Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2509384

The Riddle of Olympic Medal Table: Mathematical Model Prediction and Multi-factor Analysis

Summary

The Summer Olympics are held every four years. The Olympic medal tally and its related influencing factors are a focus of general concern. Therefore, we set up a mathematical model to analyze and forecast the relevant data of the Olympic Games.

Firstly, we build a **Hierarchical Bayesian** combinatorial **ARMA model** to predict the 2028 Olympic medal. This model comprehensively analyzes the influence of different factors and time series on the Olympic medal table. Therefore, we predict medal distribution that in 2028. Then, we combined the **Dirichlet Distribution model** to calculate the number of gold medals in the predicted medals and evaluated the analysis through the posterior distribution. The **MAE result is 3.1592**, indicating that the model has good prediction performance.

Secondly, we used the data to divide countries into two clusters (potential and non-potential) based on historical medal count and whether they have won medals. Then, we use **K-means** clustering method for analysis and prediction. The final prediction is that **seven countries** will win their first medals, and then we quantify the probability distribution and predict their odds.

Thirdly, we set up a **Multiple Linear Regression(MLR) model** to analyze and explore the impact and relationship between events and organizers on the number of medals. The model training results show that the event has a significant impact on the number of medals, and the **coefficient reaches 32.8129**. The host country also had a significant **positive impact** on the number of medals, with a **coefficient of 47.8638**. In addition, we established Lasso regression model to explore the most important sports in different countries, and the evaluation error of the model was small, indicating that the model was highly interpretable.

Fourthly, we still use **MLR model** to analyze the influence of coach factor is used to explore the effect of "great coach". The model results show that the model's fitting degree R^2 **is above 0.7**. In addition, the model indicates that the great coach has a greater contribution to the number of medals. Then, we use **Simulated Annealing(SA)** algorithm to solve a model, which employs a **dynamic penalty function** to control variables, showing the United States, Japan, Belarus three countries to hire the program and its possible impact on the number of medals in the next Olympic Games.

Finally, we explored three insights from our model and explain their reference value to country Olympic Committees.

Keywords: Hierarchical Bayesian, Dirichlet Distribution, K-Means++, MLR, Lasso, SA

Contents

1 Introduction

1.1 Problem Background

As one of the most influential comprehensive sports events in the world, the Summer Olympic Games will be held every four years, attracting the attention of people all over the world. Olympic medals are not only a platform for athletes to show themselves, challenge themselves and pursue excellence, but also reflect a country's honor, sports strength and Economic strength and many other aspects. By modeling and forecasting the distribution of Olympic medals, it can better optimize the allocation of sports resources, stimulate the athletes' competitive state and promote the development of science and technology sports.

1.2 Restatement of the Problem

To achieve these goals, we need to analyze the provided data and answer the following questions:

- Develop a prediction model to forecast the 2028 Los Angeles Summer Olympics medal table and provide the medal distribution of countries compared to the 2024 Summer Olympics.
- Build another prediction model to predict how many countries will win their first medal at the next Olympic Games. Model evaluation is also required.
- Establish a model to analyze the relationship between the number and type of events in the Olympic Games and the number of medals won by each country, the important sports of different countries and the influence of the selection of home country events on the number of medals won.
- Select three sports in which the country could benefit from a great coach and analyze the potential impact of the "great coach" effect on their country's Olympic performance.
- Explore the original insights of the model and explain its help to the Olympic Committee.

1.3 Our Work

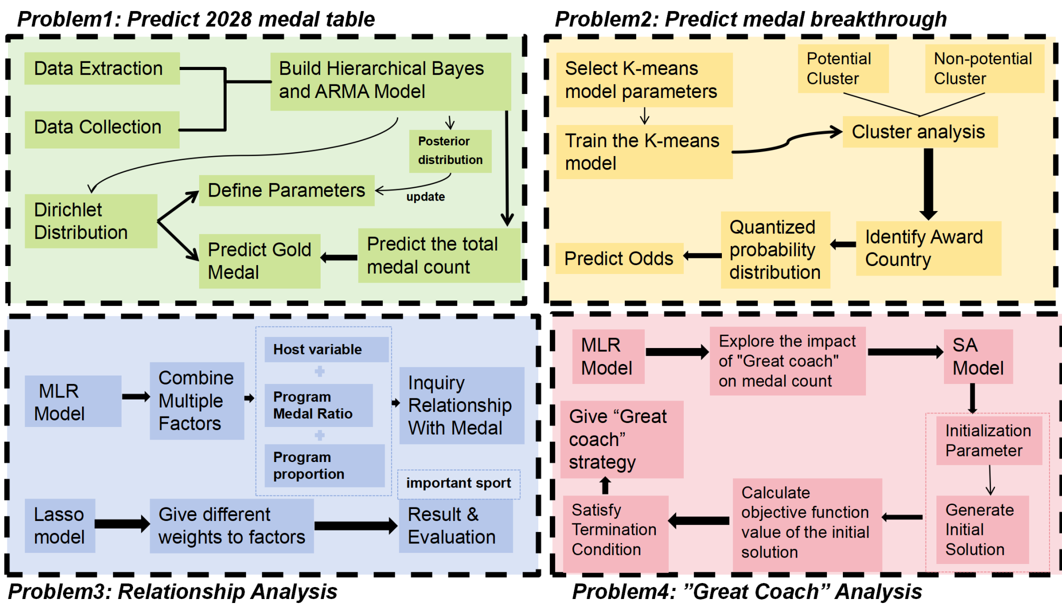

Figure 1: Our work

2 Assumption and Notations

2.1 Assumption

To simplify the problem, we make the following basic assumptions:

- Assumed that the performance of national or regional delegations in the Olympic Games has a certain degree of stability over time.
- All countries participating in the Paris 2024 Olympic Games will normally participate in the Los Angeles 2028 Olympic Games without withdrawing or refusing to participate for other reasons.
- Anomalies in historical data can be reduced by appropriate data processing to reduce their impact on predictive models.
- Assumed that athletes from each country compete in good faith, without the use of drugs (such as doping) and other cheating cases.
- Assumed that there is no advantage of the random component inherent in the average athlete's game.

2.2 Notations

Symbol	Description
t	Time series effect
β_0	Intercept
β_1	Host effect coefficient
γ_j	The impact coefficient of each project
λ_i	the total medal count
$\log(\lambda_i)$	The logarithm of the expected number of medals
$\text{ProjectAbility}_{ij}$	$\text{AthleteMedalRatio}_{ij} \times \text{ProjectPercent}_j$
α_i	Time trend for each country (AR item).
x_1	the host country
x_j	the proportion of medals in an event
z_j	item event ratio
$x_j \cdot z_j$	Interaction between medal ratio and event ratio
y_t	quantified medal count in $t$year
β_0	the intercept term of the model
ϵ_t	error term
R^2	the coefficient of determination
MAE	the Mean Absolute Error

3 Data Preprocessing

3.1 Data Processing

Based on the observation of the data given, we process it accordingly according to the following steps:

Step 1: In the stage of data cleaning, we use Python to check for missing values, outliers, and duplicate values. We measured null and outlier values for the multiple provided files and found missing values in the *summerOly_programs.csv* file. In addition, we have noticed a lack of data from the 1946, 1940 and 1944 Olympic Games. After consulting the online information, we learned that the 1906 Olympic Games were not recognized by the International Olympic Committee(IOC), and the 1940 and 1944 were during the Second World War, so we treated the relevant data as missing values. We fill all missing values with 0 to indicate thatnot included in the Olympic Games in that year.

Step 2: To make the tabular data easier to process by the model, we adjusted the horizontal and vertical coordinates of the *summerOly_programs.csv* file. The ordinate is changed to the year, and the abscissa is denoted by the code corresponding to the sport. In addition, we separately calculate the number of sports, discipline and events of each Olympic Games.

Step 3: To consider the influence of the host effect on the competition results, we created a new form file based on the data given. In this document, we use the horizontal coordinate to represent the year and the vertical coordinate to represent the participating country. For the country that is the host of the Games at that time, we mark it 1, and for the others 0.

3.2 Data profile

In order to carry out better analysis and research, we have studied the data provided were processed. Once we have the cleaned data, we use Excel and Python tools to integrate and visualize the relevant data.

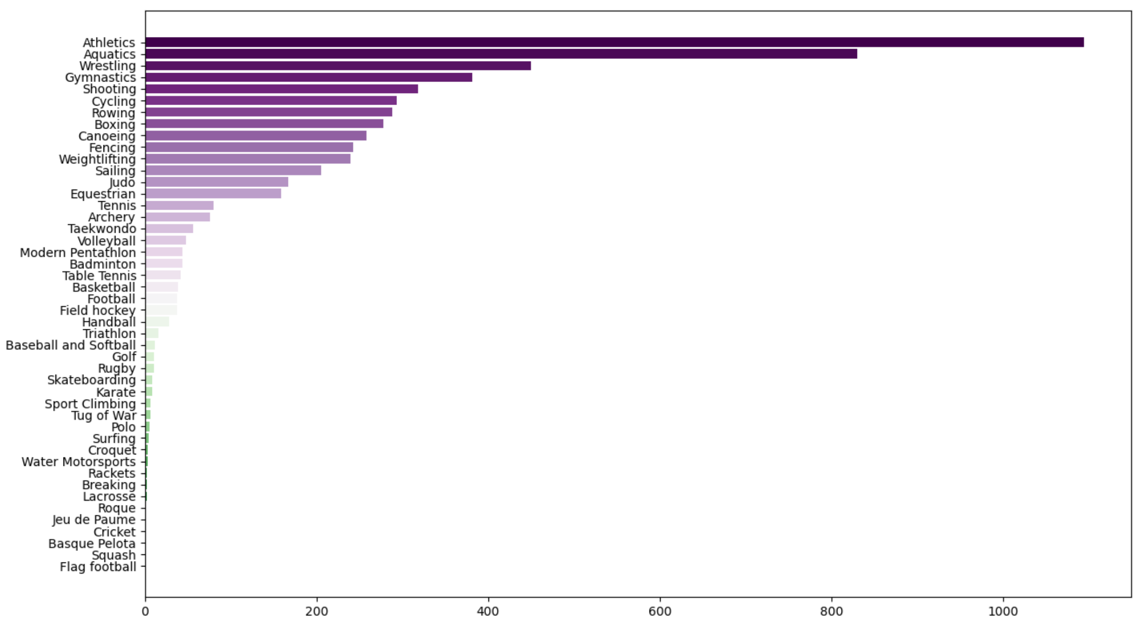

Figure 2: The number of Medals won by different sports

4 Hierarchical Bayesian Composite Model

The number of Olympic medals has a multi-layered and complex structure, including different sports, different countries and different medal types. Through the medal data provided by the research institute, in order to make a more accurate prediction, we choose to use a Hierarchical Bayesian model here. The model is able to handle this hierarchy in a natural way and capture dependencies by assigning parameters at different levels[1].

At the same time, the number of Olympic medals also has the trend of time change. We then added AR and MA models to the Hierarchical Bayesian model. In addition, our data also includes three different medal types: gold, silver and bronze. In order to better predict the medal distribution, we add the Dirichlet distribution into the model, so as to help us predict the medal distribution of 2028 Los Angeles Olympic Games in a more comprehensive and accurate way.

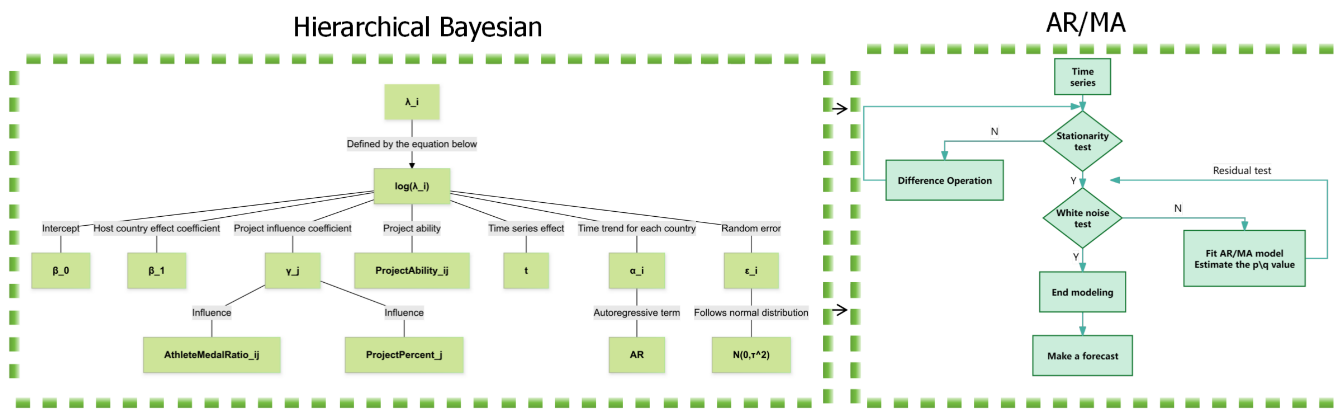

Figure 3: The algorithmic flow of Hierarchical Bayesian Composite Model

4.1 Description of Hierarchical Bayesian Algorithm

4.1.1 Hierarchical Bayesian model structure

We first estimate the total number of gold medals, and then we model with negative binomial distribution, including multiple linear regression estimate λ and Bayes formula +MCMC estimate ϕ. After that, we use the Dirichlet distribution to estimate the gold, silver and copper ratio, and finally multiply the estimated total number of gold medals by the ratio to get the medal distribution

1. **Target variable**y_i: The country's total number of medals in future Olympic Games.
2. **Independent variable:**
   - **National ability** : Use the interactive item of the medal proportion of each country in each event $\times$The percentage of the event to represent the corresponding ability of each country.
   - **Host effect** : Create a binary variable (1 for the host, 0 for the non-host).
   - **Time series** : Use autoregressive (AR) and moving average (MA) models to capture time series trends.

4.1.2 Hierarchical Bayesian model formula

In this section, because the number of medals as a count data, there is a possibility of excessive dispersion, so we use a negative binomial distribution to simulate the total number of medals. The number of medals $y_i$follows a negative binomial distribution, expressed as:

$$y_i \sim \text{NegativeBinomial}(\lambda_i, \phi)$$

The likelihood function of the negative binomial distribution is:

$$p(y|\lambda, \phi) = \frac{\Gamma(y+\phi)}{y!\Gamma(\phi)} \left(\frac{\lambda}{\lambda+\phi}\right)^y \left(\frac{\phi}{\lambda+\phi}\right)^\phi \tag{1}$$

- **Using MLS and ARMA to estimate λ_i**

  Firstly, we use MLS to estimate λ_i. Instead of selecting λ_i directly, we use $\log(\lambda_i)$. This is because we consider that logarithms compress the discreteness of the existing data, bringing it closer to a normal distribution. Besides, logarithmic transformation can better stabilize the variance, thereby improving the stability of the regression model and enhancing its prediction performance.

  The update equations of Hierarchical Bayesian are as follows:

$$\log(\lambda_i) = \beta_0 + \beta_1 \cdot \text{Host}_i + \sum_{j=1}^{M} \gamma_j \cdot \text{ProjectAbility}_{ij} + \alpha_i \cdot t + \epsilon_i \tag{2}$$

  Where:

  $\alpha_i \cdot t + \epsilon_i \sim \mathcal{N}(0, \tau^2)$ denotes random error, same as in the ordinary regression residual term.

  Secondly, we further build time series model(AR and MA) to have a better estimate of the residual term ϵ_i by adding time-related information.

  1. **Autoregressive (AR) model:**

     Suppose that the number of medals won by the country in this Olympic Games depends on its performance in the previous Olympic Games:

$$y_{i,t} = \alpha_i \cdot y_{i,t-1} + \epsilon_{it} \tag{3}$$

     Where $\alpha_i$is the autoregressive coefficient.

  2. **Moving Average (MA) model:**

     Suppose that the medal tally of the past few Olympic Games has an effect on the performance of this Olympic Games:

$$y_{i,t} = \frac{1}{3} \sum_{k=t-3}^{t-1} y_{i,k} + \epsilon_{it} \tag{4}$$

3. **Combined model of Autoregression and Moving Average:**
   The combination of autoregressive and moving average can be used to smooth out fluctuations in medal counts and capture long-term time trends in medal counts.
   The equation of ARMA model[2] used to describe the error term ϵ_t is as follows:

$$\epsilon_t = \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \cdots + \phi_p\epsilon_{t-p} + \theta_1\eta_{t-1} + \theta_2\eta_{t-2} + \cdots + \theta_q\eta_{t-q} \tag{5}$$

By substituting the error part (ϵ_t)of the ARMA model into the multi-level Bayesian model, we get the synthesized formula as follows:

$$\begin{aligned}\log(\lambda_i) =& \beta_0 + \beta_1 \cdot \text{Host}_i + \sum_{j=1}^{M} \gamma_j \cdot \text{ProjectAbility}_{ij} + \alpha_i t + (\phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \phi_3\epsilon_{t-3} + \cdots \\ & + \phi_p\epsilon_{t-p} + \theta_1\eta_{t-1} + \theta_2\eta_{t-2} + \cdots + \theta_q\eta_{t-q})\end{aligned} \tag{6}$$

- **Using Hierarchical Bayesian and MCMC to estimate ϕ**

  According to Bayes' theorem, we have:

$$P(\phi|Y) \sim P(Y|\phi) \cdot P(\phi)$$

  Where:

  - $P(\phi|Y)$ is a posterior distribution.
  - $P(Y|\phi)$ is the likelihood function.
  - $P(\phi)$ is the prior distribution.

  In addition, we assume that ϕ follows the inverse gamma distribution[3].

$$\phi \sim \text{InverseGamma}(\alpha, \beta)$$

  The probability density function of the inverse gamma distribution is:

$$p(\phi^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot (\phi^2)^{-(\alpha+1)} \cdot \exp\left(-\frac{\beta}{\phi^2}\right) \tag{7}$$

  Then, we use the Metropolis-Hastings algorithm for sampling. The basic idea of the Metropolis-Hastings algorithm is to sample from a complex posterior distribution by constructing a candidate distribution.

  Its formula for calculating the acceptance rate α for the given current sample ϕ^2_{old} and candidate sample ϕ^2_{new} is:

$$\alpha = \min\left(1, \frac{p(y|\phi^2_{\text{new}}) \cdot p(\phi^2_{\text{new}})}{p(y|\phi^2_{\text{old}}) \cdot p(\phi^2_{\text{old}})}\right) \tag{8}$$

  Where, $p(y|\phi^2)$ denotes the likelihood function. According to the calculation, we get a sample acceptance of 30%.

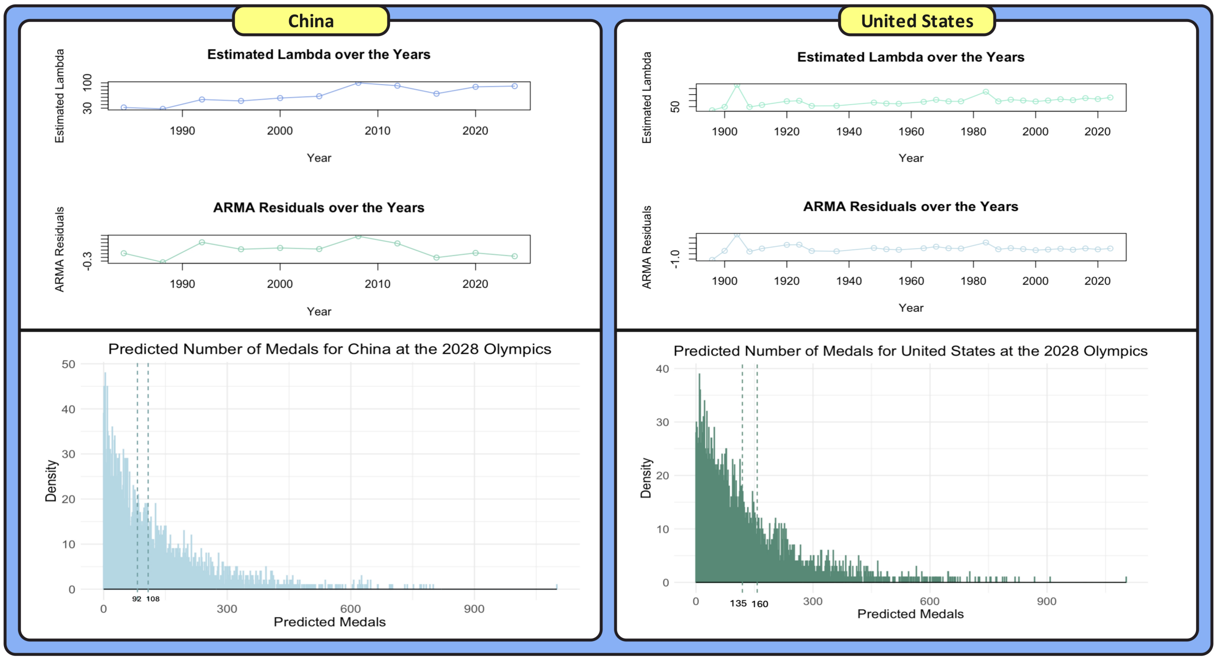

Figure 4: Predicted Result

4.2 Dirichlet Distribution

4.2.1 Description of Dirichlet distribution[4]

1. **Modeling based on the scale of existing medals:** For each country i, we modeled the proportion of gold, silver and bronze medals it won in different events:

$$p_i = \text{Dirichlet}(\alpha_i)$$

Where $p_i = (p_{i1}, p_{i2}, p_{i3})$ represents the proportion of gold, silver and bronze medals won by a country's $i$in different events, and $\alpha_i$is used as the hyperparameter of Dirichlet distribution to reflect the contribution of different sports to the total number of medals of the country.

2. **The number of medals in each event:**Once the percentage and total number of medals for each country is obtained, we calculate the specific number of medals for each country in the gold, silver and bronze events using the following formula:

$$y_{ij} = p_{ij} \cdot y_i$$

Where y_{ij} represents the number of medals won by the country i in the event j (gold, silver or bronze), p_{ij} is the proportion of medals obtained by the Dirichlet distribution, and y_i represents the total number of medals won by the country i.

4.2.2 Prior Distribution

- **Dirichlet distribution hyperparameter Settings** : $\alpha_{ij} \sim \text{Gamma}(1, 1)$, representing a prior distribution of the proportion of medals for each country on the item j.

- **Time effect** : For autoregressive coefficients and moving average coefficients, assume that their priors are normally distributed:

$$\alpha_i \sim \mathcal{N}(0, 10), \quad \beta \sim \mathcal{N}(0, 1)$$

- **Discrete parameter of negative binomial distribution** : $\phi \sim \text{Gamma}(1, 1)$.

4.3 Prediction on 2028 Los Angeles, USA summer Olympics

Here are the projected medal and gold medals for some countries at the 2028 Los Angeles Games:

Country name	Total number of medals	Gold medals
USA	128	42
CHN	109	47
FRA	37	10
GBR	52	14
AUS	53	18
JPN	52	23

Table 1: Some countries predict the number of Olympic medals

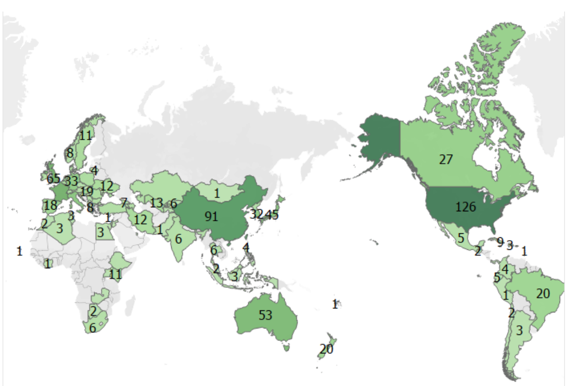

Figure 5: Distribution of medals in 2024

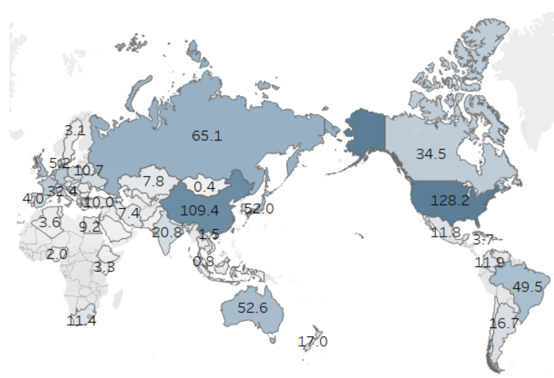

Figure 6: Distribution of medals in 2028

Based on our calculations and comparisons with 2024, we can see a noticeable increase in countries like the United States, China, and Japan. Countries such as Italy and South Korea are expected to experience a decline in the number of medals in 2028.

4.4 Evaluation of Prediction Model

4.4.1 Posterior Inference

We used MCMC (Markov Chain Monte Carlo) method to sample from the posterior distribution to obtain the medal distribution for each country. We then use these posterior samples to calculate a 95% prediction interval for future medals.

Based on the posterior distribution, we get the predicted range of medals for each country in the future Olympics as follows:

4.4.2 Evaluation Model

By comparing the predicted medal count with the actual medal count, we calculate the mean absolute error (MAE) as the model evaluation index. And the calculation method for MAE is as follows:

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{9}$$

	Mean	sd	n_eff	Rhat
α	1.00	0.02	2739	1
lp__	9408.47	0.69	3831	1

Table 2: Evaluation of the Model

- α: α has an estimate of 1 and a standard deviation of 0.02, indicating that its performance is very stable across all chains. Second, Rhat = 1 indicates that the model meets the convergence criteria. n_eff = 2913 indicates that the effective sample size of the model is very large, and the model is robust and predictable.
- **lp_:** The estimated value of $lp\_\_$ is 9408.46, indicating that the model has a good fit and sample convergence. n_eff = 3637, indicating that the effective sample size of the logarithmic posterior probability is also large enough, reflecting that the model has good stability. The convergence of the model is also represented by Rhat = 1.
- **Accuracy evaluation:** When the parameter α **is set to 30**, the mean absolute error **(MAE) obtained is 3.1592**. This indicates the accuracy of the hierarchical Bayesian composite model established by us is high.

5 K-Means++ Model

In the upcoming Olympic Games, it is expected that several countries will win their first medals. To estimate this prediction, we used the K-means clustering method to predict which countries, that have never won a medal before, are most likely to win their first medal.

5.1 K-means Clustering Steps

The K-means clustering algorithm follows these steps:

1. **Choosing the K value**: We determine how many clusters the data should be divided into, typically using methods like the elbow rule.

2. **Calculating distances**: For each country, we calculate its position in the feature space and assign it to the nearest cluster based on its distance from the cluster center.

3. **Updating the cluster centers**: We recalculate the cluster center based on the data points in the cluster

4. **Iterative optimization**: Repeat steps 2 and 3 until the cluster centers converge or change very little.

The mathematical formula for calculating the distance between a data point and the cluster center is as follows:

$$d(x_i, \mu_k) = \sqrt{\sum_{j=1}^{n} (x_{ij} - \mu_{kj})^2}$$

where:

- x_i is the data point,
- μ_k is the cluster center,
- n is the number of features in the data.

Through this process, K-means can group countries based on their historical medal data and other relevant factors, helping to identify those with significant medal-winning potential in future Olympic Games.

5.2 Predict the odds of winning a first time medal

Based on the K-means clustering results and model analysis, we predict the following seven countries are most likely to win their first medal in the next Olympic Games:

Albania	Cape Verde	Bahamas	Ecuador	Dominican Republic	Lebanon	Guinea

These predictions are based on the countries' historical medal potential and the number and types of events in the upcoming Olympic Games.

5.3 Credibility Score and Prediction Probability

- **Credibility Score:** Each country obtains a Credibility Score through analysis of clustering and capability data, with higher scores indicating a greater likelihood of winning medals. We can base predictions for first-time medals on the Credibility Score.

- **Probability Calculation:** For countries with higher Credibility Scores, we can define a threshold based on this score.

- **Probability Distribution:** Through the probability distribution, we can quantify these predictions. For example:

$$P(\text{Medal}) = \frac{\text{Credibility Score}}{\text{Maximum Credibility Score}}$$

where $P(\text{Medal})$ represents the probability of a country winning its first medal, and the Credibility Score is normalized by the highest Credibility Score.

The prediction probabilities for each country's likelihood of winning its first medal in the next Olympic Games are presented in the table below:

Country	Probability of Winning First Medal
Albania	0.764
Cape Verde	0.762
Bahamas	0.758
Ecuador	0.717
Dominican Republic	0.717
Lebanon	0.701
Guinea	0.666

Table 3: Probability of Winning First Medal in the Next Olympic Games

These probabilities reflect the likelihood of each country winning its first medal in the upcoming Olympics and provide actionable insights for national Olympic committees.

6 Consideration of Olympic Games Events

6.1 Description of MLR Algorithm

The number of medals is affected by many factors, and the multiple linear regression model can consider many factors at the same time to predict the number of medals. Therefore, we chose a multiple linear regression (MLR) algorithm to investigate the relationship between the event and the number of cards awarded in each country, and the impact of the event selected by the host country.

1. **Target variable**y_i:

   Let y_i be the total number of medals (including gold, silver and bronze) won by the i country in the t Olympic Games.

$$y_i = \text{Total medals for Country } i$$

2. **Independent variable:**

- **Host variable:** Consider x_1 as a binary variable that indicates whether the country is the host of the Olympics (1 means yes, 0 means no).

$$x_1 = \begin{cases} 1 & \text{if country } i \text{ is the host country} \\ 0 & \text{otherwise} \end{cases}$$

- **Program Medal ratio:** Let x_j be the ratio between the number of medals won by the country $i$in a certain event (such as track and field, tennis, etc.) and the total number of medals won in the event in the t Olympic Games. The number of items ranges from $j = 2$ to n.

$$x_j = \frac{\text{medals of country } i \text{ in project } j}{\text{Total medals in project } j}$$

  where n represents the total number of all items.

- **Program proportion:**Set z_j to represent the proportion of project j in all events in the first t Olympic Games, that is, the proportion of the total medal number of project $j$to the total medal number of all events.

$$z_j = \frac{\text{Total medals in project } j}{\text{Total medals in all projects in year } t}$$

- **Interaction term:**The interaction item $x_j \times z_j$is the interaction effect between the performance of the country $i$in the event j and the importance of the event in the Olympic Games, which is used to analyze the influence of different programs on the medal performance of different countries.

$$x_j \times z_j = \frac{\text{medals of country } i \text{ in project } j}{\text{Total medals in project } j} \times \frac{\text{Total medals in project } j}{\text{Total medals in all projects in year } t}$$

3. **Model representation:** The MLR model can be expressed as the following:

   MLR For Estimating Coach Effect:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_i \tag{10}$$

   where x_2 is program medal ratio and x_3 is program proportion as described before

   MLR For Deciding Which Programs Are Important to Which Country

$$y_i = \beta_0 + \beta_1 x_1 + \sum_{j=2}^{n} \beta_j x_j + \sum_{j=2}^{n} \beta_j z_j + \sum_{j=2}^{n} \beta_{j+1} (x_j \times z_j) + \epsilon_i \tag{11}$$

   Where:

Symbol	Description
β_0	constant term
β_1	the coefficient the host country variable
β_j	the coefficient of the medal proportion of the event
β_{j+1}	the coefficient of the item's proportion
β_{j+n}	the coefficient of the interaction term

Table 4: Notations used in Equation12

4. **Model assumption:**To simplify the model, we assume that the error term $\epsilon_i$is independently homodistributed and normally distributed, and that there is no strong multicollinearity between the independent variables of the model.

$$\epsilon_i \sim N(0, \sigma^2)$$

5. **Model fitting and evaluation criteria:**

- **Fitting**
  Next, we ordinary least squares (OLS) or Maximum likelihood estimation (MLE) to fit the model. The regression coefficients $\beta_0, \beta_1, \ldots, \beta_{n+2}$ are estimated by calculating the sum of squares of the minimization error terms.
  The least squares method is used to minimize the following objective functions:

$$\min_{\beta_0,\beta_1,\ldots,\beta_{n+2}} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

  Where $\hat{y}_i$ represents the total number of medals predicted based on the estimated coefficient.
- **Evaluation criteria**
  (a) **Significance of the coefficients:** We tested the significance of each regression coefficient by calculating their p-value.
  (b) **Model fit:** We then used the coefficient of determination R^2 to measure how well the model fits existing data. R^2 is calculated as follows:

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

  Where $\bar{y}$ represents the mean of the total number of medals.

R^2	Intercept	x1	x_2	x_3	$x_2 \cdot x_3$	p_value
0.715	19.7434	47.8638	240.6733	-29.6414	32.8129	<0.05

Table 5: Results of MLR analysis

6.2 MLR Model Predict Result

By using the MLR prediction model, we obtain the impact of the proportion of medals, the proportion of events and their interaction on the total number of medals. The results are as follows:

From the results, we can see that in our regression analysis, R^2 is equal to 0.715. This shows that our model can explain 70% of the total variation of the dependent variable, which means that the model can fit the data well.

In our model, the p-values of all independent variables and interaction terms are less than 0.05. In general, when the P-value < 0.05, we can reject the null hypothesis and think that the variable has a significant effect on the dependent variable. Therefore, we can conclude that the variables selected in our model have a significant impact on the dependent variable medal count.

6.2.1 Relationship between the event and the number of medals won by each country

According to the data in Table 5, we can make the following conclusions:

The medal ratio (x_j) coefficient is **positively correlated** with the total number of medals. The event ratio (z_j) coefficient is **negatively correlated** with the total number of medals. The coefficient of the interaction term ($x_j \cdot z_j$) is **positively correlated** with the total number of medals, which indicates that these two factors will interact together to determine the performance of a country in the Olympic Games.

6.2.2 Relationship between the host country's selection and the outcome of the competition

In order to explore the influence of host country on the number of medals, we use logistic regression with categorical variable–host country as the response variable and included the number of medals and above mentioned variables as independent variable in the model

According to the result of it, we get the result that total number of medals (x_1) coefficient is 47.86, which is **positively correlated** with a country being a host country. Since the total number of medals is significant to predict whether a country is a host country, this shows that when a country becomes the host of the Olympics, it usually wins more medals.

The results are also shown in figures as follows:

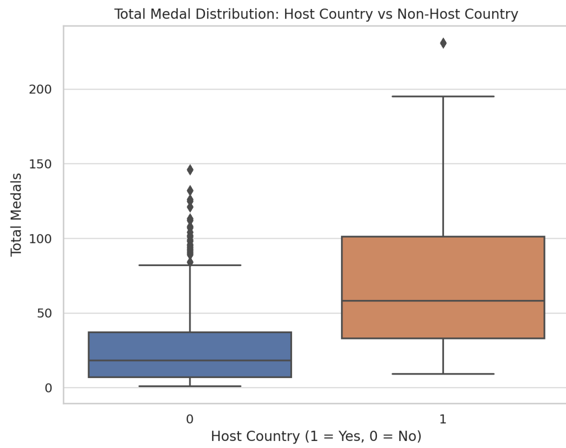

Figure 7: Relationship Between Hosts and medals

6.3 Description of Lasso Algorithm

Lasso is used to process data with a large number of potential predictors and can effectively narrow down less important variables to zero by penalizing their coefficients[5]. In this model, we use cross-validation (10x) to find the minimum mean MSE to fix the best penalty Linear-Regression-Lasso model, and then explain the important variables in it. By using the Lasso regression fitting model, we were able to identify the most important sports for different countries.

The original formula is:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = RSS+\lambda\sum_{j=1}^{p}|\beta_j|$$

6.3.1 Lasso Regression Fitting Model Building

First, we use the `glmnet` function for Lasso regression[5], while using $\alpha = 1$ to indicate the use of Lasso regularization. Next, we select the best λ value by cross-checking to avoid overfitting or underfitting the model. Then, by extracting the regression coefficient, especially the non-zero coefficient. We ranked the results in absolute terms, keeping the top 8 coefficients for each country, with the higher the coefficient indicating the more important the sport is to the country.

The Lasso regression formula we use is as follows:

$$\hat{\beta} = \text{argmin}\left(\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_1-\sum_{j=2}^{n}\beta_j x_j-\sum_{j=2}^{n}\beta_j z_j-\sum_{j=2}^{n}\beta_{j+n}(x_j\times z_j)\right)^2+\lambda\sum_{j=1}^{n}|\beta_j|\right) \tag{12}$$

Where:

- Signs have been described before.
- ϵ_i denotes the error term.

6.3.2 Lasso Regression Model Result

Here are our results and their coefficients for different national strengths:

NOC	Sport	Coefficient
USA	Athletics	63.185192606
CHN	Diving	1.493975932
FRA	Sailing	59.15087518
AUS	Swimming	8.893794303
JPN	Gymnastics	1.132075476

Table 6: Coefficient of some NOCs important sports

7 Great Coach Analysis

In this section, we analyzed the historical results of the United States, China, and Romania in volleyball and gymnastics. We then examined the medal counts during Coach Bela Karolyi's and Coach Lang Ping's tenures with Team USA, Romania, and China. The following figures show the medal counts during their respective tenures, highlighting robustness and predictability.

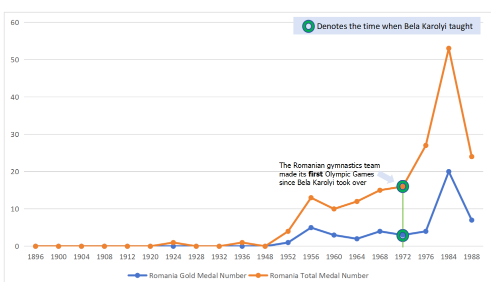

Figure 8: Number of medals won by the Romanian team

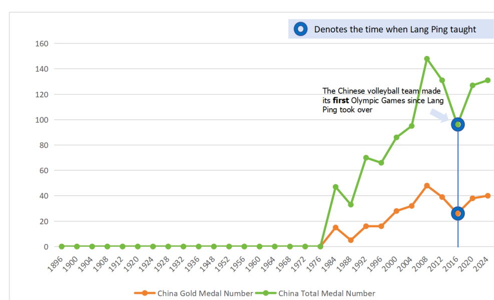

Figure 9: Number of medals won by the China team

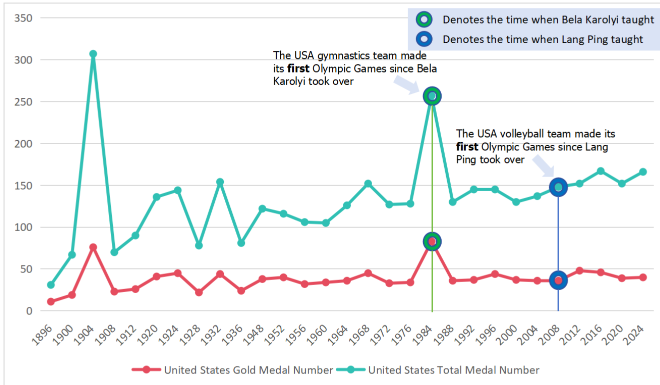

Figure 10: Number of medals won by the United States team

After that, we still use the Multiple Linear regression model (MLR) model to help us quantitatively analyze the relevant data.

7.1 Description of MLR Algorithm

1. **Target variable y_t:**

   y_t indicates the number of quantified medals in the year t. We give a combined medal score for the different medal types (gold, silver, bronze) given the weight (gold is 3, silver is 2, bronze is 1).

2. **Independent variable:**

   - **Coach:** Consider β_1 as the coefficient of the Coach variable to indicate the degree of influence of the name effect on the number of quantified medals.

   $$\beta_1 = \begin{cases} 1 & \text{if event } i \text{ has a coach} \\ 0 & \text{otherwise} \end{cases}$$

   When $\beta_1$is positive, it means that a coach will increase the number of quantitative medals. When β_1 is negative, it means that the influence of famous coach is negative, which may lead to a reduction in the number of medals.

   - **HostCountry:** Consider β_2 as the coefficient of HostCountry variable to represent the influence of host country effect on quantified medal count.

   $$\beta_2 = \begin{cases} 1 & \text{if country } i \text{is the host country} \\ 0 & \text{otherwise} \end{cases}$$

   When β_2 is positive, it means that becoming the host country will increase the number of quantified medals. When β_2 is negative, it indicates that the host country effect brings about a decline in medals.

- **ProjectRatio:** Consider β_3 as the coefficient of the ProjectRatio variable, representing the proportion of medals won by a specific event. When β_3 is positive, it means that a higher proportion of medals in a sport leads to an increase in the total number of medals for that sport.

The MLR model can be represented as the following formula:

$$y_t = \beta_0 + \beta_1 \cdot \text{Coach}_t + \beta_2 \cdot \text{HostCountry}_t + \beta_3 \cdot \text{ProjectRatio}_t + \epsilon_t \tag{13}$$

7.2 MLR Model Result

Based on the analysis of Tables 7, 8, 9, and 10, we can conclude that the "great coach" effect significantly positively impacts the medal performance in both gymnastics and volleyball teams.

For the USA Gymnastics team, the positive and substantial coefficient β_1 indicates a strong influence of renowned coaches on the number of medals won. Similarly, for the Romanian gymnastics team, the coefficient of 5.011 underscores the significant positive contribution of famous coaches to medal outcomes.

In volleyball, the coefficient of 2.361 for the US women's volleyball team and 3.052 for the Chinese women's volleyball team both suggest a substantial positive impact of famous coaches on medal performance, further supporting the significance of this effect in enhancing team success.

R^2	β_1	β_2	β_3	p_value
0.745	0.334	0.116	1.425	<0.05

Table 7: Results of US Gymnastics

R^2	β_1	β_2	β_3	p_value
0.813	5.011	0.285	3.455	<0.05

Table 8: Results of Romania Gymnastics

R^2	β_1	β_2	β_3	p_value
0.703	2.361	1.028	4.897	<0.05

Table 9: Results of US volleyball

R^2	β_1	β_2	β_3	p_value
0.813	3.052	0.613	2.354	<0.05

Table 10: Results of China volleyball

Evaluation of MLR Result

It can be seen from the results that R^2 is both positive and large in our regression analysis. This shows that our model can fit the data well.

In addition, in our model, the p-values of all independent variables are less than 0.05. Therefore, we can reject the null hypothesis and think that the "great coach" effect has a significant impact on the number of medals and contributes a lot.

7.3 Description of SA Algorithm

In order to find out which projects in various countries need to hire "great coach" to help improve their performance, we established a Simulated Annealing(SA) algorithm to solve it.

Here are the process we built SA model:

Step 1: We set up a formula to calculate the correlation. Based on the independent variables we presented in 7.1 section, we multiply them by different weights. For the quantified ability coefficient, we multiply it by the correlation between the medals won in two Olympic Games. For hosts or not, we multiplied their correlation by the difference in the number of medals between each of the last ten Olympic Games. For the weight of whether to use the coach, we are constructed by 1 minus the weight of the above two coefficients.

$$\sum_{i=1}^{n}(\text{Ability}_i \cdot w_i + \text{HostCountry}_i \cdot w_2 + (1 - w_1 - w_2) \cdot x_i)$$

Step 2: Create a dynamic penalty function[6] to control no more than 5 sports that can be coached.

$$\alpha \cdot \max\left(0, \sum_{j=1}^{n} x_i - 5\right)$$

Step 3: Construct the objective function f(x) by subtracting the penalty function from the correlation function

$$\sum_{i=1}^{n}(\text{Ability}_i \cdot w_i + \text{HostCountry}_i \cdot w_2 + (1 - w_1 - w_2) \cdot x_i) - \alpha \cdot \max\left(0, \sum_{j=1}^{n} x_i - 5\right) \quad (14)$$

Where, f(x) denotes the medal improvement.

7.4 SA Model Result

In this study, we used the simulated annealing model to analyze three countries— the United States, Japan, and Belarus. The results are as follows:

Program	**United States of America**	**Japan**	**Republic of Belarus**
Program 1	Ice Hockey	Baseball	Road Cycling
Program 2	Roque	Baseball and Softball	Modern Pentathlon
Program 3	Football	Gymnastics	Table Tennis
Increment of medals	8	4	2

Table 11: Selection of Programs and Increment of medals for Different Countries

The medal increment was calculated using the simulated annealing model. Results show that the "celebrity coach effect" significantly boosts medals. Specifically, the United

States sees the largest increase, especially in ice hockey and football, while Japan's increment is 4, with a smaller effect due to its strong foundation in sports like gymnastics. Belarus has a modest increase of 2, as the coach effect is limited by a weaker foundation in these sports.

8 Insights and Information Support

8.1 Insights

Here are some interesting insights based on our models:

- By analyzing the data of previous Olympic Games with multi-level Bayesian composite model, we can identify the correlation between the distribution of Olympic medals and the development of national industrialization level. A country with a higher level of industry will invest more resources in developing sports and promoting sports for all, thus helping the country to achieve better results in the Olympic Games. From the analysis, it can be seen that the number of medals won by the United States and China are more and still showing an upward trend. The United States is a traditional industrialized country, and China, as a developing country, is also improving its industrialization level.
- We use MLR model to analyze the performance of host countries in hosting the Olympic Games, from which we find that host countries generally win more Olympic medals than before. This shows how important home advantage is for the Olympics. In addition, the host country may host more events related to the country's sporting expertise to help it win more medals.
- We used the multiple linear regression model to analyze the relationship between great coaches and the number of medals. We find that great coaches generally have a positive impact on the sport. This means that we can look at whether a coach is a great coach by analyzing whether he has a positive or negative impact on his team's medals over the course of his coaching cycle. For a country that wants to develop the sport, this kind of analysis can be used to consider bringing in this great coach to help it break new ground in the Olympics.

8.2 Useful Information Support

Based on these insights, I think we can provide the following useful information for country Olympic Committees:

- **Invest in countries with development prospects:** For countries with less developed industrialization, the Olympic Committee can provide funds to support the development of sports in that country, so that the Olympic Games can truly become a world sports event.
- **Optimize the strategic allocation of sports resources:** Through the analysis and identification of "great coach", the Olympic Committee can better allocate sports

resources in various countries. By encouraging great coaches to coach in different countries, the Olympic Committee can help identify more "sports talents" and better promote the common development of the sport in multiple countries.

- **Develop a long-term development plan:** Through long-term planning for the host country, the IOC can help the country introduce more sports resources, develop the country's sports industry, and improve its Olympic performance.

9 Model Sensitivity Analysis

For the sensitivity analysis of the Hierarchical Bayesian Composite Model, we first delete random five program terms to formula (2). Secondly, we revise the initial assumption of the Inverse Gamma distribution for the variance. The plotted figure shows an example for Japan (The plots are interpolated for better visualization of the differences).

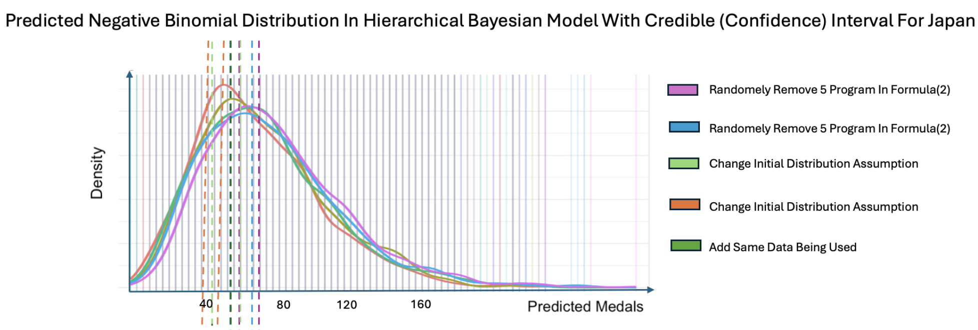

Figure 11: Sensitivity of Hierarchical Bayesian Composite Model

The visual results show some variations, but they remain relatively stable, as the credible (confidence) intervals largely overlap.

10 Model Evaluation and Further Discussion

10.1 Strengths

- **Hierarchical Bayesian Algorithm:** The hierarchical Bayesian algorithm in Section 4 has several advantages when used for prediction tasks. Hierarchical Bayesian models can effectively process data with multiple layers and complex structures, such as Olympic medal data. The data includes different sports, different countries and regions, and different types of medals, reflecting a multi-layered and complex structure. In contrast, hierarchical Bayesian models can not only handle this complex hierarchy naturally than traditional Bayesian models, but also capture the dependencies between data by assigning parameters at different levels.

- **AR and MA:**The number of Olympic medals has the characteristics of time trend, and AR and MA models can effectively capture the dynamic characteristics of the number of medals over time.

- **MLR:**MLR algorithm can consider multiple factors at the same time, which is helpful to analyze the multiple influencing factors of the number of medals, and more accurately predict the impact of different factors on the number of medals.
- **Lasso:**Lasso regression enables automatic variable selection by regularizing terms to continuously reduce the coefficients of unimportant variables, and better identify the sports that contribute the most to each country's medal.

10.2 Weaknesses and Further Discussion

- The Olympic Games are held every four years and not all countries participate in every Olympic Games, which may affect the accuracy and generalization ability of the impact model.
- Lasso regression model is sensitive to outliers and feature relationships. If a dependent variable is too correlated with the number of medals, its results may be affected and errors may occur in the recognition of important features.
- Russia has won a large number of medals in the previous Olympic Games, but due to the impact of the international situation, Russia failed to participate in the 2024 Summer Olympic Games in Paris, so the medals it had the ability to win will be distributed to other countries, which may affect the prediction results of the model.

References

[1] J. N. Rouder and J. Lu, "An introduction to bayesian hierarchical models with an application in the theory of signal detection," *Psychonomic Bulletin & Review*, vol. 12, pp. 573–604, 2005. DOI: 10.3758/BF03196750.

[2] S. J. Koopman, M. A. van der Wel, and A. Lucas, "Multi-step time series analysis and forecasting strategy using arima models," *Journal of Econometrics*, vol. 225, no. 2, pp. 255–277, 2019.

[3] A. Llera and C. F. Beckmann, *Estimating an inverse gamma distribution*, 2016. arXiv: 1605.01019 [stat.ME]. [Online]. Available: https://arxiv.org/abs/1605.01019.

[4] T. T. Wong, "Generalized dirichlet distribution in bayesian analysis," *Applied Mathematics and Computation*, vol. 97, no. 2-3, pp. 165–181, 1998.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[6] A. I. Pereira and E. M. G. P. Fernandes, "Constrained multi-global optimization using a penalty stretched simulated annealing framework," in *AIP Conf. Proc.*, vol. 1168, Melville, New York: American Institute of Physics, Sep. 2009, pp. 1354–1357. DOI: 10.1063/1.3241336.

11 Report onUse of AI

- OpenAI ChatGPT(Feb 26, 2025 version, ChatGPT-4)
  - Query 1: How is the K-Means model specifically used
  - Output: Steps to Use K-Means:Choosing K (Number of Clusters): First, you decide how many clusters (K) you want the data to be grouped into. This can be based on prior knowledge, domain expertise, or methods like the "Elbow Method" to find an optimal K. Initialize Centroids: Randomly select K points in the data as the initial centroids (center points of the clusters). These centroids represent the center of each cluster. Assign Data Points to Clusters: For each data point, calculate the distance between the point and each of the K centroids. Assign the data point to the cluster with the nearest centroid. Recompute Centroids: After assigning all data points to clusters, recompute the centroid of each cluster. This is done by calculating the mean of all the points within the cluster. Repeat: Repeat the assignment of points and recomputing of centroids until the centroids no longer change significantly (i.e., convergence). At this point, the algorithm has completed, and the data points are grouped into clusters.
- OpenAI ChatGPT(Feb 27, 2025 version, ChatGPT-4)
  - Query 2: What is the structure of the Hierarchical Bayesian Composite Model and how is it used?
  - Output: The Hierarchical Bayesian Composite Model (HBCM) combines hierarchical modeling and Bayesian inference to handle data organized at multiple levels. The structure includes: Hierarchical Levels: Data is organized at different levels, such as individual units (e.g., athletes), groups (e.g., teams), and higher-level parameters governing overall trends. Bayesian Inference: Prior distributions are updated using observed data to form posterior distributions, capturing uncertainty about model parameters. Composite Modeling: HBCM integrates different sub-models, including linear, non-linear, and latent variable models, to account for various data sources and uncertainty at each level. How it works: Priors: Represents initial beliefs about parameters before data is observed. Likelihood: Describes the probability of observed data given the model parameters. Updating: Bayesian inference updates the priors with data to form posteriors. Prediction: The model makes predictions using the estimated parameters, incorporating uncertainty. Uncertainty Quantification: The model provides uncertainty estimates by working with posterior distributions. HBCM is useful in contexts like analyzing the impact of coaches on team performance, where data exists at multiple levels (individual and team), and uncertainty needs to be accounted for in the predictions.