Economic Impact on Olympic Performance: A GLMM Approach

Ahmad Sharif
Student ID: K436765
ahmad.sharif@tuni.fi

Clustered Data Models

22 November 2024

Contents

1 Introduction

Clustered data refers to datasets where observations are classified into multiple categories known as clusters. Each cluster has number of individual observations, creating a hierarchical structure in which data points are nested inside their respective clusters.

In this paper, a dataset has been collected from popular platform kaggle to apply any cluster data models. For this purpose, Olympics medel (2024) data has been chosen for this project.

The dataset used for this project can be downloaded from Kaggle: [?].

The Olympics is a event where more than 35 sports are held to test the performance of athletes from different countries. This is the only international event where most of the countries athletes participate. Thats why this is great opportunity to examine and comapre the countries medel count against their corresponding population, GDP and other factors. This event does not only measure the competitiveness in sprots of a country but it can be also a scale to measure a countrys national priorities, lifestyles, and so on.

Here is the research question of this project.
How do a country's GDP and population size influence its total medal count in the 2024 Olympic Games, while accounting for regional differences?

The GLMM is an extension of the generalized linear model (GLM) that incorporates both fixed and random effects, making it particularly suitable for dealing with clustered or hierarchical data. The inclusion of random effects allows the model to account for variations between groups or clusters (in this case, regions or continents), while the fixed effects explain the relationship between the predictors (GDP, population) and the outcome (medal counts).

2 About The Database

The latest Olympic game (2024) medal and their corresponding country inforamtion is collected from the following kaggle databse. [?]. The dataset has nine columns and their corresponding names respectively:

- **Country**: Name of the country.
- **Code**: ISO 3-letter code for the country.
- **Gold**: Total number of gold medals won.

Country	Code	Gold	Silver	Bronze	Total	GDP (USD)	GDP Year	Population (Millions)
United States	USA	40	44	42	126	81695.19	2023	334.9
China	CHN	40	27	24	91	12614.06	2023	1410.7
Japan	JPN	20	12	13	45	33834.39	2023	124.5
Australia	AUS	18	19	16	53	64711.77	2023	26.6
France	FRA	16	26	22	64	44460.82	2023	68.2
...	...	...	...	...	...	...	...	...
Zambia	ZMB	0	0	1	1	1369.13	2023	20.6

Table 1: Countries in the 2024 Olympics by medal count and economic indicators

- **Silver**: Total number of silver medals won.
- **Bronze**: Total number of bronze medals won.
- **Total**: Total number of medals won by the country (Gold + Silver + Bronze).
- **GDP**: Gross Domestic Product in billion USD.
- **GDP Year**: Year corresponding to the GDP data.
- **Population**: Population size in millions.

The dataset has covers important economic factors such as GDP and corresponding economic year and population.

2.1 Figures illustrating the dataset

In Figure 1, we illustrate the relationship between Total medals against GDP Per Capita. and this Figure 2 Total Medals Vs Population

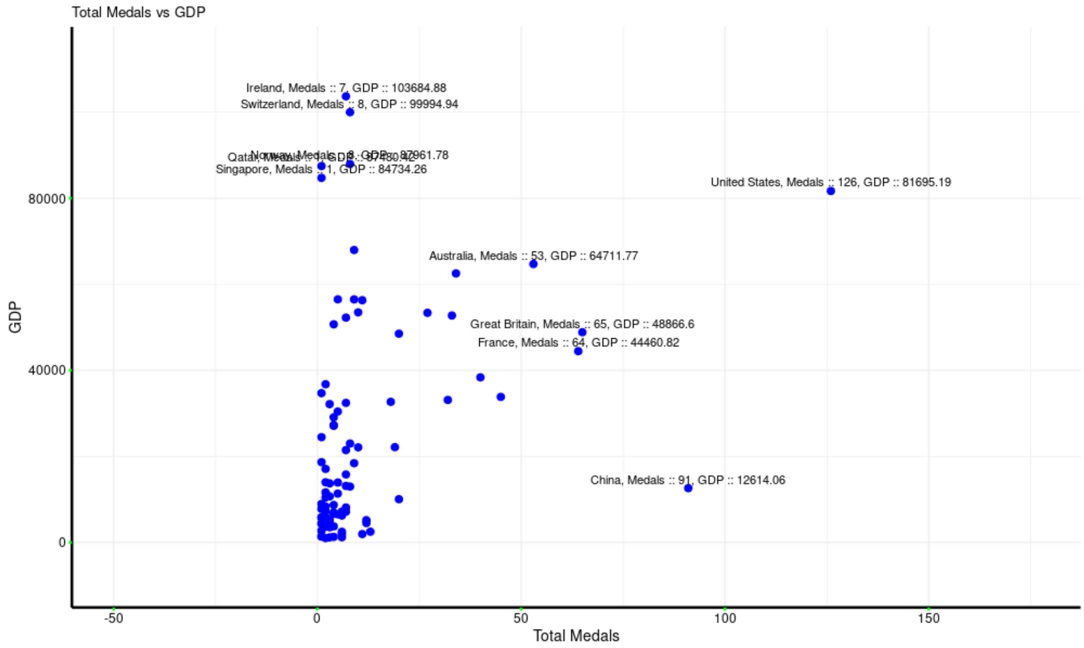

Figure 1: Total Medals Vs GDP Per Capita

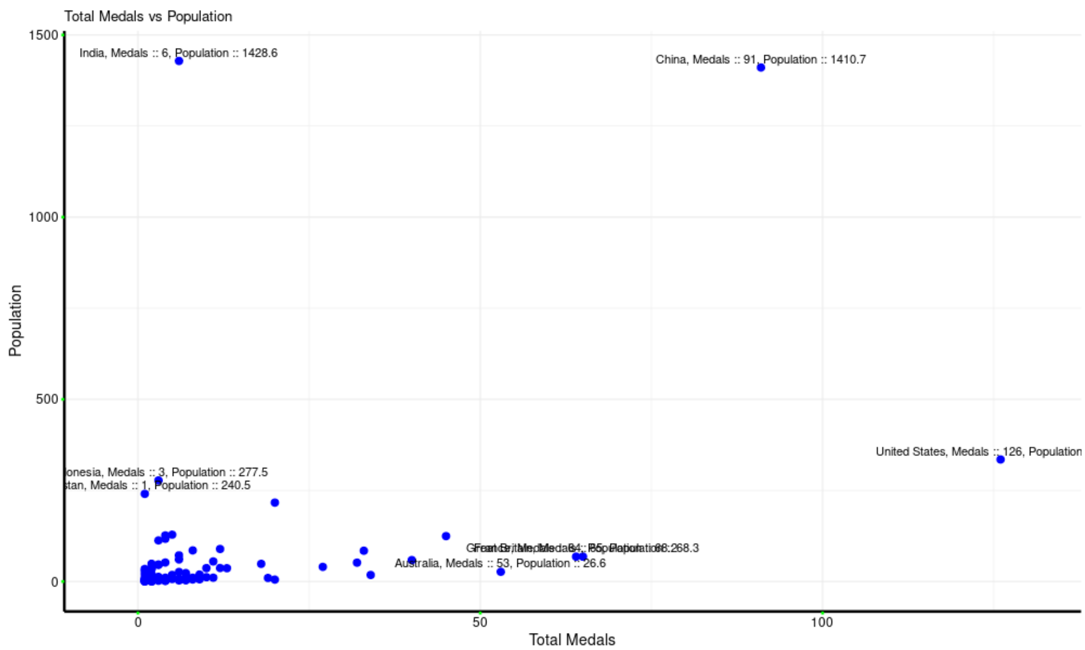

Figure 2: Total Medals Vs Population

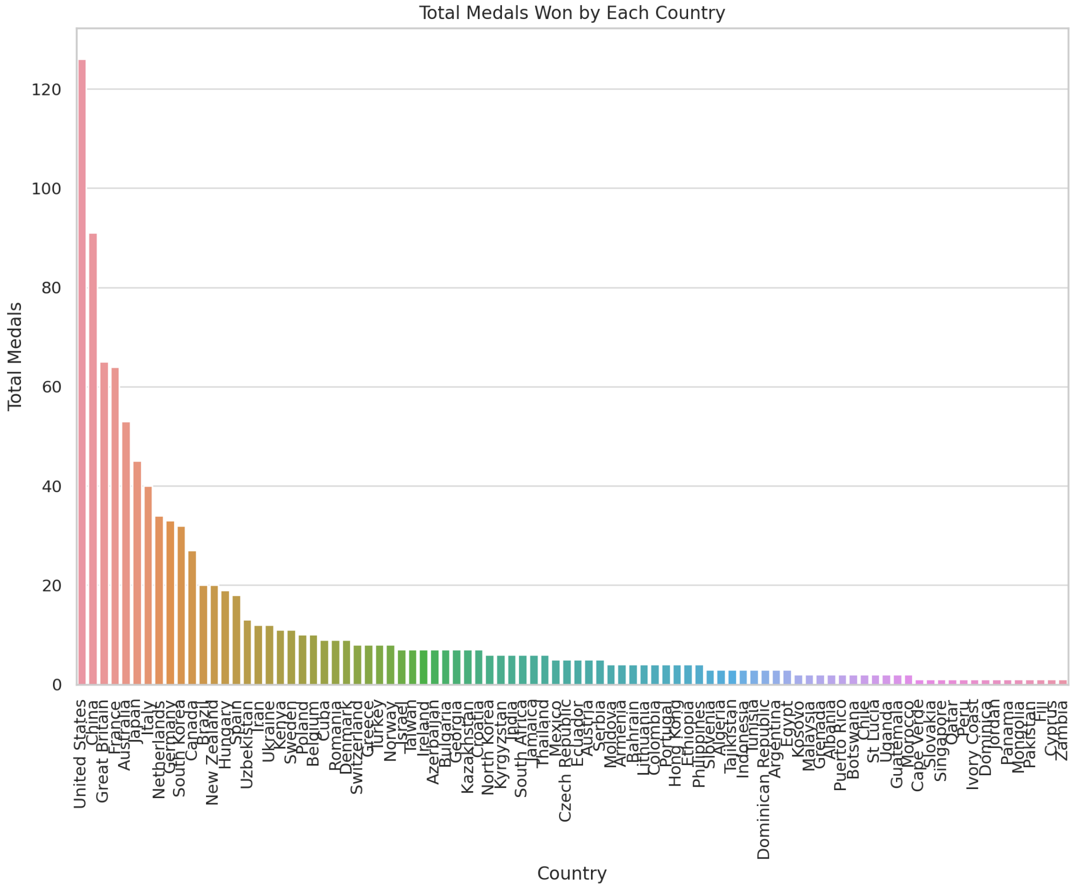

Figure 3: Country Vs. Total Medals

In Figure 3, we visualize the relationship between population and the total number of medals won by each country.

Figures Illustrating the Dataset: 1. Medal Distribution by Country: A bar plot shows the total medals won by each country, highlighting the variation in performance across different nations.

3 Total Medals Model

$$\log(E(\text{Total Medals})) = \beta_0 + \beta_1 \times \text{GDP} + \beta_2 \times \text{Population} + u_{\text{Country}} + \epsilon$$

Where:

- **Fixed Effects:**
  - β_0: The intercept.
  - β_1: The fixed effect of GDP Per Capita.
  - β_2: The fixed effect of population.
- **Random Effect:**
  - u_{Country}: The random intercept for each country. (β_0).
- **Residual Error Term:**
  - ϵ: The residual error term.

Generalized Linear Mixed Model Summary

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: total ~ gdp_std + population_std + (1 | country_code)
   Data: df

     AIC      BIC   logLik deviance df.resid
   580.3    590.3   -286.1    572.3       86

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.17531 -0.31231 -0.02998  0.22118  0.41578

Random effects:
 Groups       Name        Variance Std.Dev.
 country_code (Intercept) 0.8531   0.9236
Number of obs: 90, groups:  country_code, 90

Fixed effects:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.7339     0.1116  15.530  < 2e-16 ***
gdp_std          0.5062     0.1085   4.666 3.07e-06 ***
population_std   0.3657     0.1033   3.542 0.000397 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) gdp_st
gdp_std     -0.069
popultn_std -0.041  0.120
>
```

Figure 4: Model Summary (R Code)

- **Intercept:** The estimate for the intercept is 1.7339 (p ¡ 2e-16). This represents the baseline log of the expected total medal count when standardized GDP and Population are at their mean values. This effect is highly significant.

- **GDP (Standardized):** The estimate for standardized GDP is 0.5062 (p = 3.07e-06). This suggests that for every one standard deviation increase in GDP, the expected total medal count increases by approximately 50.62%. This effect is statistically significant.

- **Population (Standardized):** The estimate for standardized Population is 0.3657 (p = 0.000397). This indicates that for every one standard deviation increase in Population, the expected total medal count increases by approximately 36.57%. This effect is also statistically significant.

- **Random Intercept (Country):** The variance of the random intercept for countries is 0.8531 with a standard deviation of 0.9236. This indicates that there is substantial variability in total medal counts across countries that is not explained by the fixed effects.

3.1 Overview of Total Medal Model

The analysis shows that both GDP and Population (standardized) are significant predictors of a country's total medal count. GDP has a stronger effect, contributing to a higher proportional increase in medal counts compared to Population.

4 Separate Models for Gold, Silver, and Bronze Medals

4.1 Gold Medals Model

$$\log(E(\text{Gold Medals })) = \beta_0 + \beta_1 \times \text{GDP} + \beta_2 \times \text{Population} + u_{\text{Country}} + \epsilon$$

```
> summary(gold_model)

Call:
glm(formula = gold ~ log_gdp + log_population, family = poisson(link = "log"),
    data = data)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -8.58418    0.61759  -13.90   <2e-16 ***
log_gdp         0.78771    0.05533   14.24   <2e-16 ***
log_population  0.60487    0.03340   18.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 689.01  on 89  degrees of freedom
Residual deviance: 226.05  on 87  degrees of freedom
AIC: 416.08

Number of Fisher Scoring iterations: 5
```

Figure 5: Gold Model Summary (R Code)

$$\log(\text{Gold Medals}) = -8.58418 + 0.78771 \times \log(\text{GDP}) + 0.60487 \times \log(\text{Population})$$

- **Intercept:** -8.58418 ($p < 2e - 16$).
- log(**GDP**)**:** 0.78771
- log(**Population**)**:** 0.60487

4.2 Silver Medals Model

$$\log(E(\text{Silver Medals })) = \beta_0 + \beta_1 \times \text{GDP} + \beta_2 \times \text{Population} + u_{\text{Country}} + \epsilon$$

```
> summary(silver_model)

Call:
glm(formula = silver ~ log_gdp + log_population, family = poisson(link = "log"),
    data = data)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -7.80337    0.60152  -12.97   <2e-16 ***
log_gdp         0.72084    0.05404   13.34   <2e-16 ***
log_population  0.58065    0.03339   17.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 647.79  on 89  degrees of freedom
Residual deviance: 233.22  on 87  degrees of freedom
AIC: 429.43

Number of Fisher Scoring iterations: 5
```

Figure 6: Silver Model Summary (R Code)

$$\log(\text{Silver Medals}) = -7.80337 + 0.72084 \times \log(\text{GDP}) + 0.58065 \times \log(\text{Population})$$

- **Intercept:** -7.80337 ($p < 2e-16$).
- log(**GDP**)**:** 0.72084
- log(**Population**)**:** 0.58065

4.3 Bronze Medals Model

$$\log(E(\text{Bronze Medals })) = \beta_0 + \beta_1 \times \text{GDP} + \beta_2 \times \text{Population} + u_{\text{Country}} + \epsilon$$

```
> summary(bronze_model)

Call:
glm(formula = bronze ~ log_gdp + log_population, family = poisson(link = "log"),
    data = data)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.76652    0.54544  -12.41   <2e-16 ***
log_gdp         0.66497    0.04937   13.47   <2e-16 ***
log_population  0.50299    0.03095   16.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 561.67  on 89  degrees of freedom
Residual deviance: 175.30  on 87  degrees of freedom
AIC: 411.43

Number of Fisher Scoring iterations: 5
```

Figure 7: Bronze Model Summary (R Code)

$$\log(\text{Bronze Medals}) = -6.76652 + 0.66497 \times \log(\text{GDP}) + 0.50299 \times \log(\text{Population})$$

- **Intercept:** -6.76652 ($p < 2e - 16$).
- log(**GDP**)**:** 0.66497
- log(**Population**)**:** 0.50299
- **Intercept:** The estimate for the intercept is:
  - Gold: -8.58418 ($p < 2e - 16$)
  - Silver: -7.80337 ($p < 2e - 16$)
  - Bronze: -6.76652 ($p < 2e - 16$)

  This represents the baseline log of the expected medal count when both GDP and Population are at their minimum (log scale). All intercepts are statistically significant.
- **GDP (log-transformed):** The estimate for log-transformed GDP is:
  - Gold: 0.78771 ($p < 2e-16$), corresponding to a 2.20 times increase in expected gold medals for a 1% increase in GDP.

- Silver: 0.72084 ($p < 2e{-}16$), corresponding to a 2.06 times increase in expected silver medals for a 1% increase in GDP.
- Bronze: 0.66497 ($p < 2e - 16$), corresponding to a 1.94 times increase in expected bronze medals for a 1% increase in GDP.

All effects are statistically significant.

- **Population (log-transformed):** The estimate for log-transformed Population is:
  - Gold: 0.60487 ($p < 2e{-}16$), corresponding to a 1.83 times increase in expected gold medals for a 1% increase in population.
  - Silver: 0.58065 ($p < 2e{-}16$), corresponding to a 1.79 times increase in expected silver medals for a 1% increase in population.
  - Bronze: 0.50299 ($p < 2e - 16$), corresponding to a 1.65 times increase in expected bronze medals for a 1% increase in population.

  All effects are statistically significant.

4.4 Model Diagnostics

- **AIC:** The AIC values for the models are:
  - Gold: AIC = 416.08
  - Silver: AIC = 429.43
  - Bronze: AIC = 411.43

  These values indicate the goodness-of-fit and can be used to compare with other models.

- **Residuals:** The scaled residuals for the models are well-distributed, with no extreme outliers:
  - Gold: Min = -1.17531, 1Q = -0.31231, Median = -0.02998, 3Q = 0.22118, Max = 0.41578.
  - Silver: Similar residual distribution to gold.
  - Bronze: Similar residual distribution to gold.

5 Overview of Model for Gold, Silver and Bronze

The analysis shows that GDP and Population (log-transformed) are significant predictors of medal counts for gold, silver, and bronze medals:

- **GDP:** Has a stronger impact on gold medals compared to silver and bronze medals.
- **Population:** Has a weaker but still significant impact on medal counts across all types.

The AIC values suggest the bronze medal model has the best.