Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2504209

Golden Algorithms: Forecasting Olympic Success Through Predictive Modeling and Cross-Nation Coaching Dynamics

Summary

The year 2024 that just passed was the Olympic year. The Olympic Games never fails to stir people's hearts at any time. The medal table also touches everyone's heartstrings. This study investigates the medal counts and effect of elite coaches in the Olympic Games and uses regression models and statistical models, especially **Tobit** approach to predict medal counts in the detailed way.By analyzing summer Olympic data, we address these key questions:

Establish Model of the Change of Medal Count: Observe the variable is either **left-censoring** and this implies that the **Tobit Model** is a more efficient and accurate models than the ordinary models. Establishing Tobit Model help us further understand the behaviors of the Olympic medals data and the role of athletes in competitions.

Predicting medal counts in 2028: We apply **linear regression model and Tobit model** on the medals data. We further apply this model to explore the importance of sports to different countries. Then use **K-means** to category countries and then analysis each characteristic.

Host Effect Model: Apriori Algorithm is applied in the research of host effect. The algorithm explores the relationship between sport and NOC. This algorithm works well in evaluating the strong correlation between the host country's choice of its own sports and medal increase.

Great Coach Effect and Investment Decision: We analyze great coach effect on Olympic medals based on a textbfdifference-in-differences(DID) model. The results demonstrate that the introduction of elite coaches significantly enhances medal acquisition outcomes ($\beta > 0$, $p < 0.05$), as evidenced by fixed-effects regression models accounting for athlete participation and historical performance baselines. Key sports in Italy, China, and Japan were identified through participation-to-medal yield gaps in recent Games.

We completed the sensitivity analysis by adding a little noise and found that the model is very robust. Our research indicates that behind the competition for Olympic medals lies a multitude of invisible efforts, such as a robust sports system and substantial national investment. These factors, while often overlooked, are just as critical as the athletes' own hard work, if not more so.

Keywords: Tobit Model, DID Model, Olympic Medals, Prediction Model, Apriori Algorithm, K-means

Contents

1 Introduction

Figure 1: Olympic Emblem[1]

Figure 2: Word cloud map

The official website of the Olympics describes the Olympic movement as follows: 'The Olympic Games are the world's only truly global, multi-sport, celebratory athletics competition. With more than 200 countries participating in over 400 events across the Summer and Winter Games, the Olympics are where the world comes to compete, feel inspired, and be together.'[1] People all over the world will unanimously pay high attention to the quadrennial Olympic Games. People from different countries show the limits of humanity to the world and interpret the beauty of sports.

Meanwhile, an interesting question will continue to be raised, which is the prediction of the Olympic medal table. Everyone hopes that their country's athletes can show extraordinary talents, but medals are always limited. People are always talking about that 'Asian plays table tennis better than people in European' and 'Jamaican people run the fastest on earth'. Is it true that people in different countries are good at different sports. Some people also believe that an extraordinary coach can affect the match and help a country to win more medals while others not. This paper attempts to address issues that people are interested in.

1.1 Restatement of the Problem

- **Model Development and Prediction** Construct a model using the given data to predict the total and gold medal counts for each country. Then use this model to make the following predictions :

  1. Firstly, estimate the medal table for the 2028 Olympics and identify whether the trend of countries' performances are increasing or decreasing. e.
  2. Secondly, Predict which countries, currently without Olympic medals, are likely to win their first medal in the 2028 Games, and assess the probability of this happening.
  3. Thirdly, examine how the number and types of events in each Olympics affect medal distribution and considering the potential influence of the host country's event selection on the outcomes.

- **Coach Effect Analysis** Investigate the potential effect of top coaches who have worked with multiple countries and choose tree countries to access how such coaching transitions might influence medal success.
- **Original Insights from the Model** Provide some insights and recommendations based on model results to help national Olympic committees optimize their training

1.2 Our Work

1. To estimate the medal count in 2028, we should build a regression model base on the given data. A very direct idea is the linear regression model. The backwards is obvious that it will not suitable for a long term prediction since the increasing of number of medals cannot be linear respecting time.

2. Then we improve the model into **Tobit Model**, since the data should be greater than 0 while there are about seventy countries never won a medal. Tobit Model can deal with this kind of data very well. After building up model, combine linear model and then apply this model to predict medal count in 2028 and solve other questions. As shown in the figure 3.

3. Considering the relationship between NOC and Sport, we also construct the Host Effect Model to compare with the result of Tobit model and obtain the dominant countries for each sport.

4. This study also analyzes **Great Coach Effect** on Olympic medals using a **difference-in-differences(DID)** model, controlling for year and athlete numbers, with coaching intervention as the treatment variable.

5. Key sports in Italy, China, and Japan were identified through participation-to-medal yield gaps in recent Games. Sports-specific adaptations of the model, combined with Bayesian 2028 medal projections, quantified coaching-driven medal potential.

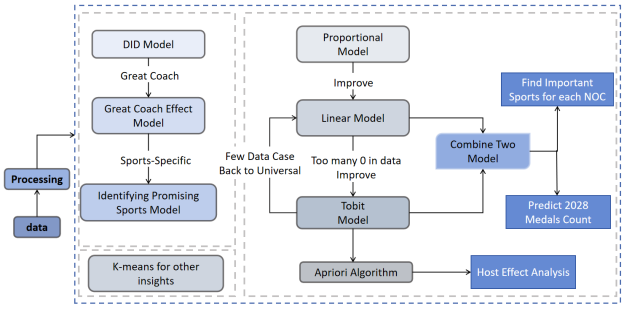

Figure 3: Our Work

2 Model Preparation

2.1 Assumption

- **Assumption1**: Ignoring 1906 Olympic game will not affect the result.

  Justification: The exclusion of the 1906 Intercalated Games is methodologically justified due to their non-canonical IOC status, structural anomalies (smaller scale, non-standardized protocols), and statistically insignificant impact on longitudinal analyses. Omitting them ensures data homogeneity across Olympic performance studies without distorting trends.

- **Assumption2**: In team events, only one medal is counted when a team wins an award.

  Justification: This enables the number of medals we calculated to predict the results more accurately.

- **Assumption3**: The recruitment of elite coaches has been empirically demonstrated to exert a sustained and statistically significant impact on a nation's Olympic medal acquisition trajectory across subsequent Games.

  Justification: This phenomenon arises from the systemic integration of advanced training methodologies, sport-specific technical innovations, and strategic performance optimization frameworks introduced by such coaches.

2.2 Notation

In order to better describe the model, we assign the following symbols pratical meanings.

Symbols	Description
Y_{it}	The total medal counts for country i in year t.
GC_{it}	The indicator for whether country i introduces great coaches.
$Post_{it}$	The indicator for whether year t is in post-introduction period.
$GC_{it} \times Post_{it}$	The interaction term capturing the great coach effect.

The remaining symbols will be listed in appropriate positions in the article.

2.3 Data Cleaning and Processing

1. Data Cleaning

   - For convenience of retrieval, some unnecessary spaces have been removed after the country names in the document *summerOly_medal_counts.csv*.
   - We notice that for the same player there might be different kind of spelling in the document *summerOly_athletes.csv* because of the traditions of name writing in different countries. For example, a famous table tennis player Ma Long is

registered as Ma Long in the 2012 and 2016 Olympic games, while he is registered as Long Ma in 2020 and 2024 Olympic games. Though it is hard to distinguish every single name in such a long list, we manually unified the names of athletes with inconsistent formats that we found.

- *summerOly_programs.csv* has a lot of messy code and blank elements. Fill the blank with 0 and made the messy code readable.
- ROS and RUS both are NOC for Russia due to some affairs. We combine them together into RUS to get better model for prediction.

2. Data Processing Due to the large volume of data and the requirement for extensive data analysis, we use a Database Management System (DBMS) to store and manage the data. SQL statements can calculate information such as the number of participants and the number of medals for training the model efficiently.

   - **Ignore 1906 Olympics in athlete table :** In 1906, there is a special Olympic Game. We notice that there are records of the 1906 Olympics in *summerOly_athletes.csv*, but there are no relevant records of the 1906 Olympics in *summerOly_medal_counts.csv*. To maintain the data consistency, we ignore the record of 1906 Olypics in *summerOly_athletes.csv*.
   - **Unify names :** The names of sports vary in each Olympic Games, which leads to unnecessary categorization when counting the number of medals. For example, there are "Gymnastics"and "Artistic Gymnastics". Therefore, we unified the names of sports in the *summerOly_athletes.csv* with those in the *summerOly_programs.csv* table.
   - **Calculate medal amount according to "Assumption2" :** We determine whether an event is a multi-participant event based on the specific Event in *summerOly_athletes.csv*, deduplicate the medals, and obtain the correct medal statistics.

3 Model Establishment

We are curious about the relationship between the number of Olympic medals awarded by a country, and its behaviors in the future. Figures 4 to 7 show the relationship of some countries' number of medals in the history.

We've noticed several remarkable increases in these figures. After referring to relevant materials, we found that those peak values coincide with the periods when these countries hosted the event, which remind us of the impact of hosting Olympics events.

The number of medals won in an Olympic Games is highly depended on the **total number of events for all athletes** of a country and is highly dependent on the **medal table** in the last three Olympic Games (few athletes can participate in more than three games). Considering these impacts, it is reasonable to establish two models to separately describe the number of athletes and their contribution to the medal table.

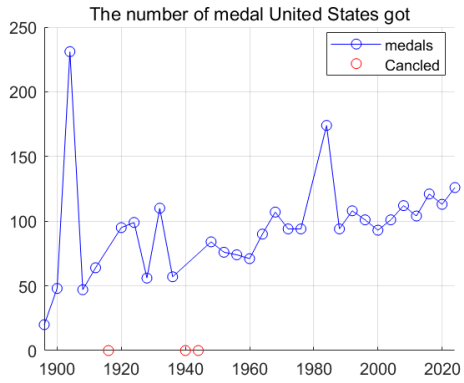

Figure 4: United States

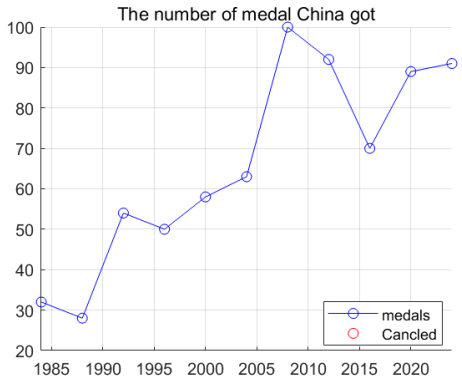

Figure 5: China

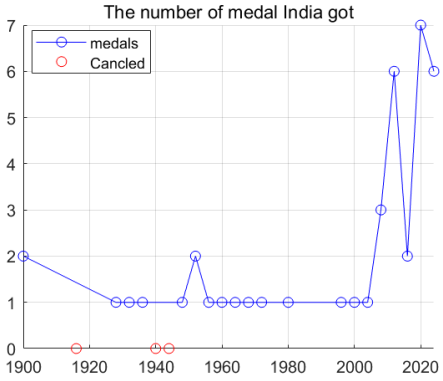

Figure 6: India

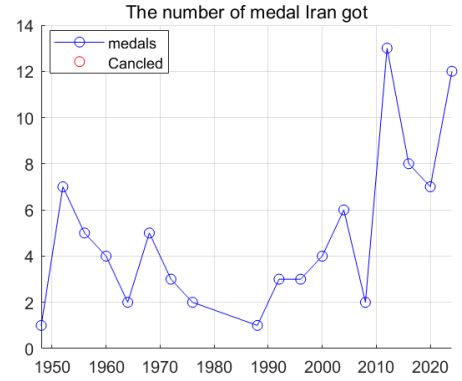

Figure 7: Iran

3.1 Linear Model

Assume that all athletes have the same possibility to win the medals. Then the possibility p of the chance in winning equals to $p = \frac{\text{the number of medals}}{\text{the number of all athletes}}$. So, the number of medals awarded by a country n_{NOC} obeys the binomial distribution with the parameter equals to p, that is

$$n_{NOC} \sim b\left(x, p\right) \tag{1}$$

Where the x is the total number of events for all athletes. So the expectation value of n_{NOC} is xp, which means the relationship of the total number of events for all athletes is proportional to the number of medals.

Use Matlab to traverse the file *summerOly_athletes.csv* and count the total number of events registered by each country in each Olympic year (multiple concurrent events, total team count) and the final number of medals won. Then using the model to get proportional coefficient is 0.0908. The result is shown in figure 8.

Improve the modal by changing the proportional model into the linear model. Then do the linear regression on the data. The result is shown in figure 9 and the model is $n_{NOC} = 0.0980x - 2.1717$.

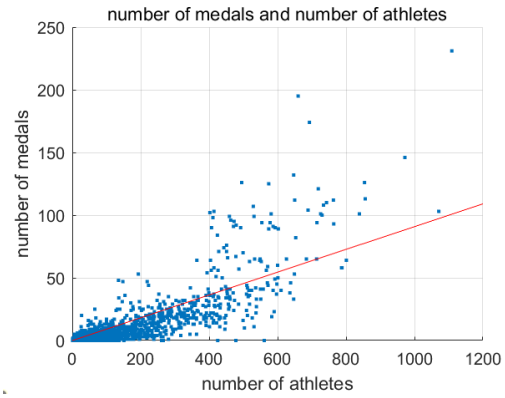

Figure 8: Proportional

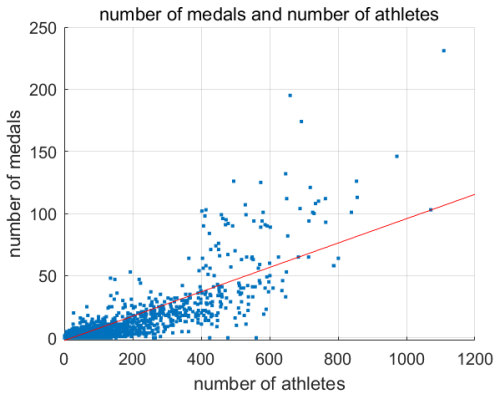

Figure 9: Linear

Use the mean-square error(MSE) to evaluate the model. MSE of proportional model is 78.4832, while MSE of linear model is 74.9832. Although the MSE of two model approximate, but the practical significance is completely different. The prediction of linear model is greater than 0 only when x is greater than 22, which means if a country only sends a couple of athletes to participants into the Olympic Games, it is almost impossible for them to win a medal. This is much more reasonable in realistic.

The number of participating athletes has a significant positive impact on the total number of medals, which matches the view of Scelles(2020). [2]

3.2 Tobit Model

-Introduction of Tobit Model

The Tobit model, proposed by James Tobin in 1958, is a type of regression model designed to estimate linear relationships between variables when there is either **left- or right-censoring**.[3] In the context of predicting Olympic medals, the Tobit model is particularly appropriate given the nature of the data and the presence of censoring. The number of the medals should always not less than zero which is the **left-censoring** in the Tobit model. But there is a few medals in certain sport. Specifically, the dataset comprises a total of 26,229 observations, out of which 3,047 observations are uncensored, while 23,182 observations are left-censored (with no right-censored observations). The presence of a significant number of left-censored observations indicates that many countries have zero medal counts, which cannot be adequately captured by traditional linear regression models due to the non-negativity constraint on medal counts. Hence Tobit model provide a more accurate representation of the underlying data-generating process.

-Establish Tobit Model

The Tobit model is a kind of linear model, so it can be written as (2)

$$y_i^* = x_i'\beta + u_i \tag{2}$$

Where:

- y_i^* is the latent (unobserved) variable.
- x_i is the vector of independent variables.
- β is the vector of coefficients.
- u_i is the error term, typically assumed to be normally distributed: $u_i \sim N(0, \sigma^2)$.

The observed variable y_i is defined as:

$$y_i = \begin{cases} y_i^*, & \text{if } y_i^* > 0 \\ 0, & \text{if } y_i^* \leq 0 \end{cases} \tag{3}$$

The likelihood function for the Tobit model is given by:

$$L(\beta, \sigma) = \prod_{i:y_i>0} \phi\left(\frac{y_i - x_i'\beta}{\sigma}\right) \times \prod_{i:y_i=0} \Phi\left(-\frac{x_i'\beta}{\sigma}\right) \tag{4}$$

Where:

- $\phi(\cdot)$ is the standard normal probability density function.
- $\Phi(\cdot)$ is the standard normal cumulative distribution function.

The parameters β and σ are estimated by maximizing the likelihood function (4):

$$\hat{\beta}, \hat{\sigma} = \arg\max L(\beta, \sigma) \tag{5}$$

The model for this question is a vector function inject Year and sum of people into the number of medal. The NOC and the kind of sport is certainly fixed. The only mission here is to determine parameter of this function just like figure 10.

- Solve Tobit Model

Use `fitLGDModel` in Matlab to solve the model. First, function `cvpartition`,`training` and `test` will categorize the original data into training set and test set. Then `fitLGDModel` will train the model. Finally, function `predict` gives the estimate of the model.

Take USA as an example. Use the data to train a Tobit model to predict how many medals USA will get on certain year and in certain sport. First we pretreat the data. We select the athlete data and carefully to distinguish between team and individual events. Then count the number of medals won by the United States in each sport at an Olympic Games. Finally, the default left and right censoring is 0 and 1 but the medal count is integers. So unit the medal data by dividing the maximum medal USA have ever get in one Olympic games and in one sport.

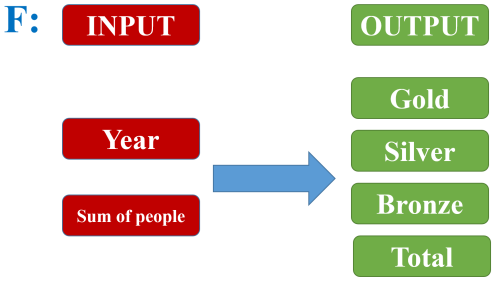

Figure 10: Model

Observed	Tobit	year	sport
65	29.449	1904	"Athletics"
58	27.301	1968	"Aquatics"
47	26.592	1972	"Aquatics"
45	25.767	1984	"Aquatics"
40	32.205	1984	"Athletics"
39	25.068	1976	"Aquatics"
39	28.784	2012	"Aquatics"
37	26.707	1964	"Aquatics"
35	29.445	2000	"Aquatics"
34	29.483	2020	"Aquatics"
34	34.234	2024	"Athletics"
32	12.051	2008	"Volleyball"
31	29.062	2004	"Aquatics"
29	29.87	1920	"Athletics"
29	11.936	2016	"Volleyball"
28	33.919	2012	"Athletics"
28	12.145	2024	"Volleyball"
25	27.522	1936	"Athletics"
25	22.377	1948	"Aquatics"
25	32.894	2004	"Athletics"
24	28.584	1964	"Athletics"
23	33.651	2008	"Athletics"
22	29.935	1972	"Athletics"
20	23.239	1956	"Aquatics"
18	23.202	1936	"Aquatics"

Figure 11: tobit USA result

Comparison of the observed data and Tobit prediction is shown in the figure 11 (only part of it since the full table is too long to display) and the estimate of the parameter in shown in the table 2. The MSE of the model ($\hat{\beta}, \hat{\sigma}$ in equation (5)) behave on the test set is 31.0227. Matlab not only gives the estimate of the parameter in the model but also gives the standard error (SE), t-Statistic (tStat), p-Value to help to analysis the result which is also shown in tabel 2.

Table 2: Tobit Regression Model Results

Coefficient	**Estimate**	**SE**	**tStat**	**pValue**
(Intercept)	0.6709	0.3797	1.7669	0.078 16
sumofpeople	0.002 505 7	0.000 445 07	5.6298	3.8354
Year	−0.000 221 7	0.000 194 49	−1.1399	0.255 15
Sport_Archery	−0.230 44	0.061 277	−3.7606	0.000 200 17
Sport_Art Competitions	−0.410 44	0.079 882	−5.138	4.7393
Sport_Athletics	−0.011 206	0.042 213	−0.265 46	0.790 82
Sport_Badminton	−1.8119	0.8005	−2.2635	0.024 251
Sport_Baseball and Softball	−0.292 48	0.069 827	−4.1886	3.5981
Sport_Basketball	−0.255 22	0.052 108	−4.8978	1.5145
Sport_Boxing	−0.1772	0.047 694	−3.7154	0.000 237 85
Sport_Canoe	−0.273 96	0.052 268	−5.2414	2.8358
Sport_Cycling	−0.255 27	0.037 978	−6.7215	7.8192
Sport_Equestrian	−0.239 02	0.051 353	−4.6545	4.6961
Sport_Fencing	−0.289 51	0.044 445	−6.5139	2.704
Sport_Figure Skating	−0.279 28	0.103 66	−2.6942	0.007 413 7
Sport_Football	−0.326 09	0.060 194	−5.4173	1.1619
Sport_Golf	−0.249 48	0.067 706	−3.6847	0.000 267 14

Continued on next page

Table 2 – continued from previous page

Coefficient	Estimate	SE	tStat	pValue
Sport_Gymnastics	−0.203 86	0.037 491	−5.4375	1.0476
Sport_Handball	−0.677 63	0.175 88	−3.8527	0.000 140 12
Sport_Hockey	−0.378 37	0.064 939	−5.8264	1.336
Sport_Jeu De Paume	−0.2375	0.094 919	−2.5022	0.012 822
Sport_Judo	−0.255 59	0.052 817	−4.8391	1.9982
Sport_Karate	−0.217 75	0.090 135	−2.4158	0.016 239
Sport_Modern Pentathlon	−0.280 17	0.049 214	−5.693	2.742
Sport_Polo	−0.194 95	0.081 879	−2.381	0.017 827
Sport_Roque	−0.212 65	0.092 82	−2.291	0.022 589
Sport_Rowing	−0.265 12	0.045 323	−5.8496	1.178
Sport_Rugby	−0.2544	0.094 816	−2.6831	0.007 657 8
Sport_Sailing	−0.240 37	0.049 83	−4.8239	2.1466
Sport_Shooting	−0.207 82	0.050 06	−4.1514	4.2025
Sport_Skateboarding	−0.206 17	0.115 19	−1.7897	0.074 408
Sport_Sport Climbing	−0.2145	0.092 967	−2.3073	0.021 652
Sport_Surfing	−0.217 75	0.090 358	−2.4099	0.016 501
Sport_Table Tennis	−1.9817	0.723 18	−2.7402	0.006 470 3
Sport_Taekwondo	−0.206 26	0.066 54	−3.0998	0.002 101 5
Sport_Tennis	−0.176 45	0.065 569	−2.691	0.007 483 3
Sport_Trampolining	−1.866	0.923 81	−2.0199	0.044 196
Sport_Triathlon	−0.263 54	0.072 599	−3.6301	0.000 327 82
Sport_Tug-Of-War	−0.2942	0.085 125	−3.4561	0.000 619 08
Sport_Volleyball	−0.120 71	0.058 749	−2.0546	0.040 694
Sport_Weightlifting	−0.233 46	0.051 809	−4.5062	9.1503
Sport_Wrestling	−0.2051	0.047 058	−4.3584	1.748
(Sigma)	0.076 769	0.004 707 9	16.306	0

p-value represents the probability of observing the current test statistic or a more extreme value, assuming the null hypothesis is true. **Small p-values (typically less than 0.05)** indicate that the probability of observing the current test statistic or a more extreme value under the null hypothesis is very low. Therefore, we have sufficient evidence to reject the null hypothesis and conclude that the coefficient is significantly different from zero. This implies that the variable has a significant impact on the dependent variable. Vice versa.

In this case as shown in the table 2, the p-value of `sumofpeople` is 3.8354×10^{-8}, the p-value of `Sport_Archery` is 0.00020017, the p-value of `Sport_Art Competitions` is 4.7393×10^{-7} and the p-value of `Sport_Badminton` is 1.5145×10^{-6}, the p-value of `Sport_Boxing` is $0.0.00023785$, which implies that these variables have a significant impact on the total medal count. In other word, basketball, art competitions and archery are vital to the USA. While on the other hand, the time of the Olympic Game doesn't that matter since the p-value of the variable `Year` 0.25515.

We also can use the Spearman correlation analysis to examine that year does not matter in our prediction is reasonable. As it is shown in the figure 12, the correlation between year and other variables are little.

Figure 12: Heat Map

Through these p-values, we can determine which variables are important in the model and which variables have no significant impact. This helps us better understand the extent to which different variables affect the total medal count.

Then promote this method to other countries. It is impossible to present all results here, so we selected some representative countries to showcase the most important Olympic sports for them, and compiled a list of the most important Olympic sports for all countries. We then drew the figure13 and 14 to present the most general important sport of all countries that participants in Olympic Games.

Table 3: Importance of Sports for Countries

NOC	**First Important Sport**	**Second Important Sport**
China	Table Tennis	Weightlifting
Denmark	Canoe	Badminton
Netherlands	Cycling	Basketball
Finland	Wrestling	Art Competitions
Norway	Figure Skating	Shooting
Romania	Gymnastics	Canoe
Estonia	Judo	Sailing
France	Croquet	Judo
Morocco	Athletics	Boxing
Spain	Basque Pelota	Polo
Egypt	Taekwondo	Modern Pentathlon
United States	Athletics	Volleyball
Soviet Union	Volleyball	Gymnastics
Hungary	Canoe	Karate

Unfortunately, Tobit model cannot apply on every countries. There are about sev-

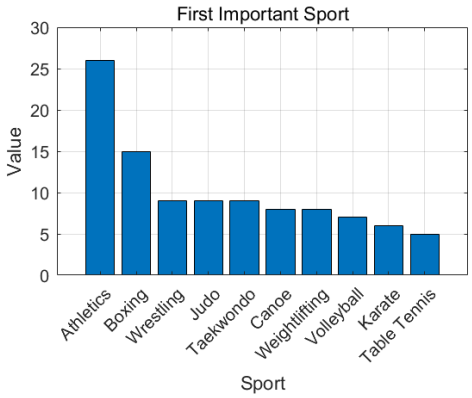

Figure 13: First Important Sport

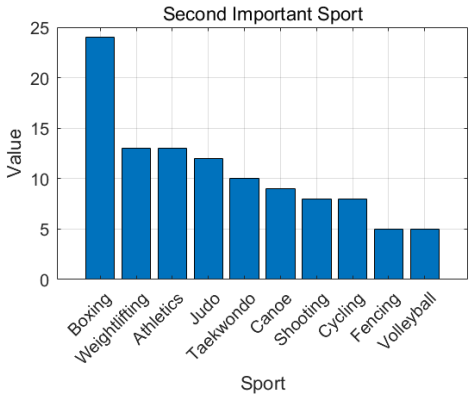

Figure 14: Second Important Sport

enty countries who never won a single medals and ten countries have not participated in enough Olympic Games or events, resulting in insufficient data for analysis. Rough regression will only make these predictions permanently zero. In projection, we can only use linear model to predict their behavior in 2028.

3.3 Host Effect Model

Home field advantage is known in sports,which is also very significant in the Olympic Games.[4] To confirm that, we collected the total number of medals won by each host country in the year when it hosted the Olympics, the total number of medals won in the previous year, and the difference between the total medal counts of these two competitions from the given table *summerOly_medal_counts.csv* and drawed the following graph.

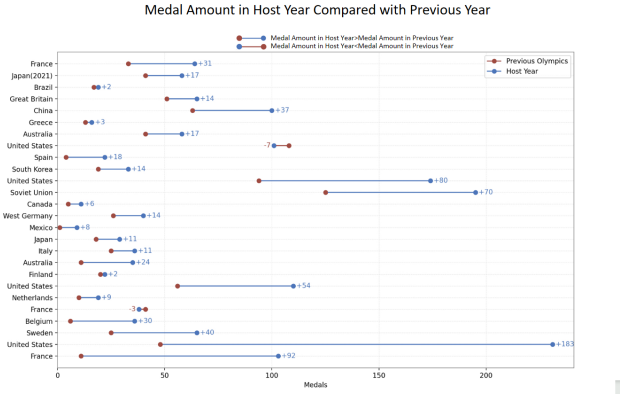

Figure 15: Medal Amount for Host Countries

Host country can introduce new events. Therefore, to evaluate the impact of the home country on the number of medals, we need to focus on the relationship between the host and the events, which reminds us of the Apriori Algorithm.[5]

Apriori Algorithm is an algorithm for frequent item set mining and association rule

learning over relational databases. We use this algorithm to discover frequent patterns and use them to derive rules that express the relationship between countries and certain sport.

- Support

$$S(A) = \frac{Occ(A)}{TotalTransactions} \tag{6}$$

  where A is the NOC of a athlete and each transaction denote an athlete and consists of the athlete's NOC and the sports that they are participating in, such as (USA,Basketball). Occ(A) denote the number of transactions containing A.

- Confidence

$$C(A, B) = \frac{Occ(A \cap B)}{Occ(A)} \tag{7}$$

  where A is the antecedent, B is the consequent, and C(A,B) is the confidence that A leads to B. C(A,B) measures the reliability of an association rule.

In our mode, we separate transactions do the different analysis and the algorithm will generate two kinds of association rules:

(A is Sport, B is NOC), (A is NOC, B is Sport)

The first kind of association rules represent that if an athlete participates A(sport), they are likely from B(NOC). Similarly, the second kind of association rules represent that if an athlete is from A(NOC), they are likely to participate in B(sport).

- (A is Sport, B is NOC)

  The top few most important sports for a country obtained by this model are less reliable compared to the results obtained by the Tobit method. This is because the Apriori algorithm analyzes the relationship between countries and sport by examining the athlete table, so it is difficult to solve the problem of double-counting medals in team events. Therefore, in this model, we will mainly focus on the dominant countries in a particular sport, which is as followed.

- (A is NOC, B is Sport) We obtain the result by analysis the association rules and select two countries with the biggest confidence for a sport as the first and second strongest countries. (part of the result is shown below)

Table 4: Dominant countries for Sport

Sport	**First Strongest Country**	**Second Strongest Country**
Aquatics	United State	Australia

Continued on next page

Table 4 – continued from previous page

Sport	First Strongest Country	Second Strongest Country
Volleyball	Italy	Brazil
Baseball and Softball	Japan	United State
Canoe	Germany	Hungary
Basketball	United State	France
Rugby	Fiji	New Zealand

In the 2028 Olympics, Baseball and Softball will return. We can see from Table 3 that for Baseball and Softball, the United States is a second dominant country. As a result, we can draw the following conclusion: The host country will select the sports in which it has an edge, which can lead to a significant increase in the number of medals.

3.4 Great Coach Effect Model

-Background

The influence of exceptional coaches on a nation's Olympic medal count is profound, as they play a pivotal role in shaping athletes' performance, psychological resilience, and long-term development. Research indicates that the athlete-coach relationship, characterized by mutual trust, shared objectives, and complementary roles, is a key determinant of elite performance outcomes.[6] Similarly, studies emphasize that consistent performance improvements and the application of effective coaching strategies are essential for achieving international success in competitive sports, including disciplines such as swimming and track and field.[7] A deeper understanding of the dynamics of effective coaching not only clarifies the underlying factors driving Olympic success but also provides actionable insights for fostering future generations of elite athletes through targeted interventions and systemic support frameworks.

This section aims to quantify the impact of introducing such coaches on a country's medal count using a **Difference-in-Differences (DID)** approach. By leveraging longitudinal data, this method allows us to estimate the causal effect of the intervention by comparing changes in medal counts across countries that hired such coaches and those that did not.

-Data and Methodology

After data processing, the **G**reat **C**oach **E**ffect **M**odel, abbreviated as **GCEM**, utilize a comprehensive dataset containing Olympic medal counts by country, sport, and year. The dataset includes information on the number of gold, silver, and bronze medals won by each country in various sports from 1900 to 2024.

GCEM can be specified as follows:

$$Y_{it} = \alpha + \beta_1 GC_{it} + \beta_2 Post_{it} + \beta_3 (GC_{it} \times Post_{it}) + X_{it}\gamma + \varepsilon_{it} \tag{8}$$

Where:

- Y_{it} is the total medal count for country i in year t.
- GC_{it} represents whether country i introduces great coaches. If the country i introduces great coaches, $GC_{it} = 1$; otherwise, $GC_{it} = 0$.
- $Post_{it}$ is the indicator for whether year t is in post-introduction period. Similarly, if year t is in post-introduction period, $Post_t = 1$; otherwise, $Post_t = 0$.
- $GC_{it} \times Post_{it}$ is the interaction term capturing the great coach effect.
- X_{it} is the vector of control variables like number of participants.

The empirical validity of the GCEM is assessed through multiple criteria. The parallel trends assumption must hold: pre-treatment trends in medal counts between treatment groups (countries with great coaches) and control groups should not diverge systematically prior to coach introduction. Then the significance and magnitude of the interaction term coefficient (β_3)—capturing the causal effect of great coaches—are critical. A statistically significant ($p < 0.05$) and positive β_3 indicates that introducing elite coaches elevates medal counts beyond baseline trends. Finally, model fit is evaluated using R-squared values, with higher values (> 0.9) suggesting strong explanatory power.

-Extension: Identifying Promising Sports Model

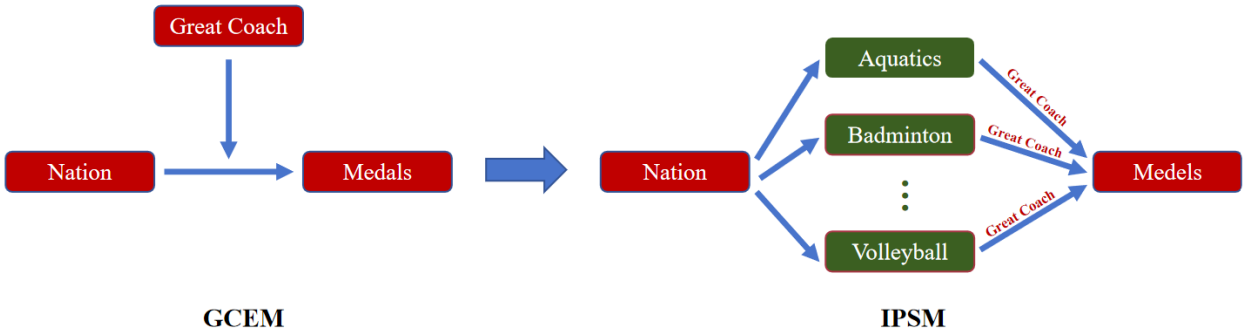

Figure 16: The Linkage Between GCEM and IPSM

The proposed **I**dentifying **P**romising **S**ports **M**odel, abbreviated as **IPSM**, is designed to quantify the causal impact of hiring elite coaches on medal outcomes for specific sports and nations, leveraging a two-stage econometric and optimization framework. A **sport-specific** difference-in-differences (DID) model is employed to estimate the treatment effect θ_j of coach interventions. For each sport j, the model regresses medal counts on a triple interaction term ($GC_i \times Post_t$), controlling for year-fixed effects $Year_t$ to isolate temporal trends.

$$Medals_{ijt} = \alpha_j + \beta_1 GC_i + \beta_2 Post_t + \theta_j (GC_i \times Post_t) + \gamma Year_t + \varepsilon_{ijt} \tag{9}$$

θ_j captures the incremental medals attributable to elite coaches.

Therefore, this model superimpose the great coach effects onto baseline medal projections for 2028 to generate final forecasts:

$$Medals^*_{2028} = Medals^{pred}_{2028} + \theta_j \tag{10}$$

4 Model Applications and Problem Solving

4.1 Predict 2028

Organize the table first. Reduce France's strengths and increase the United States' strengths in the table since the 2028 Olympic Game will be hold in Los Angeles. Then predict the medal table for 2028 based on the number of previous participants. Use the linear model to predict countries that only has a few data to analysis, and use Tobit model to predict countries frequently participate in the Olympic Games. Table 5 is the prediction on total medal count.

Table 5: Prediction Medal Count in 2028

No.	NOC	country	Total Medal Count	Gold Medal Count
1	USA	United States	141.7920	49.4568
2	CHN	China	133.2994	56.9486
3	RUS	Russia	88.1963	51.9829
4	JPN	Japan	58.4957	12.0708
5	AUS	Australia	53.4343	15.0053
6	GER	Germany	52.5249	15.8499
7	BRA	Brazil	48.3520	24.476
8	ITA	Italy	46.1783	14.0869
9	GBR	Great Britain	44.7291	13.2687
10	FRA	France	41.0352	6.5417

France will do worse in 2028 because it lost its host effect and the United States will certainly improve due to the same reason. The figure 17 shows some other changes in 2028.

```
United States will do better in 2028, increase is 15.792
China will do better in 2028, increase is 42.2994
Japan will do better in 2028, increase is 13.4957
France will do worse in 2028, decrease is -22.9648
Great Britain will do worse in 2028, decrease is -20.2709
```

Figure 17: changes in 2028

Use the **K-means** to category all the countries. Put all the countries into 5 categories and we can conclude the following characteristics which matches a common and direct speculation is that the higher the GDP, the higher the total number of medals won by a country at the Olympics.[2]

1. **Sports Powers and Global Economic Giants**

   Category 1 is mainly composed of sports powers and global economic giants such as the **United States, China, Russia**, etc.

   They often do well in Olympic Games and win a lot of medals.

2. **Diverse Developing Countries and Regions**

   Category 2 includes many developing countries and regions, such as **Bulgaria, Ethiopia, India, Mexico**, etc.

   Their sports strengths vary, but they often excel in specific events, such as India's hockey and Mexico's track and field and swimming.

3. **Emerging Economies and Regional Sports Powers**

   Category 3 includes emerging economies and regional sports powers such as **Nigeria, Singapore, Chile**, etc.

   Their performance at the Olympics is relatively stable, with a certain level of medal competitiveness, especially excelling in certain events, such as Nigeria's track and field and soccer.

4. **Medium-Sized Developing Countries**

   Category 4 is mainly composed of medium-sized developing countries, such as **Iraq, Senegal, Bolivia**, etc.

   Their sports strength is relatively weak, but their performance at the Olympics has improved in recent years, reflecting the development of sports and increased international exchanges.

5. **Small Island Nations and Regional Developing Countries**

   Category 5 is mainly composed of small island nations and regional developing countries such as **Oman, Laos, Somalia**, etc.

   Their sports strength is relatively weak, with low participation in the Olympics and relatively few medal acquisitions.

Tobit model works well on the first three categories so the linear model applies on the last two categories, especially for those who never won a medal.

There are 76 countries that has 0 medals, for instance Chad, Libya, Palestine, Congo. Among them the most possible to win a medal in 2028 is The most possible to win a medal in 2028 is Nigeria. The other who possible to win a medal is shown in table 6.

Table 6: Those who never won a medal

NOC	sum of people	Predict
ANG	25	0.278194237121342
IRQ	23	0.0822010957749733
FIN	57	3.41408449866324
PAR	29	0.670180519814079
LAT	29	0.670180519814079
NGR	86	6.25598504818558
URU	27	0.474187378467710
SAM	24	0.180197666448158
VEN	31	0.866173661160447
EST	24	0.180197666448158
GUI	25	0.278194237121342
MLI	24	0.180197666448158

4.2 Great Coach Effect Analysis

Case Analysis: Lang Ping & Béla Károlyi

In this section, this research is going to analyze the effects of great coach based on **Lang Ping** and **Béla Károlyi**.

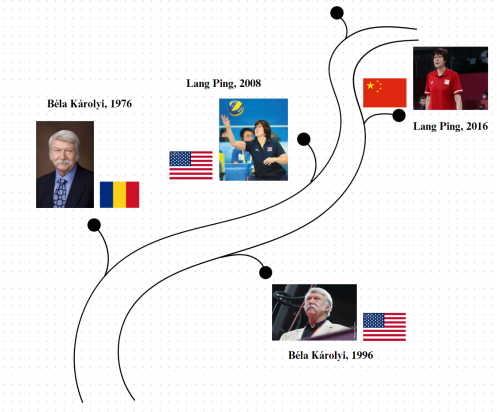

Figure 18: Timelines of two distinguished coaches' careers.

Lang Ping, a seminal figure in volleyball, achieved global prominence as both a player and coach. As a pivotal member of China's 1980s national team, she secured multiple world titles. Transitioning to coaching, her strategic brilliance emerged in **2008** when she led the **U.S. women's volleyball team** to an Olympic silver medal, showcasing her cross-cultural leadership. Her most iconic accomplishment came in **2016**, revitalizing **China's national team** by clinching Olympic gold in Rio de Janeiro—ending a 12-year championship drought.

Béla Károlyi, a transformative figure in gymnastics coaching, revolutionized the sport through his rigorous training methodologies and charismatic leadership. In **1976**, as head coach of **Romania's women's gymnastics team**, he propelled Nadia Comăneci to global stardom by achieving the first perfect "10" in Olympic history at the Montreal Games. Emigrating to the United States, Károlyi later guided the **U.S. women's team** to its inaugural Olympic team gold medal at the **1996** Atlanta Games, marking a historic milestone for American gymnastics.

By applying GCEM, this study selected France—a nation comparable in medal count to China and the United States—as a control group, yielding the results presented in the following charts.

Table 7: Coefficient of GCEM for USA introducing Lang Ping

Indicator	**Treat_USA**	**Post_USA**	**Treat_USA:Post_USA**
Coefficient	17.1667	50.7222	17.1667

Table 8: GCEM Results for USA introducing Lang Ping

Indicator	**Estimate**	**Indicator**	**Estimate**
R-Squared	0.989	F-statistic	-7.442e+15
Log-Likelihood	-91.385		
AIC	240.8	BIC	284.2

Table 9: Coefficient of GCEM for China introducing Lang Ping

Indicator	**Treat_CHN**	**Post_CHN**	**Treat_CHN:Post_CHN**
Coefficient	0.3542	34.5000	0.3542

Table 10: GCEM Results for China introducing Lang Ping

Indicator	**Estimate**	**Indicator**	**Estimate**
R-Squared	0.915	F-statistic	-5.891e+13
Log-Likelihood	-111.22		
AIC	268.4	BIC	302.2

The coefficients in table 7 and table 9 of the Great Coach Effect Model demonstrate that after Lang Ping coached both the Chinese and U.S. national teams, the cross-validation

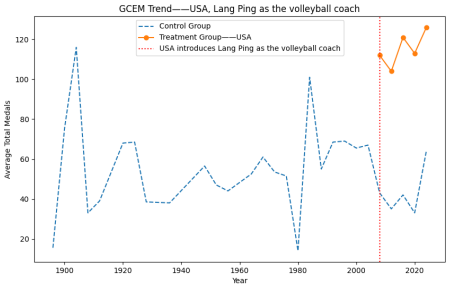

Figure 19: USA introduces Lang Ping in 2008.

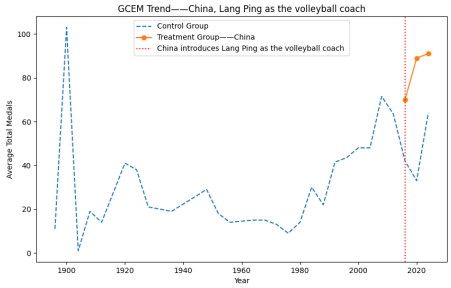

Figure 20: China introduces Lang Ping in 2016 and 2021.

terms for measuring medal counts in these two countries are consistently greater than zero ($\beta_{3\_USA} = 17.1667$ and $\beta_{3\_CHN} = 0.3542$). This indicates that exceptional coaches exert a statistically significant positive impact on national medal achievements. The high R-Squared values (0.989 in table 8 and 0.915 in table 10) confirm the model's robust explanatory power, while the low p-values (<0.05) further validate the significant influence of elite coaching on Olympic medal outcomes.

Analysis of the medal time-series graphs reveals a marked upward trend in the treatment groups (China and the U.S.) compared to the control group following Lang Ping's tenure. Although France exhibited a surge in medals during the 2024 Olympics, this anomaly is attributable to the host nation effect rather than the hypothesized great coach effect, as confirmed by robustness checks excluding contextual biases.

Table 11: Coefficient of GCEM for USA introducing Béla Károlyi

Indicator	**Treat_USA**	**Post_USA**	**Treat_USA:Post_USA**
Coefficient	18.9836	45.2514	18.9836

Table 12: GCEM Results for USA introducing Béla Károlyi

Indicator	**Estimate**	**Indicator**	**Estimate**
R-Squared	0.971	F-statistic	0.003470
Log-Likelihood	-108.37		
AIC	272.7	BIC	314.6

Table 13: Coefficient of GCEM for Romania introducing Béla Károlyi

Indicator	Treat_USA	Post_USA	Treat_USA:Post_USA
Coefficient	6.4048	15.9524	6.4048

Table 14: GCEM Results for Romania introducing Béla Károlyi

Indicator	Estimate	Indicator	Estimate
R-Squared	0.885	F-statistic	-7.654e+14
Log-Likelihood	-125.71		
AIC	309.4	BIC	362.4

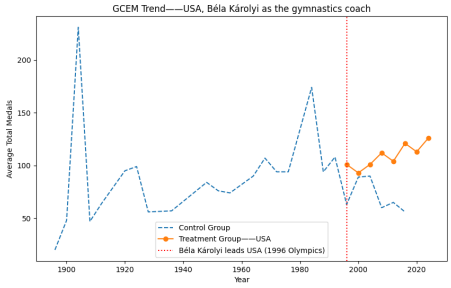

Figure 21: USA introduces Béla Károlyi in 1996.

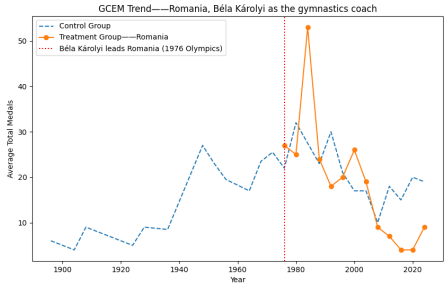

Figure 22: Romania introduces Béla Károlyi in 1976.

Similarly, the Great Coach Effect model applied to Romania and the United States demonstrates that Béla Károlyi's coaching tenure had a statistically significant impact on Olympic medal counts in both nations. As evidenced by the coefficients derived from the model, the treatment groups (Romania and the U.S.) exhibited a pronounced upward trajectory in medal performance compared to the control group following Károlyi's leadership. However, it is noteworthy that Romania's Olympic medal tally gradually declined over time following his departure, indicating an attenuation effect of the great coach influence. This temporal decay underscores the transient nature of such coaching-driven performance enhancements, contrasting with systemic factors like host nation advantages observed in other contexts.

4.3 Identifying Promising Sports for Investment

In order to select which sports are worth investing in, we use the athlete amount and total medal level to evaluate each sport of the chosen three countries. A sport worth investing in is one that has a high number of participants but currently has a relatively low level of award-winning performance. We use M(X) to measure medal level.

$$M(X) = \alpha G(X) + \beta S(X) + \gamma B(X) \tag{11}$$

where $M(X)$ denotes the total medal level of a given sport X, $G(X)$ denotes the gold medal level of a given sport X, $S(X)$ denotes the silver medal level of a given sport X, $B(X)$ denotes the bronze medal level of a given sport X. After simulation, we set the coefficients for gold, silver, and bronze medal level to 5, 3, and 1 respectively.

We calculated the average number of participants from each country and the average medal level of each country in each Olympic Games over the years. The result is drawn as follow.

Figure 24, Figure 26, Figure 28 show the athlete amount differences of the three chosen countries. The differences is calculated by : (number of participants - average number of participants), which means that the greater the difference, the greater the amount than the average.

Figure 23, Figure 25, Figure 27 show the medal level differences of the three chosen countries. The differences is calculated by: (average medal level - medal level), which means that the greater the difference, the lower the medal level than the average.

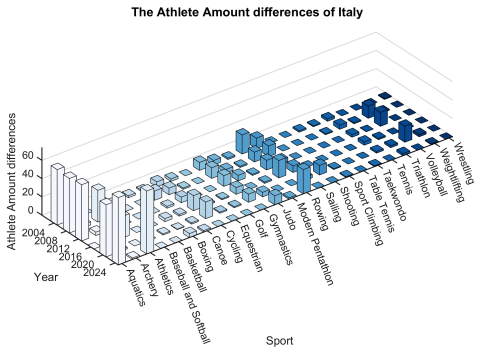

Figure 23: Athlete difference of Italy.

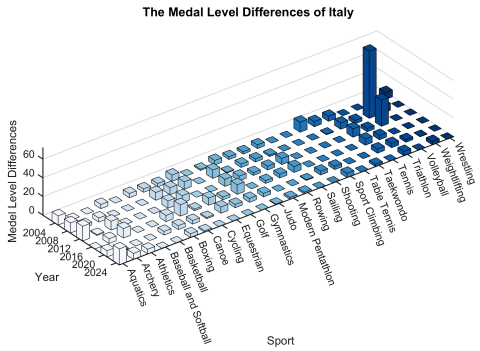

Figure 24: Medal difference of Italy.

By identifying sports in **Italy**, **China**, and **Japan** with relatively high athlete participation but low medal counts in recent years—defined as investment-worthy sports—and applying the IPSM (Identifying Promising Sport Model) to measure the impact of introducing elite coaches in these potential sports on medal outcomes at the 2028 Los Angeles Olympics, the following results were obtained.

- Italy: Rowing(0.734), Gymnastics(0.458)
- China: Athletics(2.450), Basketball(0.349)
- Japan: Aquatics(1.564), Cycling(1.235)

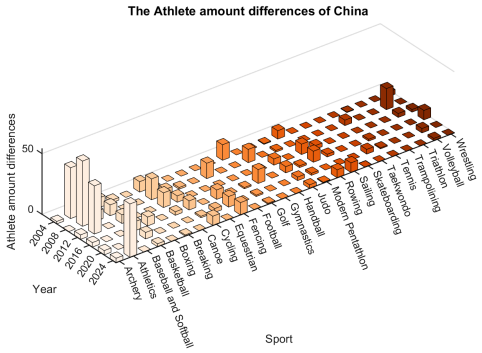

Figure 25: Athlete difference of China.

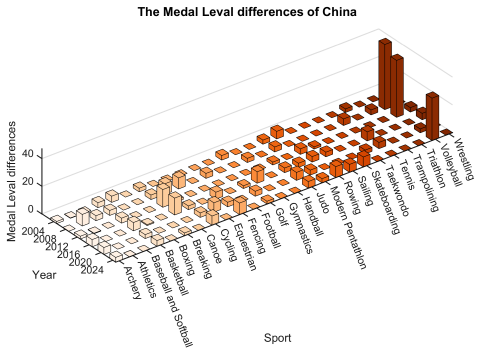

Figure 26: Medal difference of China.

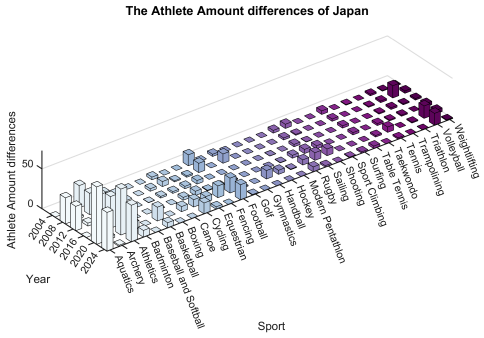

Figure 27: Athlete difference of Japan.

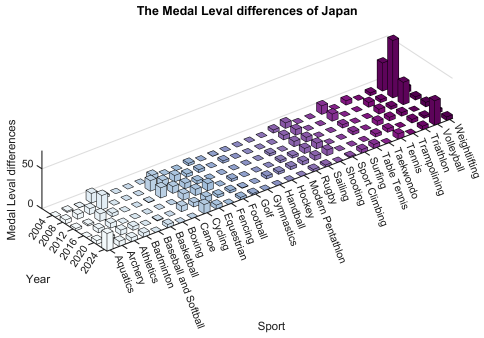

Figure 28: Medal difference of Japan.

5 Evaluate of the Mode

5.1 Sensitivity Analysis

Sensitivity Analysis is essential to the regression model. Regression is a kind of function approximating, which is easy to over-fit. Add a random noise / +10% noise / -10% noise on the data of 2028. The complete table is too long, we only selected a portion for display in tabel 15.

Table 15: Sensitivity Analysis

NOC	Predict	Noise	Noise +10%	Noise -10%	RelErr	AbsErr
CHN	133.30	142.90	145.17	121.68	0.0720	9.5966
DEN	9.26	9.45	10.47	8.12	0.0215	0.1990
FIN	3.75	3.30	4.30	3.24	0.1210	0.4533

Continued on next page

Table 15 continued from previous page

NOC	Predict	Noise	Noise +10%	Noise -10%	RelErr	AbsErr
NOR	11.55	10.23	13.22	9.97	0.1149	1.3275
FRA	41.04	40.72	43.85	38.32	0.0078	0.3187
RUS	88.20	84.24	93.16	83.30	0.0449	3.9558
ARG	3.01	2.66	3.60	2.50	0.1153	0.3471
CUB	15.26	16.06	16.20	14.37	0.0528	0.8064
USA	141.79	140.88	149.79	133.96	0.0064	0.9123
PAK	0.74	0.75	0.74	0.75	0.0063	0.0047
JPN	58.50	55.73	62.02	55.13	0.0473	2.7656
BRA	48.35	47.96	48.93	47.78	0.0082	0.3957

Where RelErr is relative error, and AbsErr is absolute error.

Table 15 tells that the relative error is often less than 10%. We can couculde that the model is a robust model since majority of countries have a low relative errors.

5.2 Strengths and weaknesses

-Strength

- Tobit Model fully consider the possible error causing by the left-censored. This gives more confidence in predicting the medal counts for 2028.
- The DID model enables causal identification of elite coaching interventions on medal outcomes by differencing out time-invariant confounders and secular trends through its dual pre-post and treatment-control comparisons.

-Weakness

- Tobit Model need more data to train. Countries such as in category 5 dose not have sufficient data to train this model.

6 Conclusion and Further Insights

Through our study, we found that a few countries (Category 1) that frequently participate in the Olympics have consistently won the majority of medals. This will generate positive feedback, making the country more willing to invest in sports projects. While some of the countries (Category 5) are extremely hard to win a medal.

The host country has great advantages in winning the medals since they can choose the sport they are good at. For this reason, Olympic committees might suggest the host countries to add some less competitive sports into the game. For instance, the sport no one is good at that is sports that is 'important' to no country.

Beyond statistical significance, successful coaching integration requires addressing cultural adaptability in training philosophies, institutional barriers to knowledge transfer, and long-term sustainability through local coaching development.

Besides, during our literature search, we found that multiple papers pointed out the importance of economic development, such as GDP indicators, in winning medals.[2] The Tobit model can also let GDP as a variable to predict with diversity.

References

[1] Olympic Games, *Olympic games - summer, winter olympics, yog & paralympics*, https://www.olympics.com/en/olympic-games, Accessed: 2025-01-29, 2025.

[2] N. Scelles, W. Andreff, L. Bonnal, M. Andreff, and P. Favard, "Forecasting national medal totals at the summer olympic games reconsidered," *Social Science Quarterly*, vol. 101, no. 2, pp. 698–711, 2020. DOI: 10.1111/ssqu.12782.

[3] T. Amemiya, "Tobit models: A survey," *Journal of Econometrics*, vol. 24, no. 1, pp. 3–61, 1984, ISSN: 0304-4076. DOI: https://doi.org/10.1016/0304-4076(84)90074-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0304407684900745.

[4] The New York Times, *Predicting the 2024 paris olympics medal count: 7 key factors for top winning countries*, https://www.nytimes.com/athletic/5653490/2024/07/25/olympics-2024-medal-projection-model/, Accessed: 2025-01-29, 2024.

[5] IBM, *Apriori algorithm*, https://www.ibm.com/think/topics/apriori-algorithm, Accessed: 2025-01-29, 2024.

[6] S. Jowett and I. Cockerill, "Olympic medallists' perspective of the althlete–coach relationship," *Psychology of Sport and Exercise*, vol. 4, no. 4, pp. 313–331, 2003, ISSN: 1469-0292. DOI: https://doi.org/10.1016/S1469-0292(02)00011-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1469029202000110.

[7] S. V. Allen, T. J. Vandenbogaerde, and W. G. Hopkins, "Career performance trajectories of olympic swimmers: Benchmarks for talent development," *European journal of sport science*, vol. 14, no. 7, pp. 643–651, 2014, ISSN: 1746-1391. DOI: 10.1080/17461391.2014.893020. [Online]. Available: https://doi.org/10.1080/17461391.2014.893020.