Statistical Approach to Basketball Players' Skill Level

Jiajun Wu

School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China
Email: 2021311014@email.cufe.edu.cn

Abstract

In basketball, each player's skill level is the key to a team's success or failure, the skill level is affected by many personal and environmental factors. A physics-informed AI statistics has become extremely important. In this article, a complex non-linear process is considered by taking into account the average points per game of each player, playing time, shooting percentage, and others. This physics-informed statistics is to construct a multiple linear regression model with physics-informed neural networks. Based on the official data provided by the American Basketball League, and combined with specific methods of R program analysis, the regression model affecting the player's average points per game is verified, and the key factors affecting the player's average points per game are finally elucidated. The paper provides a novel window for coaches to make meaningful in-game adjustments to team members.

Keywords

Physics-Informed Statistics, Multiple Linear Regression, Average Score per Game, R Program Analysis

1. Introduction

Basketball, as a competitive sport, is not only one of the core events of the Olympic Games, but also has a high popularity worldwide. As a team sport, basketball matches are scored by each team based on the cumulative score of their players on the field. Therefore, the score of each player on the field directly affects the victory or defeat of the team. The selection of players on the field is a very challenging task for coaches. This is because the scoring performance of each player on the field is influenced by various factors, such as playing time, shooting percentage, number of assists, etc. It is a complex nonlinear process. During the

game, the issue of how to choose players to play often relies on the coach's on-site experience, and the statistical theory behind it is a meaningful research topic.

In recent years, domestic and foreign scholars have conducted in-depth research and analysis on the average score of players and the total score of teams in basketball from different perspectives. In 2015, Chen *et al.* used simple correlation analysis, descriptive statistics and other statistical methods to study the correlation between total salary of basketball clubs and team performance and player performance [1]. They concluded that player salary is directly influenced by five factors: average score per game, number of assists, rebounds, number of appearances, and average assists per game. In 2020, Song *et al.* conducted post-game technical statistics and analysis on five teams in the Xizang Women's Basketball League through expert interviews, cluster analysis and other methods. The results showed that high altitude also had a certain impact on women's basketball scores [2]. In the same year, Li *et al.* conducted a statistical analysis of the 2018-2019 CBA league season using multiple linear regression methods [3]. Using 10 technical indicators such as average rebounds and assists per game as independent variables, they established a regression model to find the key factors affecting the team's victory or defeat. Inspired by the research work of Li *et al.*, in 2022, Tan took Guangzhou team of Era China as the research target, and used various data indicators of the team's scores in the 2020-2021 season as independent variables, and the game results as the dependent variable for multiple linear analysis. The results have certain theoretical significance for the Guangzhou team of Era China to improve the winning rate of matches [4]. It can be seen that multiple linear regression, as a statistical analysis technique, is an important tool for theoretical research on the winning situation of modern competitive sports.

As a statistical analysis method for studying nonlinear problems, multiple linear regression models play a very important role in statistics, and are widely used in fields such as data analysis, prediction, and modeling [5] [6] [7]. Its main functions include modeling and prediction, variable screening and impact analysis, control variable method, and model diagnosis. For example, Huang *et al.* used a linear regression algorithm to conduct a statistical analysis of local tourism data [8]. Liu *et al.* estimated the demolition rate using multiple linear regression based on statistical data of building scale [9]; Zhang *et al.* conducted data statistics and analysis on the general equilibrium model of environmental dynamics using multiple linear regression, and identified important parameters of environmental patterns [10]. In 2020, Li *et al.* conducted an empirical statistical analysis of agricultural e-commerce data in China using this model [11]. In addition, in the process of implementing nonlinear regression models, R language is a method for handling nonlinear relationships between variables, which involves optimizing the parameters of nonlinear functions. It can better help researchers understand and sort out the nonlinear relationships between multiple variables, and is a common important tool used in combination with multiple linear regression models [12] [13].

As a typical nonlinear process, which key factors have a significant impact on the average score of basketball players per game? Can a detailed statistical analysis of player post-game technical data be conducted based on a nonlinear regression model to obtain theoretical results? There is currently no relevant research on the analysis process of game statistics in basketball. This article intends to start from official data provided by basketball matches, use data analysis techniques of multiple linear regression models, and combine with specific implementation methods of R language to analyze many factors that affect the average score of players on the field, providing theoretical support for basketball coaches to choose playing players.

2. Basic Theory of Multiple Linear Regression Models

The multiple linear regression model is an extended form of linear regression analysis technique commonly used to handle the impact of two or more factors on the variables to be analyzed or predicted results. Formally, it is a statistical method used to study the multilinear relationship between the dependent variable and the multivariate explanatory variable [14]. The general form of a regression expression involving k explanatory variables is:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k + \mu, \tag{1}$$

where, Y is the dependent variable, X_i is the explanatory variable, β_i is the regression coefficient, and μ is the error term. Here, the error term is a variable factor that is related to the dependent variable Y but cannot be explained by the previous k explanatory variables as well as their linear combinations.

In the process of applying multiple linear regression models to analyze practical problems, the following basic assumptions need to be met:

1) The error term, as a random variable, has an expected value of 0. Namely: $E(\mu) = 0$.

2) For all observed values of k explanatory variables $X_1, X_2, \cdots, X_k$, the variance value $D(\mu) = \sigma^2$ of the error term should be constant.

3) The error term follows a normal distribution $\mu \sim N(0, \sigma^2)$ and different errors are independent of each other.

In order to obtain a clear relationship between the dependent variable Y and various explanatory variables, it is necessary to estimate the regression coefficients β_i by observing the sample values, and establish an estimated regression expression:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots \hat{\beta}_k X_k. \tag{2}$$

The most commonly used methods for estimating regression coefficients are the least squares estimation method and the maximum likelihood estimation method. Among them, the least squares estimation method has advantages such as unbiasedness, consistency, and effectiveness. However, when there is multicollinearity between independent variables, its solution may not be unique

or exist. Maximum likelihood estimation has the characteristics of consistency and asymptotic normal distribution, but in practical applications, its computational complexity will rapidly increase with the size of the problem and the increase of observation data, and there may be problems such as non-unique estimation values. Therefore, in different application scenarios, other types of parameter estimation methods have emerged, such as instrumental variable estimation method, Bayesian estimation method, etc. Among them, Bayesian estimation method is a method based on Bayesian statistical theory, which estimates parameter values by introducing prior and posterior distributions. In Bayesian parameter estimation, different levels of significance can be set according to actual needs. For example, the significance level P-value can be set to 0.05 [15]. That is, when the statistical hypothesis test value is less than or equal to 0.05, the statistical result is considered significant; otherwise, it is considered that the statistical results are not significant.

3. Data Sources and Variable Selection

3.1. Data Sources

Due to the fact that the National Basketball Association (NBA) in the United States almost gathers basketball elites from various countries around the world, this article will analyze the data of NBA players in 82 regular season rounds (excluding playoffs) of the 2022-2023 season, with detailed data sourced from Sina.com [16]. As is well known, the NBA has 30 teams in the entire league, each with 15 players, and only about 10 players can play in each game. Due to other 5 players playing during garbage time, their playing time is very limited, so their statistical data is meaningless. Therefore, in this article, we select the top 10 players per game in each team for analysis. Therefore, the analysis of the average score per game for players in this article is based on the data of players playing with a total sample size of 300.

3.2. Selection of Interpretable Variables

In order to apply multiple linear regression models to analyze the average score of players on the field, it is necessary to analyze and screen multiple factors that affect the average score of players on the field [16]. Based on the basic data information of basketball players published on the website, first select the 11 feature attributes shown in Table 1 as candidate interpretable variables, and use the field average score as the dependent variable Y influenced by these interpretable variables.

The following is an analysis of the sample indicators of 300 players based on the R program. In order to clearly see the relationship between different random variables, a scatter plot matrix is constructed using the pairs() and boxplot() functions in the R program (see Figure 1). The scatter plot matrix can intuitively observe the distribution, correlation, and possible correlation between different random variables, while also discovering outliers in the sample data.

Table 1. Candidate interpretable variable indicators.

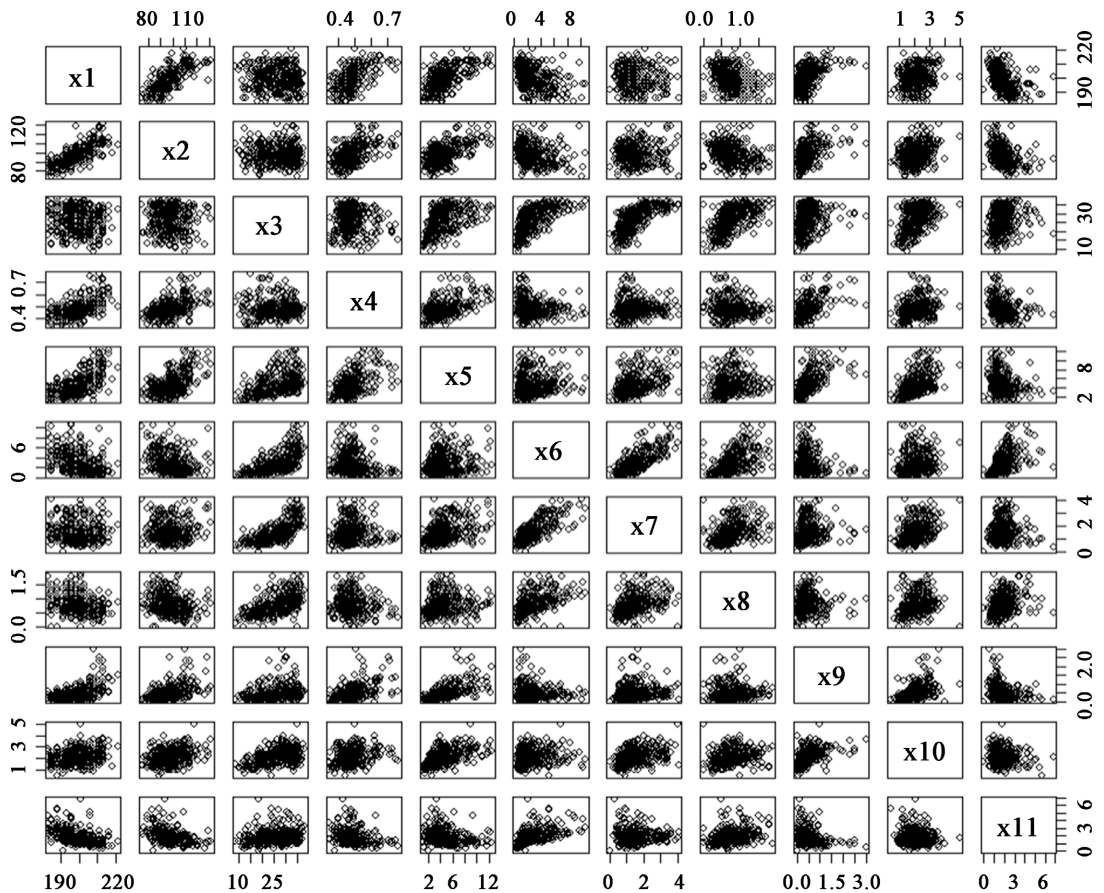| Serial Number	Symbols	Symbolic Representation
1	X_1	Height
2	X_2	Weight
3	X_3	Minutes Played per Game
4	X_4	Field Goal Percentage
5	X_5	Rebounds per Game
6	X_6	Assists per Game
7	X_7	Turnovers per Game
8	X_8	Steals per Game
9	X_9	Blocks per Game
10	X_{10}	Personal Fouls per Game
11	X_{11}	Assist Loss Ratio

Figure 1. Scatter plot matrix between different independent variables.

In the R program, the function lm()is used to fit a linear regression model, which is used to verify the linear relationship between variables, the error term of normal distribution, and homoscedasticity. When using the lm() function in

practice, it is necessary to carefully test and evaluate whether the sample data to be analyzed conforms to these assumptions. In addition, when analyzing data, the generalized linear model glm() function cannot be used for descriptive estimation due to the fact that the explainable variables do not follow a normal distribution, the variance of the error term is not constant, the binary countability of the explainable variables, and multi classification. Through analysis, it was found that using the multiple linear regression model lm() for scanning analysis is more reasonable, while paying attention to the logistic regression and constructing a linear relationship between the dependent variable Y and different interpretable variables.

To this end, now, we use the function lm() to obtain descriptive statistical information of the data, and then uses the summary() function to generate statistical information of the object, including descriptive statistics and model fitting results. Finally, using the ggplot2 library, draw a correlation coefficient graph between independent variables to determine whether the model assumptions are met and whether there are multicollinearity problems. Based on the sample data of 300 players, a correlation coefficient graph between different interpretable variables is shown in Figure 2.

3.3. Adjustment of Interpretable Variables

The descriptive statistical information of the full model logistic regression is
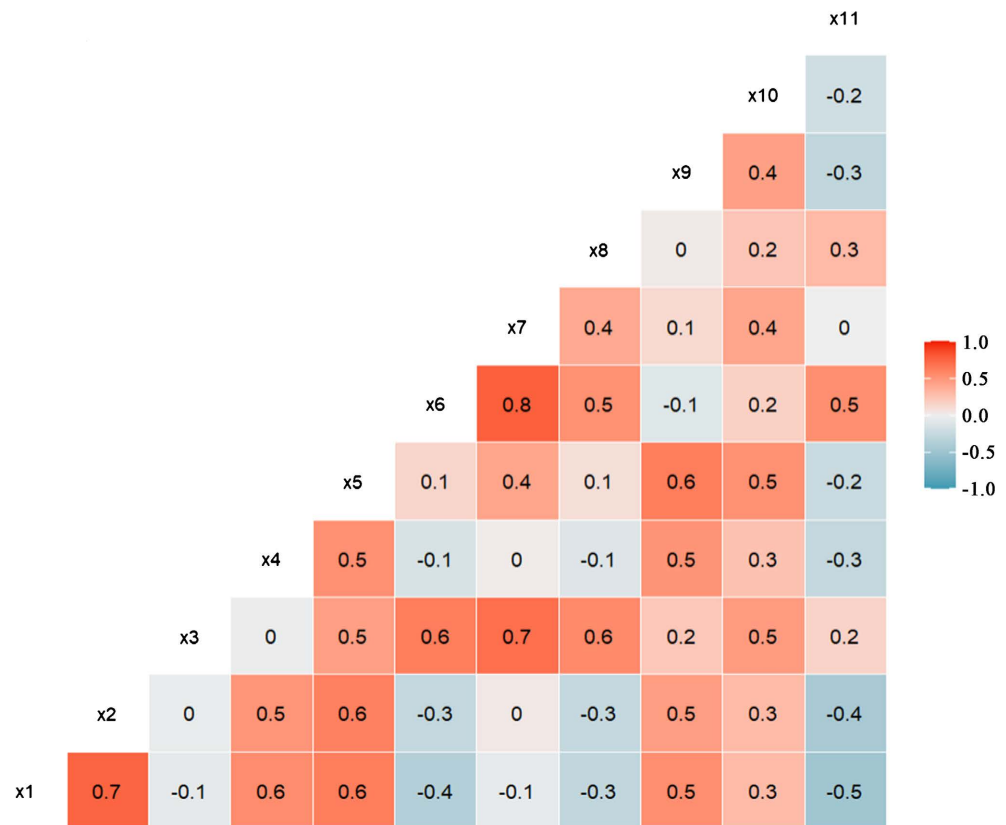

Figure 2. Correlation coefficient between independent variables.

calculated below, and the statistical information of 11 interpretable random variables can be obtained as shown in **Figure 3**. In multiple linear regression analysis, the *P*-value (*i.e.* the last column of **Figure 3**) is usually used to determine the degree of support of the observed data for the original hypothesis. When the *P* is large, it indicates that there is not enough evidence to support the observed effect, especially when the *P* is greater than the significance level (the significance level value set in this article is 0.05), we usually choose not to reject the original hypothesis. According to the regression coefficient p-value analysis results shown in **Figure 3**, it can be seen that X_1, X_2, X_5, X_8, X_9 and X_{11} have no significant contribution to the field average score Y of the dependent variable, and the hypothesis of parameter 0 cannot be rejected. Therefore, these independent variables will be removed to optimize the model and test its fit. In addition, X_7 is the "number of mistakes", and the characteristic of basketball is that the core players control the ball to organize the attack. The data item "number of mistakes" has little research significance, and will be removed during the experimental process.

After the above analysis, the autocorrelation coefficients and descriptive statistical information of logistic regression between the selected interpretable random variables X_3, X_4, X_6, and X_{10} are shown in **Figure 4** and **Figure 5**, respectively. By analyzing the value of P, it can be concluded that the four retained random variable values are the key factors affecting the value of the dependent variable Y. Therefore, X_3: average playing time per game, X_4: shooting percentage, X_6: average rebounds per game, and X_{10}: average fouls per game were selected as explainable random variables in the multiple linear regression model, and regression analysis was conducted based on the sample values of these random variables.

```
Call:
lm(formula = Y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7242 -1.9841  0.1106  1.5082  9.3928

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.905750   6.930367  -0.996  0.31987
x1           0.009593   0.037828   0.254  0.80000
x2          -0.018931   0.025691  -0.737  0.46181
x3           0.601878   0.043577  13.812  < 2e-16 ***
x4           7.325942   2.969753   2.467  0.01421 *
x5          -0.036848   0.140806  -0.262  0.79374
x6          -0.676724   0.259566  -2.607  0.00961 **
x7           4.976979   0.595682   8.355 2.80e-15 ***
x8          -0.428248   0.607145  -0.705  0.48116
x9          -0.261849   0.539320  -0.486  0.62768
x10         -1.943013   0.329926  -5.889 1.08e-08 ***
x11         -0.015945   0.349756  -0.046  0.96367
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.877 on 288 degrees of freedom
Multiple R-squared:  0.8108,    Adjusted R-squared:  0.8035
F-statistic: 112.2 on 11 and 288 DF,  p-value: < 2.2e-16
```

Figure 3. Descriptive statistical information of full model logistic regression.
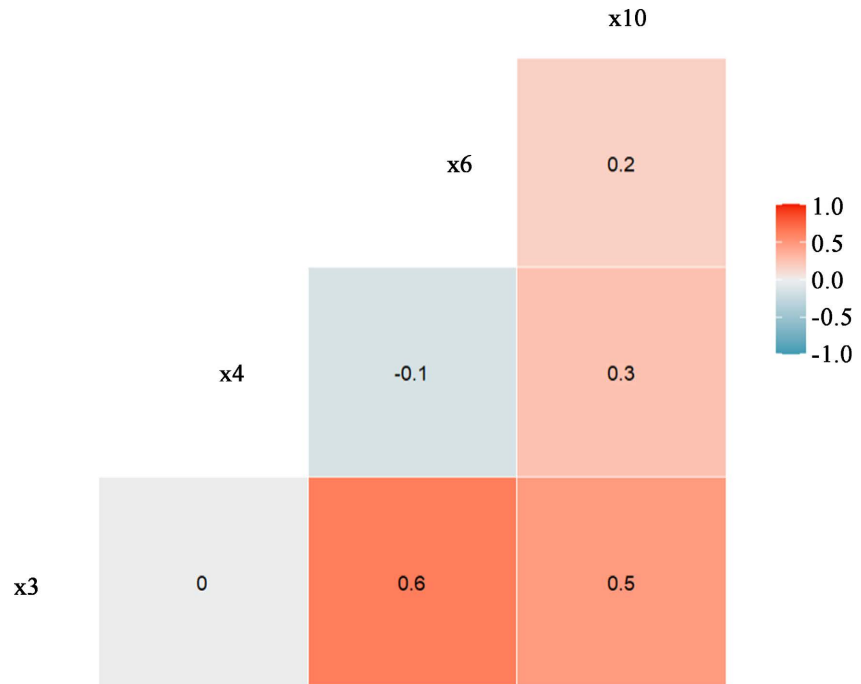
Figure 4. Correlation coefficient graph between independent variables after deleting other independent variables.

```
Call:
lm(formula = Y ~ x3 + x4 + x6 + x10, data = data)

Residuals:
     Min      1Q  Median      3Q     Max
-10.1462  -2.5043  -0.0954  2.3237  13.3374

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.7267     1.6013   -6.074 3.84e-09 ***
x3            0.7277     0.0449   16.208  < 2e-16 ***
x4            9.0963     2.9483    3.085 0.002227 **
x6            0.5082     0.1317    3.860 0.000139 ***
x10          -0.9122     0.3787   -2.409 0.016616 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.601 on 295 degrees of freedom
Multiple R-squared:  0.6963,    Adjusted R-squared:  0.6922
F-statistic: 169.1 on 4 and 295 DF,  p-value: < 2.2e-16
```

Figure 5. Descriptive statistical information of logistic regression in a new model after removing other insignificant independent variables.

Finally, the function vif() in the R program can be used to calculate the Variable Inflation Factor (VIF). Here, VIF is an indicator used to detect the degree of multicollinearity between independent variables in multiple linear regression models. According to the theoretical knowledge of multiple linear regression, the larger the VIF, the stronger the collinearity relationship between independent variables, which may lead to unstable estimation of regression coefficients. Usually, when the VIF is greater than 10, it is considered not an ideal estimation model. Calculate the VIF values for four random variables X_3, X_4, X_6, and X_{10}, and obtain the computation results given in Table 2. After determining these four

Table 2. VIF results for explainable random variables.

Explainable Random Variables	X_3	X_4	X_6	X_{10}
VIF Value	2.1392	1.1145	1.7356	1.4392

variables X_3, X_4, X_6, X_{10}, use the lm() function to obtain descriptive statistical information of the data, and then use the summary() function to generate statistical information of the object. The data.frame() function can be used to create a data frame object.

Based on the above analysis, it can be concluded that among the selected interpretable variables, each regression coefficient is highly significant ($P < 0.05$), and the VIF value of each random variable is less than 10. This means that the VIF method can be used to determine that the new model does not have multicollinearity issues. Therefore, the four selected random variables can ensure the stability and reliability of the regression model.

4. Modeling and Testing of Multiple Linear Regression on the Average Score of Basketball Players

4.1. Regression Modeling for the Problem of Average Score per Game

After selecting the explainable variable, the following will conduct multiple linear regression modeling analysis on the average score of NBA players based on the actual data obtained. To this end, the following multiple linear regression equation is proposed to be constructed:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4, \tag{3}$$

where, x_1, x_2, x_3, x_4 are four interpretable variables selected from the previous section. That is, x_1 is "Minutes Played per Game", x_2: "Field Goal Percentage", x_3: "Assists per Game" and x_4: "Personal Fouls per Game". The constant term β_0 and the undetermined coefficients $\beta_1, \beta_2, \beta_3, \beta_4$ need to be fixed by combining the sample values of four interpretable variables.

4.2. Parameter Estimation and Testing of the Problem of Average Score per Game

In this paper, the Bayesian method is proposed to estimate the undetermined coefficients. Bayesian method is a method for calculating the probability of hypotheses. This method is based on the assumed prior probability, the probability of observing different data under a given assumption, and the observed data itself to obtain an estimate of the coefficient. The core idea is to integrate prior information about unknown parameters with sample information, then use Bayesian formulas to obtain posterior information, and finally infer unknown parameters based on the posterior information.

The steps to estimate parameters using Bayesian methods in this article are as follows:

Step (1). **Define a prior distribution:** Determine the problem to be solved and select an appropriate prior distribution based on the loss function.

Step (2). **Bayesian inference:** By using Bayesian theorem, the loss function is combined with prior knowledge to construct a Bayesian model.

Step (3). **MCMC sampling:** In order to obtain samples of parameters from the posterior distribution, the MCMC method can be used to determine the MCMC algorithm for sampling based on the model.

Step (4). **Analysis results:** Analyze the posterior distribution obtained by the MCMC algorithm, and obtain the mean and standard deviation of each parameter.

After the 4-step analysis of Bayesian estimation above, the estimated density value of the constant term β_0 and other parameters $\beta_1, \beta_2, \beta_3, \beta_4$ can be obtained (as shown in **Figure 6**).

The mean() function is used to calculate the average value of a numerical vector, and the sd() function is used to calculate the standard deviation of a numerical vector. After calculating the mean and standard deviation of the above 5 coefficients, the posterior estimation results can be obtained as shown in **Table 3**.

From the above calculation results, it can be seen that x_2 is the most significant on the overall impact of the model. That is, the better the shooting percentage, the higher the score. At the same time, it is noted that the corresponding coefficient is negative, which is also in line with our understanding that the fewer fouls a player has, the better. In addition, the remaining two parameters are
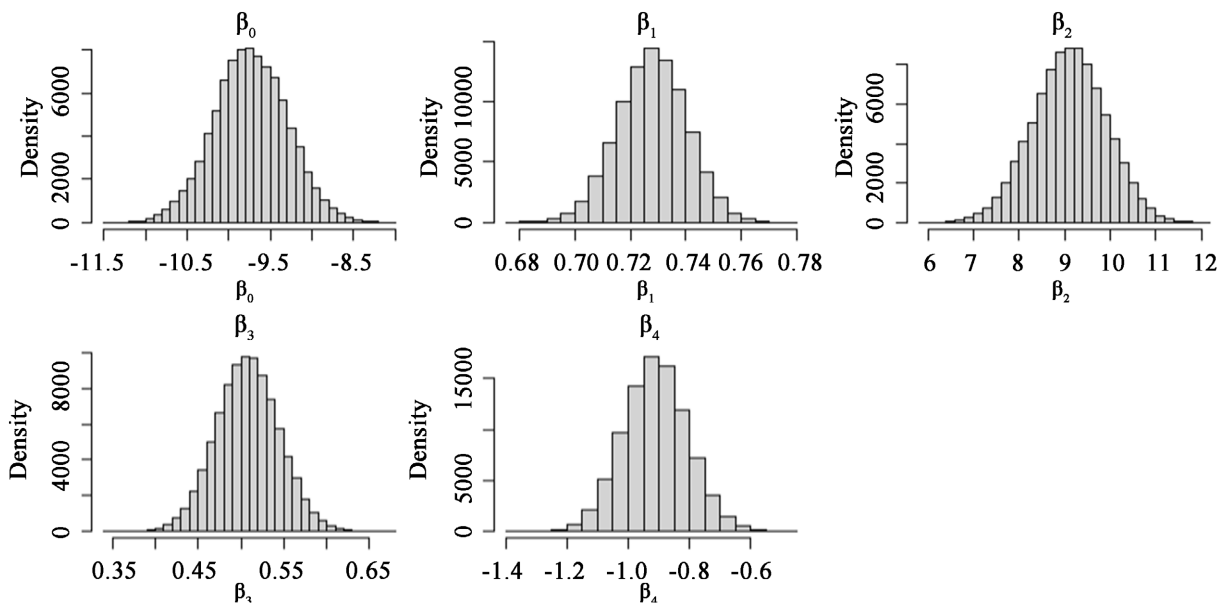

Figure 6. Bayesian estimates for 5 coefficients.

Table 3. Posterior estimation distribution values for 5 coefficients.

Parameters	β_0	β_1	β_2	β_3	β_4
Mean Values	−9.7334	0.7279	9.0988	0.5077	−0.9113
Standard Deviation	0.4384	0.0124	0.8031	0.0363	0.1042

both positive, which also indicates that "Minutes Played per Game" and "Assists per Game" have a positive effect on the dependent variable value (*i.e.* average score per game) of the model.

Based on the above data analysis, the impact of a player's shooting percentage on the average score per game is very significant. In order to achieve a higher average score per game, basketball players need to put in more effort, focus on practice, and improve their shooting accuracy under competitive conditions during regular training. In addition, strong physical fitness is the basic guarantee for effective "Minutes Played per Game". When physical fitness is insufficient, "Minutes Played per Game" will be greatly reduced. Therefore, high-intensity endurance training is also essential for basketball players in their daily training process. In addition, the number of assists reflects more on the practice of basic basketball skills, basketball skills and the observation of teammates in the game are both important. In basketball games, fouls usually result in an exchange of ball rights, and sometimes the opponent is directly given free throws. Players who commit a lot of fouls may also be sent off the court, so it is important to minimize the number of "Assists per Game".

5. Conclusion

Basketball has a strong mass foundation in our country, but the level of our national team is relatively backward. The NBA league almost gathers basketball elites from various countries around the world. By studying the game data of NBA players and combining multiple linear regression models, key factors affecting the average score of NBA players per game have been identified. Based on these factors, it can be concluded that the basketball skills required for different positions vary, and their average height is basically arranged from low to high according to big data statistics. If further optimization of the model is needed, players in different positions can be classified to analyze the data of players in each position, and then highlight which data is more important for players in the corresponding positions. This idea can enable professional athletes in our country to develop skill packages that suit their own characteristics as early as possible based on their height (including increasing to average weight, etc.), strengthen some skills, and have a targeted approach, thereby improving the overall level of basketball. Due to the crucial role of starting players in basketball games, every player wants to start the game, and the selection of starting players is a very challenging task for coaches. Subsequent experiments can also use data analysis to determine which players are more suitable for starting matches, providing coaches with algorithmic selection support, and striving to make their own scores higher than the opponent's scores, in order to win the game.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Chen, S.T., Cheng, C. and Chai, Y.J. (2015) A Study on the Correlation between Salary, Team Performance, and Player Performance. *Sports*, **123**, 33-35. (In Chinese)

[2] Song, H.F. and Han, G.L. (2020) Technical Statistics and Analysis of Women's Basketball League in the Third Xizang Workers' Games. *Vocational Education Research*, **6**, 249-251. (In Chinese)

[3] Li, Y.X. and Xiu, C.G. (2020) Analysis of Factors Influencing the Victory and Defeat of the 2018-2019 CBA Regular Season—Based on Multiple Linear Regression. *Statistics and Management*, **7**, 113-116. (In Chinese)

[4] Tan Z.L. (2022) Based on the Multiple Regression Model CBA, the Impact of Scoring Methods of China Guangzhou Team on Game Results during the 2020-2021 Season. Master's Thesis, Guangzhou University, Guangzhou. (In Chinese)

[5] Akakuru, O., Adakwa, C., Ikoro, D., *et al*. (2023) Application of Artificial Neural Network and Multi-Linear Regression Techniques in Groundwater Quality and Health Risk Assessment around Egbema, Souastern Nigeria. *Environmental Earth Sciences*, **82**, Article No. 77. https://doi.org/10.1007/s12665-023-10753-1

[6] Ravichandran, C. and Padmanaban, G. (2024) Estimating Cooling Loads of Indian Residences Using Building Geometry Data and Multiple Linear Regression. *Energy and Built Environment*, **5**, 741-771. https://doi.org/10.1016/j.enbenv.2023.06.003

[7] Aziz, A. and Anwar, M.M. (2024) Assessing the Level of Urban Sustainability in the Capital of Pakistan: A Social Analysis Applied through Multiple Linear Regression. *Sustainability*, **16**, Article 2630. https://doi.org/10.3390/su16072630

[8] Huang, Z.P., Ma, X., Chen, X., *et al*. (2024) Analysis and Application of Colorful Guizhou Tourism Data Based on Linear Regression Algorithm. *Soft Engineering*, **27**, 63-66. (In Chinese)

[9] Liu, D.B., Jin, Z.Y., Ke, Z.F., *et al*. (2023) Regression Analysis of Building Scale Data and Estimation of Demolition Rate. *Acta Scientiarum Naturalium Universitatis Pekinensis*, **59**, 547-554. (In Chinese)

[10] Zhang, J. and Xue, Y. (2022) Environmental DSGE Models' Important Parameters: Research Based on Multiple Regression Analysis. *Construction Economy*, **43**, 840-843. (In Chinese)

[11] Li, Y., Zhu, F.Y., Chen, J.Y., *et al*. (2020) An Empirical Study on Cross-Broder E-Commerce Development of Agricultural Products in China Based on Multiple Linear Regression Analysis. *Mathematics in Practice and Theory*, **50**, 299-310. (In Chinese)

[12] Bhargavi, N.and Poornima, T. (2024) Radiative Impact on Jeffery Trihybrid Convective Nanoflow over an Extensible Riga Plate: Multiple Linear Regression Analysis. *Contemporary Mathematics*, **5**, 1036-1053. https://doi.org/10.37256/cm.5120244058

[13] Wang, S., Monjurul, H. and Ming, L. (2024) Global Sensitivity Analysis Methodology for Construction Simulation Models: Multiple Linear Regressions versus Multilayer Perceptions. *Journal of Construction Engineering and Management*, **150**, Article ID: 04024035. https://doi.org/10.1061/JCEMD4.COENG-14059

[14] Rekabi, S., Garjan, H., Goodarzian, F., *et al*. (2024) Designing a Responsive-Sustainable-Resilient Blood Supply Chain Network Considering Congestion by Linear Regression Method. *Expert Systems with Applications*, **245**, Article ID: 122976. https://doi.org/10.1016/j.eswa.2023.122976

[15] Mao, S.S., Cheng, Y.M. and Pu, X.L. (2019) Course in Probability Theory and Mathematical Statistics. Higher Education Press, Beijing, 319. (In Chinese)

[16] Sina. https://slamdunk.sports.sina.com.cn/teams