Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2524070

From Data to Podiums: A Study for Olympic Medal Forecasting

Summary

This research examines the forecasting of Olympic medal counts by analyzing the competitive outcomes of various nations since the inception of the Olympic Games in 1896. The study addresses five primary concerns:

Predict Medal Standings for the 2028 LA Olympics: To effectively model the number of medals awarded to countries, we propose a **zero-inflated negative binomial regression model (ZINB)**. Adopting a **Bayesian framework**, we employ the **Markov Chain Monte Carlo (MCMC)** method to derive the posterior distribution of the model parameters. Subsequently, we conduct **random simulations** to predict the medal standings for the 2028 Olympics. Comprehensive testing of all parameters yielded **Rhat** values of 1, with both **Bulk Effective Sample Size (ESS)** and **Tail ESS values** reaching ten thousand, thereby demonstrating the model's exceptional predictive stability and precision.

Overcoming the Historical Lack of Medals: We introduced the concepts of **sampling zero and structural zero** within the ZIBN model. By evaluating the probability of generating structural zeros in various countries, we predicted which nations might secure their first medal at the 2028 Los Angeles Olympics. The findings indicate that Mauritania has a **24.8%** chance of achieving this milestone!

Factors Influencing Medal Totals: In this section,we employed an **Association Rule Model** utilizing **GRI algorithm**. Using the United States as a case study,we developed a 'type-num' comprehensive indicator φ_i as the antecedent and the medal count as the consequent, which led to the identification of **five strong association rules**. Furthermore, the model reinforced the finding that "the host's choice contribute to its medal success", resulting in **two strong rules**. The analysis revealed that **'Athletics'** plays a vital role for all nations. The support for these rules was as high as **80.56%**, with confidence levels generally exceeding **80%**, and lift values **all above 1**, confirming the validity of the findings.

Great Coach Effect: To investigate the presence of the "great coach" effect while accounting for the influence of both the coach and the country they represent, we developed a **mixed-effects model**. In this model, the pairing of coach and country is treated as a random block, while the "great coach" factor is considered a fixed two-level factor. Our analysis **confirmed the existence of the "great coach" effect** and provided a point estimate of **3.74**. For the countries we analyzed, hiring a "great coach" increases the likelihood of winning a silver or gold medal to **over 50%**.

Original Insights: Based on the results of the regression parameter estimates and the established association rules, we found that the number of events has a minimal impact on performance and can be considered negligible. Therefore, we recommend **targeted resource allocation**, **prioritization of qualifications based on medal probability**, and **precise talent scouting** to synergistically enhance medal efficiency.

Keywords: Zero-Inflated Negative Binomial Regression; Association Rules; GRI Algorithm; Mixed-effects Model; Prediction of the LA Olympics.

Contents

1 Introduction

1.1 Background

During each Olympic Games, spectators not only follow the thrilling sports competitions but also keep an eye on the national medal standings. Figure 1 illustrates the fluctuations in the total medal counts of seven nations, including the United States and China, over the past five Summer Olympics.

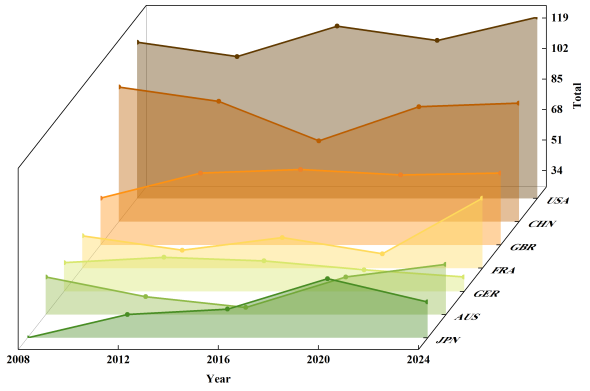


Figure 1: Total Medal Count of the Last 5 Summer Olympics – 7 Countries

For these relatively strong nations, fans are particularly interested in ranking changes, such as whether they have made it into the top 5 or top 3. They also pay attention to variations in the total number of medals, especially gold medals. Conversely, for some countries, securing even a single medal is a noteworthy accomplishment.

Given this context, forecasting the medal counts for each nation in the upcoming Olympic Games is valuable. This article utilizes historical data from the Summer Olympics to predict the number of medals anticipated for the 2028 Games and seeks to identify factors that may influence medal outcomes based on past data.

1.2 Problem Restatement

Through an in-depth analysis of the background, while considering the existing constraints, we can only use the given datasets to build the model. We need to address the following issues:

• Develop a predictive model for the medal count for the 2028 Summer Olympics. This model should estimate the number of gold, silver, and bronze medals for each participating country, as well as the overall medal tally, and should include a prediction interval. Furthermore, an analysis of the results should be conducted to determine whether the performance of each country is comparatively better or worse, with particular emphasis on including predictions for nations that have not previously secured any medals.

• Create a model to investigate the correlation between various sporting events and the medal outcomes of different countries. This model should account for both the quantity of events and

the diversity of disciplines, thereby identifying which sports are significant for multiple nations. Additionally, it should consider the influence of event additions motivated by the host country's objectives in prior Olympic Games on the resulting medal outcomes.

• Formulate a model to evaluate the influence of prominent coaches. This model must ascertain the existence of such an impact exists and, if present, assess its magnitude. Subsequently, three countries should be identified for analysis regarding which sports would benefit from investment in elite coaching, along with an indication of the potential advantages.

• Utilize the model previously developed to uncover additional distinctive insights and communicate these findings to the Olympic Committee.

1.3 Our Work

Framework of our work are as follows.

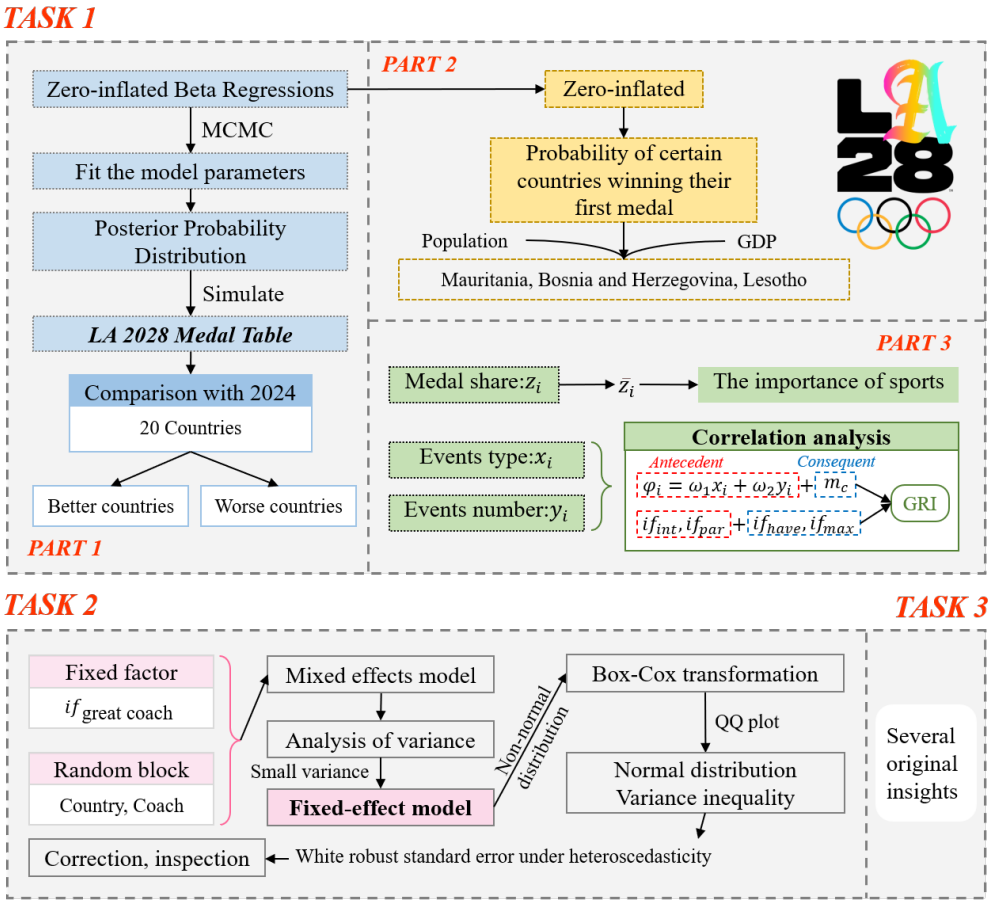


Figure 2: Framework of Our Work.

2 Assumptions and Notations

Concerning the assumptions of the model, we have:

Assumption 1: Countries that have never won a medal at the Olympics are those that participated in the Olympics but have never won any awards in 'summerOly_athletes.csv'.

Assumption 2: The number of medals won by the country follows a negative binomial distribution.

Assumption 3: A country's performance in the next Olympic Games can be predicted based on its performance in previous Olympic Games, whether it is the host country, and the number of events that year.

With respect to the notation, we define the following:

Symbol	Explanation
$n_{gold,i}$	number of gold medals in the ith most recent Olympic Games
$n_{silver,i}$	number of silver medals in the ith most recent Olympic Games
$n_{bronze,i}$	number of bronze medals in the ith most recent Olympic Games
n_{events}	number of events in that Olympic Games
$host$	binary variable: whether that country is the host country
$country$	random effect of different countries
σ_c^2	variance of the random effect $country$
p_0	probability of a structural zero occurring
Y	random variable of the number of medals
ϕ	overdispersion of the negative binomial regression model
α_i	regression coefficients of the zero-inflated model
β_i	regression coefficients of the negative binomial regression model
a_{ijk}	value of the response variable for the k-th replicate in the j-th block at the i-th level.

Table 1: List of Symbols

3 Data Preprocessing

In the analysis of the provided datasets, specifically 'data_dictionary.csv', 'summerOly_athletes.csv', 'summerOly_hosts.csv', 'summerOly_medal_counts.csv', and 'summerOly_programs.csv' (hereafter referred to as dic, $athletes$, $hosts$, $medals$, and $programs$), the following processing steps were undertaken:

1. The $medals$ dataset did not include countries that did not win any medals. Therefore, we retrieved data on countries that participated in the Olympics but did not win any medals from the $athletes$ dataset.

2. Within the $athletes$ dataset, the Team column contained notations such as -1 and -2, and it is noted that early Olympic Games featured teams represented by cities. For the purpose of data screening, only the National Olympic Committee (NOC) was considered, excluding the Team column.

3. Data pertaining to the year 1906, which is classified as an unofficial competition, was entirely removed from the datasets.

4. An analysis of the NOC revealed that certain countries have undergone changes due to historical and political developments. As a result, all data associated with the original countries of the Soviet

Union, Yugoslavia, and other nations that have subsequently fragmented have been removed. Furthermore, data from countries such as East Germany and West Germany, which have ceased to exist due to historical circumstances, have also been excluded.

5. The issue of missing values within the provided datasets was systematically addressed. The *athletes* and *medals* datasets contained no missing values. For the *programs* dataset, if related events were recorded in the *athletes*, the occurrences were counted to supplement the missing data; if no related events were recorded, all occurrences for that particular year were recorded as 0.

6. The treatment of question marks within the *programs* dataset was conducted as follows: for '?0', it was interpreted that the event did not occur during that year for unspecified reasons. For '? non-0', after verification, if the event was not recorded in the athletes data, it was also noted as zero. The '??' designation, which appeared in the athletes table, was determined to refer to informal competitions, such as exhibitions, which do not contribute to medal counts; consequently, data pertaining to these informal competitions was deleted. Furthermore, the items Jeu de Paume and Roque, which appeared with '?' in the Code, were found through research to have been briefly included as official events in the early Olympic Games but are no longer recognized; therefore, this data was also removed.

7. A garbled text issue was identified in the *programs* dataset, where 'Baseball**nd Softball' was recorded, while the *athletes* dataset correctly listed it as 'Baseball/Softball'. Consequently, the *programs* dataset was subsequently amended to align with the format used in the *athletes* dataset.

4 Who Will Rule the Medal Race at the 2028 LA Olympics?

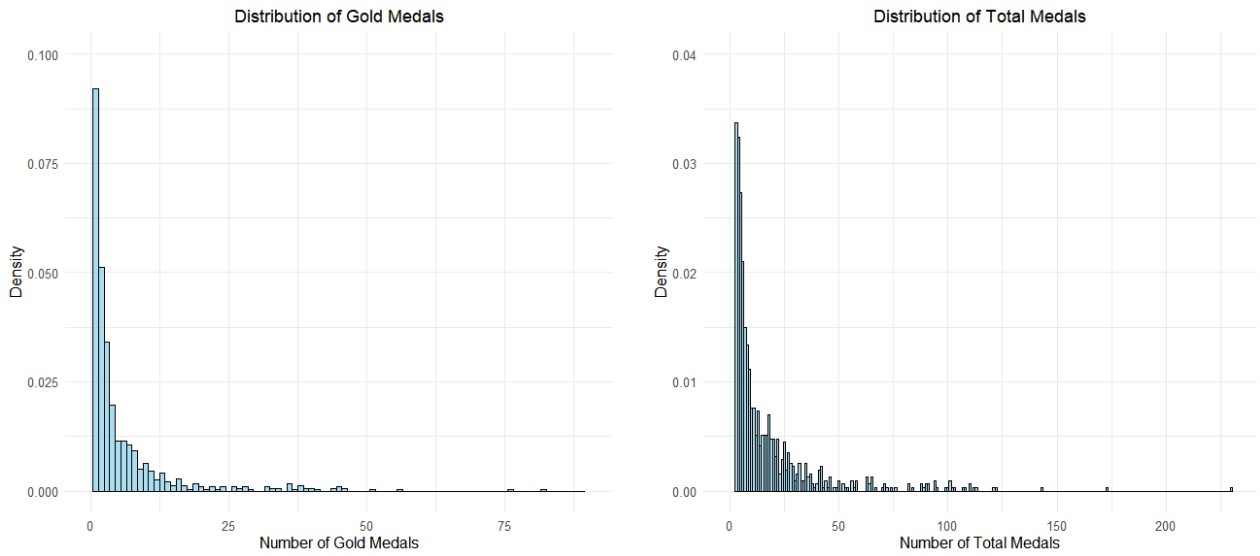
4.1 Count Regression Model

Problem 1 requires us to predict the medal outcomes for various countries at the 2028 Los Angeles Olympics. Since the dependent variable—the number of medals—is a count variable, meaning it can only take on non-negative integer values, we have opted to use a count regression model instead of a general regression or machine learning model.

In counting problems, there are two primary types of regression models: Poisson regression and negative binomial regression. Poisson regression is predicated on the assumption that the data follows a Poisson distribution, which necessitates that the mean and variance are approximately equal. Upon examining this assumption, we found the variances for the number of gold medals and the total number of medals to be 28.57 and 202.05, respectively, while the means were only 1.64 and 5.11. The variance of the total number of medals significantly exceeds the mean, indicating that the data is overdispersed. Consequently, we chose to employ negative binomial regression for our analysis.

Type	Expectation	Variance
Gold Medals	1.64	28.57
Total Medals	5.11	202.05

Table 2: Expectation and Variance for Gold Medals and Total Medals



(a) Distribution of Gold Medals (b) Distribution of Total Medals

Figure 3: Histograms of Gold and Total Medals

We believe that a country's performance in a specific Olympic Games can be predicted by considering several factors: the number of events that year, whether the country is the host, and its performance in the previous three Olympic Games. This reasoning is intuitive: more events provide greater opportunities to win medals; host countries often benefit from a home advantage and typically secure more medals; and past performance reflects the country's overall strength. Based on these factors, we developed the following model[1]:

$$P(Y = y) = \binom{y + \phi - 1}{y} \left(\frac{\mu}{\mu + \phi}\right)^{y} \left(\frac{\phi}{\mu + \phi}\right)^{\phi} \tag{1}$$

where ϕ is the overdispersion parameter and μ is the mean number of medals, which can be calculated as follows:

$$\ln(\mu) = \beta_0 + \sum_{i=1}^{3} \beta_i n_{gold,i} + \beta_4 n_{events} + \beta_5 host \tag{2}$$

4.2 Zero-Inflated Model

The Olympic Games, a prominent international sporting event, attract a significant number of athletes from around the world for each iteration. However, it is important to note that in each competition, typically only three nations are awarded medals, and instances of tied results are relatively rare. Historically, since the inception of the Olympic Games in 1896, numerous countries have yet to secure any medals. As a result, the dataset contains a substantial number of zero values. The prevalence of these zeros presents a challenge, as the negative binomial distribution does not adequately account

[1]Using the count of gold medals as an illustration, the scenario for all medal totals is similar, with the only difference being that in the explanatory variables, the count of gold medals from the previous three years is replaced with the overall medal count from those same three years.

for the frequent occurrence of zeros. This necessitates the use of the **Zero-Inflated Model**, which incorporates an additional framework specifically designed to address the zeros within the dataset, thereby providing a more accurate representation of the phenomenon of high-frequency zeros.

We propose that **there are two distinct scenarios that may result in a value of zero**. The first scenario occurs when a nation has the potential to secure a medal but ultimately fails to do so due to random factors; we categorize this type of zero value as a **sampling zero**, signifying that it is merely a random outcome derived from sampling. The second scenario arises when a nation lacks the requisite strength in a particular sport and, as a result, has no realistic chance of winning a medal; we designate this type of zero value as a **structural zero**. Subsequently, we will model the occurrence of zero values as a Bernoulli distribution process, which will be executed through logistic regression, as follows:

$$p_0 = P(Y = 0) = \frac{1}{1 + e^{-\lambda}} \tag{3}$$

$$\lambda = \alpha_0 + \sum_{i=1}^{3} \alpha_i n_{gold,i} + \sum_{i=4}^{6} \alpha_i n_{silver,i-3} + \sum_{i=7}^{9} \alpha_i n_{bronze,i-6} + country \tag{4}$$

$$country \sim \mathrm{N}(0, \sigma_c^2) \tag{5}$$

In this analysis, we incorporate a random effect for the variable 'country' to examine the variations in medal counts among various countries. This random effect is denoted as 'country'.

4.3 Zero-Inflated Negative Binomial Model

We integrate the zero-inflated model with the negative binomial model to create the **Zero-Inflated Negative Binomial model, ZINB**:

$$P(Y = y) = \begin{cases} p_0 = \dfrac{1}{1 + e^{-\lambda}}, & y = 0 \\ (1 - p_0)P(Y = y|Y > 0) = (1 - p_0)\binom{y+\phi-1}{y} \left(\frac{\mu}{\mu+\phi}\right)^y \left(\frac{\phi}{\mu+\phi}\right)^\phi, & y > 0 \end{cases} \tag{6}$$

This model typically consists of two components: one part employs a binomial distribution to model the generation of zeros, while the other utilizes a negative binomial distribution to characterize the non-zero values. This dual structure enables the model to effectively address both actual zero counts and non-zero count data simultaneously.

4.4 Fitting the Model Using Bayesian Inference and MCMC

In this problem, we utilized **Bayesian inference** in conjunction with **Markov Chain Monte Carlo(MCMC)** methods to fit the model and derive its posterior distribution. Bayesian inference allows for the integration of prior knowledge or beliefs regarding the parameters, which can subsequently be updated in light of the observed data. The application of MCMC techniques enables efficient sampling from the posterior distribution, particularly in scenarios where direct computation is challenging or unfeasible. The detailed procedural framework of the algorithm is illustrated in the following diagram.

Algorithm 1 MCMC Algorithm for Bayesian Model Fitting with Multiple Chains

Input: Data $\{y_i, X_i\}$, initial parameter values $\theta^{(0)}$
Output: Samples from the posterior distribution of θ
Initialization: Set initial values for the model parameters:
$\theta^{(0)} = \{\beta_0, \beta_1, \ldots, \beta_5, \alpha_0, \alpha_1, \ldots, \alpha_9, \mu, \phi\}$
where β are the negative binomial regression coefficients, α are the zero-inflation coefficients, and μ and ϕ are the parameters of the negative binomial distribution.
Initialize chains: Run C parallel chains, each starting with random initial values of $\theta_c^{(0)}$ for chain $c \in \{1, 2, \ldots, C\}$
for each chain $c = 1, \ldots, C$ **do**
  for each iteration $t = 1, \ldots, T$ **do**
    Step 1: Compute the log-likelihood of the model with current parameter values:

$$\text{log-likelihood}_c = \sum_{i=1}^{N} \log(p(y_i|\theta_c))$$

where $p(y_i|\theta_c)$ is the likelihood of the data given the parameters for chain c.
    Step 2: Calculate the posterior distribution using Bayes' Theorem:

$$p(\theta_c|\text{data}) \propto p(\text{data}|\theta_c) \cdot p(\theta_c)$$

where $p(\text{data}|\theta_c)$ is the likelihood, and $p(\theta_c)$ is the prior distribution for chain c.
    Step 3: Propose new values for the parameters $\theta_c^* = \{\beta_0^*, \beta_1^*, \ldots, \mu^*, \phi^*\}$ using a proposal distribution.
    Step 4: Calculate the acceptance probability for the new parameter set:

$$A(\theta_c^* \rightarrow \theta_c) = \min\left(1, \frac{p(\theta_c^*|\text{data})}{p(\theta_c|\text{data})}\right)$$

    if a random number $u \sim U(0, 1)$ is less than the acceptance probability A **then**
      Accept the new parameter set: $\theta_c^{(t)} = \theta_c^*$
    else
      Reject the new parameter set: $\theta_c^{(t)} = \theta_c^{(t-1)}$
    end if
  end for
end for
Return: The concatenated set of accepted parameter samples from all chains:

$$\theta = \left\{\theta_1^{(1)}, \theta_1^{(2)}, \ldots, \theta_1^{(T)}, \ldots, \theta_C^{(1)}, \theta_C^{(2)}, \ldots, \theta_C^{(T)}\right\}$$

4.5 Problem Solutions

We have previously derived the posterior distributions of the parameters associated with the zero-inflated negative binomial regression model using MCMC methods. To determine the medal standings of various nations for the year 2028, we employ **simulation techniques** to estimate the medal outcomes for each country. The detailed algorithmic procedure is outlined as follows:

Algorithm 2 Bayesian Zero-Inflated Negative Binomial Model Prediction

Input: Trained Bayesian model, new data X_{new}
Output: Point estimate and prediction interval
1. Get posterior samples: $\theta^{(s)} \sim p(\theta|X, Y)$
2. Make predictions for new data:
for $s = 1$ **to** S **do**
    Compute $p_0^{(s)}$ and $\mu^{(s)}$
    Sample $y_{\text{new}}^{(s)} \sim \text{NegBin}(\mu^{(s)}, \theta_{\text{disp}}^{(s)})$
    With probability $1 - p_0^{(s)}$, set $y_{\text{new}}^{(s)}$
end for
3. Point estimate:
Compute point estimate $\hat{y}_{\text{new}} = \frac{1}{S}\sum_{s=1}^{S} y_{\text{new}}^{(s)}$
4. Prediction interval:
Calculate the 2.5
5. Output: Return $\hat{y}_{\text{new}}$ and the prediction interval

Based on the construction of the previous zero-inflated negative binomial model, we choose to use the zero-inflation factor to evaluate the likelihood of these nations achieving their first medal at the 2028 Los Angeles Olympics. The relevant formula can be expressed as follows:

$$P_{first}^{(i)} = 1 - \frac{1}{1 + e^{-\lambda_i}}$$

$$\begin{aligned}\lambda_i &= \alpha_0 + \sum_{i=1}^{3} \alpha_i n_{gold,i} + \sum_{i=4}^{6} \alpha_i n_{silver,i-3} + \sum_{i=7}^{9} \alpha_i n_{bronze,i-6} + country_i \\ &= \alpha_0 + country_i\end{aligned}$$

4.6 Results

Based on analytical assessments, the anticipated medal rankings for the 2028 Los Angeles Olympics are as follows.

NOC	Gold	95%HDI	Silver	95%HDI	Bronze	95%HDI	Total	95%HDI
USA	48	43~53	45	40~49	41	36~45	123	110~135
CHN	33	28~38	24	20~29	21	16~25	82	70~94
GBR	18	13~23	20	15~24	22	18~27	60	47~72
JPN	19	13~24	11	7~15	15	10~19	48	36~60
FRA	12	7~17	18	14~23	16	11~20	45	34~58
AUS	14	9~19	12	7~16	16	11~20	44	31~55
ITA	10	5~15	10	6~15	13	9~18	34	22~46
GER	11	7~17	10	6~15	11	6~15	33	21~45
NED	11	6~16	8	4~13	10	6~15	31	20~44
KOR	9	5~15	6	1~10	9	5~14	25	14~38
CAN	7	2~12	5	1~10	11	7~16	24	11~36
ROC	6	1~11	9	4~13	7	2~11	22	9~33
NZL	7	2~12	7	2~11	5	0~9	20	9~32
BRA	5	0~10	6	2~11	8	4~13	19	7~31
HUN	6	1~11	6	1~11	5	1~10	17	5~29
ESP	5	0~10	5	1~10	7	3~12	16	5~29
UKR	2	0~7	5	1~9	6	2~11	13	2~26
RUS	5	0~9	4	0~9	4	0~8	12	0~24
POL	2	0~7	4	0~8	5	1~10	11	0~23
CUB	4	0~9	2	0~6	5	1~10	11	0~23
KEN	4	0~9	3	0~8	3	0~8	11	0~24
UZB	5	0~10	1	0~6	3	0~8	11	0~23
DEN	2	0~7	3	0~8	5	1~10	11	0~22
SWE	3	0~8	5	0~9	2	0~7	10	0~22
KAZ	1	0~6	2	0~7	6	2~11	10	0~21
TUR	1	0~6	3	0~7	6	1~10	9	0~21
AZE	1	0~6	3	0~8	5	0~9	9	0~22
IRI	3	0~8	3	0~8	3	0~8	9	0~20
SUI	2	0~7	3	0~7	5	0~9	9	0~21
CZE	3	0~8	2	0~6	3	0~8	9	0~21
JAM	3	0~8	2	0~7	3	0~7	8	0~19
NOR	3	0~8	1	0~6	3	0~7	8	0~21
CRO	3	0~8	3	0~7	3	0~7	8	0~20
BEL	3	0~8	1	0~6	4	0~8	8	0~20
TPE	2	0~6	1	0~6	5	0~9	8	0~20
SRB	3	0~8	2	0~6	3	0~7	8	0~19

To evaluate the performance of countries in 2028 relative to 2024, we identified the twenty countries predicted to win the most medals in 2028 and the twenty countries that ranked highest in the 2024 medal standings for our analysis. Here are the findings.

Figure 4: Comparison of the top twenty predicted medal-winning countries in 2028 with the highest-ranked countries in the 2024 medal standings.Russia, Poland, and Cuba are anticipated to break into the top twenty, while Uzbekistan, Iran, Kenya, and Sweden are expected to fall short. The United States, Japan, Germany, and Ukraine are projected to earn more medals than they did in 2024, whereas China, Great Britain, France, Australia, Italy, the Netherlands, South Korea, Canada, Brazil, Hungary, and Spain are expected to win fewer medals. New Zealand's medal count is anticipated to remain unchanged.

The anticipated medal distribution for the leading six countries in the 2028 medal standings is projected to be as follows.

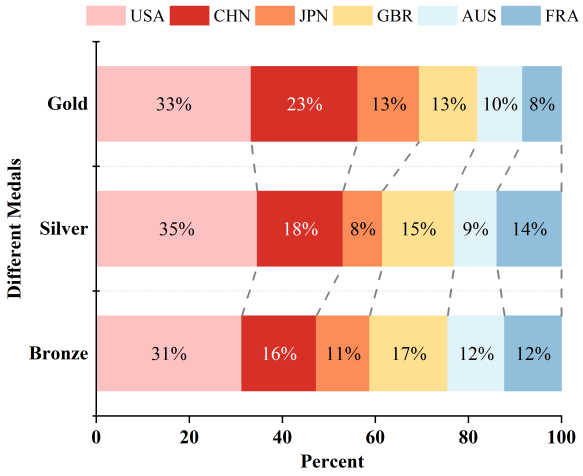


Figure 5: Predicted Medal Distribution for Top 6 Nations at LA28.

The following six nations have the potential to secure their inaugural Olympic medal at the 2028 Los Angeles Games.

NOC	Country	Est. prob	Population	GDP
MTN	Mauritania	0.248	5,022,441	10,651,709,411
BIH	Bosnia and Herzegovina	0.163	3,185,073	27,514,782,476
LES	Lesotho	0.145	2,311,472	2,117,962,451
AND	Andorra	0.139	80,856	3,785,067,332
CAY	Cayman Islands	0.136	73,038	7,139,428,558
COM	Comoros	0.136	850,387	1,352,380,971

Figure 6: The top six nations poised to win the inaugural medal at the Los Angeles Olympics.

4.7 Model Assessment

Due to space limitations, only the performance evaluation and uncertainty measurement of the gold medal count prediction model are presented. The analysis process for the other models is similar, and the results are comparable.

According to the findings from MCMC, the parameter estimates for the zero-inflated negative binomial model and the highest density interval (HDI) are as follows: the standard errors for each parameter are quite small, indicating minimal uncertainty associated with each estimate.

The table also presents the Rhat and Effective Sample Size (ESS) values for each parameter. In this model, all Rhat values are 1.00, indicating that all parameters have converged, which reflects excellent model convergence. The Bulk ESS values for all parameters are substantial, suggesting that the MCMC chains provide a sufficient number of effective samples, thereby enhancing the reliability of the results. Additionally, the Tail ESS values are also high, indicating that the model possesses an adequate number of effective samples in the tail region, which ensures that the estimation of extreme predictions is equally reliable.

	Estimate	**95% HDI**	**Rhat**	**Bulk ESS**	**Tail ESS**
β_0	-0.05	-0.49-0.38	1.00	24043	10826
β_1	0.41	0.38-0.44	1.00	11050	10315
β_2	0.28	0.24-0.31	1.00	10853	9464
β_3	0.22	0.19-0.26	1.00	11291	10474
β_4	0.00	-0.00-0.00	1.00	25202	10858
β_5	10.47	9.55-11.36	1.00	13949	11491
σ_c^2	2.53	2.46-2.60	1.00	13927	11230

Table 3: Model Estimates with Uncertainty Intervals, Rhat, and ESS values

5 How Events of Various Sports Impact the Results?

5.1 Establishment of Indicators

This section aims to investigate the influence of events on the outcomes of the Olympic Games. Specifically, we seek to determine whether the number of events within each discipline for each Olympic game, as well as the types of events, affect the performance of participating countries, thereby impacting the overall medal standings.

To analyze this phenomenon, it is essential to establish several key indicators. For a given country (designated as A) participating in a specific Olympic Games, let p denote the total number of events conducted, q represent the total number of events in which country A participated, and s signify the total number of events in which country A secured medals (i.e., the total number of medals awarded to country A). Furthermore, for a particular discipline i, let a indicate the total number of events held, b represent the number of events in discipline i that country A participated in, and c denote the total number of medals won by country A in that discipline. To facilitate this analysis, it is necessary to filter the *athletes* and *programs* data. It is noteworthy that discrepancies were identified in the notation of *sport* within the *athletes* dataset, which correspond to *discipline* outlined in the *programs* dataset.

Discipline	Indicators	Explanation	Method of Calculation	Meaning
i	x_i	Type influencing factor	b/q	Proficiency level
	y_i	Number influencing factor	b/p	Popularity level
	z_i	Discipline importance factor	c/s	Award winning ability

Table 4: Relevant Indicators

Consequently, by considering both the quantity and type of events, we can derive a comprehensive indicator reflecting the competitive advantage of discipline i for country A. We propose the following indicators:

$$\varphi_i = \omega_1 x_i + \omega_2 y_i \tag{7}$$

In this context, ω_1 and ω_2 are designated weights, and a higher value of $\varphi_i$indicates a greater advantage of discipline i for country A.

5.2 Events-Medals Association Rule Model Based on GRI Algorithm

The correlation between events and medal counts can be examined through the application of association rule models. Association analysis reveals causal relationships among sets of items by categorizing various data item sets as either antecedents or consequents. In summary, the formal representation of association rules is as follows: (m_c denotes the counts of medals).

$$if\ antecedent \rightarrow consequent$$
$$if\ \varphi_i \rightarrow m_c$$

Association analysis can be conducted using various algorithms, with the GRI algorithm and the Apriori algorithm being among the most prevalent. This study employs the GRI algorithm due to its advantages over the Apriori algorithm, particularly in minimizing the generation of redundant rules through the incorporation of concepts such as interaction entropy and information gain, thereby enhancing the efficiency of rule discovery. Furthermore, given the extensive range of disciplines involved, the GRI algorithm demonstrates greater adaptability in exploring rules within large item sets compared to the Apriori algorithm.

The fundamental principle of the GRI algorithm can be expressed through a specific formula, which delineates the disparity between the probability of m_c occurring in the presence of φ_i and the overall probability of m_c.

$$J(m_c\,|\varphi_i) = p(\varphi_i)\left(p(m_c\,|\varphi_i)log\frac{P(m_c\,|\varphi_i)}{p(m_c)} + (1-p(m_c\,|\varphi_i))log\frac{1-p(m_c\,|\varphi_i)}{1-p(m_c)}\right) \tag{8}$$

Utilizing the aforementioned algorithm, we conducted an analysis of data from the 29 Summer Olympic Games in which the United States has participated, spanning the years 1896 to 2024. In accordance with Formula 7, we assigned a value of 0.5 to both ω_1 and ω_2, reflecting the equal significance attributed to the type factor and the number factor of events. Following this, we computed φ_i for 65 disciplines based on the established formula, with select data presented in Table 5.

Year	**Wrestling**(φ_1)	**Athletics**(φ_2)	**Golf**(φ_3)	**Rowing**(φ_4)	**Fencing**(φ_5)	**Swimming**(φ_6)	.....	m_c
1896	0	0.617	0	0	0	0.061		20
1900	0	0.648	0.068	0.011	0.034	0.023		48
1904	0.160	0.611	0.042	0.097	0.097	0.153		231
1908	0.052	0.554	0	0	0	0.073	.....	47
1912	0.009	0.597	0	0	0.036	0.071		64
1920	0.085	0.339	0	0.024	0.036	0.145		95
				.....				
2016	0.045	0.236	0.010	0.035	0.035	0.182		121
2020	0.047	0.218	0.012	0.028	0.040	0.165	.....	113
2024	0.051	0.250	0.009	0.038	0.041	0.161		126

Table 5: Events-Medals Data Matrix

Association rules necessitate the classification of data. For the variable φ_i , classification is conducted using the threshold $Q_{0.5}$; values exceeding $Q_{0.5}$ are designated as 1, signifying that the discipline is advantageous, while values that are less than or equal to $Q_{0.5}$ are assigned a value of -1, indicating a normative status. In relation to the variable mc, its quantitative representation is illustrated in Figure 4. This variable is categorized into three segments: high, low, and normal, based on the prevailing conditions in the United States, with corresponding values of 1, 0, and -1, respectively.

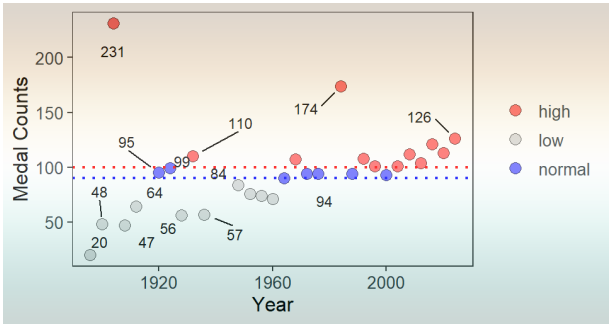
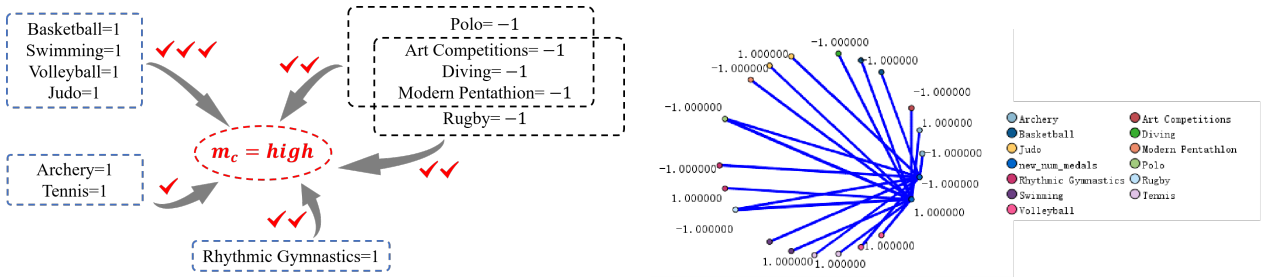


Figure 7: Total Medal Count of the USA in Past Summer Olympics

Consequently, we construct a data matrix of dimensions 29 by 66 and employ the GRI algorithm to extract the following robust association rules:



(a) Illustration of 5 Strong Rules (b) Network Graph Visualization

Figure 8: The Result of Association Rule Model(USA)

According to the aforementioned criteria, it can be inferred that Basketball, Swimming, Volleyball, Judo, Archery, Tennis, and Rhythmic Gymnastics hold considerable significance for the United States. Increased participation in these sports may enhance the likelihood of the U.S. securing additional medals.

Employing a similar analytical framework, it is feasible to derive insights for other nations and perform comparative assessments to identify disciplines of importance across multiple countries. Furthermore, the previously defined indicator z_i warrants examination. By calculating the average value of z_i for all participating countries over the last seven Olympic Games, one can derive $\overline{z_i} = \frac{1}{7}\sum_{i}^{7} z_i$ (the average medal-winning proportion for each event), which serves as a metric of importance. Visualization through a heatmap indicates that Athletics is of considerable significance for numerous countries.

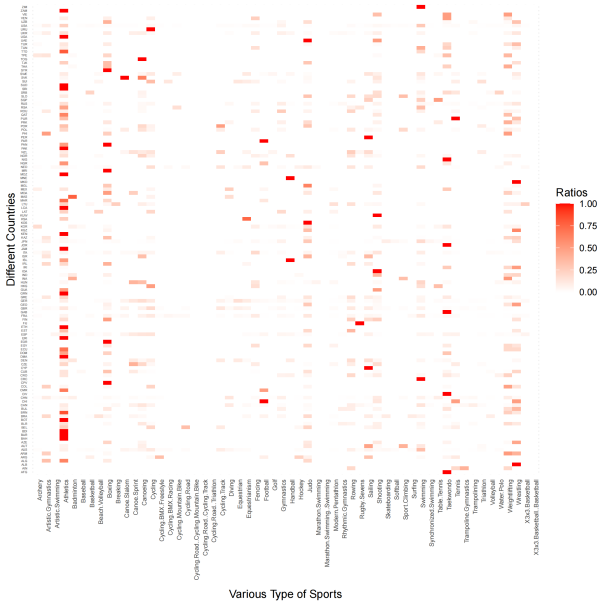


Figure 9: Importance of Sports(disciplines) to Various Countries

Moreover, through the application of association rules, we can investigate whether the introduction of new disciplines by the host country contributes positively to the increase in medal counts. We construct item sets as delineated in the table..., where if_{int} indicates whether the new discipline was primarily introduced at the behest of the host country, if_{par} denotes the host country's participation in this discipline, if_{have} indicates whether the host country secured an award, and if_{max} signifies whether the host country achieved the highest number of awards in this event. Each row of the data matrix encapsulates information regarding the new disciplines introduced in previous Olympic Games:

Symbol	if_{int}	if_{par}	if_{have}	if_{max}
Sign	1/-1	1/-1	1/-1	1/-1

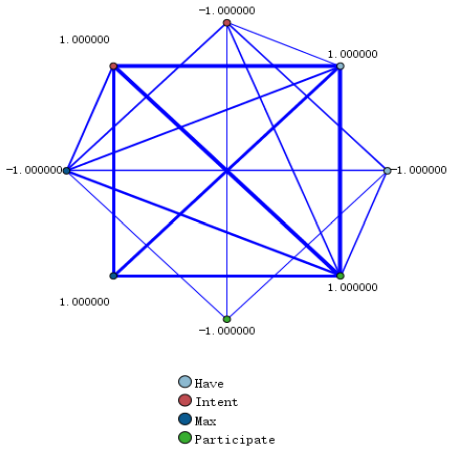
Table 6: Relevant indicators of Host Countries

We designate if_{int} and if_{par} as the antecedents, while if_{have} and if_{max} serve as the consequents, and we perform association analysis utilizing the GRI algorithm to derive the following rules.It can be posited that the inclusion of events at the behest of the host country is likely to be advantageous for that country, potentially facilitating an increase in their medal count:

$$if \quad if_{int} = 1.0 \text{ and } if_{par} = 1.0 \Rightarrow if_{have} = 1.0$$

$$if \quad if_{int} = 1.0 \text{ and } if_{par} = 1.0 \Rightarrow if_{max} = 1.0$$

(a) Illustration of 2 Strong Rules



(b) Network Graph Visualization

Figure 10: The Result of Association Rule Model(Host Country)

5.3 Results Analysis

To assess the efficacy of a rule, it is essential to examine it through the lens of validity indicators pertinent to association rules. For instance, consider the association analysis where the antecedent is denoted as φ_i and the consequent as m_c:

- **Support**

$$S_{\varphi_i \to m_c} = \frac{N(\varphi_i \bigcap m_c)}{N} \tag{9}$$

$N(\varphi_i \bigcap m_c)$ signifies the number of transactions that include both φ_i and m_c, while N denotes the total number of transactions. The support metric of the rule indicates its generalizability; a higher support value suggests a greater universality of the rule. In general, the minimum level for support can be established at a fairly low point.

- **Confidence**

$$C_{\varphi_i \to m_c} = \frac{N(\varphi_i \bigcap m_c)}{N(\varphi_i)} = \frac{S_{\varphi_i \to m_c}}{S_{\varphi_i}} \tag{10}$$

$N(\varphi_i)$ represents the number of transactions that include φ_i. The confidence of the rule can be interpreted as a conditional probability; a higher confidence level indicates a stronger reliability of the rule. In general, the minimum confidence threshold should be established at a fairly high level.

- **Lift**

$$L_{\varphi_i \to m_c} = \frac{N(\varphi_i \bigcap m_c)}{N(\varphi_i)} \Big/ \frac{N(m_c)}{N} = \frac{C_{\varphi_i \to m_c}}{S_{m_c}} \tag{11}$$

The lift of a rule is defined as the ratio of the rule's confidence to the support of the consequent. This metric illustrates the extent to which the antecedent influences the consequent in comparison to the overall context. Consequently, when $L_{\varphi_i \to m_c} > 1$ holds true, the antecedent exerts a positive influence on the consequent, with a greater lift indicating a stronger degree of positive influence. Conversely, if $L_{\varphi_i \to m_c} < 1$ is true, even with high support and confidence, the rule still not be deemed effective.

- **Deployment Capability**

$$D_{\varphi_i \to m_c} = S_{\varphi_i} - S_{\varphi_i \to m_c} \tag{12}$$

Deployment capability measures the proportion of instances where the antecedent is true but the consequent is not. A higher deployment capability suggests greater potential for improvement in the rule; therefore, a lower deployment capability is generally preferred.

In conclusion, we have conducted a validity analysis of the rule outcomes from the preceding section, and the findings indicate that all our rules are valid and of considerable quality.

Antecedent	Consequent	Supp.	Conf.	Lift	Deploy.	Rules Strength
Basketball=1 Swimming=1 Volleyball=1 Judo=1	m_c=high	34.48	90.0	3.175	2.448	•••
Rhythmic Gymnastics=1	m_c=high	31.03	88.89	2.148	3.447	••
Archery=1 Tennis =1	m_c=high	34.48	80.0	1.933	6.896	•
Art Competitions=-1 Diving=-1 Modern Pentathlon=-1 Polo=-1	m_c=high	41.38	83.33	2.014	6.898	••
Art Competitions=-1 Diving=-1 Modern Pentathlon=-1 Rugby=-1	m_c=high	41.38	83.33	2.014	6.898	••

Table 7: Analysis of the Validity of Association Rules 1

Antecedent	Consequent	Supp.	Conf.	Lift	Deploy.	Rules Strength
if_{int}=1 if_{par}=1	if_{have}=1	80.56	100	1.2	0	•••
if_{int}=1 if_{par}=1	if_{max}=1	80.56	68.97	1.241	24.998	•

Table 8: Analysis of the Validity of Association Rules 2

6 Could a Great Coach Boost Your Country's Medal Count?

In competitive sports, achieving success depends not only on the skills and training of athletes but also significantly on the influence of coaches. The term "great coach" refers to those who achieve remarkable achievements with various teams or countries due to their outstanding tactical knowledge, extensive expertise, and ability to inspire their teams. These coaches are not limited by their nationality; they can transcend borders, sharing their coaching experience and strategies in diverse sports environments, which can enhance the overall performance of their teams. However, this is merely a perception. Is the "great coach" effect genuinely real? In the following discussion, we will explore the existence of the "great coach" effect and attempt to assess its impact on the medal tallies of different nations.

To examine the "Great Coach" phenomenon, we initially selected four prominent coaches:

1. Lang Ping, who led the U.S. women's volleyball team to a silver medal at the 2008 Beijing Olympics and guided the Chinese women's volleyball team to a gold medal at the 2016 Rio Olympics;

2. Bela Karolyi, the head coach of the women's gymnastics team, led the Romanian team to three gold medals, one silver medal and two bronze medals at the 1976 Montreal Olympics. He later coached the U.S. team to a gold medal at the 1996 Atlanta Olympics.

3. Anastasia Bliznyuk, who began coaching in China in 2022, secured a gold medal in the Group All-Around Rhythmic Gymnastics event at the 2024 Paris Olympics;

4. Erwann Le Péchoux, who has been coaching the Chinese fencing team since 2021 and won a gold medal in Men's Foil Team fencing.

6.1 Analysis of Variance, Mixed-effects Model

We aim to investigate the role of a "great coach" as a variable influencing a country's sports performance. This variable has two levels: a value of 1 indicates the presence of a "great coach," while a value of 0 indicates their absence. However, we recognize that the effectiveness of a "great coach" may depend on the specific conditions within a country. Even with an outstanding coach, if the athletes are unable to collaborate effectively or lack the necessary skills, the country's sports performance may not improve. Therefore, it is essential to consider each country's unique circumstances alongside the "great coach" factor to accurately assess the coach's impact.

To facilitate this analysis, we will categorize our units of analysis into distinct "blocks", each comprising a specific coach and the country they represent. For example, Lang Ping coaching the U.S. team would represent one block, while her coaching of the Chinese team would constitute another. This approach will enhance our understanding of the relationship between coaches and countries and its impact on performance.

In our analysis, the presence of a "great coach" will be treated as a fixed effect, enabling us to concentrate on the specific impact of each coach. Conversely, the "block effect" will be regarded as a random effect, given that our study involves a limited selection of coaches and countries.

In summary, a country's performance can be decomposed as follows:

$$Score = \varphi_0 + \varphi_1 \omega + \delta + \varepsilon \tag{13}$$

$$\omega = \begin{cases} 1, & \text{if a "great" coach is present} \\ 0, & \text{if no "great" coach is present} \end{cases} \tag{14}$$

$$\delta \sim \mathrm{N}(0, \sigma_\delta^2), \quad \varepsilon \sim \mathrm{N}(0, \sigma^2) \tag{15}$$

In this context, φ_1 denotes the constant effect of the "great coach", δ signifies the random effect associated with the block, and ε stands for the error term.

6.2 Definition of the Response Variable

How should we define the response variable? We believe that using the total number of medals earned by teams led by "great coaches" is not appropriate. Gold, silver, and bronze medals should not be treated as equal; for some nations, achieving their first bronze medal is a significant milestone, while for already successful countries, progressing from silver to gold represents a substantial advancement. Selecting only one type of medal fails to capture the true impact of "great coaches". Furthermore,

different types of medals cannot be simply aggregated to reflect the influence of "great coaches" on a country's performance in a specific sport. After thorough consideration, we have decided to assign weights to each type of medal, and the final weights are presented in the table below:

	Gold	**Silver**	**Bronze**	**No Medal**
Weights	10	5	2	0

Table 9: Weights of Different Medals

The definition and calculation formula for the response variable value is as follows:

$$Score = 10 \times n_{gold} + 5 \times n_{silver} + 2 \times n_{bronze} \tag{16}$$

6.3 Problem Solutions

We believe that the influence of a "great coach" can have a significant but short-lived impact on the countries they coach. To analyze this phenomenon, we examine the year a country hired the "great coach" and review the outcomes of the last three Olympic Games in which they participated, as well as the results of the subsequent two Olympic Games, totaling six Olympic Games. For each experimental point, we conduct three repetitions and perform a variance analysis of these results using a mixed-effects model with equal repetitions. If the fixed effect is both significant and positive, it indicates the existence of the "great coach" effect, allowing us to estimate this effect based on the point estimate of the effect of the fixed factor in the mixed-effects model.

	Block$_1$	**Block**$_2$	**Block**$_3$	**Block**$_4$	**Block**$_5$
0	$a_{011}, a_{012}, a_{013}$	$a_{021}, a_{022}, a_{023}$	$a_{031}, a_{032}, a_{033}$	$a_{041}, a_{042}, a_{043}$	$a_{051}, a_{052}, a_{053}$
1	$a_{111}, a_{112}, a_{113}$	$a_{121}, a_{122}, a_{123}$	$a_{131}, a_{132}, a_{133}$	$a_{141}, a_{142}, a_{143}$	$a_{151}, a_{152}, a_{153}$

Table 10: Repeated Measures Mixed-effects Experimental Data Table

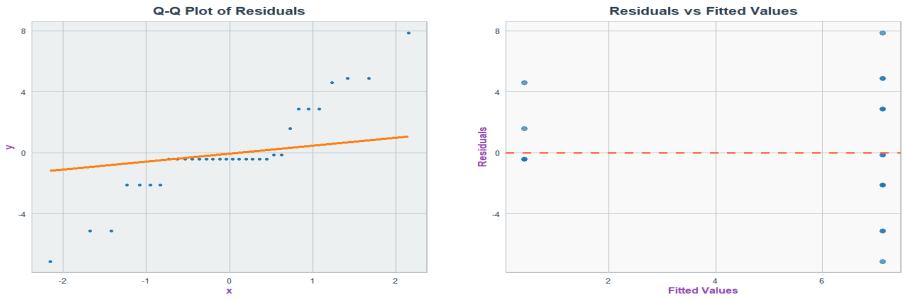
6.4 Fitting the Model

We have conducted a preliminary fitting of the model and obtained the following results.

The table below indicates that **the variance of the random factor is quite low and can be ignored**. We create a **residual plot** and a **QQ plot** to assess if the data satisfy the assumptions of normality and homoscedasticity.

Random Effects	Variance	Std. Dev.			
Block	7.220e-14	2.687e-07			
Residual	9.395	3.065			
Number of observations: 32, groups: block, 5					
Fixed Effects	**Estimate**	**Std. Error**	**df**	**t value**	**Pr(>|t|)**
(Intercept)	0.4118	0.7434	30.0000	0.554	0.584
Fixed Factor	6.7216	1.0858	30.0000	6.190	8.2e-07 ***
*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*					

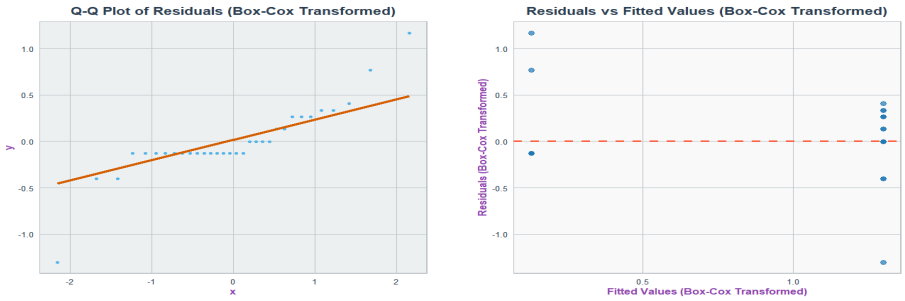
Table 11: Random and Fixed Effects Summary



(a) Distribution of Gold Medals (b) Distribution of Total Medals

Figure 11: QQ Plot and Residuals Plot for Model Diagnostics

The QQ plot indicates that the data fails to satisfy the normality assumption, and the residual plot reveals a noticeable trend. To address this issue, we apply the **Box-Cox transformation** in an effort to adjust the parameters to align with the Gauss-Markov assumptions. Since the Box-Cox transformation requires the data to be positive and our dataset contains zeros, we have adjusted the values by adding one to each. The calculated coefficient for the Box-Cox transformation is 0.38. After performing the transformation and refitting the data, we generate new QQ and residual plots.



(a) Distribution of Gold Medals (b) Distribution of Total Medals

Figure 12: QQ Plot and Residuals Plot for Model Diagnostics(After Box-Cox Transformation)

At this stage, the QQ plot shows a slanted straight line, with data points scattered around it, indicating that the data appears to satisfy the normality assumption. However, an analysis of the residual plot reveals that there is still a noticeable trend in the data following the Box-Cox transformation, suggesting that the residuals do not meet the homoscedasticity requirement. Consequently, we applied **White's robust standard errors** to adjust and evaluate the model fitting results, which are presented in the table below. The effect of the "great coach" factor, obtained through the Box-Cox inverse

Variable	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	df
(Intercept)	0.1289	0.08992	1.434	0.1619	-0.05471	0.3126	30
Fixed Factor	1.1719	0.14391	8.143	4.329e-09	0.87800	1.4658	30

Table 12: Model Estimates with Confidence Intervals and t-Statistics

transformation formula, is **3.74**.

6.5 Results

We conducted research to identify countries eager to achieve breakthroughs in specific sporting events. We found that Australia is particularly focused on advancing in basketball and tennis, with tennis being a traditional sport. However, in recent years, both men's singles tennis and men's basketball have failed to secure any medals. A similar situation is observed in France with men's 100m sprint athletics and in Canada with men's basketball.

Based on the estimated "Great Coach" effect and the conversion relationship between effect size and medal weight, if these nations choose to invest in a "great coach", the following outcomes are expected:

1. For countries that currently lack the capacity to secure a medal, there is a **74.8%** probability of winning a silver medal following investment;
2. For countries that can achieve a bronze medal but not a silver, there is a **57.4%** probability of winning a gold medal following investment;
3. For countries that are capable of winning a silver medal but not a gold, there is an **87.4%** probability of winning a gold medal following investment.

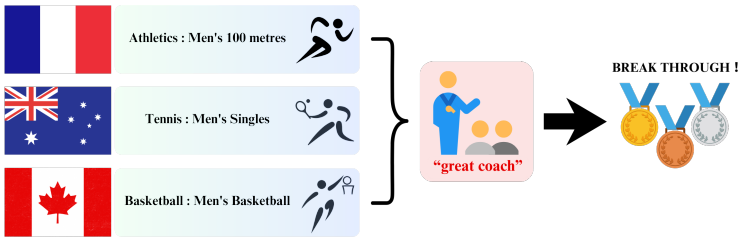


Figure 13: If these countries invest in great coaches, they will reap the rewards of medals.

7 What Else Is Included in Our Model?

While conventional wisdom holds that "more events lead to higher medal chances", our Bayesian regression model shows: The regression coefficient for event count (β_4) has a 95% Highest Density Interval (HDI) that fully overlaps with zero. This means **increasing the number of events entered does not have a statistically significant impact on a nation's total medal count.**

Furthermore, the findings from the Events-Medals Association Rule Model in the United States show a significant link between restricted involvement in specific sports and an increased total medal tally for the nation. This suggests that **it is crucial for a country to carefully select the events it competes in, rather than taking part in a wide range of competitions**.

Strategic Implications for Olympic Committees:

1. **Resource Reallocation Opportunity**: Concentrating resources on dominant events rather than dispersing them across multiple disciplines may create a "focus effect" for higher returns.
2. **Qualification Strategy Overhaul**: Transition from a "quantity-driven" participation model to a "medal conversion rate" evaluation system to prioritize high-potential events.
3. **Talent Development Shift**: Targeted investments in athletes/projects with podium potential yield better results than a broad but shallow approach to talent cultivation.

8 Model Evaluation

8.1 Advantages

1. Zero-Inflated Negative Binomial Model.
   (a) This model generally has two elements: one uses a binomial distribution to represent the creation of zeros, and the other applies a negative binomial distribution to describe the non-zero values. This two-part framework allows the model to effectively handle both true zero counts and non-zero count data at the same time.
   (b) Unlike regression models or machine learning algorithms designed for continuous outcomes, the Zero-Inflated Negative Binomial model is specifically tailored for count data. It ensures that predictions are non-negative integers, thus avoiding the problem of predicting continuous values or negative numbers, which would be unrealistic in count data scenarios.
2. GRI Association Rule Model.
   (a) Association rule mining is an automated analysis method that does not rely on pre-set hypotheses but directly extracts relationships from data, avoiding biases from manual assumptions.
   (b) This model can handle large-scale datasets. The historical data of the Olympics involves a vast and complex array of data from multiple countries, spanning various years, and numerous events, and using association rules can efficiently reveal the intricate relationships between the data.

(c) The output of association rules is typically presented in the form of "if... then...", making the interpretation of the rules intuitive and easy to understand.

(d) GRI can discover multi-level associations. For example, it can reveal not only the relationship between a single event and the total number of medals but also the combined impact of multiple events on the number of medals.

3. Mixed-effects Model. Mixed-effects models are highly effective for assessing significant differences among various levels of factors and accurately estimating the effects associated with fixed factor levels. This model facilitates the investigation of the "great coach" effect, showcasing both efficacy and clarity in their application.

8.2 Limitations

1. The resolution of the zero-inflated negative binomial model is accomplished through the implementation of the Markov Chain Monte Carlo (MCMC) algorithm, which necessitates considerable computational resources and is marked by lengthy processing times.

2. Forecasting medal counts utilizes a simulation approach that also demands significant computational resources and entails extended solving durations.

3. When mining rules, GRI may encounter false positive issues (irrelevant rules mistakenly identified as having strong associations). For instance, we may not be particularly concerned about strong rules where mc equals 0, but such rules can still occur.

References

[1] Csurilla, Gergely, and Imre Fertő. 2024. "How to win the first Olympic medal? And the second?". Social Science Quarterly 105: 1544–1564. https://doi.org/10.1111/ssqu.13436.

[2] Olympic Games France Delegation.

[3] How patience helps French sprinter Lemaitre to aim for a third Olympic podium.

[4] World Bank.