Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 11
TPCEE 2025

A Variable Association Modeling and Dynamic Optimization Approach for Complex System Prediction

Wenqian Zhong [#], Jingrui Xu [#], Jiabao Luo [#, *]

China University of Petroleum (East China), Qingdao, China

* Corresponding Author Email: LuoJiabao2025@163.com

[#]These authors contributed equally.

Abstract. In this paper, a two-stage modeling framework integrating multivariate statistical analysis and dynamic correction is proposed to address the problems of lack of variable correlation and insufficient modeling of nonlinear influence mechanism in traditional prediction models in complex systems. The key explanatory variables are first screened out based on feature engineering, and the multiple linear regression prediction model and logistic regression model are constructed respectively. To further address the interference of unobserved variables on the prediction results, the prediction model of important influencing factors is innovatively established, and the prediction results are optimized by quantifying the interference intensity of external factors. The empirical study shows that the improved combination model achieves a large improvement in prediction accuracy compared with the traditional prediction model. The modeling method provides a new technical path for accurate prediction of complex systems by establishing an interpret-able variable association system and a dynamic correction mechanism.

Keywords: anticipate; multiple linear regression; logistic regression; dynamic correction.

1. Introduction

High-profile global events rank among the most prestigious and influential gatherings worldwide, drawing considerable attention for both the outstanding accomplishments of participants and the intricate distribution of outcomes across diverse contingents. Accurately forecasting these results has become a pivotal research objective, seeking to unravel the myriad factors shaping performance outcomes and to guide strategic resource allocation.

Existing research highlights the significance of economic development, population size, historical performance, and host-related advantages as key determinants [1-2]. Analytical approaches range from classical linear regression to advanced machine learning [3-5], yet conventional models frequently overlook the interdependence of variables and the nonlinear dynamics characteristic of real-world systems.

To address these limitations, this paper proposes a two-stage modeling framework that integrates multivariate statistical analysis with a dynamic correction mechanism. First, critical explanatory variables are extracted through feature engineering, followed by the construction of multiple linear regression and logistic regression models. Subsequently, an innovative predictive model accounts for unobserved factors by quantifying their interference and refining initial forecasts. Empirical evidence demonstrates notable improvements in predictive accuracy relative to traditional methods, while also offering an interpretable variable association system and a robust correction process [6-7].

2. Model

2.1. Predictive Models Based on Multiple Linear Regression

In general, the factors affecting the outcome of an event are complex and diverse. In order to explore the relationship between each influencing factor and the outcome, and to predict the outcome, we establish a prediction model based on multiple linear regression. Multiple linear regression is a very

commonly used predictive model in mathematical modeling, and it is especially suitable for dealing with the problem of predicting multiple independent variables on one dependent variable. The core idea is to predict the dependent variable by a linear combination of multiple independent variables. Its mathematical expression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \tag{1}$$

where y is the dependent variable, $x_1, x_2, \ldots, x_n$ is independent variables, β_0 is the intercept term, $\beta_1, \beta_2, \ldots, \beta_n$ is the regression coefficient, and ε is the error term.

The following will show the specific application of the model in predicting the results of international events. Based on the linear regression model, a model for predicting the total number of medals and a model for predicting the number of gold medals were developed. They were used to predict the total number of medals and the number of gold medals for each country, respectively. The formulas are respectively:

$$M_i = \beta_0 + \beta_1 \cdot M_{i,base} + \beta_2 \cdot E_i + \beta_3 \cdot T_i + \epsilon_i \tag{2}$$

where M_i is the total number of medals for the country, $M_{i,base}$ is the historical weighted medal count, E_i is project participation benefits and T_i is the medal trend factor (which indicates the rate of growth or decline in the historical number of medals).

$$G_i = \alpha_0 + \alpha_1 \cdot G_{i,base} + \alpha_2 \cdot G_{i,eff} + \zeta_i \tag{3}$$

where G_i is the number of gold medals for the country, $G_{i,base}$ is the historical weighted gold medal count and $G_{i,eff}$ is Project Gold Medal Efficiency (which indicates the number of gold medals as a percentage of the number of medals).

Since host countries can increase their medal counts by increasing the number of events and improving event participation, it is necessary to modify the medal counts of host countries by modeling the host country effect, and Schlembach et al. [8] have also emphasized the role of host country advantages in influencing medal outcomes. Our model incorporates this perspective and adjusts the predictions accordingly. Its formula is:

$$\Delta M_{host} = \gamma_1 \cdot N_{events_{added}} + \gamma_2 \cdot P_{host} \tag{4}$$

where ΔM_{host} is the number of medals added by the host country, N_{events_added} is the number of projects added by the host country, P_{host} is Coverage of host countries in all projects.

2.2. Logistic Regression Models

Logistic regression modeling is a widely used statistical learning method for classification problems. It predicts the probability of an event occurring by mapping the output of a linear regression between 0 and 1 using a logistic function. Therefore, we build logistic regression models to predict the occurrence or non-occurrence of time. Its formula is:

$$P(y = 1|x) = \sigma(z) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)}} \quad (5)$$

where $P(y = 1|x)$ is the probability of the dependent variable given the variables.$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ is a linear combination of parts called Logit.

Based on a logistic regression model, it can be used to predict the probability of a country winning its first Olympic medal, expanding research related to the impact of first-time medalists and emerging countries [9]. The formula is:

$$P_{medal} = \frac{1}{1+e^{-(\eta_0+\eta_1 \cdot E+\eta_2 \cdot T_{participation})}} \quad (6)$$

where P_{medal} is the probability that a country will win the first medal, E is the number of project participants, $T_{participation}$ is cumulative number of participating sessions.

2.3. Significant Influencing Factor Model

Some events are characterized by certain important factors that may have a significant impact on the outcome of the event. Therefore, a model of significant influencing factors was developed.

Taking the Olympic Games as an example, it is well known that the number of medals won at the Olympics can be easily influenced by certain factors such as good coaching, athletes' form and major international events. Turgut and Mumcu [10] have documented the impact of the coaching system on Olympic success. Therefore, this model refines this analysis by quantitatively estimating the contribution of coaches to the number of national medals. Prior to building the model, we made assumptions about the model: it was assumed that the addition of good coaches increases the number of medals, and that this increase is mathematically expressed as a fixed increment; it was assumed that a country's level of competition naturally varies over time, and that this variation is mathematically expressed as a linear relationship between the number of medals and time; and it was assumed that other factors also affect the number of medals at the Olympic Games, and that this effect is uniformly expressed as the Number of medals. Thus, the great coaching model can be expressed as:

$$M = \beta_0 + \beta_1 \cdot C + \beta_2 \cdot T + \beta_3 \cdot Y_i + \epsilon \quad (7)$$

where M is number of medals in a country's games, C is the influence of great coaches, T is A time trend indicating the rate of change in the number of medals over time and ζ is an error term.

There is also a need to quantify the impact of great coaches on medal counts. According to the great coaches' model, the impact of great coaches on the number of medals can be quantified as:

Great coaches contribute incrementally:

$$\Delta M = \beta_1 \quad (8)$$

where ΔM indicates the value of the increase in the number of national medals after the introduction of a great coach.

Great coaches contribute to rate increases:

$$R_{\text{coach}} = \frac{\beta_1}{M_{\text{no-coach}}} \tag{9}$$

where R_{coach} indicates the percentage improvement in medal counts after the introduction of a great coach, $M_{\text{no-coach}}$ indicates the number of medals the country has won when no great coach was brought in.

3. Result and analysis

The data used in this study came from three authoritative platforms. First, get Olympic medal information from the Paris 2024 MEDALS page on the Olympic website olympis.com. As the official platform of the International Olympic Committee, it provides comprehensive and accurate data related to the Olympic Games. Second, the personal information of Lang Ping, coach of the Chinese Women's Volleyball team, comes from her profile page on Olympics.com, which details her career and Olympic achievements. Finally, about Coach Bela Karoly and his team from the official page of the USA Gymnastics Hall of Fame, usagym.org, which authoritatively documents the significant contributions of Coach Karloy and his team. All data collection is subject to the respective website's terms of use and privacy policy.

To more accurately present the characteristics and efficiency of the model constructed in this paper, the Olympic Games, a very representative and influential international large-scale sports event, is selected as the study case. The following will elaborate on the specific analysis carried out after the Olympic Games related data is input and substituted into the model.

3.1. Multiple Linear Regression Predictive Analysis

Disregarding the host effect, the United States will total around 120 medals secured in the 2028 Olympics. Given that Los Angeles, USA, hosted the 2028 Olympics, it's projected that the U.S. will gain significant advantages, such as advantageous competitions like baseball, and flag football, and local backing and event organization. Foreseeing the hosting nation's impact, the U.S. is poised to secure approximately 122-123 medals, with a forecast of 46 gold. Figure 1 and Figure 2 below show how the United States won.

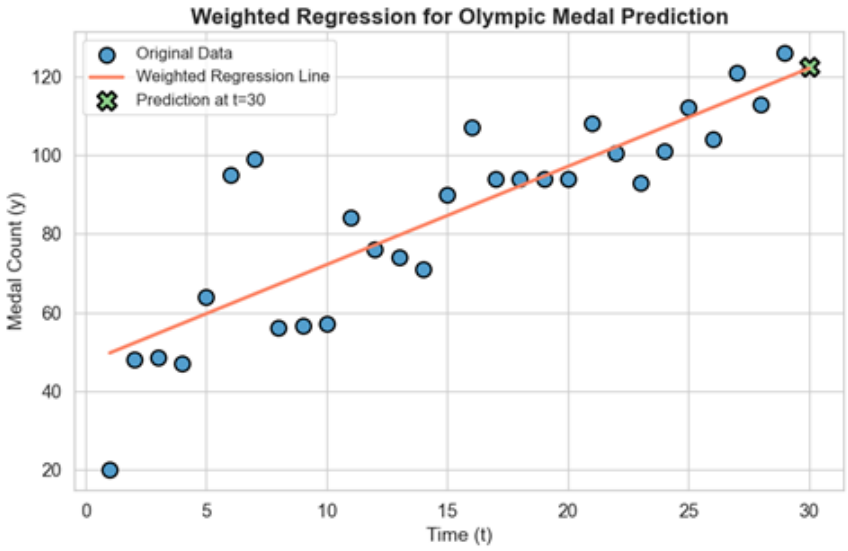

Figure 1. The prediction of America (total)

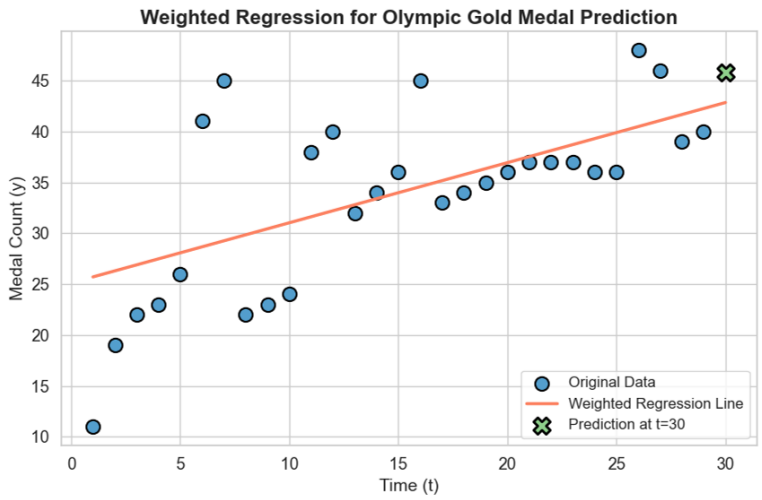

Figure 2. The prediction of America (gold)

Predictions are made for the number of gold medals and total medals for China, Australia, Japan, France and Great Britain. Their gold medal counts are: 39, 15, 15, 13, 19. Their total medal counts are: 91, 49, 43, 46, and 57.

Based on the model's prediction for the 2028 Olympic Games, Figure 3 shows our predicted medal tally.

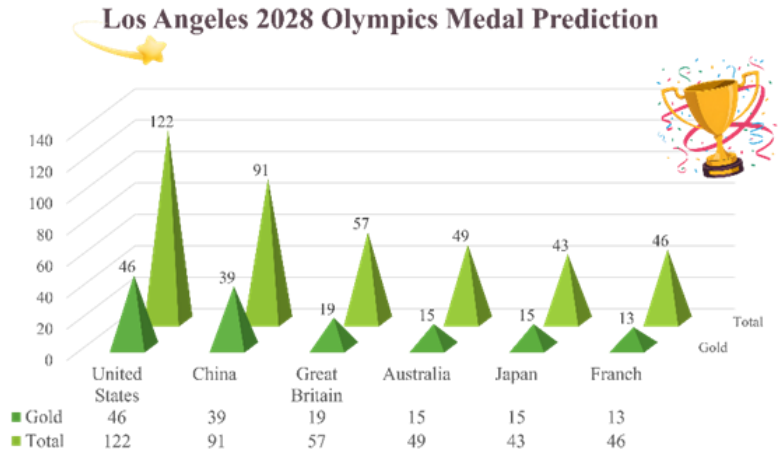

Figure 3. Prediction of the 2028 Olympic Awards

The following conclusion can be drawn:

The USA is ahead in both the gold medal count and overall medal tally, with China coming in next with 39 golds and a total of 91. The United States and China are leading, with Australia, Japan, and France experiencing a narrowing of their discrepancies in medals. Australia will probably rank fourth overall in medal tally, overtaking France.

To compare the changing trends, one can compare the achievements of various countries in previous years.:

In the 2028 Olympic Games, the total number of medals won by the United States and China will be similar to that in 2024, but the number of gold medals will increase. Although China's gold medal count is less than that of the United States, its growth rate is higher, indicating that the United States has a superior competitive level and China has increased its investment and achieved better results. For the United Kingdom, France, Japan and Australia, their total medal counts will decrease to varying degrees.

The total medal counts of the UK, France, Japan and Australia will decrease. Their gold medal numbers will also decline, indicating that more countries are increasing their investment in competitive sports and the Olympic Games, and the competition is becoming increasingly fierce.

The number of gold medals won by Japan will increase, reflecting the rising investment in sports events by Japan. The country's performance is expected to improve and there is great growth potential.

3.2. Analysis of Logistic Regression Model

Based on logistic regression models, we aim to predict which countries will achieve their own medal zero breakthrough in the future.

Figure 4 displays the proportion of countries securing their inaugural medal. Projections indicate that 93.4% of nations will not attain this milestone, whereas 6.6% have a potential to win their first Olympic medal.

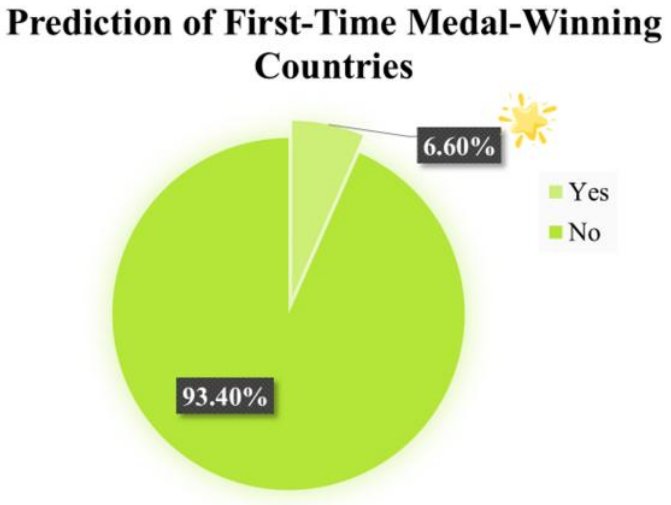

Figure 4. Forecasting First-Time Medal Winners' Percentages

It is predicted that in 2028, there will be 12 athletes from different countries, and they will win the first medal for their respective countries. Among them, the one with the highest possibility is the athlete from El Salvador, who has a 0.85 probability of becoming a dark horse in the Olympics.

The athlete from Bolivia has a 0.83 probability of winning a medal, followed by the athlete from Samoa, with a probability of 0.78.

The other members with higher probabilities can be obtained from Figure 5.

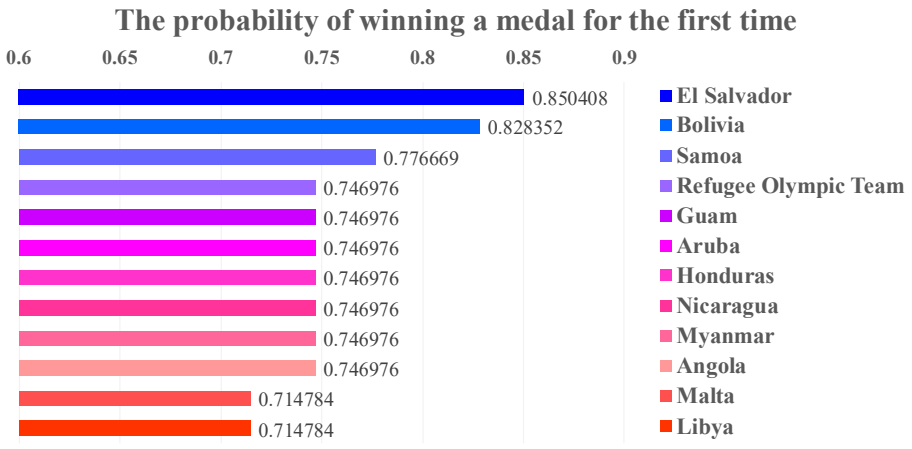

Figure 5. Predictions for first-time medal winners

3.3. Analysis of Significant Influencing Factor Model

We all know that Lang Ping is an excellent coach. Take Lang Ping's winning of Olympic MEDALS as an example to analyze. We can see the result in Figure 6.

The conclusion drawn pertains to Lang Ping's coaching effectiveness and temporal trends. Lang Ping's coaching effect, which means from the perspective of curve position, the predicted values corresponding to Lang Ping's coaching are generally higher than those without Lang Ping's coaching, suggesting that the model believes Lang Ping's coaching has a positive impact on increasing the number of medals/gold medals. In reality, Lang Ping coached the US team from 2005 to 2008. The number of medals in the US (with great coaches) has changed significantly, while that of China (without great coaches) has changed little. This might be caused by the effect of great coaches. Time trend, which means both lines slightly decline over time, but this trend is not very strong.

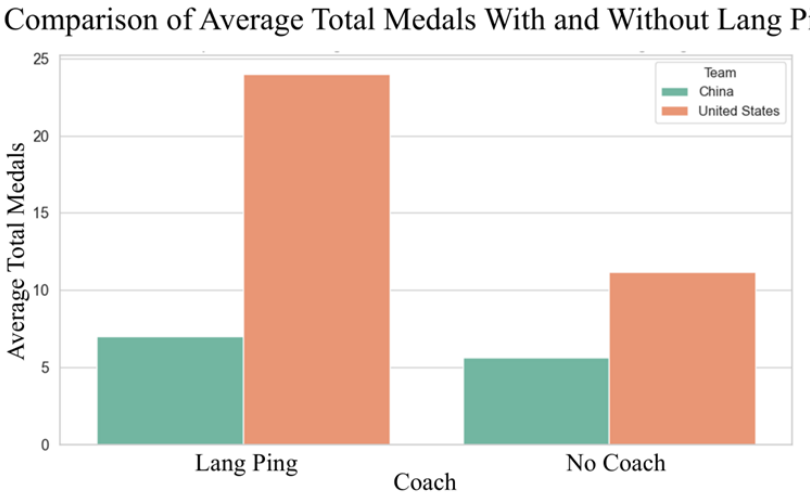

Figure 6. Analysis of influencing factors of excellent coach

4. Conclusions and outlooks

In this paper, a prediction model based on multiple linear regression is established. The core principle of the model is to establish a linear relationship between multiple independent variables and dependent variables and comprehensively consider the influence of various traditional factors on the prediction target, to achieve more accurate prediction. Based on in-depth analysis of Logistic regression and principal component analysis, the Logistic regression model is further optimized. Specifically, using principal component analysis to extract the most representative combination of variables as the input of Logistic regression not only retains the main information inherent in the original data set, but also effectively reduces the complexity of the model and enhances the generalization ability of the model. Finally, on the basis of the above analysis results, a significant impact factor model is constructed to focus on identifying and quantifying the factors that have a significant impact on the prediction results, to provide more intuitive and profound insights for decision makers. Through in-depth mining of historical data, these models not only accumulate a large number of high-quality sample data, but also carry out iterative optimization of model parameters, which significantly improves the prediction accuracy. This in turn contributes to a more comprehensive and detailed solution to the various complex forecasting tasks in practice.

This model has high practical value and broad application prospect. The prediction model based on multiple linear regression can provide a powerful reference for the relevant personnel to make competitive strategy, resource allocation and marketing plan. At the same time, this research method also provides a useful reference for other fields of predictive analysis. These models obtained a large number of samples through deep mining of historical data, effectively improving the accuracy of prediction and improving the robustness of the model.

Although this study has made some achievements, it also has some limitations. Due to data limitations, some unconventional missions lack data, which brings uncertainty to the forecast results. Future research will focus on addressing this issue and further refining the model.

In summary, the multi-dimensional regression model proposed in this paper provides a new perspective and method for predicting various complex tasks in practice. Through the continuous optimization and application of the model, it is expected to provide important reference value for predicting the future development trend of the event.

References

[1] Tchamkerten A , Chaudron P , Girard N ,et al.Career factors related to winning Olympic medals in swimming[J]. PLoS ONE, 2024,19(6): e0304444.

[2] Maszczyk A, Gołaś A, Pietraszewski P, et al. Application of neural and regression models in sports results prediction[J]. *Procedia-Social and Behavioral Sciences*, 2014, 117: 482-487.

[3] Nicolas Frevel, Daniel Beiderbeck, Sascha L. Schmidt, The impact of technology on sports – A prospective study, *Technological Forecasting and Social Change*, 2022, 182: 121838.

[4] Horvat T, Job J. The use of machine learning in sport outcome prediction: A review[J]. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2020, 10(5): e1380.

[5] Bunker R P, Thabtah F. A machine learning framework for sport result prediction[J]. *Applied Computing and Informatics*, 2019, 15(1): 27-33.

[6] Apostolou K, Tjortjis C. Sports analytics algorithms for performance prediction[C]// *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, 2019: 1-4.

[7] Wei Jiang, Julie Josse, Marc Lavielle, Logistic regression with missing covariates—Parameter estimation, model selection and prediction within a joint-modeling framework, *Computational Statistics & Data Analysis*, 2020, 145: 106907.

[8] 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.

[9] Baumer B S, Matthews G J, Nguyen Q. Big ideas in sports analytics and statistical tools for their investigation[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2023, 15(6): e1612.

[10] Turgut, Abdüsselam, Mumcu H E. The effect of couching systems on Olympics success: The case of the United States and the Russian Federation coaching education systems [J].*International Journal of Education Technology & Scientific Researches*, 2023, 8(23):1040-1069.