Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 12, No. 1, 2025

Prediction of Olympic Medal Based on Multiple Linear Regression and Logistic Regression

Zihan Lu [1,*], Songling Li [2], Jinzhou Sun [2]

[1] School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an, Shaanxi, China
[2] School of Mathematics and Statistics, Northwestern Polytechnical University, Xi'an, Shaanxi, China
*** Corresponding author:** Zihan Lu (Email: 18005259889@mail.nwpu.edu.cn)

Abstract: This paper focuses on the issue of Olympic medal distribution and conducts in-depth exploration by comprehensively applying multiple models and algorithms. A medal-counting prediction model is constructed using multiple linear regression. By considering factors such as historical medal counts and the host-country effect, the medal standings of the 2028 Los Angeles Olympics are predicted. Through the construction of indicators such as the country-project alignment degree, the impact of competition events on medal distribution is analyzed. A logistic regression model is used to predict countries that will win medals for the first time and to identify potential medal - winning countries. The research reveals the laws of medal distribution and clarifies the influence mechanisms of various factors. These models provide a scientific basis for countries to develop sports development strategies, help optimize resource allocation, enhance the competitiveness of Olympic medals, and have important reference value for sports event planning and national sports development.

Keywords: Multiple Linear Regression; Logistic Regression; Medal Prediction.

1. Introduction

The Olympic Games, as the world's top-level sports event, the medal table is a key indicator for measuring the sports strength of various countries [1]. The distribution of medals is affected by many factors, such as economic strength, population size, and the host-country advantage [2]. In-depth research on its laws is of great significance for countries to develop reasonable sports development strategies. Although previous studies have covered related factors, they lack systematic and precise comprehensive analysis. Based on this, this paper uses models such as multiple linear regression and logistic regression, combines historical data and event characteristics, and constructs a comprehensive Olympic medal analysis system. By analyzing historical medal data, quantifying the impacts of various factors, and exploring potential laws, this paper aims to provide scientific decision-making support for national Olympic committees, promote the balanced development of the global sports industry, and offer new ideas and methods for the research field of sports events.

2. Prediction of the Olympic Medal Table Based on Multiple Linear Regression

2.1. Model Explanation

Firstly, we introduce a concept called "historical medal performance": the weighted average of the number of gold medals and the total number of medals from several past Olympic Games. Now, HM_i is defined as,

$$HM_i = \frac{1}{n}\sum_{j=1}^{n} w_j \cdot M_i, j \tag{1}$$

where $M_{i,j}$ represents the number of medals for country i in the j -th edition, w_j is the weighting coefficient, and the closer the edition is to the present, the higher its weight. Then, we introduce the host country i effection H_i. This is a binary variable, meaning that if country i is the host country, then H_i is 1; if not, then H_i is 0.

By using multiple linear regression for modeling, we can derive the regression equation [3].

$$TM_i = \beta_0 + \beta_1 HM_i + \beta_2 H_i + \varepsilon_i \tag{2}$$

Where β_0 is the intercept, β_1 and β_2 are the regression coefficients. The number of gold medals is the same.

In order to analyze the impact of the host country's new projects on the distribution of medals, we will additionally include the interaction term of the host country.

$$TM_i = \beta_0 + \beta_1 HM_i + \beta_2 H_i + \beta_3 \cdot H \cdot ET_i + \varepsilon_i \tag{3}$$

2.2. Results

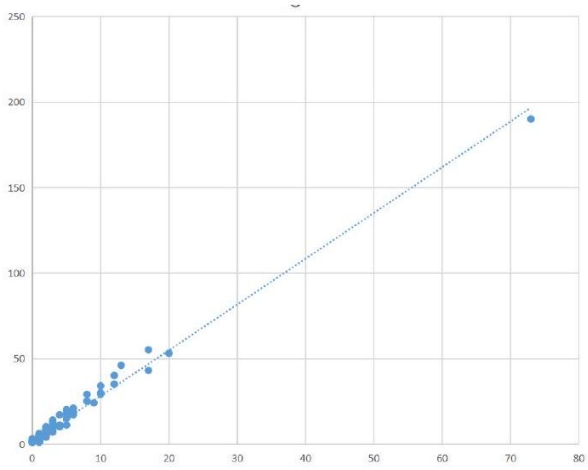

Figure 1. Prediction result

As shown in Figure 1, the horizontal axis represents the predicted number of gold medals, while the vertical axis represents the predicted total number of medals.

2.3. Sensitivity Analysis

2.3.1. Gold

Table 1. Parameter

Project	Estimate	SE	tStat	pValue
(Intercept)	0.23905	0.040128	5.9571	2.7056×10^{-9}
HistoricalGold	0.055927	0.00081261	68.823	0
Host	15.599	0.60549	25.762	3.1077×10^{-139}

According to Table.1, after calculation, we can obtain the model expression:

$$G_i = 0.23905 + 0.055927 \times HG_i + 15.599 \times H_i + \varepsilon_i \quad (4)$$

The number of observations for this model is 6300, with an error degree of freedom of 6297. The large number of observations is beneficial for making the regression results more reliable.

The F statistic of the model is 3.03×10^3, and the p-value is 0. The F statistic is used to test the significance of the overall regression model, and a p-value of 0 indicates that, at the given significance level (usually 0.05), the null hypothesis (which states that all regression coefficients of the independent variables are 0) is rejected, meaning that the overall regression model is significant, and at least one independent variable has a significant effect on the dependent variable.

The test statistic is used to examine whether the regression coefficient of a single independent variable is significantly different from zero. The calculation formula is: $t = \frac{\text{Estimate}}{SE}$.

The R^2 value of the model is 0.49, indicating that HG_i and H_i can explain approximately 49% of the variation in G. Although the model has a certain explanatory power, a considerable portion of the variation remains unexplained, suggesting that there may be other significant factors not included in the model.

The root mean square error of the model is 3.1, indicating that the model has a high level of accuracy.

2.3.2. Total Medals

Table 2. Parameter

Project	Estimate	SE	tStat	pValue
(Intercept)	0.74004	0.10513	7.0393	2.1374×10^{-12}
HistoricalTotalMedals	0.057701	0.00079905	72.212	0
Host	40.492	1.5754	25.702	1.2541×10^{-138}

According to Table.2, after calculation, we can obtain the model expression:

$$TM_i = 0.74004 + 0.057701 \times HM_i + 40.492 \times H_i + \varepsilon_i$$

The number of observations for this model is 6300, with an error degree of freedom of 6297. The large number of observations is beneficial for making the regression results more reliable.

The F statistic of the model is 3.29×10^3, and the p-value is 0. The F statistic is used to test the significance of the overall regression model [4], and a p-value of 0 indicates that, at the given significance level (usually 0.05), the null hypothesis (which states that all regression coefficients of the independent variables are 0) is rejected, meaning that the overall regression model is significant, and at least one independent variable has a significant effect on the dependent variable.

The test statistic is used to examine whether the regression coefficient of a single independent variable is significantly different from zero. The calculation formula is: $t = \frac{\text{Estimate}}{SE}$.

The R^2 value of the model is 0.511, indicating that HG_i and H_i can explain approximately 51.1% of the variation in G. Although the model has a certain explanatory power, a considerable portion of the variation remains unexplained, suggesting that there may be other significant factors not included in the model.

The root mean square error of the model is 8.08. The root mean square error of this model is relatively large, which may indicate that the error in predicting the total number of medals is greater compared to the Gold model.

3. Analysis of the Impact of Competition Events on Medal Allocation Based on the Project Alignment Degree

3.1. Model Explanation

Now considering the alignment of national projects (FitDegree_i), since the sport proportion of projects in previous Olympic Games has not changed significantly and there is no obvious trend, the insights that can be derived from the composition of past projects are limited; therefore, only the most recent edition is taken into account. The proportion of project k in the country i is:

$$\text{NationalRatio}_i(k) = \frac{\text{NationalNumberSport}_i(k)}{\text{NationalTotalSport}_i} \quad (5)$$

The total proportion of the project is:

$$\text{Ratio}(k)=\frac{\text{NumberSport}(k)}{\text{TotalSport}} \tag{6}$$

The alignment of national projects is:

$$\text{FitDegree}_i=\frac{\sum_k N^i_{\text{ationalRatio}}(k)\cdot\text{Ratio}(k)}{\max\limits_k \text{Ratio}(k)} \tag{7}$$

This is because when $\text{Ratio}(k)$ is fixed, the maximum value of the numerator $\text{NationalRatio}_i(k)$. Ratio(k) is $\max\limits_k$ Ratio(k), while FitDegree_i is only related to i, and the denominator $\max\limits_k \text{Ratio}(k)$ is not related to i; therefore, we do not perform normalization here. However, when considering more general historical cases, this operation is necessary.

Now we study the importance of sports events for a country's award achievements. Analyze the proportion of the number of events held by country i for the sport with the largest share, and sort the elements of that sport by their proportion. Using this data to characterize the importance of sports events is because it reflects the structural layout of a country's sports projects. Those with a larger proportion are more likely to win medals, while those with a smaller proportion are less likely to be considered important sports projects for a country. Typically, countries rely on a layout of major projects; even emerging projects, due to a lack of accumulated experience, find it difficult to win medals. Another approach is to consider whether the projects have achieved a significant number of awards for the country, which we did not adopt, as the number of awards is positively correlated with the number of projects. Furthermore, even if projects with a small proportion win many medals, such awards are less stable due to a lack of experience accumulation, leading to an increasing emphasis on larger proportion projects, while the occurrence of small proportion projects winning many medals will decrease.

The impact of the host country effect on the results is divided into two influencing factors: the number of gold medals and the total number of medals.

$$GoldInfluence = GoldHostCofficient \cdot GoldAdjust\,R = 15.599\times 0.49 \approx 7.64 \tag{8}$$

$$\begin{aligned} TotalMedalsInfluence &= TotalMedalsHostCof\ ficient\ \cdot TotalMedalsAdjust\,R \\ &= 40.492\times 0.511 \approx 20.69 \end{aligned} \tag{9}$$

In other words, the host country effect will bring the nation 7.64 gold medals and 20.69 medals.

3.2. Results

Table 3. Sports

Sport	Group Count	Importance
'Athletics'	142	0.606837607
'Swimming'	30	0.128205128205128
'Gymnastics'	16	0.0683760683760684
'Wrestling'	7	0.0299145299145299
'Hockey'	6	0.0256410256410256
'Shooting'	5	0.0213675213675214
'Fencing'	4	0.0170940170940171

As shown in Table.3, the group count of “Athletics” is 142, with an importance of 0.606837607, much higher than other sports, indicating that Athletics occupies a very important position in medal distribution. “Swimming, with a group count of 30 and an importance of 0.128205128205128, also has a high impact on medal winning and is an important medal producing sport. In contrast, “Wrestling”, “Hockey”, “Shooting” and “Fencing” have lower scores in both group count and importance. In contrast, “Wrestling”, “Hockey”, “Shooting” and “Fencing” have lower importance scores, indicating that these sports have less weight in the medal allocation and their contribution to the overall medal count is relatively limited. This means that when countries formulate their sports development strategies, they should focus on sports with high importance, such as athletics and swimming, and invest more resources to improve their competitiveness in these sports, so that they are more likely to win more medals.

Table 4. NOC

NOC	Sport	Percentage
'AFG'	'Wrestling'	0.304347826086957
'AHO'	'Athletics'	0.189189189189189
'AIN'	'Tennis'	0.304347826086957
'ALB'	'Shooting'	0.24390243902439
'ALG'	'Athletics'	0.237871674491393
'AND'	'Athletics'	0.25
'ANG'	'Handball'	0.361022364217252

As shown in Table.4, at the national level, the percentage of “AFG” in “Wrestling” is 0.304347826086957, indicating that wrestling is a key strength of the AFG and the country has a high medal potential in this sport. The percentage of “ALB” in “Shooting” is 0.24390243902439, and shooting is its relative advantage. Countries can increase the number of medals more efficiently by allocating resources and planning development based on their own advantageous programs. For some countries with limited resources, focusing on the development of advantageous programs is more conducive to achieving good results in the Olympic Games than distributing resources evenly.

4. Prediction of Countries Winning Medals for The First Time Based on Logistic Regression

4.1. Model Explanation

Due to the uneven distribution of E and considering the marginal effect of E on I, we applied a logistic regression model after taking the logarithm of E to obtain the probability of I being 1 for those countries where I is 0, totaling 39 countries [5,6]. As mentioned earlier, due to

missing data, only 123 countries have F, but referring to the data from previous Olympic Games, the number of countries winning the first prize does not exceed 20, which is less than 39. Moreover, the countries with missing data are mostly small and remote nations that have been established for a short time and have only recently participated in the Olympics, or their data is missing due to political issues affecting national administrative divisions. Based on this, we believe it is reasonable to look for first prize-winning countries within the scope of these 39 countries. According to the data preprocessing, we can extract the following data:

PA_i : The number of athletes in the country i.

EP_i : The number of projects in which the country i participates.

PR_i : The geographical proximity of the country i to the host country.

HP_i : The historical award records of countries with the same GDP or within the same cultural sphere.

Next, we will use a logistic regression model to predict the probability of these unawarded countries winning for the first time.

$$P\left(\text{Win}_i = 1\right) = \frac{1}{1+e^{-\left(\beta_0+\beta_1\cdot PA_i+\beta_2\cdot EP_i+\beta_3\cdot PR_i\right)}} \tag{10}$$

4.2. Results

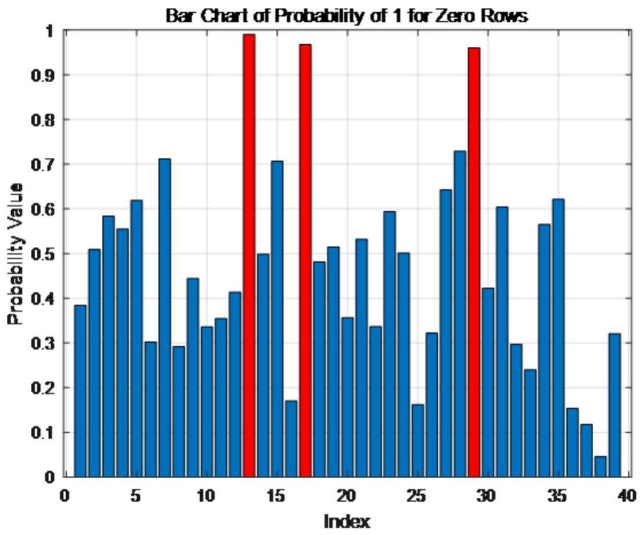

Figure 2. Possibility

As shown in Figure 2, the countries numbered 13, 17, and 29 exhibit data anomalies. Upon reviewing the table, it is found that they are the United Kingdom, Korea, and the Soviet Union, the latter of which has dissolved. Prior to its dissolution, the Soviet Union was also very strong in the Olympics, while the first two have records of winning awards. Their presence among these 39 countries is due to the inconsistency in the extracted data table, which uses Great Britain and separates South Korea and North Korea. These three anomalous data points actually indicate that our model is quite reasonable, as the predicted probabilities for these countries, which have actually won medals, are very close to 1. After excluding these three anomalous data points, an analysis of the remaining 36 countries will be conducted. According to the calculations of the aforementioned model, we can determine the country that will win the award for the first time in the next session: Côte d "Ivoire, Iraq, Papua New Guinea, Syrian Arab Republic.

The relationship between the number and types of events and the number of medals: The probability of a country winning the first prize is positively correlated with the number of medals to some extent. The degree of fit reflects the alignment between the country's event project structure and international mainstream, which is closely related to the type of event. The number of projects directly indicates the number of events. By conducting a statistical significance analysis on the model predicting the probability of the first prize country, the p-values $\left[0.0036, 0.0981, 3.3106\times10^{-6}\right]$ corresponding to the intercept, FitDegree_i, and Event $_i$ were obtained, leading to the conclusion that the number of events significantly affects the number of medals, while the type of event has a relatively small impact on the number of medals.

4.3. Sensitivity Analysis

It is now necessary to analyze the number of countries that have won the first prize Country Number, assuming that the award status of each country is independent of one another.

We set a parameter: IsPridictedGain_i .If the country i is about to receive an award, then it is equal to 1; otherwise, it is equal to 0, so

$$\begin{gathered}\text{CountryNumber} = \sum_{i=1}^{n}\text{IsPridictedGain}_i \\ \text{IsPridictedGain}_i \sim B\left(1, \text{Possibility}_i\right)\end{gathered} \tag{11}$$

According to the central limit theorem, Country Number is approximately normally distributed.

$$\begin{gathered}\text{CountryNumber} \sim N\left(\mu, \sigma^2\right) \\ \mu = \sum_{i=1}^{n}\text{Possibility}_i \approx 15 = \tilde{\mu} \\ \sigma^2 = \sum_{i=1}^{n}\text{Possibility}_i\cdot\left(1-\text{Possibility}_i\right)\end{gathered} \tag{12}$$

Considering the number of countries that have won awards in previous years and the possibility of a few countries with a history of winning not receiving awards in the next session, this data is relatively reasonable.

Let the probability density function corresponding to $N\left(\mu,\sigma^2\right)$ be $f\left(x\right)$, and the confidence level is:

$$\text{ConfidenceDegree} = \int_{\mu-0.5}^{\mu+0.5} f\left(x\right)\mathrm{d}x \approx 14.1272\% \tag{13}$$

5. Conclusion

This study systematically analyzed the distribution of Olympic medals through models such as multiple linear regression and logistic regression, and achieved valuable results. The multiple linear regression model effectively revealed the significant impacts of factors such as historical medal counts and the host-country effect on the number of medals, providing a powerful tool for predicting future medal tables. However, there is still room for optimization. In the

research on the impact of competition events on medal allocation, the innovatively constructed indicator system clearly demonstrated the differences in the importance of different events. The logistic regression model successfully predicted countries that would win medals for the first time and clarified the degree of influence of the number and type of events on the number of medals. These research results provide a scientific basis for countries to develop sports development strategies, help countries allocate resources rationally, and enhance their competitiveness in winning Olympic medals. Future research can further expand the models, incorporate more influencing factors, improve the prediction accuracy and explanatory power of the models, and provide more precise guidance for the development of the sports industry.

References

[1] Houlihan B, Zheng J. The Olympics and elite sport policy: Where will it all end? [J]. The international journal of the history of sport, 2013, 30(4): 338-355.

[2] Aygün M, Savaş Y. Analysing Winter Olympic Medals Through Economic Variables: A Comprehensive Examination [J]. Research in Sport Education and Sciences, 2024, 26(4): 197-209.

[3] Etemadi S, Khashei M. Etemadi multiple linear regression[J]. Measurement, 2021, 186: 110080.

[4] Sureiman O, Mangera C M. F-test of overall significance in regression analysis simplified[J]. Journal of the Practice of Cardiovascular Sciences, 2020, 6(2): 116-122.

[5] Pregibon D. Logistic regression diagnostics[J]. The annals of statistics, 1981, 9(4): 705-724.

[6] Stoltzfus J C. Logistic regression: a brief primer[J]. Academic emergency medicine, 2011, 18(10): 1099-1104.