multiple linear regression approach for estimating the market value of football players in forward position. arXiv. https: //doi.org/10.48550/arXiv.1807.01104
[5] Anandhi, P., & Nathiya, E. (2023). Application of linear regression with their advantages, disadvantages, assumption and limitations. International Journal of Statistics and Applied Mathematics, 8(6), 133–137.
[6] Edelman, E. R., van Kuijk, S. M. J., Hamaekers, A. E. W., de Korte, M. J. M., van Merode, G. G., & Buhre, W. F. F. A. (2017). Improving the prediction of total surgical procedure time using linear regression modeling. Frontiers in Medicine, 4, 85.
[7] Shi, Y. (2023). Application of improved linear regression algorithm in business behavior analysis. Procedia Computer Science, 228, 1101–1109.
[8] Wu, J. J. (2024). Statistical approach to basketball players' skill level. Journal of Applied Mathematics and Physics, 12, 1352–1363.
[9] Sina. (n.d.). Slamdunk. https: //slamdunk.sports.sina.com.cn/teams
[10] Yang, Z. (n.d.). Model prediction of factors influencing NBA players' salaries based on multiple linear regression. Atlantis Press. doi: 10.2991/978-94-6463-036-7_213
[11] Basketball-Reference. (n.d.). Data source. https: //www.basketball-reference.com/