|  |  |  |
| --- | --- | --- |
| **Problem Chosen** C | **2021 MCM/ICM Summary Sheet** | **Team Control Number** 2522907 |

**Olympic medal prediction model based on Zero-Inflated Poisson**

**Summary**

|  |
| --- |
| **说明：**  **本Word模版的版本编号是0.1版，是以美赛的标准创建的模版，也适用于其他用英文写作的数学建模比赛，例如亚太赛和小美赛。模版的使用方法可以查看下面这个视频：**  **数学建模清风——论文排版教程**  **https://www.bilibili.com/video/BV1Ci4y1c7Ld**  **未来当发现模版中的问题或者比赛要求有更新时，我会发布更新后的新版本。大家可以在公众号《数学建模学习交流》后台发送“美赛论文模版”获取最新的模版的信息。只要本文档的版本号和公众号后台最新的版本号一致，则说明你下载的是最新版本。**   * **不同的美赛特等奖论文写作风格差别较大，我这里给的模版是提取出来的通用框架，大家可以根据自己的需要进行修改。** * **红色字体标注的内容是对这个部分的解释，看完后请删除。** * **黄色高亮的是需要大家修改的内容，大家修改后请取消高亮显示。**     这里写美赛的摘要，美赛中摘要格外重要！！官方对于摘要的说明：  *The summary is an essential part of your MCM/ICM paper and should appear as the first page of your solution report. The judges place considerable weight on the summary, and winning papers are often distinguished from other papers based on the quality of the summary.*  翻译：摘要是美赛论文的重要组成部分，应该放在论文的第一页展示。评委们对摘要相当重视，获奖论文和其他论文的区别往往就在于摘要的好坏！  论文摘要的写法可以参考这个视频：  <https://www.bilibili.com/video/BV1Na411w7c2/>  标题下面的Summary可要可不要，如果你觉得你摘要写的不够多，页面下方留白太大的话就加上，这样看起来稍微好看点。  注意：美赛要求英文写作，很多同学英文写作水平不够，可以先写成中文的论文，然后再进行翻译！翻译可以使用有道翻译或者谷歌翻译，翻译后一定要人工修改，机器翻译的很生硬，对专有名词的翻译也不是很准确。 |

**Keywords:** keyword1; keyword2; keyword3; keyword4

Contents最后记得更新整个目录

[1 Introduction 3](#_Toc58786693)

[1.1 Problem Background 3](#_Toc58786694)

[1.2 Restatement of the Problem 3](#_Toc58786695)

[1.3 Literature Review 3](#_Toc58786696)

[1.4 Our Work 4](#_Toc58786697)

[2 Assumptions and Justifications 5](#_Toc58786698)

[3 Notations 6](#_Toc58786699)

[4 The name of model 1 6](#_Toc58786700)

[4.1 Data Description 7](#_Toc58786701)

[4.2 The Establishment of Model 1 7](#_Toc58786702)

[4.3 The Solution of Model 1 7](#_Toc58786703)

[5 The name of model 2 8](#_Toc58786704)

[6 The name of model 3 8](#_Toc58786705)

[7 Sensitivity Analysis 8](#_Toc58786706)

[8 Model Evaluation and Further Discussion 8](#_Toc58786707)

[8.1 Strengths 8](#_Toc58786708)

[8.2 Weaknesses 9](#_Toc58786709)

[8.3 Further Discussion 9](#_Toc58786710)

[9 Conclusion 9](#_Toc58786711)

[References 10](#_Toc58786712)

[Appendices 11](#_Toc58786713)

# Introduction

## Problem Background

The Olympic Games have long been the focus of global attention. It is the world's largest sporting event, which not only reflects the sports competition between countries, but also represents the excellent Olympic spirit. Since the first Summer Olympics in 1896, the competition events have become more and more diverse, and the number of participants has also grown rapidly. Olympic medals carry the hard work of athletes and represent personal glory, national spirit and national strength. As we look forward to the 2028 Los Angeles Summer Olympics, the public's call for medal prediction is getting louder and louder. Therefore, more and more scholars have begun to join the research on mathematical models for medal prediction.

In recent years, this research has gradually heated up, and a variety of prediction models have been proposed one after another, and its theoretical methods and model construction have been rapidly improved and developed. However, traditional methods for predicting medal results often rely on simple historical averages or expert opinions, which may ignore some important factors, such as home advantage, the evolution of sports disciplines, and the influence of coach expertise. In addition, in recent decades, the role of "great coaches" in shaping the success of national teams has become an increasingly important factor, and successful coaches have a profound impact on athletes' performance, thereby affecting the number of medals.

Against this backdrop, this study aims to develop a comprehensive model that takes into account the broader dynamics that influence a country’s performance. By integrating historical medal data, athlete performance indicators, and event details from past Olympic Games, the model will provide insights into which countries are most likely to improve their rankings, which countries are likely to perform worse, and the likelihood of emerging countries winning Olympic medals for the first time.

In addition, this analysis will explore the relationship between Olympic events and medal distribution, identify countries’ key events, and understand how host countries’ event selection affects medal outcomes. Ultimately, this study aims to provide a more nuanced understanding of Olympic medal trends and offer valuable recommendations for countries aiming to optimize their Olympic strategies.

需要删减

## Restatement of the Problem

Combined with the background information and constraints identified in the problem statement, our goal is to solve the following problems:

* Develop a predictive model to estimate the number of medals each participating country is likely to win. For countries that have never won a medal, predict their probability of winning their first medal in the next Olympic Games. At the same time, the model needs to explore the potential relationship between events and the number of medals won by each country, explore which sports are most valuable to countries, and the impact of countries' choice of different sports on their chances of winning medals. Finally, the accuracy and performance of the model need to be evaluated.
* Analyze the potential impact of excellent coaches on the number of medals won by countries, and select three countries to consider whether they should start a "great coach" program in a certain sport.
* Propose insightful new insights based on the model. And explain how these insights can help the Olympic Organizing Committee. These strategies should be based on data and take into account the strengths and weaknesses of each country in specific events.

## Literature Review

At present, the prediction models for this problem are mainly time series models and empirical models.

### Time series models

Time series models are statistical or prediction models based on time series data. They are established by analyzing the trend, periodicity, seasonality or randomness of data over time.

The basic idea of ​​time series prediction is to use historical patterns to predict the future changes of a phenomenon. By revealing the trends in past data, these patterns are extended to the future to predict the future state of the phenomenon. The main models used to predict Olympic medals include deterministic time series models [18], trend extrapolation [20], Markov methods [14] and grey prediction models [16,39].

In order to accurately predict the number of medals, time series prediction must combine quantitative and qualitative analysis. Qualitative analysis explores the relationship between various factors and the number of medals. By in-depth investigation and research on the factors affecting the change of the number of medals, the predicted value can be determined.

Wu Dianting and Wu Ying [16] combined the GM(1,1) model (quantitative) with host country effects (qualitative) to predict the number of gold medals that China and the United States might win at the 2008 Beijing Olympics, achieving excellent prediction results.

### mpirical models

Empirical models are models based on observed data or empirical formulas. They usually do not consider the temporal characteristics of the data, but describe the system behavior by fitting the relationship between input and output.

Empirical models based on econometric principles generally include three steps:

* First, a set of influencing factors of the target variable is determined and divided into significant constant factors, significant occasional factors, and insignificant factors.
* Then, using significant constant factors as explanatory variables, the functional relationship between the target variable and the explanatory variables is determined, including related parameters, and the probability distribution characteristics and parameters of the random disturbance terms are established.
* Finally, a comprehensive model is derived. Determining the set of influencing factors is the key to building such models.

For models predicting Olympic performance, the key influencing factors usually include population size, per capita GDP, host country advantage, political system, sports culture, and national policies. These factors can significantly affect a country's performance in the Olympics.

Bernard and Busse (2000) used population size and economic resources (measured by GDP) as inputs to the Cobb-Douglas production function to study the distribution of Olympic medals. They assumed that population and economic resources follow the principle of diminishing marginal returns. For example, if economic resources remain unchanged, an increase in athletes with athletic talent will lead to a decrease in investment per athlete, thereby limiting the availability of resources required for high-level training, thereby reducing the marginal benefit of population growth on the number of medals.

Their model uses the proportion of medals won by a country to the total number of medals as the dependent variable. In addition, factors such as host country advantages and national policies (social and economic systems) are introduced as dummy variables. They used the number of medals in previous Olympic Games as an explanatory variable and effectively predicted the number of medals in the Sydney Olympics, achieving excellent results. However, the assumption of the principle of diminishing marginal returns needs further demonstration, and the consideration of national policies cannot be deepened.

### Comprehensive Model

Johnson and Ali (2000) proposed a combined model. First, they constructed a participation model based on factors such as GDP per capita, population, host country advantage, geographical proximity to the host country, political system, and colonial history. Then, they developed a separate model to estimate the probability of an athlete winning a medal and combined these results to calculate the number of additional medals and gold medals attributed to host country advantage and socialist system.

Kuper and Sterken (2001) extended this and proposed two empirical models combining Olympic participation and performance prediction. They used economic, geographical, and demographic variables to analyze the factors that affect Olympic participation and success. Their empirical model combined quantitative variables and dummy variables in the form of exponentials. They analyzed factors such as per capita income, population size, host country advantage, and national fixed resources to calculate their respective contributions to Olympic participation and performance.

Their results show that per capita income plays a key role in determining the number of athletes a country sends to the Olympics and their performance. Countries with higher per capita income not only send more athletes, but also win more medals. Moreover, the impact of host country advantage gradually weakens as the number of athletes sent increases.

需要删改和降重

## Our Work

Our study began with analyzing and processing the data. Some of the original data had problems with coding, vacancies, redundancy, etc. We first cleaned the data.

After visualizing the data, we observed that the available data was sparse and irregular, and we tried to fit the limited data using the grey prediction model GM(1,1).

However, unsatisfactory results were produced by the grey prediction model: the predictions were unstable, and the confidence assessment results were poor for some countries (especially those that never won a medal).

Further examining the data, we observed a high incidence of zeros (indicating obvious zero inflation) and realized that external factors, such as host country advantage, strongly influenced the results. Therefore, we chose the zero-inflated Poisson (ZIP) model, which is well suited for data with excess zeros.

We then built the zero-inflated Poisson model and evaluated the model using data from 2024 and then predicted the medal distribution for 2028.

To verify the performance of the ZIP model, we trained it using historical data to predict the medal distribution for 2028 and studied the countries that improved and regressed based on the prediction interval.

Additionally, we explored the impact of “great coaching” on performance outcomes. To do this, we conducted a difference-in-difference (DID) analysis to detect structural breaks in the data while controlling for other confounding factors. This ensured that the observed effects were not attributable to general trends or common events, but to the intervention of great coaches.

图示

描述已自动生成We then trained and validated the model, evaluated it using the log-likelihood method, and finally predicted the new 2028 results for comparison.

Figure 1：

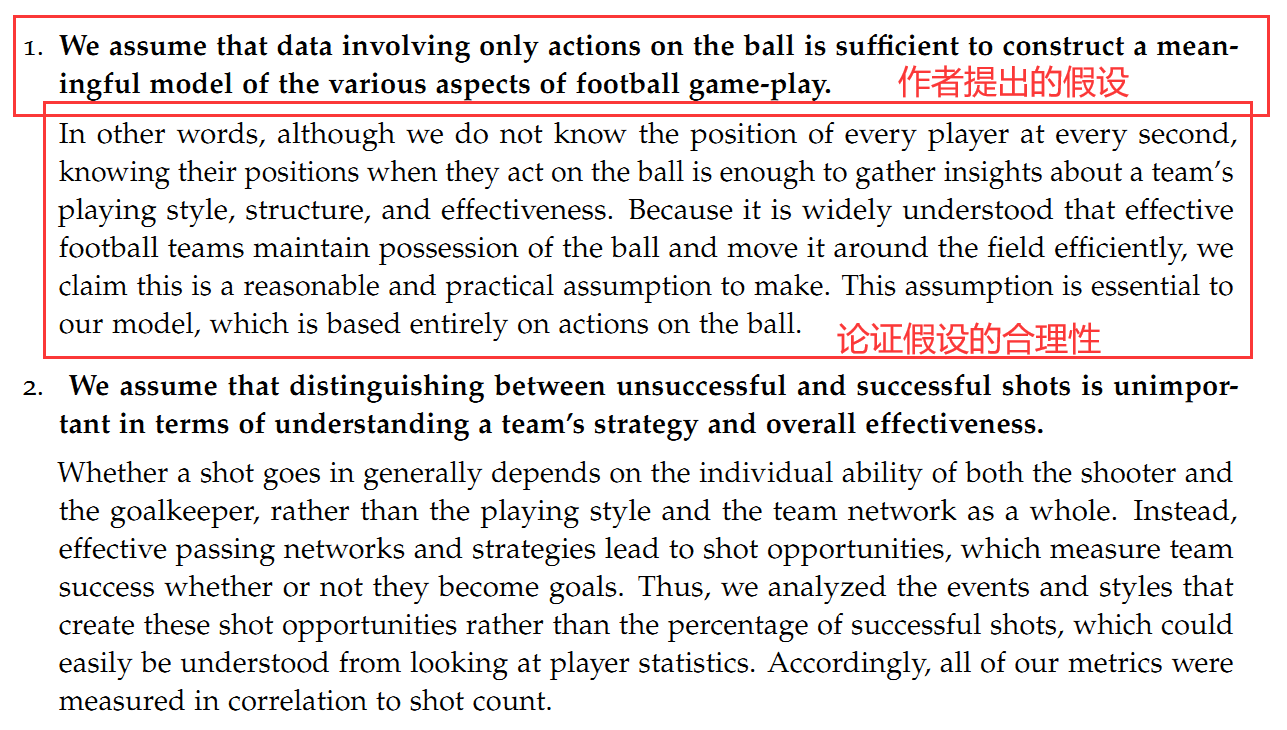
# Assumptions and Justifications

这一部分要写模型假设，并且要对假设的合理性进行论证，这一点比国赛的要求要高，请大家引起足够的注意。

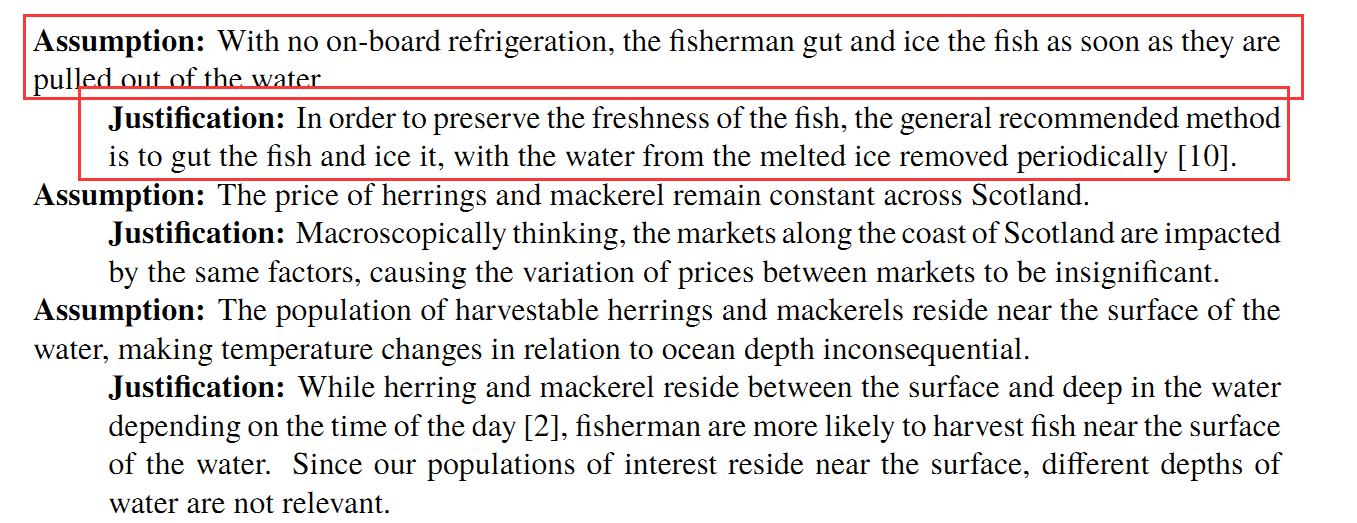
《正确书写美国大学生数学建模论文》一书中说道：无论解答什么样的赛题，参赛小组在论文中都应该明确列出所有用到的假设条件，并解释其合理性。如果对某个假设无法给出满意的解释，则应重新考虑这个假设是否合理，并进行修改，使得修改后的假设能有满意的解释。评委不但会检查论文是否列出了在建模过程中用到的所有假设，而且还会审查这些假设是否合理，以及论文对这些假设的合理性是否给出了满意的解释。

这里给大家看两篇特等奖的范例：

例1：2020特等奖论文Team # 2022868



例2：2020特等奖论文Team # 2017785



# Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

|  |  |  |
| --- | --- | --- |
| **Symbol** | **Description** | **Unit** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

本部分对应国赛论文中的符号（英文翻译为notation或者symbol）说明部分。

另外，少数论文有一个Deﬁnitions部分，这一部分主要是对文中出现的专有名词或者模糊的概念进行定义，这样可以帮助读者理解题目。



如果你需要写这一个部分的话，可以和Notations合并在一起变成一个大部分，该部分的命名为“Deﬁnitions and Notations”，写作时再把这个大的部分分成两个小的部分（用二级标题的方式）分别写即可。

# The name of model 1

这个大的部分主要介绍论文中第一个模型的建立和求解，第一个模型往往用来解决题目问的第一个问题。这个模型的标题需要结合你的题目或论文内容进行调整，我这里给的是一个通用的名称。

## Data Description

如果自己收集了数据或者题目给了数据的话，可以先对数据进行一个描述，一般将数据可视化，然后再从图形中得到一些直观的结论。如果是美赛C题（C题一般是数据分析类型的题目），可以把这个部分单独作为一个大的部分，然后进行数据预处理和数据可视化分析。当然，这部分也不是必须的，大家根据自己的需要调整。

## The Establishment of Model 1

这里可以写第一个模型的建立，模型建立是将原问题抽象成用数学语言的表达式，它一定是在先前的问题分析和模型假设的基础上得来的。因为比赛时间很紧，大多时候我们都是使用别人已经建立好的模型。这部分一定要将题目问的问题和模型紧密结合起来，切忌随意套用模型。我们还可以对已有模型的某一方面进行改进或者优化，或者建立不同的模型解决同一个问题，这样就是论文的创新和亮点。

和国赛不同，在美赛论文中，很少有论文直接用“模型建立”作为这部分的标题，一般使用模型要解决的问题作为标题。

如果需要用公式的话，可以复制下面这个隐藏的表格，并粘贴到你需要的地方，这个公式的编号会同步更新。

|  |  |
| --- | --- |
| 这里插入公式 | () |

## The Solution of Model 1

这里可以写第一个模型的求解，把实际问题归结为一定的数学模型后，就要利用数学模型求解所提出的实际问题了。一般需要借助计算机软件进行求解，例如常用的软件有Matlab, Spss, Lingo, Excel, Stata, Python等。求解完成后，得到的求解结果应该规范准确并且醒目，若求解结果过长，最好编入附录里。（注意：如果使用智能优化算法或者数值计算方法求解的话，需要简要阐明算法的计算步骤）

同样的，很少有论文直接用“模型求解”作为这部分的标题，大家可以根据得到的结论来合理设计这里的标题。

另外，很多美赛论文对于模型的建立和求解没有区分开，我这里沿用的是类似于国赛的形式，这样可以让论文框架清晰点。

|  |  |
| --- | --- |
| 这里插入公式 | () |

|  |  |
| --- | --- |
| 这里插入公式 | () |

# The name of model 2

和上一个部分类似的写法。

# The name of model 3

和上一个部分类似的写法。

注意：大多数美赛优秀论文都是对每个问题或者每个模型作为一个一级标题，就像我们上面的这种布局；也有一部分论文建立一个大的一级标题，取名为“Models and Solutions”，然后在这个大的标题下设计每个问题或者模型对应的二级标题，这一种排版布局在国赛中用的更多。

# Sensitivity Analysis

在国赛论文写作视频中（<https://www.bilibili.com/video/BV1Na411w7c2/>），我介绍过怎么写模型的分析和检验这个部分：

模型的分析 ：在建模比赛中模型分析主要有两种，一个是灵敏度(性)分析，另一个是误差分析。灵敏度分析是研究与分析一个系统（或模型）的状态或输出变化对系统参数或周围条件变化的敏感程度的方法。其通用的步骤是：控制其他参数不变的情况下，改变模型中某个重要参数的值，然后观察模型的结果的变化情况。误差分析是指分析模型中的误差来源，或者估算模型中存在的误差，一般用于预测问题或者数值计算类问题。

模型的检验：模型检验可以分为两种，一种是使用模型之前应该进行的检验，例如层次分析法中一致性检验，灰色预测中的准指数规律的检验，这部分内容应该放在模型的建立部分；另一种是使用了模型后对模型的结果进行检验，数模中最常见的是稳定性检验，实际上这里的稳定性检验和前面的灵敏度分析非常类似，等会大家看到例子就明白了。

在美赛的写作中，写的最多的就是灵敏度分析（Sensitivity Analysis），因此这里我们的标题就直接取得是灵敏度分析；如果你既要写灵敏度分析，又要写误差分析（Error Analysis），那么你可以把标题改成： Sensitivity Analysis and Error Analysis

# Model Evaluation and Further Discussion

注：本部分的标题需要根据你的内容进行调整，例如：如果你没有写进一步讨论的话，就直接把标题写成模型的评价。（优缺点一定要写）

## Strengths

这里写论文或者模型的优点

## Weaknesses

这里写缺点：缺点写的个数一般要比优点少

## Further Discussion

进行进一步的讨论，这里可以写模型的改进和拓展：

模型的改进：主要是针对模型中缺点有哪些可以改进的地方；

模型的拓展：将原题的要求进行扩展，进一步讨论模型的实用性和可行性。

# Conclusion

结论部分，这个部分在国赛论文很少见到，但在美赛中出现的频率很高。

这个部分可以是论文中心思想的重申、研究结果或主要观点的归纳，也可以是某些启示性的解释或考虑。

有些论文把“Model Evaluation and Further Discussion”的内容放到了结论部分，这也是可以的，大家可以灵活调整。

# References

参考文献：所有引用他人或公开资料(包括网上资料)的成果必须按照科技论文的规范列出参考文献，并在正文引用处予以标注。

一般新起一页列出参考文献，如果上一个部分的下面有很多空白，那么就不用新起一页了。

美赛中不要出现中文，如果引用中文文献请翻译过来。

# Appendices

|  |
| --- |
| Appendix 1 |
| Introduce: 这里放上附录1的介绍 |
|  |

|  |
| --- |
| Appendix 2 |
| Introduce: 这里放上附录2的介绍 |
|  |

本部分是附录部分，美赛对于附录不是特别看重，今年还限制了论文的页数（从第二页开始编号，不能超过25页）。

一般新起一页列出附录。

在不超过页数限制的条件下，附录中可以包括下面内容：

* 你们写的代码；
* 某一问题的详细证明或求解过程；
* 自己在网上找到的数据；
* 比较大的流程图；
* 较繁杂的图表或计算结果。