Modeling Historical Trends in Olympic Medal Counts with Mathematical and Computational Methods

MCM Problem C
Team 2518240
January 27, 2025

SUMMARY SHEET

The Olympics are a global stage where countries compete for prestige through athletic achievement. While gold has long been the symbol of Olympic success, the path to securing medals is not uniform across nations. The investment in athletic programs and the level of competition varies, and certain countries have consistently dominated the medal tally, while others struggle to secure even a single medal. This creates an interesting challenge in predicting future medal outcomes, as it involves understanding the trajectory of athletic growth and international competitiveness.

In this paper, we forecasted medal outcomes for the 2028 and 2032 Olympics using past Olympic performance data. A linear regression model was applied, considering key variables such as the host country's influence, the number of athletes participating, and historical trends in medal distribution. Historical data suggests that begin the host country can provide a significant boost in the medal count, due to anticipated factors like being able to select the events at-hand, greater support, along with greater motivation to provide resources into a local event. Our predictions took these directly into effect in our calculations, adjusting medal counts based on which country would be hosting Olympics in each of the upcoming Games. The model highlighted countries "The United States", "Russia", "China", and "The United States", "China", and "Australia" as the most likely to dominate the medal count in each of the 2028 and 2032 Olympics, respectively. Australia in particular is quite surprising but could be understood considering the host country effect.

Additionally, we implemented a logistic regression model that examined the impact of the changes in the number of events per country, the frequency of participation in past Olympics, and whether a country had previously won a medal before. This model identified Myanmar, Monaco, and Liechtenstein as countries likely to win medals for the first time in 2028. In particular, the logistic regression gives that these three countries are the only ones that passed the 0.75 probability threshold to receive a medal for the first time. Even though historical data indicates that Monaco has not received a medal in 23 Olympics, the data indicates that it is highly likely that is on the brink of getting a gold medal for the first time.
However, these predictions doe not account for evolving dynamics, such as shifting sport cultures, increased funding to African athletes from changing geopolitical influences, or global economic changes that may impact future performances. Further analysis points out that the success of emerging nations, like those from emerging countries may be influenced by increasing investment by some international government organizations (e.g. the well-known Chinese economic influence in Africa). While our models do not have a direct mathematical inclusion of these rapidly changing dynamics, they strongly emphasize the need for future models to evolve and integrate emerging trends in global athletics and the geopolitical landscape! The influence of individual countries in this increasingly neoliberal sport is vast, which requires future inclusion of these variables.

Ultimately, while our models provided a well-rounded prediction based on historical data, we must remain mindful of the evolving nature of international sports. While countries like X, Y, Z, and A, B, C remain strong contenders, the emergence of new athletic powerhouses and the influence of external variables may lead to surprising shifts in the medal standings in 2028 and 2032. We hope that our approach serves as a foundation for future predictions, while acknowledging the unpredictable nature of the Olympics. Future predictions should incorporate not only computational analysis of historical results but also fundamental analysis of factors such as geopolitical influences.

Contents

1 Introduction

It has always been a fun hobby for sports enthusiasts to predict what may occur in certain sports. This hobby led to the rise of the sports-betting industry and is a successful way for enthusiasts to engage with the sport and others in the community.

The Olympic Games in particular are a global spectacle that unite nations through competition and celebrate athletic excellence. Having been revived since their origins in Ancient Greece, the games have taken on importance on an international scale, and it is of great interests of countries to be able to predict what medals they can win in the Olympics, as this helps them gain prestige at the global level. Analyzing and predicting these medal counts is a multifaceted challenge, requiring consideration of historical performance, current trends, and numerous external factors. With the advent of powerful computational tools, artificial intelligence, and large datasets, it has become possible for computers to predict winners of sporting events based on past outcomes. In this research paper, we aim to use such computational tools and datasets to create an accurate model that accurately models the medals received in previous Olympics along with projecting the performance of countries in the 2028 Los Angeles Olympic Games. This paper develops a comprehensive framework to model Olympic medal tables, addressing key questions such as:

- How can historical trends and current data predict medal counts for the 2028 Olympics?
- Which countries are most likely to improve or decline in performance?
- Which countries with no prior medals will win for the first time in 2028?
- How do specific events and disciplines shape medal outcomes?
- To what extent does the influence of coaches or the "great coach" effect alter medal counts?

Using the datasets provided, we employ statistical modeling, data analysis, and scenario testing to develop insights and projections. Our work includes simple data science techniques to predict the count of Gold and total medals. Through this exploration, we hope to contribute some valuable insight into the incredible Olympic Games.

2 Analysis of the Problem

The analysis of Olympic medal tables involves addressing several interdisciplinary and interconnected factors. The factors influencing a country's success are quite diverse. In fact, previous studies have shown that many socioeconomic factors, such as population size, GDP, sports infrastructure, political stability, etc. are crucial to a country's success at the Olympics [1]. Our data set, however, does not have information on these facets, so the key areas of focus for this analysis will include the following:

Historical Data of Olympic Games

It has been seen in the past that the way athletes and countries perform in the past is a significant factor in the potential to perform in the future. This has multiple reasons, the primary of which being the strength of the country's athletes and coaches. Some essential factors not being explored in this paper is the influence of socioeconomic growth and having a larger population. This also includes the influence of political variables like boycotts and lacking participation, which is out of the scope of this paper. To understand historical change, it is necessary to observe trends in specific sports and disciplines that contribute significantly to medal totals.

Event-Sport Dynamics
Each Olympic Games features a unique set of events, with variations in the number and types of competitions. These factors play a critical role in shaping medal outcomes. For example, the particular introduction or removal of events can benefit certain countries with specialized strengths. Additionally, sports with multiple events, such as Swimming or Cycling, offer numerous opportunities for medals. Since the host country works with the IOC to decide many of the events, they play a strong part in deciding how many medals are given out each year, and they can possibly hold more events that favor themselves. It will be very important to analyze the influence of these host countries, who might prioritize events where they have a competitive advantage.

Home-Country Advantage
Hosting the Olympics provides several advantages that are important to keep in mind:

- Familiarity with venues, reduced travel fatigue, and strong crowd support often boost performance.
- Host nations, as discussed above, have a strong influence on the events for each sport in that year's Olympics.
- Strategic allocation of resources toward hosting and training programs amplifies this effect. The prestige factor of hosting the Olympics in their country for the year also motivates a greater national focus of resources on the Olympics for hosting years.

We will attempt to understand the extent to which this effect plays a part in the number of medals countries win.

First-Time Medal Winners
For countries without prior medals, the major thing that can influence them to get a medal is the inclusion of new events that may better align with the countries' traditional strengths. Additionally, improved training programs, often with the support of international governing bodies of some sort, can catalyze success. There are also overall historical trends in the number of countries that win first-time medals every year that we analyzed to predict the number of first-time winners in the 2028 Olympics.

Role of Coaches and International Influence
The "Great Coach" effect is a strong influence on the work and success of national teams. Coaches have always held a strong influence on the ability of the players to work hard and succeed. Examining patterns where nations hired foreign coaches or benefited from athlete migration offers insights into medal outcomes. This can be analyzed specifically through the lens of the host country effect or other factors discussed above.

3 Assumptions

To ensure the feasibility and clarity of our analysis, the following key assumptions were made:

1. The event structures of the 2028 and 2032 Olympics are assumed to be largely similar to those in 2024. That is, core sports like swimming and athletics will remain largely unchanged.

2. The performance of a country in past events correlates to the performance of the same country in future events. A country performing well we assume incentivizes them to spend more resources towards sustaining their success.

3. We assume standard resource allocation and stability without any sudden political issues giving difficulties in any countries participation in the Olympics. This makes it likely that they are to maintain or improve their medal count due to investments in training facilities, coaching, and athlete support.

4. The number of events year-by-year is increasing (not strictly). This assumption is important in statistical tests of the accuracy of a historical facto in the change of the total medal count over time.

5. We are assuming that every country that has participated before (minus the defunct/dissolved countries) and will be participating in 2028 again.

4 Model Discussion

4.1 Historical Model

The model of the Olympic Medals based on the historical results of previous Olympics considers the different factors that can be treated independently. This not only models previous medals earned but also projects the medal distribution in 2028 and 2032

Regression

To model and predict the total number of Olympic Medals for a country x, we used a line of best fit model. The equation is the following

$$M_{x(\text{type})} = (\beta_0 + \beta_1 \cdot F_x \log(E_x) + \beta_2 \cdot M_x) \cdot H_x + \xi_x$$

with β_0, β_1, β_2 being our regression coefficients. The parameters that we have in our model are

- H_x: The binary indicator for host country (1 if hosting and gold medal count, 0 otherwise)
- M_x: The number of previous medals
- F_x: The binary indicator for whether the country has previously won medals (1 if yes, 0 otherwise)
- E_x: The number of unique events the country has participated in by country x.
- ξ_x: The "noise" term or the error term accounting for unmodeled variability.

4.2 Model Performance and Predictions

After pre-processing the data, we were able to accumulate the total number of medals that a country had won in previous events, as well as the total number of unique events they had participated in. Doing so resulted in $m = 0.0676$ and $b = 0.995$ with p-value $1.55 \cdot 10^{-91}$.

Coming up with this model allowed us to predict the number of medals the countries will win during the 2028 and 2032 Olympics. The model particularly gave upon training that the countries most likely to win medals are, in 2028, "The United States", "Russia", "China", and, in 2032, "The United States", "China", and "Australia".

4.3 Host Country Effect

The host country effect is the trend where the nation hosting the Olympic Games tends to outperform its historical medal counts, relative to previous Olympic Games. However, the impact of hosting the Games we investigate may vary across different medal types (Gold, Silver, and Bronze), and our model seeks to quantify and assess this affect.

Statistical Methodology
To assess the host country effect, we used basic data science in Python Pandas to compare the number of gold medals, silver medals, and bronze won by a country in an year that it hosts the Olympic games relative to the two adjacent Olympic Games (the game they participated in before the hosting game and after). The strength of this method is that it allows us to capture the effect of the host country status while accounting for the increase in the number of events and other hidden variables resulting from data that describes the country in years both before and after the hosting year.

To be specific on the statistical test, we took the difference between the medal count of the host year and the average of the adjacent years. We then employed a z-test to compare the observed number of medals won by the host country with the expected number of medals (of each type) based on the previous and subsequent Olympic Games. The hypothesis tested in this case was whether the number of medals (for each medal type) won in the host country's year was statistically greater than the average of the medals won in the year before and after the host Games:

$$h_0 : n = 0 \quad h_a : n > 0$$

where n is the number of medals won. Now, we considered $\alpha = 0.05$ and got the following results:

- Total: Fail to reject null hypothesis ($p = 0.0597$). No significant evidence the mean is greater than 0.
- Gold: Reject the null hypothesis ($p = 0.0290$). The mean is significantly greater than 0.
- Silver: Fail to reject null hypothesis ($p = 0.0629$). No significant evidence the mean is greater than 0.
- Bronze: Fail to reject null hypothesis ($p = 0.1278$). No significant evidence the mean is greater than 0.

Host-Country Multiplier for Gold Medals
We specifically found that the host country effect is statistically significant for gold medals but not for silver or bronze medals. The increased focus on elite athletes and the heightened national pride could be the likely explanation of the host-country effect's focus on gold medalists.

To quantify this host-country effect for gold medals, we calculated the host-country multiplier as described earlier in the total model. The multiplier is the ratio of the number of gold medals won by the host country in the host year to the number of gold medals they won in the year before. This ratio provides a multiplicative factor that reflects the increased likelihood of winning gold medals when hosting the Olympics. The result:

$$\text{Host Multiplier} = \text{Avg}\left(\frac{\text{Gold}_{\text{Host}}}{\text{Gold}_{\text{Pre}}}\right) \approx 1.852$$

4.4 Countries Without Medals

In this problem we seek to project how many countries will earn their first medal in the next Olympics and to what degree of certainty we believe our model.

To start tackling this problem, we first compiled a list of the countries without medals by parsing the lists for the participating countries and subtracted the set of countries with medals already. The resulting list had several countries/polities that no longer participate or are defunct/dissolved, so those had to be removed. These were found using a page for IOC codes [2]. After removing these, the list contained 66 countries, the majority of which were island countries or African countries.

With the countries that had already participated, we compiled graphs of the number of Olympics it took before they won their first medal, using the assumption that a country's prior performance correlates with their next performances to observe some trends, namely that most countries won a medal fairly soon after competing.

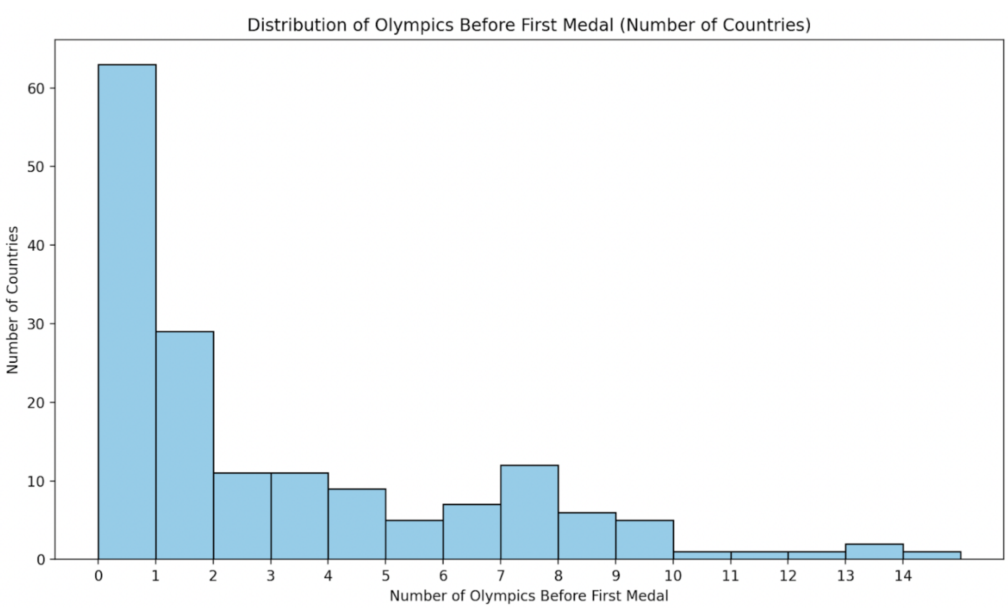

Figure 1: Bar graph of the number of Olympics before first medal

We also created a graph to see how many Olympics the countries that currently have no medals have already participated in.

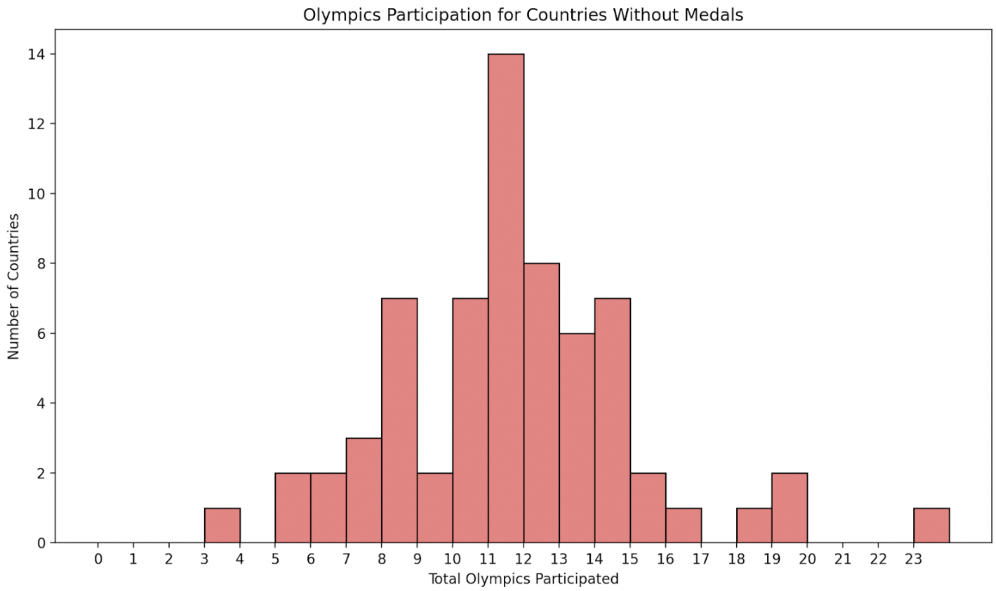

Figure 2: Number of Olympics for countries that have no medals

We observe that most of these countries have already participated in numerous Olympics, so we would not expect many countries to suddenly start winning. Similarly, we have a graph depicting the proportion of countries that win their first medal in each Olympics.

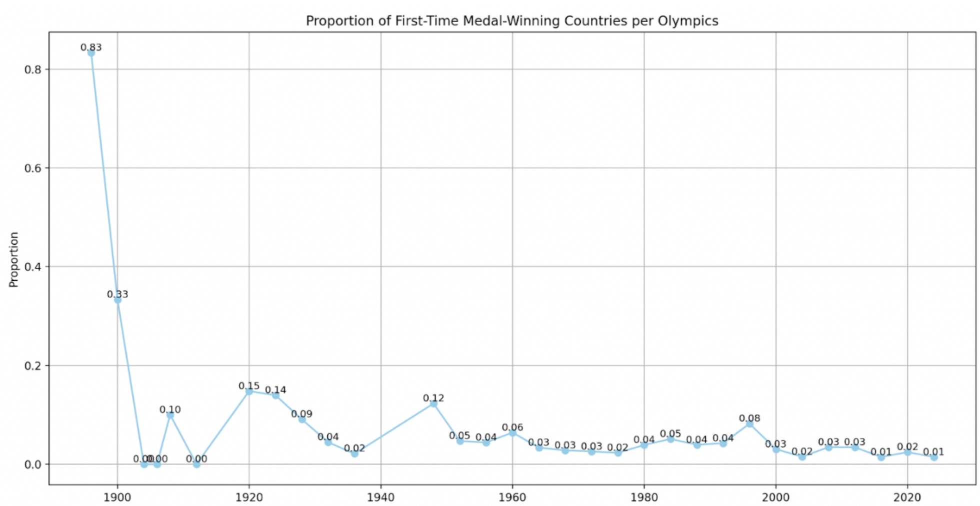

Figure 3: Proportion of countries who win their first medal per Olympics

Considering the latter half of the data points (as that's when most big countries like the United States had already won), we notice a spike in 1996. This is because 1996 is the first time every single country invited (197 total) sent athletes, and 24 countries debuted in the Olympics for the first time (these countries were all former Soviet republics) [3]. So to test linear regression and polynomial fit, we consider the data points from 1996 onward:

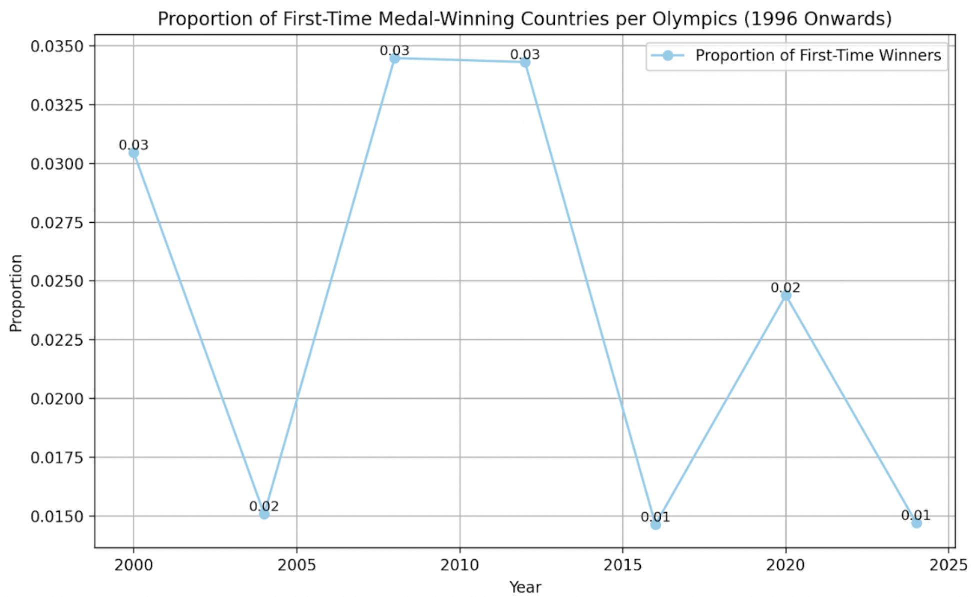

Figure 4: From 1996 onwards

However, performing linear regression and polynomial fit on these data points yielded poor r^2 values of 0.164 and 0.244, respectively, so we decided to perform a logistic regression on our data instead.

The logistic regression formula is as follows:

$$P(y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)}}$$

Where $P(y=1|X)$ is the probability of winning a medal given the conditions X,
β_0 is the intercept term representing the baseline log-odds of the event occuring (winning a medal) without the conditions,
β_1, β_2 are the coefficients of the conditions representing how each feature will affect the log-odds,
X_1, X_2 are the input features. Here, they are the change in events each country participates in and the number of Olympics a country has already participated in.

The result of the logarithmic regression is depicted in the graph below:

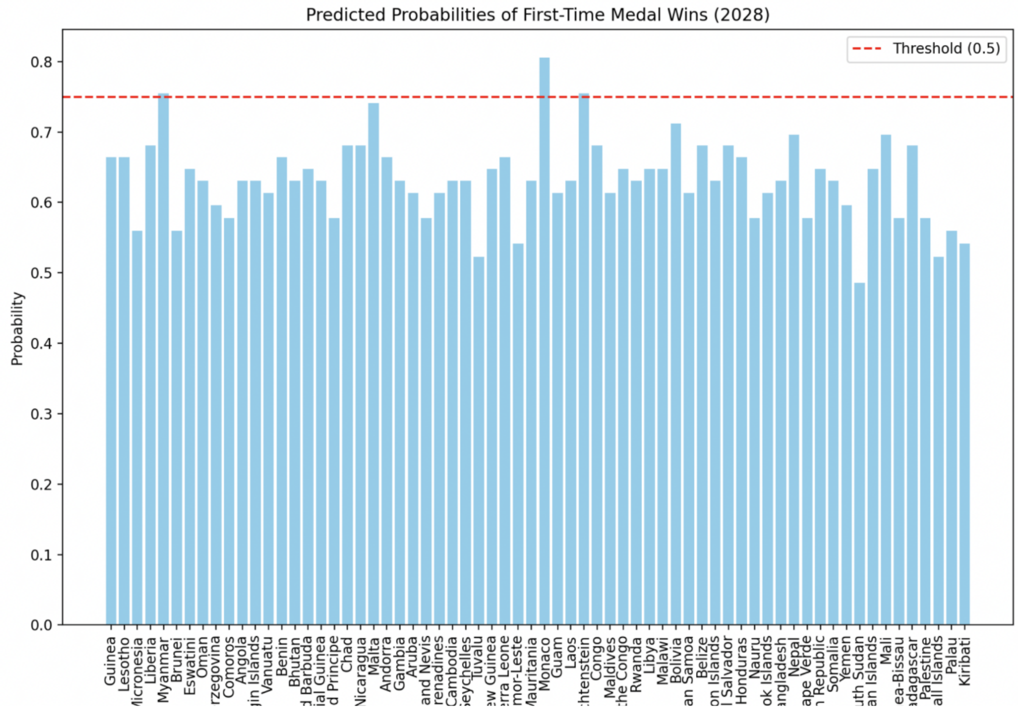

Figure 5: Logarithmic regression probability yields per country

Note that we set the threshold for the probability of a country to win a medal to be 75% to reflect the difficulties of winning a medal for the first time.

From this data, we expect 3 new countries to achieve a medal for the first time. Specifically, these countries are Myanmar, Monaco, and Liechtenstein.

This result aligns with our expectations that not many countries will win for the first time. The initial graphs of the number of Olympics the countries that haven't won yet have participated in show that these countries have generally participated in many Olympics and still haven't won which could be due to a lack of sports culture or athletics funding. However, there are some caveats to this result as it ignores many factors such as changing sports culture, GDP changes, etc. which will be discussed more in our conclusions page.

4.5 "Great Coach" / Event Types

The "Great Coach" Effect depicts the relationship between the hiring of foreign coaches and a country's Olympic medal performance.

Statistical Methodology

To assess the "Great Coach" effect, we considered that a country will hire a foreign coach when one of two factors is true: the country is hosting the Olympics or there is great change in the number of events.

For the first factor, we previously showed that a country hosting the Olympics causes a statistically significant increase in the number of medals for only the number of gold medals.

For the second factor, we found that there was not a statistically significant difference because of the following p-values:

- Gold: Reject the null hypothesis ($p < 0.01$). The mean is significantly greater than 0.
- Silver: Fail to reject null hypothesis ($p \approx 0.08$). No significant evidence the mean is greater than 0.
- Bronze: Fail to reject null hypothesis ($p \approx 0.35$). No significant evidence the mean is greater than 0.

This suggests that elite coaches have the greatest impact on athletes competing at the highest levels, that is, those who win Gold.

5 Strengths and Weaknesses

Like any prediction model, there are strength and weaknesses, and this linear regression model is no different. We analyze the strengths and weaknesses of the model below.

Strengths

- Accounting for Host Country Advantage
  Through our analysis, we found that being the host country plays a role in the number of gold medals they win. Through this linear regression model, we emphasize that the number of medals a country wins will depend on whether they are the host.
- Data Normalization
  Through the use of the logarithm, we remove the impact of outliers and normalize the data. This makes the data stronger while preserving a lot of the trends that are needed to accurately model future medal wins.
- Computational Simplicity
  The linear regression model is simple and efficient, while still providing a good prediction for the number of medals that'll be won in the future.

Weaknesses

- Overfitting to historical patterns This model focuses heavily on the number of events that a country has participated in the past, and correlates that to the number of medals won. However, there is no term that accurately models the improvements or deterioration of a country's performance in a sport. Therefore, a country that has performed poorly in the past but is beginning to improve will not see that improvement accurately predicted by this model.
- Over emphasis on participation This model takes into participation only. There is no metric by which to judge how competitive a certain sport is, or how close a certain team was to winning or losing a medal.
- Lack of analysis on individual sports Although modeling the total number of medals that a team may win directly is viable, it may be more accurate to model sports individually and predict the outcomes of sports. This type of modeling might lend itself more to analysis of recent improvements and growth.

- Interdependence of various variables There may be a more complex relationship between E_x, F_x, and M_x than this model assumes, as countries with strong performances in the past are likely to have higher amounts of participation in future events.

6 Conclusions

After running our models, we came up with the following conclusions regarding the various subquestions. The overall predictions following the linear regression of the historical data were "The United States", "Russia", and "China" in 2028 and "The United States", "China", and "Australia" in 2032.

Host Country Effect:
Although we expected the host country to perform better in the Olympics, as they have the "home-court advantage," and the unique opportunity to influence what events take place in the Games, particularly events that give them the competitive edge, the true extent of this effect was pretty surprising.
Running z-tests on our datasets, considering $\alpha = 0.05$, we concluded that the host country effect only matters significantly with Gold medals, but the effect is an astonishing 1.852x multiplier.

Countries Without Medals:
We concluded that we expect 3 countries, specifically Myanmar, Monaco, and Liechtenstein, to achieve medals for the first time in their countries' histories at the 2028 LA Olympics. This number, however, should be taken with a grain of salt, for it doesn't take into account the fact that Africa is investing more money into Olympic athletes [4], which will certainly cause an impact, as several of the countries without medals so far are African, and none of the countries this model predicts to win are in Africa.
Additionally, it doesn't take into account the overall increase in events in the 2028 LA Olympics, which are not included in our given datasets. Specifically, the 2028 Olympics will feature the debut of several sports like flag football and squash, as well as the return of old sports like lacrosse, cricket, baseball, and softball [5].
Since the 2028 LA Olympics will feature over 40 sports [6], which is more than any previous Olympics [7], it is quite possible that there will actually be more than 3 countries that wins a medal for the first time.

Great Coach Effect
Considering that countries will hire foreign coaches when they are hosting or the number of events changes, we performed more statistical tests and concluded that the only statistical significant effects of elite coaches are on athletes who are the best in the world. That is, the "Great Coach Effect" helps significantly with winning Gold medals.

7 Further Directions

This paper does provide a strong framework for predicting Olympic medal counts, but there are lots of areas in which future research can enhance on the 4 day's work of this model's accuracy and applicability

There are a few ways that the accuracy and the precision of the projections could be improved in the model. For example, more data could be used, such as the Athlete-Level Data or the events of the 2028/2032 Olympics, which have already been ordained. This can significantly enhance the accuracy and precision

of the model. Additionally, there could be some event-specific analysis of how the particular event lineups impact country-specific medal outcomes. Out model reduces it to simply just the change in the events and the count of events, but the existence of newer events may disproportionately affect emerging nations.
Inclusions of further fundamental factors within the models are possible, such as a more granular approach to the inclusion of the "Great Coach" effect by incorporating coach tenures, and performance improvements across multiple disciplines, and cross-national coaching exchanges. Other places for fundamental analysis that could be included are the influence of geopolitical events and economic fluctuations on economic fluctuations on Olympic performance.

We also believe that the analysis could have been done in a stronger computational sense. For example, there could have been a use of computational tools like LSTM or ARIMA in order to model the data. However, the lack of enough time-series data makes it difficult to account for temporal trends, such as increasing globalization of sports or evolving training methodologies.

8 Bibliography

[1] Smith, A. (2022). Evaluating predictive models in applied settings: The case of logistic regression. Journal of the Royal Statistical Society Series A: Statistics in Society, 13(3), 22–34. https://doi.org/10.1093/jrssig/13.3.22

[2] International Olympic Committee, "Olympic Data Feed: NOC Codes," Olympic Games Beijing 2022. Available: https://odf.olympictech.org/2022-Beijing/codes/HTML/og_cc/NOC.htm. [Accessed: Jan. 26, 2025].

[3] Olympics.com. (n.d.). *The history of the Olympic Games*. Retrieved January 27, 2025, from https://www.olympics.com/en/news/the-history-of-the-olympic-games/

[4] BBC Sport. (n.d.). New sports and changes at the LA 2028 Olympics. Retrieved from https://www.bbc.com/sport/olympics/articles/c4gdwy72ee5o.

[5] Newsweek. (2023). New sports added to LA 2028 Olympics. Retrieved from https://www.newsweek.com/new-sports-la-olympics-2028-1936988.

[6] NBC Olympics. (n.d.). Where and when are the next Olympics? LA 2028 dates, venues, and more. Retrieved from https://www.nbcolympics.com/news/where-and-when-are-next-olympics-la-2028-dates-venues-and-more.

[7] Topend Sports. (n.d.). Number of sports at the Summer Olympics. Retrieved from https://www.topendsports.com/events/summer/sports/number.htm.

9 AI Use Report

Throughout the process of writing this report, AI, specifically ChatGPT 4o, was used primarily to gather information about the structure of Olympics, general information about countries, which countries are no longer participating or defunct. Additionally, AI was also used to gain a high-level overview of what tools and methodologies could be used for the data processing.

1. OpenAI *ChatGPT* (ChatGPT-4)
Query: What do you think is the best choice to solve the problem of seeing whether a country without previous medals will win its first medal in the 2028 Olympics given datasets of previous medal counts: logistic regression or polynomial fit?
Response: Logistic regression is the right choice for this problem because it is specifically designed for binary classification problems, such as predicting whether an event belongs to one of two categories (e.g., success vs. failure, pass vs. fail). It models the relationship between input features and the probability of the outcome, providing interpretable coefficients and clear decision boundaries. Its simplicity and efficiency make it ideal when the dataset is relatively small or when the relationships between features and the outcome are approximately linear.

2. OpenAI *ChatGPT* (ChatGPT-4)
Query: Why is logistic regression better than polynomial fit?
Answer: Logistic regression is better than a polynomial fit for binary classification because:
Appropriate for Binary Outcomes: Logistic regression is specifically designed to model probabilities for binary outcomes, ensuring the predicted values are between 0 and 1. Polynomial fits, on the other hand, are regression models suited for continuous outcomes and may produce values outside the valid probability range.
Interpretable Decision Boundary: Logistic regression defines a clear and interpretable decision boundary (e.g., at a probability of 0.5), whereas a polynomial fit may create unnecessarily complex and oscillating boundaries that do not generalize well to new data. Avoids Overfitting: Polynomial fits can overfit the data, especially with high-degree polynomials, capturing noise instead of the true underlying pattern. Logistic regression avoids this by focusing on a simpler, more robust model.
Efficiency and Simplicity: Logistic regression is computationally efficient and straightforward to implement, while polynomial fits may require careful tuning of the degree to balance bias and variance.
In short, logistic regression is purpose-built for classification and produces reliable, interpretable, and generalizable results compared to a polynomial fit.

3. GitHub CoPilot (Jan 23, 2025 version)
Auto-completions for code used in preparing our models.

4. OpenAI *ChatGPT* (ChatGPT-4)
Query: What are some socioeconomic factors that affect a country's Olympic performance
Response: Several socioeconomic factors influence a country's Olympic performance, including:

1) **Economic Resources** - **GDP and Funding**: Wealthier countries with higher GDPs can allocate more resources to sports infrastructure, athlete training, and support systems.
2) **Population Size** - **Talent Pool**: Larger populations provide a broader talent pool, increasing the

likelihood of finding elite athletes.
3) **Government Support and Policies** - **National Investment**: Investment in grassroots programs and high-performance centers fosters talent development.
4) **Cultural and Historical Factors** - **Sports Traditions**: Countries with a strong sports culture or historical success in specific disciplines tend to excel.
5) Education and Health** - **Physical Education**: Access to quality physical education programs in schools builds a foundation for athletic success.
By addressing these factors strategically, countries can enhance their Olympic performance and build a sustainable sports ecosystem.

10 Appendix

Python code for the logistical regression model of predicting which countries without medals will win their first medal in 2028. This data has been cleaned up with comments to replace big chunks of code that are unnecessary:

```
#importing packages and modules

noc_to_country = {
 #list of noc codes to country names to make the datasets more consistent
}


medal_data = pd.read_csv("2025_Problem_C_Data/summerOly_medal_counts.csv")
athlete_data = pd.read_csv("2025_Problem_C_Data/summerOly_athletes.csv")

medal_data["NOC"] = medal_data["NOC"].str.strip()
athlete_data["NOC"] = athlete_data["NOC"].str.strip()
athlete_data["Year"] = athlete_data["Year"].astype(int)
athlete_data["NOC"] = athlete_data["NOC"].replace(noc_to_country)

medal_data["HasMedal"] = medal_data[["Gold", "Silver", "Bronze"]].sum(axis=1) > 0

first_medal_year = medal_data[medal_data["HasMedal"]].groupby("NOC")["Year"].min()

first_medal_df = first_medal_year.reset_index()
first_medal_df.columns = ["NOC", "FirstMedalYear"]


def calc_oly_before_medal(row):
    noc = row["NOC"]
    first_year = row["FirstMedalYear"]
    # Filter athlete data for the specific country and years before their first
        medal
    participated_years = athlete_data[athlete_data["Team"] == noc]

    participated_before = participated_years[participated_years["Year"] <
        first_year]["Year"].drop_duplicates()

    return len(participated_before)

participating_countries = athlete_data["NOC"].unique()
countries_with_medals = first_medal_df["NOC"].unique()
not_real_countries = {"URS", "ANZ", "WIF", "AIN", "YUG", "EUN", "FRG", "SCG",
    "AHO", "CRT", "ROT", "BOH", "NBO", "YAR", "TCH", "EOR", "IOA", "SAA", "GDR",
    "YMD", "NFL", "UAR", "UNK"}
countries_without_medals = set(participating_countries) -
    set(countries_with_medals) - set(not_real_countries)

```

```
athlete_data = athlete_data[~athlete_data["NOC"].isin(not_real_countries)]

#function for how many olys they've already done was here

countries_without_medals_participation = {
    noc: calc_olys_participated_in(noc) for noc in countries_without_medals
}
# Linear regression code was here

# Polynomial fit code was here

#Logistical Regression logic here:
def calc_change_in_events(noc):
    country_data = athlete_data[athlete_data["NOC"] == noc]
    events_by_year = country_data.groupby("Year")["Event"].nunique()
    return events_by_year.diff().fillna(0).tolist()

no_medals_df["ChangeInEvents"] = no_medals_df["NOC"].apply(calc_change_in_events)

# Created the training dataset by creating another dataframe, iterating over each
    row in the new dataframe, and adding the information to the training_data
    dataframe
training_data = pd.merge(
    athlete_data,
    medal_data[['NOC', 'Year', 'HasMedal']],
    on=['NOC', 'Year'],
    how='left'
)
#code for creating the training dataset
# Features and target variable
X = training_data[["TotalOlympicsParticipated", "ChangeInEvents"]]
y = training_data["HasWonMedal"]
imputer = SimpleImputer(strategy="mean")
X_imputed = imputer.fit_transform(X)
# Split into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X_imputed, y, test_size=0.2,
    random_state=42)
# Train the model
logistic_model = LogisticRegression(class_weight='balanced')
logistic_model.fit(X_train, y_train)

# Evaluate the model
y_pred = logistic_model.predict(X_test)
countries_2028 = no_medals_df.copy()
countries_2028["TotalOlympicsParticipated"] += 1
events_by_year = athlete_data.groupby("Year")["Event"].nunique()
# Filter for years 2012 and onwards
events_since_2012 = events_by_year[events_by_year.index >= 2012]
```

```python

# Calculate the change in events between consecutive Olympics
change_in_events = events_since_2012.diff()
# Estimate the change for 2028 by using the average change since 2012
avg_change_since_2012 = change_in_events.mean()
estimated_change_2028 = avg_change_since_2012

countries_2028["ChangeInEvents"] = avg_change_since_2012;
X_2028 = countries_2028[["TotalOlympicsParticipated", "ChangeInEvents"]]
# Predict probabilities for 2028
probs_2028 = logistic_model.predict_proba(X_2028)[:, 1]
threshold = 0.75
countries_2028["PredictedMedalWin"] = (probs_2028 > threshold).astype(int)
# Count predicted first-time winners
predicted_first_time_winners = countries_2028["PredictedMedalWin"].sum()
print("Predicted number of first-time medal-winning countries in 2028:",
    predicted_first_time_winners)
#Code for plotting the probabilities in a neat graph was here
```

Listing 1: Python Code for Logistical Regression