World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3042

FORECASTING OLYMPIC MEDAL COUNTS: A MULTIPLE LINEAR REGRESSION MODEL

YiFan Guo
College of Physics, East China University of Science and Technology, Shanghai 200237, China.
Corresponding Email: 18217316473@163.com

Abstract: With the successful conclusion of the 2024 Summer Olympics in Paris, the Olympic medal table rankings have been finalized. The medal table is not only the personal honor of athletes, but also a symbol of the comprehensive strength and national cohesion of countries. Therefore, the research on the prediction of Olympic medal list has received wide attention. However, the current Olympic performance prediction research mainly focuses on macro factors and ignores micro variables, while multiple linear regression can deal with the relationship between multiple independent variables and one dependent variable, which becomes an ideal solution to this complex prediction problem. First and foremost, this paper develops multivariate linear regression models to predict the number of gold medals and the total number of medals for athletes, respectively. These models are used to predict their performance in the 2028 Los Angeles Olympic Games. After obtaining the predicted value of medals won by athletes in 2028, the predicted medal counts of athletes from each country are summed up to initially obtain the medal predictions of each country. In addition, considering that the actual number of national medals will be affected by the host country and the type of program, so this paper establishes a multiple linear regression model to predict the interference of the host country and the type of program on the actual number of medals of each country, and thus, constructs a more accurate medal prediction model. The final prediction result is: the total number of medals and the total number of gold medals is ranked first by the United States as the host country, followed by the United Kingdom, Germany, China and so on, of which France and Australia have the same number of medals and are tied for the fifth place. According to the above rankings, the United Kingdom, France and Germany have improved compared with the previous Olympic Games, while China, Australia and Japan have declined compared with the previous Olympic Games.
Keywords: Multiple linear regression; Host effect; Olympic medal prediction; F-tests

1 INTRODUCTION

As the world's most influential comprehensive sports event, the Olympic Games are not only a concentrated display of the competitive strength of various countries, but also an important window for a country's overall image and international status. With the continuous expansion of the social influence of the Olympic Games, the prediction research on its medal table has gradually become the focus in the fields of sports science and management decision-making. However, the existing prediction models still have significant limitations, which restrict their application value:

First of all, over-reliance on macro indicators while neglecting micro dynamics. Traditional models mostly construct prediction frameworks based on macroeconomic variables such as gross domestic product (GDP) and population size. Although such methods can reflect a country's overall resource endowment, a country's performance in the Olympic Games cannot be fully equivalent to its total GDP. At the same time, it is also difficult to explain the performance fluctuations between some sports powers and small countries. Especially for countries that lack economic advantages but have specialized competitive advantages (such as Jamaica and Kenya), the prediction accuracy is significantly limited[1].

Secondly, there is a disconnection between micro and macro data, and the model hierarchy is fragmented. Existing national-level prediction models usually take the total number of national MEDALS as an independent sample, ignoring the dynamic characteristics of the career trajectories of individual athletes. For example, simply adding up the number of MEDALS won in previous years as an input variable not only fails to capture the phased changes in an athlete's career but also makes it difficult to quantify the potential of the new generation of athletes, resulting in the accumulation of prediction errors.

Finally, static modeling is difficult to adapt to dynamic effects. Most traditional linear regression methods adopt the assumption of fixed parameters and are unable to effectively describe the nonlinear amplification mechanism of the host effect (such as home audience support and facility adaptability training) or the dynamic adjustment of event events within the Olympic cycle. Such limitations make the model insufficiently adaptable to emergent variables[2].

In response to the above problems, this paper proposes the following innovative solutions:

1.Dynamic sample embedding mechanism: At the individual athlete level, a historical performance dataset spanning three Olympic cycles (1992-2024) is constructed. By analyzing the performance history of athletes including the three Olympic cycles, the stage transitions in their careers can be precisely captured, thereby precisely predicting their potential to win MEDALS in the next Olympics.

2. Quantitative modeling of the host country effect: Introduce a binary discrete variable (0/1) to identify the identity of the host country, and combine historical data to construct a multiple regression model to quantify the contribution of the host country's identity to the number of MEDALS.
3. Microscopic-macro integrated multiple linear regression framework: In the national-level model, micro-variables such as the individual prediction results of athletes (Formulas 2-4), the host country effect, and the number of event events are integrated to break through the limitations of the traditional single-layer model.
This study achieved the organic connection between micro individual data and macro national variables through the above-mentioned methods: The individual achievements of athletes were summarized and input into the national medal prediction model (Formulas 10 and 12), and the host country effect/sports event variable was introduced to provide macro correction. This innovation not only provides more refined decision support for the formulation of Olympic strategies, but also offers a methodological reference for multivariate modeling in the field of sports competition prediction.

2 INDIVIDUAL ATHLETE MEDAL PROJECTIONS

The data of this study comes from https://olympics.com, in which the awards of athletes of various countries after 1992 are downloaded and organized into the number of gold medals won and the total number of medals won by athletes of various countries, respectively, and this is used as the data sample[3].
In order to predict the number of gold medals won by each individual athlete in the next Olympic Games, it may be useful to assume that the number of medals won by each individual athlete in the next Olympic Games is correlated with the number of medals won by the athletes in the previous three Olympic Games, and to set up a multiple linear regression model[4].

$$y_{gold} = \alpha_1 x_{gold} + \alpha_2 x_{silver} + \alpha_3 x_{bronze} + b_{gold} + \varepsilon_{gold} \tag{1}$$

To simplify the expression, the matrix form is used.

$$\begin{aligned} X_{gold} &= (x_{gold} \quad x_{silver} \quad x_{bronze} \quad 1)^T \\ \alpha_{gold} &= (\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad b_{gold}) \\ y_{gold} &= \alpha_{gold} X_{gold} + \varepsilon_{gold} \end{aligned} \tag{2}$$

Where x_{gold}、x_{silver}、x_{bronze} denote the total number of gold, silver and bronze medals won by the athlete in the past three competitions, b_{gold} is the constant term of the multiple linear regression, ε_{gold} is the error term, and y_{gold} is used to measure the number of gold medals won by the athlete in the next competition. Similarly, a multiple linear regression model was developed to predict the number of medals won by each athlete in the next Olympic Games.

$$y_{total} = \beta_1 x_{gold} + \beta_2 x_{silver} + \beta_3 x_{bronze} + b_{total} + \varepsilon_{total} \tag{3}$$

Similarly, the matrix form is used to simplify the expression.

$$\begin{aligned} X_{total} &= (x_{gold} \quad x_{silver} \quad x_{bronze} \quad 1)^T \\ \beta_{total} &= (\beta_1 \quad \beta_2 \quad \beta_3 \quad b_{total}) \\ y_{total} &= \beta_{total} X_{total} + \varepsilon_{total} \end{aligned} \tag{4}$$

Where $x_{gold}、x_{silver}、x_{bronze}$ represent the total number of gold, silver and bronze medals won by each athlete in the past three competitions, b_{total} is the constant term of the multiple linear regression, ε_{total} is the error term, y_{total} is used to measure the number of medals to be won by each athlete in the next Olympic Games.
Next, starting with the 1992 dataset, historical data on national athletes was extracted as a sample. The specific format of a single sample is x_1, x_2, x_3, y. Where y represents the number of gold medals won by the athlete in a particular year, and x_1, x_2, x_3 represent the total number of gold, silver, and bronze medals won by the athlete in the past three competitions, respectively. The total number of samples extracted for this study is N ,which is put into the sample input matrix X_{in}, and the output is put into the sample output matrix Y_{out}:

$$\begin{aligned} X_{in} &= \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & x_{N3} \end{pmatrix}_{N\times 3} \\ Y_{out} &= (y_1 \quad y_2 \quad y_3 \quad \cdots \quad y_N) \end{aligned} \tag{5}$$

In order to find the value of each parameter in $(\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad b_{gold})$, the principle of least squares is applied to solve the problem with the following formula[5]:

$$L(\alpha_1,\alpha_2,\alpha_3,b_{gold})=\sum_{i=1}^{N}[y_i-(\alpha_1x_{i1}+\alpha_2x_{i2}+\alpha_3x_{i3}+b_{gold})]^2$$
$$\nabla L=\left(\frac{\partial L}{\partial\alpha_1}\ \frac{\partial L}{\partial\alpha_2}\ \frac{\partial L}{\partial\alpha_3}\ \frac{\partial L}{\partial b_{gold}}\right)=\vec{0} \tag{6}$$

The values of each parameter are obtained by calculation as shown in Table 1:

Table 1 On the Parameters of the Athlete's Gold Medal Prediction model

Parameters	Values
α_1	0.228
α_2	0.204
α_3	-0.070
b_{gold}	-0.012
MSE	0.148

The model is then subjected to a joint hypothesis test, the F-test[6].

$$H_0:\alpha_i=0$$
$$H_1:\alpha_i\neq 0 \tag{7}$$

The formula for the F-statistic is as follows, where k is the number of independent variables and, N is the sample size.

$$F=\frac{ESS/k}{ESS/(N-k-1)}\sim F(k,N-k-1) \tag{8}$$

The model was then subjected to an F-test and the results of the test are shown in Table 2 for a given confidence level of 0.05 ($\alpha=0.05$):

Table 2 F-statistics on Athletes' Gold Medal Prediction Models

Statistical quantities	Values
F	37.76
F_α	2.67

Therefore, the null hypothesis is rejected, indicating that the individual athlete gold medal prediction model is significant.

Similarly, in order to predict the number of medals won by each individual athlete in the next Olympic Games, this study establishes a multiple linear regression model with the values of each parameter, as shown in Table 3:

Table 3 On the Parameters of the Model for Predicting the Total Medals of Athletes

Parameters	Values
β_1	0.303
β_2	0.234
β_3	-0.021
b_{total}	-0.228
MSE	0.350

The model was then subjected to an F-test and the results of the test are shown in Table 4 for a given confidence level of 0.05 ($\alpha=0.05$):

Table 4 F-statistics on Athlete Total Medal Prediction Models

Statistical quantities	Values
F	27.76
F_α	2.67

Therefore, the null hypothesis is rejected, indicating that the individual athlete total medal prediction model is significant.

3 MEDAL PROJECTIONS AT THE NATIONAL LEVEL

Firstly, the sample was extracted by screening the dataset table for athletes from each country after 1992. In the case of specific years and specific countries, the datasets of historical award-winning performance of national athletes in the last three years are extracted. Substituting these datasets into the above, a multiple linear regression model is built to predict the awards of individual athletes. When the y_{gold} and y_{total} won by athletes in 2028 are solved, the predicted medal counts of the athletes from each country are summed to calculate y_{sg} and y_{st}, which represent the total number of gold medals and total number of medals won by each athlete from each country in that year, respectively. In addition, this study considers the parameter $host$ to represent the host country effect and 0, 1 to indicate whether it is the host country or not (0 for no, 1 for yes)[7]. Finally, parameter num represents the number of programs in the Olympic Games.

In order to predict the number of gold medals won by each country at the next Olympic Games, it may be assumed that the number of medals won by a country is related to the overall number of medals won by the athletes representing their country, the host effect (whether or not they are a host country), and the number of events at that particular Olympic Games, and that the correlation between these factors is weak. Therefore, a multiple linear regression model can be developed[8].

$$\begin{aligned} y_{sg} &= \sum_{i=1}^{m} y_i \\ y_{gos} &= l_1 y_{sg} + l_2 host + l_3 num + b_{gos} + \varepsilon_{gos} \end{aligned} \tag{9}$$

To simplify the expression, the matrix form is used.

$$\begin{aligned} X_{gos} &= (y_{sg} \quad host \quad num \quad 1)^T \\ L_{gos} &= (l_1 \quad l_2 \quad l_3 \quad b_{gos}) \\ y_{gos} &= L_{gos} X_{gos} + \varepsilon_{gos} \end{aligned} \tag{10}$$

Where y_i is the number of gold medals won by individual athletes in each country predicted by the model, and the sum is obtained as y_{sg}, which is used to indicate the overall gold winning situation of athletes in each country; $host$ refers to whether the country is the host country of the session; num is the number of events of the session; b_{gos} is a constant; and ε_{gos} is the error term.

Similarly, in order to predict the total number of medals to be won by each country in the next Olympic Games, a multiple linear regression model was developed in this study.

$$\begin{aligned} y_{st} &= \sum_{i=1}^{m} y_i \\ y_{tos} &= k_1 y_{st} + k_2 host + k_3 num + b_{tos} + \varepsilon_{tos} \end{aligned} \tag{11}$$

The matrix form is used to simplify the expression.

$$\begin{aligned} X_{tos} &= (y_{st} \quad host \quad num \quad 1)^T \\ K_{tos} &= (k_1 \quad k_2 \quad k_3 \quad b_{tos}) \\ y_{tos} &= K_{tos} X_{tos} + \varepsilon_{tos} \end{aligned} \tag{12}$$

Where y_i is the total number of medals won by individual athletes in each country predicted by the model, and y_{st} is obtained after summation, which is used to indicate the total number of medals won by athletes in each country; $host$ refers to whether or not the country is the host country of the session; num is the number of events in the session; b_{tos} is a constant; and b_{tos} is the error term.

Next, starting from the 1992 dataset, the prizes won by all athletes from all countries in all previous years, the host country, and the number of Olympic events are extracted, and the specific format of individual samples is y_{sg}, $host$, num, y_{gos}. The total number of samples extracted in this study is M, which is placed into the sample input matrix X_{in} and the outputs are placed into the sample output matrix Y_{out}:

$$\begin{aligned} X_{in} &= \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ \vdots & \vdots & \vdots \\ x_{M1} & x_{M2} & x_{M3} \end{pmatrix}_{M\times 3} \\ Y_{out} &= (y_1 \quad y_2 \quad y_3 \quad \cdots \quad y_M) \end{aligned} \tag{13}$$

The values of each parameter are obtained by calculation as shown in Table 5:

Table 5 Parameters of the Gold Medal Forecasting Model for Each Country

Parameters	Values
l_1	0.527
l_2	15.593
l_3	-0.024
b_{gos}	13.707
MSE	0.794

The model was then subjected to an F-test and the results of the test are shown in Table 6 for a given confidence level of 0.05 ($\alpha = 0.05$):

Table 6 F-statistics on Countries' Gold Medal Forecasting Models

Statistical quantities	Values
F	19.094
F_α	2.790

Therefore, the null hypothesis is rejected, indicating that the model of countries receiving gold medal predictions is significant.

Similarly, in order to predict the total number of medals won by each country in the next Olympic Games, this study establishes a multiple linear regression model with the values of each parameter, as shown in Table 7:

Table 7 Parameters of the Model for Predicting Total Medals by Country

Parameters	Values
k_1	0.726
k_2	31.865
k_3	0.195
b_{tos}	-46.896
MSE	0.896

The model was then subjected to an F-test and the results of the test are shown in Table 8 for a given confidence level of 0.05 ($\alpha = 0.05$):

Table 8 F-statistics on the Model for Predicting Total Medals for Each Country

Statistical quantities	Values
F	18.889
F_α	2.790

Therefore, the null hypothesis is rejected, indicating that the model predicting the total medals won by each country is significant.

Based on the above prediction model, this study can predict the ranking of the medal table of the 2028 Olympic Games as shown below:

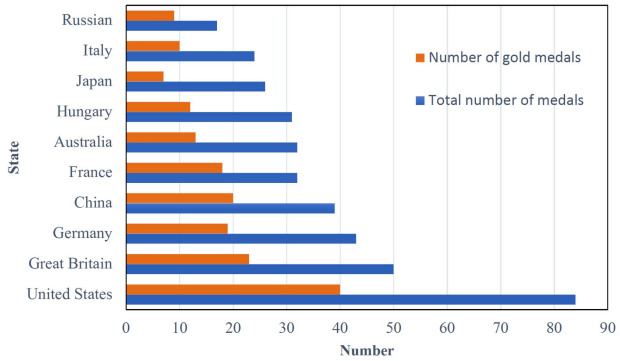

Figure 1 Ranking of Countries in Terms of Medals at the 2028 Olympic Games

According to Figure 1, this study finds that the top ten countries in terms of medal count in 2028 are all sports powerhouses, with the United States of America as the host country ranking first in terms of total number of medals and total number of gold medals, followed by the United Kingdom, Germany, China and so on, with equal numbers of medals for France and Australia, both with 32 medals, and according to the above rankings, the United Kingdom, France, and Germany improved compared to the previous Olympics, while China, Australia, and Japan According to the above ranking, Great Britain, France and Germany have improved compared to the last Olympics, while China, Australia and Japan have decreased compared to the last Olympics.

Finally this study performs error analysis on the model by dividing the sample data into training set and test set in the ratio of 8:2, after constructing the multiple linear regression model based on the training set, the predicted values are generated on the test set, and then the predicted values are compared with the true values to measure the predictive performance of the model.

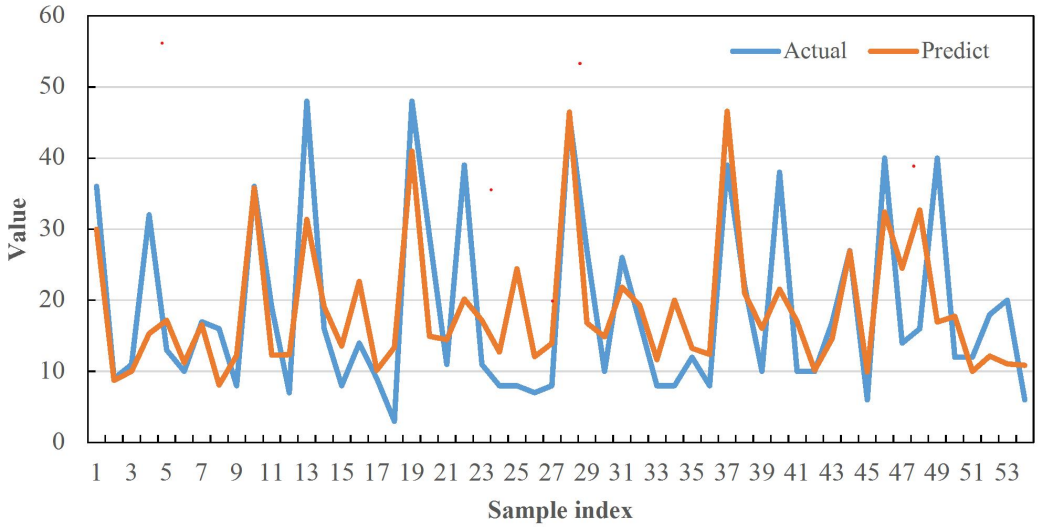

Figure 2 Error Analysis of Gold Medal Forecasting Models for Each Country

According to Figure 2, the model errors show systematic deviations; overall, the predicted values and the actual values fit tightly in the low sample index interval, but as the sample index increases, the predicted values gradually deviate from the actual values, and the magnitude of the deviation significantly expands with the increase in the order of the number of medals. In addition, the error dispersion of high medal intervals is significantly higher than that of low intervals, reflecting that the model is less stable in predicting extreme values, which may be related to the sensitivity of linear regression to high variance data.

4 CONCLUSIONS

By constructing the prediction model of multiple linear regression model, this paper analyzes the influence of individual athlete's historical awards on his next awards; the influence of each athlete's awards, the host effect, and the number of events in each country on each country's next Olympic Games medals, and both types of models have successfully passed the F-test and predicted the ranking of each country's medals in the 2028 Olympic Games. However, this paper still has limitations in some aspects, firstly, the influencing factors considered in the model are only from the existing data set, which cannot guarantee the comprehensiveness of the model, and the possibility of incomplete factors makes the model's prediction have a certain bias, secondly, the multivariate linear regression model is used in both the individual and the national medal prediction models, and the model selection is not diversified enough. In view of the above shortcomings, more data will be collected in the future to refine the gender characteristics and age characteristics of individual athletes to enrich the existing data set and better optimize the model, so as to achieve more accurate prediction, in addition, a variety of prediction models will be used to compare the prediction results to obtain the best prediction scheme.

This paper provides a research idea and framework applied to the field related to the management of sports statistics, and proves the feasibility of multiple linear regression models to predict the number of medals in the Olympic Games. By establishing multiple linear regression models, it is possible to predict an individual's performance in the next Olympic Games, and based on the historical performance of athletes in each country, the host effect, and the type of Olympic program each factor, it can be predicted that each country will have the number of medals in the next Olympic Games. Medal counts.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Zhao Xin, Xue Ye, Niu Chonghuai. Correlation Analysis of the Total Number of Olympic MEDALS of Various Countries and the Total GDP. Sports Culture Guide, 2013, (08): 1-4.

[2] Zhang Yuhua. Medal Count Prediction of China's 31st Olympic Games Based on Linear Regression Dynamic Model. Journal of henan normal university (natural science edition), 2013, 9(02): 24-26+60.

[3] Raja M, Sharmila P, Vijaya P, et al. Olympic Games Analysis and Visualization for Medal Prediction//2025 International Conference on Artificial Intelligence and Data Engineering (AIDE). IEEE, 2025, 822-827.

[4] Vagenas G, Vlachokyriakou E. Olympic medals and demo-economic factors: Novel predictors, the ex-host effect, the exact role of team size, and the "population-GDP" model revisited. Sport Management Review, 2012, 15(2): 211-217.

[5] Nagpal P, Gupta K, Verma Y, et al. Paris Olympic (2024) Medal Tally Prediction//International Conference on Data Management, Analytics & Innovation. Singapore: Springer Nature Singapore, 2023, 249-267.

[6] Griffiths W E, Hill R C. On the Power of the F-test for Hypotheses in a Linear Model. The American Statistician, 2022, 76(1): 78-84.

[7] TIAN Hui, HE Yiman, WANG Min, et al. Medal Prediction and Participation Strategy of Chinese Athletes in the 2022 Beijing Winter Olympics-Based on the Analysis of Olympic Home Field Advantage Effect. Sports Science, 2021, 41(02): 3-13+22.

[8] Luo Yubo, Cheng Yanfang, Li Mengyao, et al. Forecasting the number of medals and overall strength ofChina in the Beijing Winter Olympics - Based on host effect and gray prediction model. Contemporary Sports Science and Technology, 2022, 12(21): 183-186.