International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 6491-6502

https://doi.org/10.70517/ijhsa464551

Olympic Medal Prediction Based on Multiple Regression and Time Series Analysis

Zizheng Wang[1,*] and Wenjie Yang[2]
[1] Ulster College, Shaanxi University of Science & Technology, Xi'an, Shaanxi, 710021, China
[2] School of Mathematics and Data Science, Shaanxi University of Science & Technology, Xi'an, Shaanxi, 710021, China
Corresponding authors: (e-mail: 202215020118@sust.edu.cn).

Abstract In order to predict the distribution of medals at the 2028 Olympic Games in Los Angeles scientifically, this study constructs models like multiple linear regression and ARIMA logistic regression for analyzing the issues of the quantity of national medals in the Olympic medal table and its trend change, the possibility of the country's first medal, and the effect of the selection of Olympic sports on the number of medals. Firstly, the paper establishes a prediction model for the quantity of medals in the Olympic Games based on multiple linear regression, and obtains a high linear correlation between data characteristics and the number of medals. Secondly, in order to predict the medal development trend of each country in the 2028 Olympic Games, this paper adopts the ARIMA model for prediction, obtaining the medal trend of each country in the 2028 Olympic Games, and verifying that the ARIMA model is smooth and effective. Finally, for the countries that have not yet won medals, the relevant characteristics of the countries that have not yet won medals are collected, and the 10 countries with the highest probability of winning medals for the first time are obtained, among which Azerbaijan has the highest probability of 62.5%. The present study found that the quantity of events is positively correlated with the quantity of medals, which provides a reliable basis for the resource allocation of the National Olympic Committee. The model is both explanatory and flexible, thus providing a novel perspective on the prediction of Olympic medals and the strategic planning of national sports.

Index Terms Multiple Linear Regression, ARIMA Model, Olympic Medal Prediction

I. Introduction

The Olympic Games represent the pinnacle of global sporting competition, extending beyond competition to signal geopolitical influence, cultural capital, and national prowess [1]. Furthermore, they offer critical frameworks for strategic international sports governance decision-making, especially regarding resource allocation, athlete development, and long-term competitive planning. The proliferation of advanced computational methodologies—specifically the maturation of machine learning algorithms and data science paradigms—has catalyzed a paradigm shift in Olympic studies, fostering heightened scholarly and institutional interest in deploying predictive modeling to anticipate medal allocations with unprecedented sophistication [2]. As the 2028 Los Angeles Olympiad approaches with accelerating immediacy, the operational imperatives confronting National Olympic Committees (NOCs) intensify considerably. These bodies face the complex tripartite challenge of: (1) generating statistically robust projections for probable national medal yields; (2) identifying emergent sporting nations poised for competitive breakthroughs; and (3) evaluating the systemic implications of event modifications introduced within the Olympic programme [3]. Such multidimensional intelligence constitutes the foundational bedrock for evidence-based policymaking—directly informing nuanced strategy formulation, budget prioritization, talent identification protocols, and the strategic deployment of technical resources across Olympic quadrennials [4].

Empirical analysis of historical performance metrics reveals persistent structural hierarchies within the Olympic medal economy. Traditional sporting hegemons—notably the United States and China—maintain statistically significant dominance through systemic advantages including institutionalized training infrastructures, substantial financial investments, and deep athlete pools. Conversely, discernible shifts in the global sporting landscape emerge through the sustained advancement of developing nations such as Kosovo and Azerbaijan, whose incremental ascendance manifests through targeted sport specialization and strategic resource allocation. This dynamic dialectic between sporting continuity and disruption is mediated through a confluence of multifaceted determinants: historical performance trajectories, sport-specific technological innovations, differential national investment patterns, host-nation advantages, and crucially, the strategic composition of the Olympic sporting programme itself. The introduction or elimination of disciplines—whether driven by technological relevance, broadcast appeal, or gender

equity imperatives—creates cascading effects across national competitive matrices. Consequently, the development of sophisticated computational frameworks capable of synthesizing these complex variables becomes methodologically indispensable. Only through systematic, model-driven analysis can stakeholders achieve the requisite analytical resolution to navigate the increasingly fluid terrain of global sporting competitiveness and translate predictive insights into actionable strategic advantage.

A substantial corpus of research on Olympic medal prediction is characterized by a multidisciplinary approach. Schlembach [5] proposed a machine learning model based on socio-economic variables using a two-stage random forest algorithm. The model demonstrated high accuracy in predicting total, non-zero and zero medals, a finding that was confirmed by several metrics. Makiyan [6] proposed respective prediction models from economic perspectives, emphasising the application of multidisciplinary intersections in medal prediction. Meanwhile, Forrest [7] and Zhao [8] provide comparisons of multiple prediction methods from the perspectives of statistics and econometrics. However, single models are the focus of most extant studies, and the local factors lack a comprehensive analysis of medal statistics and award trends.

The present paper employs models such as multiple linear regression and ARIMA to construct a multi-model forecasting framework, with a view to addressing the following key issues:

(1) forecasting the medal table of the 2028 U.S. The following research was conducted for the Summer Olympic Games in Los Angeles. Firstly, historical data was analysed using ARIMA time-series analysis in order to predict and analyse the changes in the medal counts by country.

(2) Secondly, country-specific medal probabilities for the first time was predicted.

(3) Thirdly, the basis for the first-time rise of medal-winning countries was provided.

(4) Finally, the results of the model were synthesised, the medal distribution law was distilled, and strategic recommendations for the National Olympic Committee were provided.

II. Materials and Methods

II. A. Data Acquisition and Preprocessing

II. A. 1) Data acquisition

The research scope encompasses medal counts awarded, athlete participation figures, the event programme settings, and the host information of each country from 1984 to 2024. The data have been sourced from two primary sources: Olympic movement central database and National Olympic Committees' public reports.

II. A. 2) Data preprocessing

In order to ensure the accuracy and stability of the modelling and the high quality and reliability of the data, this paper preprocesses the data.

The identification of potential outliers in the data can be facilitated by the interquartile range (IQR) method [9]. This method has been demonstrated to effectively eliminate a small number of extreme events (such as newly added or canceled events) from distorting the overall data. Consequently, this ensures that the analysis results are more accurate and representative. Following the completion of the outlier detection process, the decision was taken to replace the outliers with the median value for each individual sport. The replacement of outliers with medians has been demonstrated to be an effective method of mitigating the impact of outliers on data distribution and modelling, whilst preserving the overall structure and stability of the data. Subsequently, the quartile distance method was applied for secondary testing, and the corresponding box-and-line plots showed that all the processed data points were within the normal range [10]. The results of the data processing are illustrated in Figures 1 and 2.

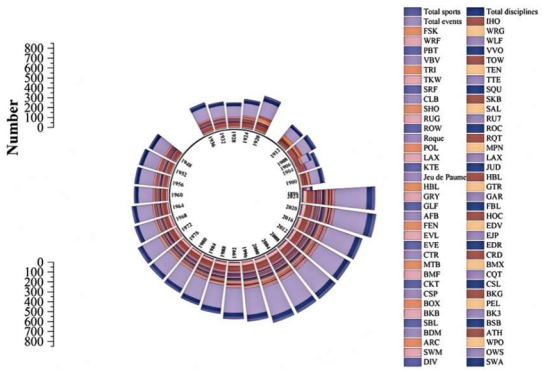
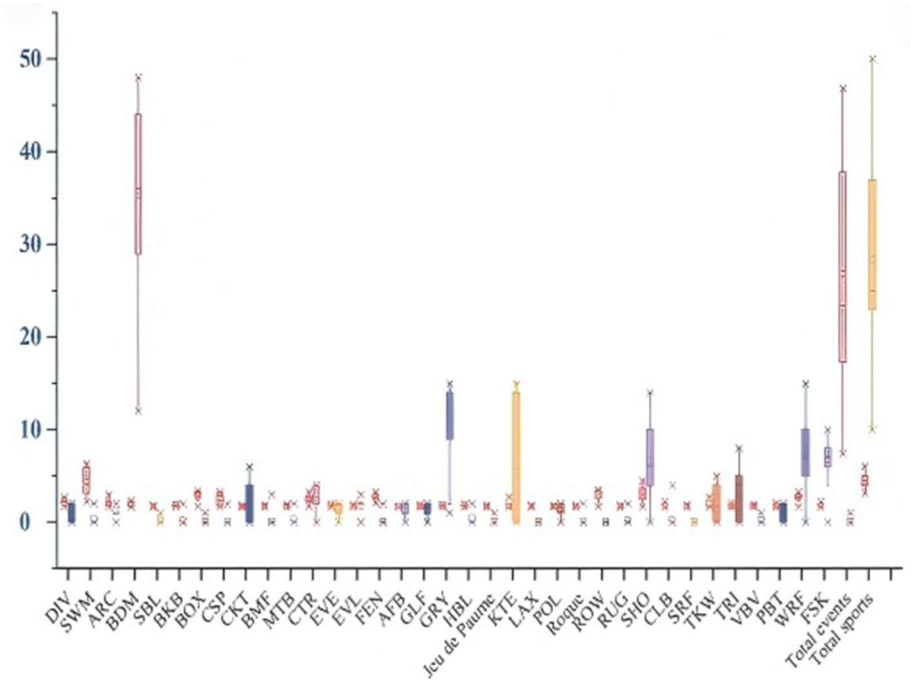
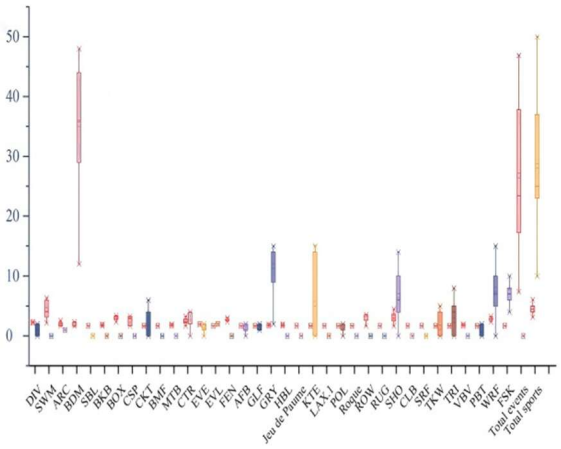


Figure 1: Spiral bar chart after missing values are filled in



(a) Boxplot for detecting outliers



(b) Boxplot for verifying data validity after outliers are processed

Figure 2: Box plots before and after data preprocessing

II. A. 3) Data preprocessing Correlation analysis of characteristic variables

In light of the plethora of data features that have the potential to influence the number of medals in this study, it is imperative that the correlations between these features are meticulously examined prior to the construction of the model and subsequent solutions. This methodological approach guarantees the selection of the most appropriate model to address the correlations between features, thereby facilitating a more nuanced and informed analysis.

During the course of the analysis, a number of characteristic variables were identified as having the potential to exert an influence on the total number of medals awarded, including Rank, Gold, Silver, Bronze, Total, and Year. Prior to the formalisation of the analysis, a test of normality was conducted on these variables. The statistical method of quantile-quantile plots was utilised in this test, which yielded the conclusion that the observed values of all the characteristic variables exhibited a low degree of overlap with their predicted values. This finding indicates that non-normality was observed in the dataset. Consequently, the Spearman's correlation coefficient was selected for the analysis of correlation [11].

Spearman correlation calculation requires variable ranking. First, the data for each variable is ranked, and then the difference between the rankings of the two variables is calculated. The required mathematical expression:

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)} \tag{1}$$

where r_s is the Spearman rank correlation coefficient; d_i is the difference in the rankings of the two variables on each pair of data points; n is the sample size.

The visualized heat map of the results obtained according to the Spearman's correlation coefficient is shown in Figure 3. Considering the presented heat map, it is evident that the number of gold, silver, and bronze medals exhibits a robust positive correlation with the total number of medals, with the three categories typically increasing in tandem. Furthermore, a pronounced negative correlation is observed between the ranking and the aggregate medal tally, revealing that higher rankings associate with greater medal accumulation. The relationship between the year and the number of medals is less pronounced. Through correlation analysis, it is evident that gold, silver, and bronze medal counts serve as primary characteristic variables in prediction models. To anticipate national medal tallies in future competitions, multiple linear regression serves as a robust modeling tool [12].

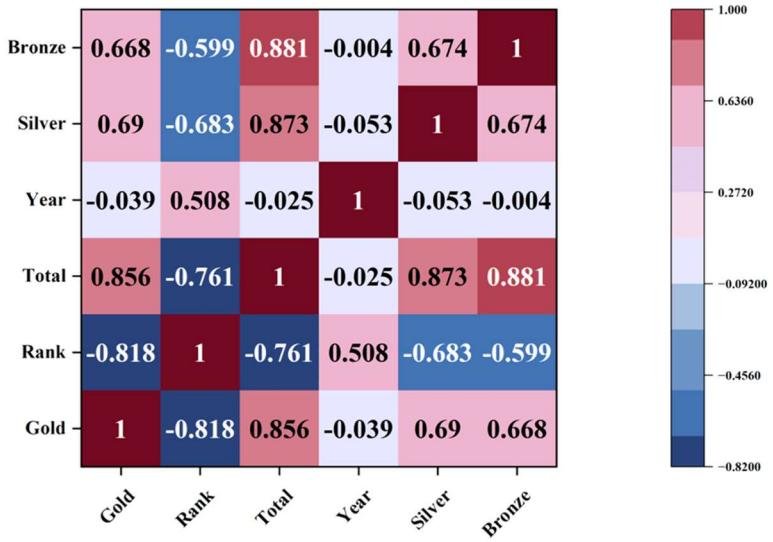


Figure 3: Heat map of correlation analysis results of data features

II. B.Methodology

II. B. 1) Multiple linear regression model

The multiple linear regression model boasts lots of advantages, including its simplicity and intuitive nature, as well as its high computational efficiency. Data preprocessing findings reveal a linear relationship connecting variables that influence medals awarded to actual medals received. To predict the number of gold medals and the total number of medals for each country in 2028, a linear regression model was selected for analysis.

Olympic sports play a significant role in determining a country's medal distribution and total medal count. Increasing sports tends to raise medal totals, while the addition of new sports may change the medal pattern. For example, traditional sports powerhouses may dominate in new events, or the type of event chosen by the host country may affect its medal count. Again, this problem can be analyzed by creating a multiple linear regression analysis, which can quantify the impact of characteristics such as the number and type of events on gold and total medals, thus providing the reference point for the National Olympic Committees to make strategic decisions.

II. B. 2) ARIMA model building and solving

To comprehensively address the temporal dependencies inherent in longitudinal medal data, the Autoregressive Integrated Moving Average (ARIMA) model was deployed. This advanced time-series methodology excels in identifying and extrapolating latent trends, cyclical fluctuations, and seasonal variations embedded within historical records—attributes critically necessary for forecasting medal trajectory evolution toward the 2028 Games. The ARIMA framework, widely validated in econometric and operational research contexts, is expressly suited for non-stationary continuous data exhibiting serial correlation.

ARIMA model is a widely used model for time series analysis and is suitable for dealing with continuous data with temporal correlation. ARIMA model consists of three main parts:

Autoregression (AR): the model's current-past value linear relationship;

Difference (I): differencing smooths the time series through trend removal;
Sliding Average (MA): predicts the current value by a weighted average of past errors.

II. B. 3) Logistic regression model building and solving

With the intention of predicting the probability that non-medal-winning country will win a medal for the first time at the 2028 Olympic Games in Los Angeles, a logistic regression model can be used in the classification model. The construction of a binary model with "whether to win a medal" as the target variable allows for the prediction of a country's probability of achieving a medal win for the first time in 2028.

The goal of this model is of learning regression coefficients given the training data. Maximizing the likelihood function allows the learning process of the regression coefficients to make the probability values predicted by the model as close as possible to the actual observed labels. This classifier adjusts internal values until predicted probabilities most closely match actual results. Here, the likelihood function represents the probability of observing the current data given the feature data.

II. C.Model Evaluation Metrics

The number of Olympic sports has been shown to have a significant impact on the distribution and aggregate medals for each country. An increase in the number of sports has been shown to result in an increase in the total number of medals awarded, while the introduction of new sports has been demonstrated to effect a change in the medal pattern. For instance, traditional sports powerhouses may dominate in new events, or the type of event chosen by the host country may affect its medal count. Once more, the problem can be analysed through the creation of a multiple linear regression analysis, which has the capacity to quantify the impact of characteristics such as the number and type of events gold medals counts and the tally medals. This, in turn, supports National Olympic Committees in developing evidence-informed policies [13].

When evaluating linear regression models, the coefficient of determination (R^2) and the mean square error (MSE) are two commonly used evaluation metrics that reflect the model's fit and prediction accuracy, respectively, from different perspectives.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}_i\right)^2} \tag{2}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{3}$$

R^2 is 0.86 and MSE is 16.58, calculated from the formulas. Given the larger medal data values, R²near 1 and MSE near 0 show the model fits well. Despite potential non-linearity, the multiple linear regression model provides accurate medal predictions.

III. Results and Analysis

III. A. Least squares estimation and multiple linear regression solving

III. A. 1) Model establishment

The regression model is configured as follows:

$$\mathrm{Y}=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n \in \tag{4}$$

where, Y is the predicted gold and total medal tallies; $X_1, X_2, \ldots, X_n$ is the dependent or characteristic variable, which may include factors such as historical medal count, national rankings, number of athletes, number of events and so on;β0 is the intercept term, which represents the baseline number of medals when all the characteristic variables have a value of zero; $\beta_1, \beta_2, \beta_3, \ldots, \beta_n$ is the regression coefficient; $\in$ is the error term.

The magnitude and sign of the regression coefficients indicate the relative importance of every characteristic variable in predicting the number of medals and its specific direction of influence on the dependent variable. By performing regression analysis on the training data set and estimating the best values of the regression coefficients, the main goal of this study is to minimize the discrepancy between the predicted values of the model and the actual observed values for optimizing the fitting effect of the model and then obtain an accurate and reliable regression model [14].

III. A. 2) Least squares estimation and multiple linear regression solving

This study applies ordinary least squares (OLS) to estimate parameters in a multiple linear regression model predicting medal counts. The specific objective function is:

$$J(\beta) = \sum_{i=1}^{N}\left(y_i - \left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_n x_{ni}\right)\right)^2 \tag{5}$$

where, y_i is the actual number of medals; $\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_n x_{ni}\right)$ is the predicted value of the model.

By minimizing the above objective function, it is able to estimate $\beta_0, \beta_1, \ldots, \beta_n$.

To minimize cumulative Squared Residuals, the objective function must be minimized by $J(\beta)$ by taking the derivative of the regression coefficient β and setting it to zero to obtain a closed-form solution for the regression coefficient.

The solution for the coefficient of regression is obtained as follows:

$$\beta' = \left(X^T X\right)^{-1} X^T Y \tag{6}$$

where, $\beta' = (\beta_0, \beta_1, \ldots, \beta_n)$, T is the estimate of the regression coefficient; X is an $N \times (n+1)$ matrix containing all the sample's eigenvariables, where the first column is 1 (used to represent the intercept term) and the remaining columns are the individual eigenvariables; Y is an $N \times 1$ vector containing the actual number of medals for all training samples; X^T is the transpose matrix of X.

With the above equation, it can directly calculate the regression coefficient β and thus harness regression analysis to anticipate medal distributions. For example, use the new feature data X' to predict the new medal count Y'. The formula is:

$$Y' = \beta X' \tag{7}$$

III. A. 3) Least squares estimation and multiple linear regression solving

Figure 4 shows the 2028 LA Summer Olympics medal standings predictions. The top 20 countries in overall and gold medal rankings are displayed, along with predicted intervals for all countries' medal and gold medal counts [15].

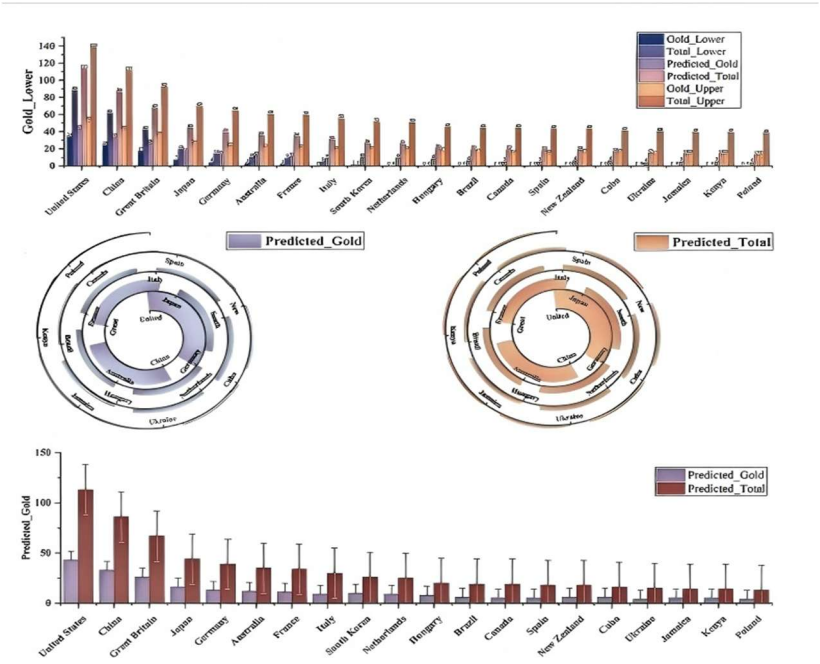


Figure 4: Projected results of the number of medals, gold medals and their projection intervals

For the possible nonlinear part in the multivariate linear regression mentioned in the previous model evaluation, logarithmic transformation or polynomial regression model can be considered to be added to the linear regression in the subsequent model improvement, which can further improve the accuracy of the result prediction.

III. B. Least squares estimation and multiple linear regression solving

III. B. 1) Model establishment

The ARIMA model can be written as:

$$Y_t = c + \sum_{i=1}^{p} \phi_i y_{t-1} + \sum_{i=1}^{q} \theta_j \in_{t-j} + \in_t \tag{8}$$

where, Y_t denotes the observed value at time t (in this problem, the number of medals); p is the order of the autoregressive (AR) term, capturing the linear dependence of the current value on its past values; q is the order of the moving average (MA) term, which is a linear combination of the error terms; Autoregressive coefficients ϕ_i

and moving average coefficients θ_j parameterize the model; $\in_t$ is called white noise, which denotes unpredictable random fluctuations; c is the constant term.

In the ARIMA model, autocorrelation and random fluctuations in the time series are captured by adjusting the values of p and q to provide a more accurate prediction of future medal counts.

A model smoothness test is also required during the model fitting process [16]. If the p-value is less than the significance level (usually 0.05), the data are considered smooth; otherwise, the data are considered non-smooth and further differencing is required. When a differencing operation is used, it is usually first-order differencing to smooth the data. The ADF test was repeated after differencing until the data became smooth.

First order differencing (FOD) is computed as:

$$Y_t^{'} = Y_t - Y_{t-1} \tag{9}$$

III. B. 2) Visualize changes in medal count results for the 2028 Olympics

As illustrated in Figure 5, the top 20 countries' medal tallies for the 2028 Olympics are presented, with a comparison between these predicted tallies and historical averages. The area in the figure denotes the discrepancy between these two tallies. Kosovo demonstrates the most significant improvement, while Spain exhibits the most regression. An increase in a country's medal count compared to its historical average is indicative of potential enhanced performance in 2028. Conversely, a decreasing trend suggests the possibility of regression.

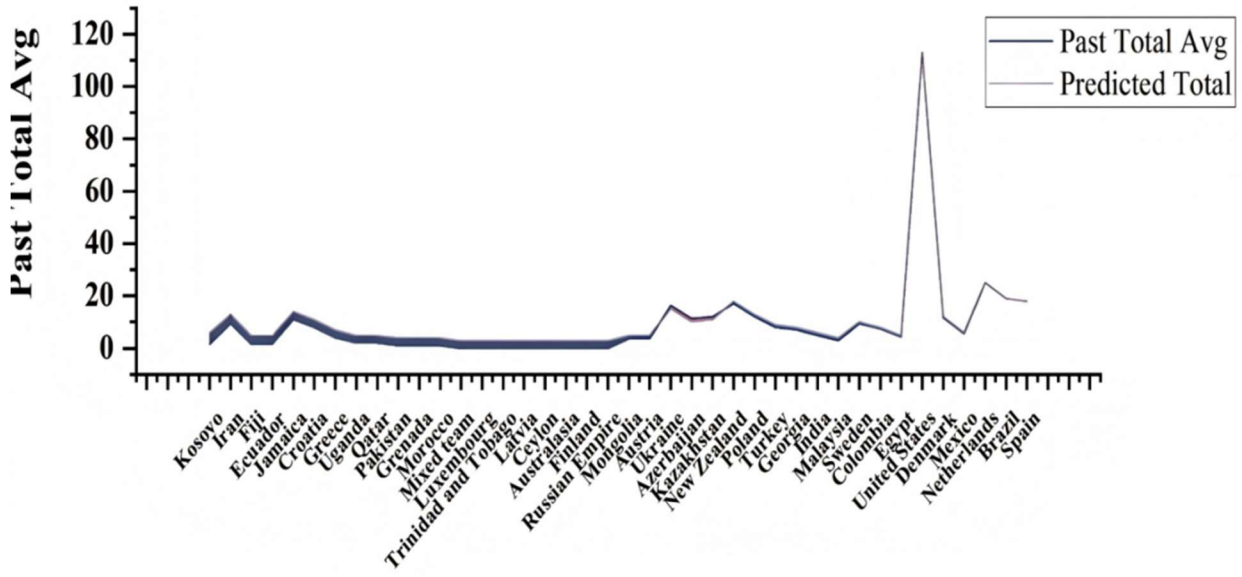


Figure 5: Two-color plot of national medal trends

III. C. Logistic regression model building and solving

III. C. 1) Model establishment

Logistic regression model makes prediction by calculating the probability of occurrence of an event and its expression is as follows:

$$P(Winning\ a\ medal) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots+\beta_n x_n)}} \tag{10}$$

where, $P(Winning\ a\ medal)$ indicates the probability that a country will win a medal in the 2028 Olympics; $X_1, X_2, \ldots, X_n$ are the characteristic variables that affect whether the country will win a medal, and together these characteristics determine the probability that the country will win a medal; $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients that gauge each factor's impact on medal outcomes. By training the model, the values of these coefficients can be estimated; e is the base of the natural logarithm, which is used to ensure that the output probability value of the model is always between 0 and 1, according to the definition of probability.

Logistic regression aims to determine the optimal coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ given the training data. Maximizing the likelihood function allows the learning process of the regression coefficients to make the probability values predicted by the model as close as possible to the actual observed labels [17].

III. C. 2) Maximum likelihood estimation and solution of regression coefficients

In logistic regression, the regression coefficients are estimated by maximizing the likelihood function. The likelihood function represents the probability of observing the current data given the feature data. Assuming there are n samples with labels y_i and features X, the likelihood function can be expressed as:

$$L=(\beta_0,\beta_1,\beta_2,\ldots,\beta_n)=\prod_{i=1}^{n}P(y_i\mid X_i) \tag{11}$$

where, $P(y_i\mid X_i)$ denotes the probability that the label y_i of sample i is 1. For the binary classification problem considered in this question, the label y_i takes the value 0 or 1, indicating whether the country won a medal or not. Therefore, $P(y_i\mid X_i)$ can be written as

$$P(y_i\mid X_i)=(Wm)^{y_i}(1-Wm)^{(1-y_i)} \tag{12}$$

If the label $y_i=1$ for sample i (i.e., the country won a medal), then the probability for that sample is $P(Winning\ a\ medal)$; if $y_i=0$ (i.e., the country did not win a medal), then the probability is $1-P(Winning\ a\ medal)$.

To maximize the optimization of the log-likelihood function and improve the stability of the value, the likelihood function is logarithmically transformed to obtain the log-likelihood function. The log-likelihood function is expressed as:

$$\ell(\beta_0,\beta_1,\beta_2,\ldots,\beta_n)=\sum_{i=1}^{n}\left[y_i\log(Wm)+(1-y_i)\log(1-Wm)\right] \tag{13}$$

where in equations (12) and (13), $Wm=1-P(Winning\ a\ medal)$.

By maximizing the log-likelihood function, one can solve for the regression coefficients $\beta_0,\beta_1,\beta_2,\ldots,\beta_n$, to make the model prediction probability as close to the actual label as possible. Gradient descent is a commonly used optimization method that adjusts the regression coefficients along the direction of increase of the log-likelihood function by computing the partial derivatives of each regression coefficient and iteratively updating them until the log-likelihood function converges and the regression coefficients stabilize [18].

III. C. 3) Optimization and Prediction of Logistic Regression Models

After training to obtain regression coefficients, these coefficients can be used to make predictions for new samples. For a given sample of feature $X_1,X_2,\ldots,X_n$, the probability of winning a medal can be calculated from the learned regression coefficients. Concretely, the computations are:

$$P(Winning\ a\ medal\mid X_1,X_2,\ldots,X_n)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\ldots+\beta_nx_n)}} \tag{14}$$

It should be noted that the accuracy and performance of the model depends on the quality of the training data and the optimization effect of the regression coefficients.

III. C. 4) Results of visualizing the probability of winning a medal for the first time

In the implementation of the logistic regression model for predicting inaugural medal attainment, the analytical procedure commenced with the computation of a critical longitudinal metric: the ratio of a nation's historical medal acquisitions to its total Olympic participations. This covariate served as a key predictor variable capturing sustained competitive engagement. Subsequent validation was provided through visual analysis in Figure 6, which confirmed the metric's discriminative power in distinguishing potential medal-winning nations. Leveraging this foundation, the model generated probabilistic forecasts for the 2028 Los Angeles Olympics, identifying ten nations exhibiting the highest likelihood of securing their first Olympic medals.

As depicted in Figure 7, Azerbaijan emerges as the foremost contender with a predicted probability of 62.5% - substantially exceeding probabilities assigned to other candidates. This projection derives from the nation's progressive improvement in athlete competitiveness and targeted sports investment, factors quantified within the regression framework. The statistically significant margin distinguishing Azerbaijan from subsequent candidates underscores its unique positioning for breakthrough performance. These findings operationalize the logistic model's capacity to translate complex historical patterns into actionable strategic intelligence. Such probabilistic rankings equip National Olympic Committees with empirically grounded insights for prioritizing developmental initiatives toward historically non-medaling nations demonstrating emergent potential.

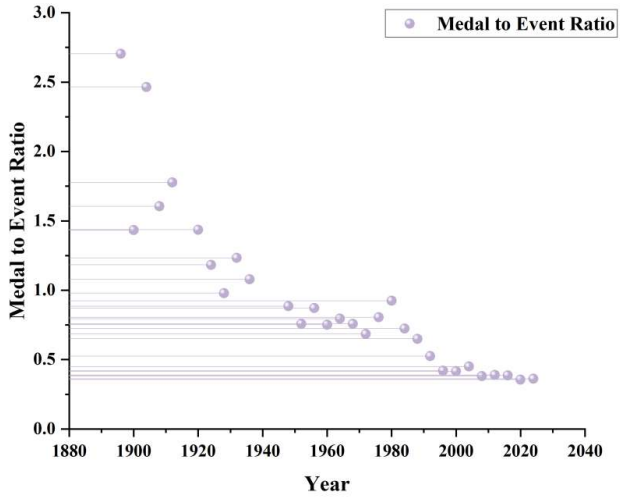


Figure 6: Scatterplot of ratio of number of medals to number of events

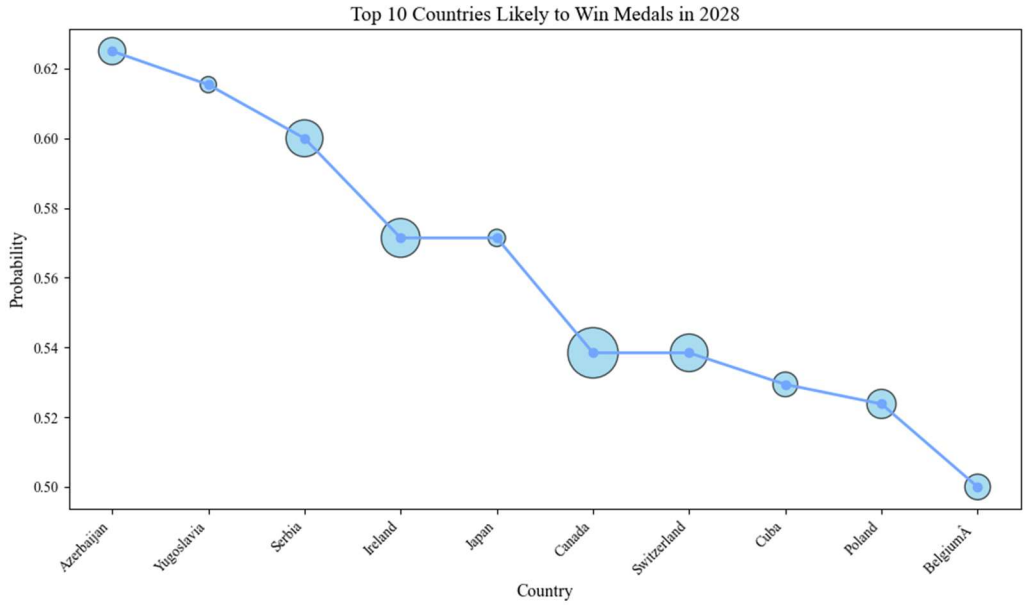


Figure 7: Countries likely to win medals in the 2028 Olympics

III. D. Multiple linear regression modeling and solving

III. D. 1) Model establishment

The present study has developed a multivariate linear regression model for the purpose of analyzing the relationship between the quantity of competition events and the quantity of medals. This model is designed to predict the number of gold medals or the total number of medals that a country will win at the Olympic Games. The employment of regression analysis permitted the quantification of the effect of various characteristics (e.g. number of competition events, type of competition, etc.) on the number of medals. The fundamental structure is outlined as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n + \in \tag{15}$$

where, Y is the target variable, indicating the predicted number of gold medals or total medals; $X_1, X_2, \ldots, X_n$ are characteristic variables related to the setting of the program, which may include the number of events in the Olympics, the type of events, the difficulty coefficient of the events, and so on [19]. Each characteristic can affect the variation of the number of medals; β_0 is the regression coefficient, which indicates the degree of influence of each characteristic on the number of medals; $\beta_1, \beta_2, \beta_3, \ldots, \beta_n$ is the regression coefficient, which indicates the degree of influence of each characteristic on the number of medals; $\in$ is the error term, which includes all random factors not captured by the characteristic variables [20].

In this model, this study considers the number of medals Y as a linear combination of all characteristic variables $X_1, X_2, \ldots, X_n$ in a linear combination. To better understand the degree of influence of each characteristic, it also need to estimate the regression coefficients in the model.

III. D. 2) Least squares estimation and multiple linear regression solving

The optimal regression coefficients are obtained by minimising the sum of squared residuals through least squares optimisation. Given m samples, the predicted value of the model is

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} \tag{16}$$

where, Y is the predicted number of medals for the ith sample.

$$\in_i = Y_i - \hat{Y}_i \tag{17}$$

$$F(\beta) = \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_n x_{ni} \right) \right)^2 \tag{18}$$

The estimates of the regression coefficients can be expressed in the following matrix form:

$$\hat{\beta} = \left(X^T X \right)^{-1} X^T Y \tag{19}$$

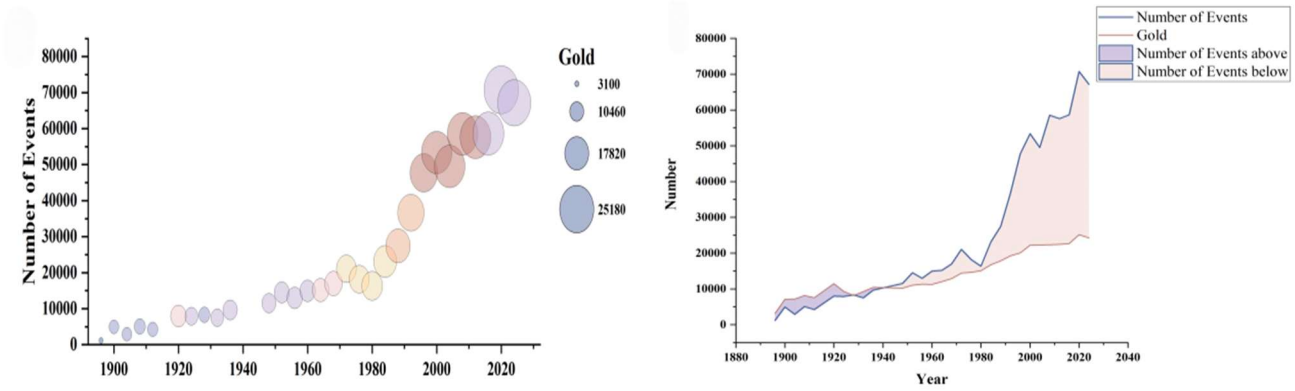
where, $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n \right)$, T is the estimate of the regression coefficient; X is an $N \times (n+1)$ matrix containing all the sample's eigenvariables, where the first column is 1 (used to represent the intercept term) and the remaining columns are the individual eigenvariables; Y is an $N \times 1$ vector containing the actual number of medals for all training samples; X^T is the transpose matrix of X.

According to equation 20, the estimated values of the regression coefficients can be obtained. The final regression model obtained can effectively predict the impact of different program settings on medal counts and provide suggestions for further decision making.

III. D. 3) Visualization and analysis of results

Figure 8 visualizes the correlation between gold medal counts and event volumes, which demonstrates a positive correlation between the quantity of events and medal acquisition outcomes.

This may be especially true for some traditionally strong countries, which tend to have more opportunities to win gold medals in a greater number of events, leading to an increase in the number of medals.



(a) Bubble plot

(b) Two-color filled plot

Figure 8: Plot of Number of Gold Medals vs. Number of Events

A critical nuance emerging from the analysis of Olympic medal distributions pertains to the differential impact exerted by novel sporting disciplines integrated into the Games programme. While the aggregate expansion of events generally correlates with increased total medal allocations, specific emergent sports—such as sport climbing, skateboarding, and surfing—demonstrate markedly limited influence on reshaping established hierarchies, particularly among traditional sporting powerhouses [21]. This constrained effect manifests in two primary dimensions: firstly, dominant nations often lack specialized infrastructure or deeply developed talent pipelines for such niche activities during initial Olympic cycles; secondly, the inherently smaller participant pools and medal opportunities in nascent disciplines inherently cap their contribution to aggregate national medal tallies. Consequently, the strategic value of programme modifications cannot be uniformly generalized; it necessitates

granular, nation-specific evaluation predicated on existing athletic proficiencies, historical performance datasets, and targeted investment trajectories.

Therefore, structural alterations to the Olympic sporting portfolio—whether involving discipline inclusion, exclusion, or format modification—demand meticulous ex-ante analysis calibrated to individual national contexts. National Olympic Committees (NOCs) must undertake systematic audits of their competitive advantages across the revised sporting matrix. Through such diagnostic exercises, NOCs can probabilistically identify: (a) disciplines where marginal resource investments could yield disproportionate medal returns; (b) high-risk events misaligned with national strengths warranting divestment; and (c) emerging sports offering viable pathways for competitive breakthroughs [22]. This evidence-based approach transforms programme changes from exogenous uncertainties into strategic opportunities, enabling optimized athlete development prioritization, facility investment targeting, and technical staff deployment—ultimately enhancing medal efficiency per unit of resource expenditure.

Complementarily, athletic performance optimization extends beyond sport selection to encompass human capital factors, most saliently the influence of elite coaching. High-caliber coaches function as performance multipliers through multidimensional interventions: they drive technical refinement via biomechanical analysis and skill drills; cultivate psychological resilience under competitive duress; and devise bespoke tactical frameworks exploiting opponents' vulnerabilities [23]. This triad of competencies—technical, psychological, tactical—collectively translates into measurable competitive advantages demonstrable through podium outcomes. Quantifying this "coaching effect" thus constitutes a vital analytical frontier.

Our methodological approach employs multivariate regression techniques to isolate and quantify the marginal contribution of elite coaching to national medal fluctuations. Control variables include athlete pool depth, funding levels, and sport-specific training infrastructure. The analysis further examines heterogeneity in coaching efficacy across sporting domains—contrasting, for instance, the highly individualized technical coaching vital in gymnastics or swimming against the complex system-based coaching essential in team sports like hockey or volleyball. This granular understanding informs NOC strategies for coach recruitment, retention, and continuous professional development, positioning coaching excellence as a deliberate performance strategy rather than an incidental variable.

IV. Conclusions

In this study, a multi-model fusion framework was constructed for the purpose of conducting a comprehensive analysis of the trend of medal counts, the probability of countries' first-time winners, and the impact of programs in the 2028 Olympic Games. The multiple linear regression model is characterized by its high interpretability, which allows for the quantification of the impact of historical results, events, and other factors. In contrast, the ARIMA model is capable of accurately capturing the trend of medal counts and identifying potential and risky countries. The logistic regression model provides a scientific assessment of the likelihood of first-time winners for non-medal-winning countries. The study found that the host effect is significant, and the expansion of the event programme contributes to the increase in the number of medals. It is evident that emerging countries have the potential to achieve significant success in new events. Subsequent research and the extension of the model will concentrate on the prediction of Paralympic medals and the financial consequences of sporting events. This will assist in the effective allocation of global sports resources and the promotion of fair competition.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

Data Sharing Agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

[1] MacAloon J J. Olympic Games and the theory of spectacle in modern societies[M]//The Olympics. Routledge, 2023: 80-107.

[2] Sahu M, Gupta R, Ambasta R K, et al. Artificial intelligence and machine learning in precision medicine: A paradigm shift in big data analysis[J]. Progress in molecular biology and translational science, 2022, 190(1): 57-100.

[3] Mair J, Chien P M, Kelly S J, et al. Social impacts of mega-events: A systematic narrative review and research agenda[J]. Methodological Advancements in Social Impacts of Tourism Research, 2023: 140-162.

[4] Lopes dos Santos G, Gonçalves J. The Olympic Effect in strategic planning: insights from candidate cities[J]. Planning Perspectives, 2022, 37(4): 659-683.
[5] Schlembach C, Schmidt S L, Schreyer D, et al. Forecasting the Olympic medal distribution–a socioeconomic machine learning model[J]. Technological Forecasting and Social Change, 2022, 175: 121314.Johnson, D. (2024). Olympic Medal Prediction Model. Colorado College.
[6] Makiyan S, Rostami M. Economic determinants of success in Olympic Games[J]. Turkish Journal of Sport and Exercise, 2021, 23(1): 33-39.
[7] Forrest D, Tena J D, Varela-Quintana C. The influence of schooling on performance in chess and at the Olympics[J]. Empirical Economics, 2023, 64(2): 959-982.
[8] Zhao S, Cao J, Steve J. Research on Olympic medal prediction based on GA-BP and logistic regression model[J]. F1000Research, 2025, 14: 245.
[9] Alabrah A. An improved CCF detector to handle the problem of class imbalance with outlier normalization using IQR method[J]. Sensors, 2023, 23(9): 4406.
[10] Wang H, Li J, Li Z. AI-generated text detection and classification based on BERT deep learning algorithm[J]. arXiv preprint arXiv:2405.16422, 2024.
[11] Ali Abd Al-Hameed K. Spearman's correlation coefficient in statistical analysis[J]. International Journal of Nonlinear Analysis and Applications, 2022, 13(1): 3249-3255.
[12] Öztürk O B, Başar E. Multiple linear regression analysis and artificial neural networks based decision support system for energy efficiency in shipping[J]. Ocean Engineering, 2022, 243: 110209.
[13] Das A. Logistic regression[M]//Encyclopedia of quality of life and well-being research. Cham: Springer International Publishing, 2024: 3985-3986.
[14] Valentini M, dos Santos G B, Muller Vieira B. Multiple linear regression analysis (MLR) applied for modeling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the state of Rio Grande do Sul—Brazil[J]. SN Applied Sciences, 2021, 3: 1-11.
[15] Huimin S H I, Dongying Z, Yonghui Z. Can Olympic Medals Be Predicted? Based on the Interpretable Machine Learning Perspective[J]. Journal of Shanghai University of Sport, 2024, 48(4): 26-36.
[16] Gil-Alana L A, Yaya O O S. Testing fractional unit roots with non-linear smooth break approximations using Fourier functions[J]. Journal of Applied Statistics, 2021, 48(13-15): 2542-2559.
[17] Vrugt J A, de Oliveira D Y, Schoups G, et al. On the use of distribution-adaptive likelihood functions: Generalized and universal likelihood functions, scoring rules and multi-criteria ranking[J]. Journal of Hydrology, 2022, 615: 128542.
[18] Haji S H, Abdulazeez A M. Comparison of optimization techniques based on gradient descent algorithm: A review[J]. PalArch's Journal of Archaeology of Egypt/Egyptology, 2021, 18(4): 2715-2743.
[19] Arpna N, Dalal S. Optimizing Linear Regression model[J]. Trends in Mechatronics Systems: Industry 4. 0 Perspectives, 2024: 73.
[20] Gulati S, Bansal A, Pal A, et al. Estimating PM2. 5 utilizing multiple linear regression and ANN techniques[J]. Scientific Reports, 2023, 13(1): 22578.
[21] Gill R, Greenland S. Branding and the Olympics[J]. Asia Pacific Public Relations Journal, 2024, 26.
[22] Kumar R. Building a healthy nation: A white paper on Olympic sports and the Indian education system[J]. Journal of Family Medicine and Primary Care, 2024, 13(8): 2805-2818.
[23] Mu'ammal I, Muzakki A, Fakhri E A, et al. The competence of a coach in sports: How does it correlate with athlete motivation?[J]. Journal Sport Area, 2022, 7(3): 396-404.