National Olympic Performance: Analyzing Economic, Demographic, and Historical Determinants of Summer Olympics Medal Counts

David Almona

Centre College

Fall Semester 2024

I. Introduction

The Summer Olympics is a major international multi-sports event hosted every four years that stretches back to 1896 in Athens, Greece. The traditions of awarding medals began in 1904 as gold medals are awarded for first place, silver medals for second place, and bronze medals for third place. However, what factors play a role in the success of nations at the Olympic games, as defined by the number of medals won? This paper looks at the relationship between a country's medal count at the Summer Olympic Games and its economic factors, as well as some other factors.

II. Literature Review

Several research papers have examined the factors influencing a nation's success at the Summer Olympic Games using quantitative analysis. While all sources acknowledge the importance of athletic excellence, they focus on identifying structural determinants of medal outcomes, primarily using regression analysis to explore these relationships.

Economic resources emerge as the most consistent and powerful predictor of Olympic success across the research literature. A positive relationship between economic resources, whether GDP, GDP per capita, or GDP share, is frequently observed. This suggests that wealthier nations have more resources to invest in athletic development, like equipment and facilities, athlete support, etc. Bernard and Busse (2004) emphasize the importance of considering both population size and GDP per capita, as a large population alone is insufficient for Olympic success without adequate resources per individual.

Research has identified several key socioeconomic indicators beyond GDP that influence Olympic success. Hosein et al. (2013) identified literacy levels as a potentially significant factor, suggesting that education plays a role in developing athletes and fostering a culture that values sports. Moosa

and Smith (2004) investigate the impact of health expenditure and education expenditure as indicators of a nation's commitment to overall well-being, finding it to be an important predictor of Olympic success. Also, Forrest et al. (2010) suggests that public spending on recreation is a relevant factor, as it reflects a nation's broader support for sports and leisure activities, which may contribute to a stronger athletic foundation.

Johnson and Ali (2004) looked at the impact of political systems on Olympic performance. They observe that communist and single-party regimes historically achieved greater success in the Olympics. Possible explanations for this phenomenon include greater state control over resource allocation to sports, the ability to prioritize athletic development over individual freedoms, and the use of sports to enhance national prestige. However, Forrest et al. (2010), who considered the Soviet bloc, cautioned that this trend may diminish as former communist nations transition to more democratic systems.

Lui and Suen (2008) and other sources recognize that host nations tend to outperform expectations at the Olympics. This can be attributed to increased investment in athletic infrastructure leading up to the Games, home crowd support, and familiarity with the competition venues. Forrest et al. (2010) argues that even the anticipation of hosting future Games can incentivize nations to elevate their athletic programs, leading to improved performance in preceding Olympics.

These sources use diverse econometric techniques, like ordinary least squares (OLS) regression, Tobit and Poisson regressions, and Extreme Bounds Analysis (EBA), highlighting the complexity of modeling Olympic success. They also acknowledge limitations in data availability and suggest areas to improve future research.

Overall, this paper, along with the cited articles, provides a comprehensive analysis of the factors contributing to a nation's medal count at the Olympic Games. This understanding can inform policy decisions regarding resource allocation to sports and the promotion of athletic development.

III. MODEL SPECIFICATION

Dependent Variable:

MEDALCOUNT$_{it}$ = the total number of medals won by country i in the Summer Olympics of year t.

Independent Variables: (include variable modifications???)

GNIPPP$_{it}$ = Gross National Income per capita for country i four years prior to the games, adjusted for purchasing power parity (PPP), in thousands of current international dollars.

GDPPPP$_{it}$ = Gross Domestic Product per capita for country i four years prior to the games, adjusted for purchasing power parity (PPP), in thousands of current international dollars. (not used in the final model, **GNIPPP$_{it}$** proved to be a better fit).

POPULATION$_{it}$ = Population of country i four years prior to the Olympic event, in millions of people.

URBPOPPCT$_{it}$ = Percentage of the population living in urban areas in country i four years prior to the Olympic event.

HOSTNOW$_{it}$ = 1 if country i is hosting the Summer Olympics in year t, 0 otherwise.

HOSTINFOUR$_{it}$ = 1 if country i will host the next Summer Olympics in four years, 0 otherwise.

HOSTLASTFOUR$_{it}$ = 1 if country i hosted the last Summer Olympics four years ago, 0 otherwise.

TIMESHOST$_{it}$ = Number of times country i has hosted the Summer Olympics prior to current year t.

FAILEDBID$_{it}$ = Number of unsuccessful bids by country i to host the Summer Olympics, including most recent year, from 2000 – 2024, excluding withdrawn bids.

NUMATHLETES$_{it}$ = Number of athletes representing country i in the Summer Olympics of year t.

MEDALLASTFOUR$_{it}$ = Number of medals won by country i in the last Summer Olympics, four years ago.

HEALTHEXPENDPPP$_{it}$ = Current health expenditure per capita, adjusted for purchasing power parity (PPP), in hundreds of current international dollars.

ASIA$_i$ = 1 if country i is in Asia, 0 otherwise. (not used in the final model)

AFRICA$_i$ = 1 if country i is in Africa, 0 otherwise. (not used in the final model)

NORTHAMERICA$_i$ = 1 if country i is in North America, 0 otherwise. (not used in the final model)

EUROPE$_i$ = 1 if country i is in Europe, 0 otherwise. (not used in the final model)

SOUTHAMERICA$_i$ = 1 if country i is in South America, 0 otherwise. (not used in the final model)

OCEANIA$_i$ = 1 if country i is in Oceania, 0 otherwise. (not used in the final model)

YEAR = Categorical variable: year t of Summer Olympics

Variable	Obs	Mean	Std. dev.	Min	Max
Year	716	2015.006	5.915486	2008	2024
medalcount	716	5.296089	14.28468	0	126
gnippp	716	16.71994	19.09643	.46	152.93
gdpppp	716	17.20408	20.01232	.4740153	163.5428
population	716	38.06982	141.5762	.009791	1411.1
urbpoppct	716	55.98212	22.90374	9.139	100
healthexpe~p	716	11.7338	15.47988	.2054935	117.5842
hostnow	716	.0055866	.0745865	0	1
hostinfour	716	.0055866	.0745865	0	1
hostlastfour	716	.0055866	.0745865	0	1
timeshost	716	.1424581	.4982056	0	4
failedbid	716	.1815642	.6306802	0	5
numathletes	716	58.21229	103.1498	1	619
medallastf~r	716	5.132682	13.9063	0	113
asia	716	.2458101	.4308674	0	1
africa	716	.2765363	.4475978	0	1
northamerica	716	.122905	.3285577	0	1
europe	716	.2150838	.4111677	0	1
southamerica	716	.0614525	.2403264	0	1
oceania	716	.0782123	.2686931	0	1

TABLE 1: SUMMARY STATISTICS

IV. Expected Signs of Coefficients

GNIPPP$_{it}$: Gross National Income is the total amount of money earned by a nation's people and businesses. It is used to measure and track a nation's wealth from year to year. As a result, the sign of the coefficient is expected to be positive because high GNI means the nation's people can afford good food, access to water, and resources to train for the Summer Olympics like access to equipment and facilities, coaches and training clubs/camps.

POPULATION$_{it}$: A nation with a greater population has a larger talent pool for discovering exceptional athletes and allows for focused training across different sports rather than specializing in one. Therefore, the sign of the coefficient is expected to be positive.

URBPOPPCT$_{it}$: A nation with high urban population percentage often has more sports infrastructure, training centers, and coaching resources which can leads to more medals at the

Summer Olympics. However, a very high urban population percentage might cause overcrowding problems that limit access to quality training for some and might also mean the nation prioritizes academics or work over sports engagements. Therefore, the sign of the coefficient is ambiguous (+/-).

HOSTNOW$_{it}$: Hosting the Summer Olympics is a major boost to the total medal won at the Games. Firstly, being the host country means you have well equipped sports infrastructure boosting athlete development. Secondly, the host nation is selected about seven years prior to the Games, which means they would heavily invest in searching for and training younger talents in sports. Thirdly, athletes representing the host nation have "home" advantages as they are familiar with the facilities and climate and do not have to worry about traveling, jetlag and other inconveniences that come from going elsewhere to compete. Lastly, the host nation can field way more athletes (almost double or triple that of non-host nations) because they do not have to worry about travel costs leading to a greater chance to win more medals. Wildcard entries mean they can qualify into more events. As a result, the sign of the coefficient is expected to be positive.

HOSTINFOUR$_{it}$: As the host nation for the Summer Olympics is chosen about seven years prior, a nation would begin investing heavily in their scouting and training of younger talents, sports facilities improvements, coaching resources, etc. Therefore, the sign of the coefficient is expected to be positive.

HOSTLASTFOUR$_{it}$: A nation that host the last Olympics edition four years prior must still have some of the benefits that came with hosting the last edition. They still have their improved facilities and resources and the younger athletes they had developed should have developed more or could even be in their prime. Therefore, the sign of the coefficient is expected to be positive.

TIMESHOST$_{it}$: A nation's history of hosting the Olympics can speak about the sporting culture that exists within that nation and the amount of investment the government of that nation has put into sports development. As a result, the sign of the coefficient is expected to be positive.

FAILEDBID$_{it}$: To bid to host the Olympic Games, a city must meet several criteria. These include development plans, sustainability, funding, support, infrastructure, safety and experience hosting sports events. The number of unsuccessful bids can show a nation's investment into sports leading to higher medals. However, high numbers of unsuccessful bids could mean misallocated resources at the detriment of training and athlete development. Therefore, the sign of the coefficient is ambiguous (+/-).

NUMATHLETES$_{it}$: More athletes representing a nation at the Olympics means more chances to win a medal. The sign of the coefficient is expected to be positive.

MEDALLASTFOUR$_{it}$: This is a lagged form of the variable **MEDALCOUNT$_{it}$**. Winning more medals at the prior Olympics should be a good indicator that a nation would win about the same number of medals at the current Olympics because they already have the momentum. As a result, the sign of the coefficient is expected to be positive.

HEALTHEXPENDPPP$_{it}$: High health expenditure means a nation is investing in the health and wellness of its citizens (and athletes). This means their athletes have access to good medical care, injury prevention programs, recovery support and nutrition. Therefore, the sign of the coefficient is expected to be positive.

The null and alternative hypothesis for each of these variables are:

Positive expected coefficients: $GNIPPP_{it}$, $POPULATION_{it}$, $HOSTNOW_{it}$, $HOSTINFOUR_{it}$, $HOSTLASTFOUR_{it}$, $TIMESHOST_{it}$, $NUMATHLETES_{it}$, $MEDALLASTFOUR_{it}$, $HEALTHEXPENDPPP_{it}$

$$H_0: \beta \leq 0$$

$$H_A: \beta > 0$$

Negative expected coefficients: there are no variables with negative expected sign.

$$H_0: \beta \geq 0$$

$$H_A: \beta < 0$$

Ambiguous expected coefficients: $URBPOPPCT_{it}$, $FAILEDBID_{it}$

$$H_0: \beta = 0$$

$$H_A: \beta \neq 0$$

V. Data Collection

The dataset contains 716 observations across the 2008, 2012, 2016, and 2024 Olympic Games. Specifically, there are 178 observed countries from 2008, 180 observed countries from 2012, 179 observed countries from 2016, and 179 observed countries from 2024. The 2020 Olympics (that was postponed to 2021) is excluded because it was severely affected by the COVID-19 pandemic. The data for the variables $GNIPPP_{it}$, $POPULATION_{it}$, $HEALTHEXPENDPPP_{it}$, and $URBPOPPCT_{it}$ were sourced from World Bank Open Data which provides free and open access to global development data. Initially, the dataset contained over 800 observations but as more data were collected, World Bank Open Data had no recorded value for some variables thereby excluding that observation nation entirely. The remaining variables were sourced from Wikipedia.

As this dataset combines both time-series and cross-sectional data by observing the same countries over different Olympic time periods, it is classified as panel data.

VI. Estimating the Equation

This is an analysis of four different model specifications.

First, a linear regression model was estimated treating the dataset as a cross-section. The regression equation:

$$MEDALCOUNT_i = \beta_0 + \beta_1 GNIPPP_i + \beta_2 POPULATION_i + \beta_3 URBPOPPCT_i + \beta_4 HEALTHEXPENDPPP_i + \beta_5 HOSTNOW_i + \beta_6 HOSTINFOUR_i + \beta_7 HOSTLASTFOUR_i + \beta_8 TIMESHOST_i + \beta_9 FAILEDBID_i + \beta_{10} NUMATHLETES_i + \beta_{11} MEDALLASTFOUR_i + \beta_{12} ASIA_i + \beta_{13} AFRICA_i + \beta_{14} NORTHAMERICA_i + \beta_{15} SOUTHAMERICA_i + \beta_{16} EUROPE_i + \varepsilon_i$$

Source	SS	df	MS	Number of obs	=	716
Model	140205.308	16	8762.83174	F(16, 699)	=	1076.13
Residual	5691.92122	699	8.14294882	Prob > F	=	0.0000
				R-squared	=	0.9610
Total	145897.229	715	204.052069	Adj R-squared	=	0.9601
				Root MSE	=	2.8536

medalcount	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
gnippp	-.0151902	.0112163	-1.35	0.176	-.0372119	.0068315
population	.0023382	.00096	2.44***	0.015	.0004534	.0042231
urbpoppct	-.0061237	.006622	-0.92	0.355	-.0191251	.0068777
healthexpendppp	.0419771	.0157801	2.66***	0.008	.0109949	.0729592
hostnow	20.22254	1.657479	12.20***	0.000	16.96831	23.47678
hostinfour	9.405838	1.513194	6.22***	0.000	6.434889	12.37679
hostlastfour	-5.834023	1.527912	-3.82	0.000	-8.83387	-2.834177
timeshost	2.757301	.4456818	6.19***	0.000	1.882266	3.632337
failedbid	-.6510768	.221116	-2.94***	0.003	-1.085208	-.2169457
numathletes	.010427	.0030595	3.41***	0.001	.0044201	.016434
medallastfour	.818112	.0205527	39.81***	0.000	.7777596	.8584644
asia	.8563723	.4648359	1.84	0.066	-.0562696	1.769014
africa	.3416881	.4384575	0.78	0.436	-.5191635	1.20254
northamerica	.5616713	.4924416	1.14	0.254	-.4051706	1.528513
europe	-.7504699	.4818345	-1.56	0.120	-1.696486	.1955463
southamerica	-.1854534	.5989708	-0.31	0.757	-1.361451	.9905441
_cons	-.1035019	.4800059	-0.22	0.829	-1.045928	.8389242

TABLE 1: Linear Regression Model (as cross-sectional data)

The model shown in TABLE 1 appears to be a real model as shown by the adjusted R^2 of 0.9601, indicating that 96% of the variance in the dependent variable, $MEDALCOUNT_i$, is explained by the independent variables. Also, eight variables are statistically significant at the 1% level, which include $HOSTNOW_i$, $HOSTINFOUR_i$, and $TIMESHOST_i$. Here, the coefficient of $HOSTNOW_i$ means that if $country_i$ hosts the Olympics, their medal count increases by 20, ceteris paribus.

However, treating the model as a cross-section, rather than panel data, does not account for differences that exist between countries that might influence medal counts such as culture, history, and attitude towards sports. Many of these differences are constant or evolve slowly over time and cannot be directly measured. As a result, omitting these relevant variables from the model forces them into the error term, introducing omitted variable bias and putting the validity of the model into question.

To address these problems, dummy variables for the countries and years were generated and included in the following re-estimated models. This approach now treats the dataset as panel data, accounting for unobserved differences, and improving the model's ability to explain variations in medal count while mitigating omitted variable bias.

Second, a linear regression model was estimated treating the dataset as panel data. The regression equation:

$$MEDALCOUNT_{it} = \beta_0 + \beta_1 GNIPPP_{it} + \beta_2 POPULATION_{it} + \beta_3 URBPOPPCT_{it} + \beta_4 HEALTHEXPENDPPP_{it} + \beta_5 HOSTNOW_{it} + \beta_6 HOSTINFOUR_{it} + \beta_7 HOSTLASTFOUR_{it} + \beta_8 TIMESHOST_{it} + \beta_9 FAILEDBID_{it} + \beta_{10} NUMATHLETES_{it} + \beta_{11} MEDALLASTFOUR_{it} + \alpha_1 CNTRY_1 + \ldots + \alpha_{184} CNTRY_{184} + \rho_1 YR_1 + \ldots + \rho_3 YR_3 + \varepsilon_{it}$$

```
Linear regression, absorbing indicators                     Number of obs     =    716
Absorbed variable: Country                                  No. of categories =    185
                                                            F(14, 517)        =  16.33
                                                            Prob > F          = 0.0000
                                                            R-squared         = 0.9823
                                                            Adj R-squared     = 0.9756
                                                            Root MSE          = 2.2334

------------------------------------------------------------------------------
     medalcount | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
----------------+-------------------------------------------------------------
         gnippp |  -.0043211   .0223763    -0.19   0.847    -.0482808    .0396386
     population |  -.0067608   .0104269    -0.65   0.517     -.027245    .0137235
      urbpoppct |    .047426   .0482282     0.98   0.326    -.0473212    .1421733
healthexpendppp |   .0736413   .0269294     2.73***   0.006     .0207369    .1265457
        hostnow |   7.516541   2.296161     3.27***   0.001     3.005588    12.02749
     hostinfour |   4.846409   1.587834     3.05***   0.002      1.72701    7.965808
   hostlastfour |   2.559551    1.56764     1.63*   0.103    -.5201775    5.639279
      timeshost |   5.137346   2.042372     2.52***   0.012     1.124977    9.149714
      failedbid |   .4288276   .5159609     0.83   0.406    -.5848101    1.442465
    numathletes |   .0363354   .0067353     5.39***   0.000     .0231035    .0495673
  medallastfour |   .0624778   .0425517     1.47*   0.143    -.0211178    .1460734
            Yr1 |   .2454455    .436205     0.56   0.574    -.6115067    1.102398
            Yr2 |   .1852562   .3600357     0.51   0.607    -.5220567    .8925691
            Yr3 |  -.0837492   .3040845    -0.28   0.783    -.6811423     .513644
          _cons |  -1.308965   2.940569    -0.45   0.656    -7.085898    4.467967
------------------------------------------------------------------------------
F test of absorbed indicators: F(184, 517) = 3.562            Prob > F = 0.000


------------------------------------------------------------------------------
          Model |          N   ll(null)  ll(model)      df        AIC        BIC
----------------+-------------------------------------------------------------
              . |        716  -1605.784  -1474.718      15   2979.436   3048.041
------------------------------------------------------------------------------
```

TABLE 2: Linear Regression Model (as panel data)

By including dummy variables for countries (absorbed indicators), the model in TABLE 2 accounts for country-specific unobservable characteristics. These characteristics, if omitted, could bias the estimates due to their influence on the dependent variable, MEDALCOUNTit. The inclusion of these fixed effects lessens the severity of omitted variable bias, ensuring a more accurate estimation of the independent variables' effects on medal counts.

This model shows improvement compared to the cross-sectional model, as shown by an adjusted R^2 of 0.9756, indicating that 97.6% of the variance in the dependent variable, $MEDALCOUNT_{it}$, is explained by the independent variables, including the absorbed country effects. The small difference between the R^2 and the adjusted R^2 shows that the independent variables seem very important. Here, the coefficient of $HOSTNOW_i$ means that if $country_i$ hosts the Olympics, their medal count increases by approximately 8, ceteris paribus.

The statistical significance of the variables is marked in each table with asterisks denoting the extent of their significance; 10% with one asterisk (*), 5% with two asterisks (**), and 1% with three asterisks (***). Seven variables are statistically significant, with five of them being significant at the 1% level. Two variables expected to be relevant to the model turned out to be statistically insignificant – $GNIPPP_{it}$ and $POPULATION_{it}$. Besides just being statistically insignificant, the coefficients of these variables have a negative sign. This interprets that as $GNIPPP_{it}$ and $POPULATION_{it}$ increase, the medal count for a nation at the Olympics decreases by 0.0043211 and 0.0067608 respectively.

The generated dummy variables for the years (Yr1, Yr2, and Yr3) appear to be statistically insignificant. This hints that the country-specific differences and independent variables constant or evolve slowly over time. To be certain, a joint hypothesis test was run on the coefficients of these variables. The null hypothesis, alternative hypothesis, and results are shown below:

$$H_0: \beta_{Yr1} = \beta_{Yr2} = \beta_{Yr3} = 0$$

H_A: At least one of the coefficients is not zero.

```
F(  3,   516) =    0.36
  Prob > F =    0.7825
```

Since this p-value (0.7825) is much larger than typical significance levels (0.10, 0.05 or 0.01), we fail to reject the null hypothesis. So, there is no evidence to suggest that the variables Yr1, Yr2, and Yr3 jointly have a statistically significant effect on the dependent variable, $MEDALCOUNT_{it}$, in the model.

Third, a left-hand side (LHS) semi-log regression model with the estimated regression equation:

$$LnMEDALCOUNT_{it} = \beta_0 + \beta_1 GNIPPP_{it} + \beta_2 POPULATION_{it} + \beta_3 URBPOPPCT_{it} + \beta_5 HEALTHEXPENDPPP_{it} + \beta_6 HOSTNOW_{it} + \beta_7 HOSTINFOUR_{it} + \beta_8 HOSTLASTFOUR_{it} + \beta_9 TIMESHOST_{it} + \beta_{10} FAILEDBID_{it} + \beta_{11} NUMATHLETES_{it} + \beta_{12} MEDALLASTFOUR_{it} + \alpha_1 CNTRY_1 + \ldots + \alpha_{184} CNTRY_{184} + \rho_1 YR_1 + \ldots + \rho_3 YR_3 + \varepsilon_{it}$$

```
Linear regression, absorbing indicators                Number of obs     =    323
Absorbed variable: Country                             No. of categories =    106
                                                       F(14, 203)        =   1.94
                                                       Prob > F          = 0.0241
                                                       R-squared         = 0.9187
                                                       Adj R-squared     = 0.8710
                                                       Root MSE          = 0.4467

------------------------------------------------------------------------------
  lnmedalcount | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
---------------+--------------------------------------------------------------
         gnippp |   -.003873   .0068163    -0.57   0.571    -.0173129    .0095669
     population |   -.001766   .0022692    -0.78   0.437    -.0062402    .0027082
      urbpoppct |   .0146249   .0181536     0.81   0.421    -.0211688    .0504186
healthexpendppp |  -.0024372   .0075135    -0.32   0.746    -.0172517    .0123772
        hostnow |  -.5492644   .4712627    -1.17   0.245    -1.478462    .3799332
     hostinfour |   .0960023   .3222107     0.30   0.766    -.5393066    .7313112
   hostlastfour |  -.0881118     .31945    -0.28   0.783    -.7179773    .5417538
      timeshost |  -.1494517    .413121    -0.36   0.718    -.9640102    .6651068
      failedbid |  -.1483904   .1088436    -1.36   0.174    -.3629993    .0662186
     numathletes |    .003546   .0014299     2.48***  0.014     .0007266    .0063653
  medallastfour |   .0062588   .0086554     0.72   0.470    -.0108071    .0233248
            Yr1 |  -.3178235   .1603688    -1.98**  0.049    -.6340258   -.0016213
            Yr2 |  -.2021082   .1252943    -1.61   0.108    -.4491534    .0449369
            Yr3 |  -.1851479   .1012487    -1.83*   0.069    -.3847819    .0144862
          _cons |   .7317246    1.24558     0.59   0.558    -1.724209    3.187658
------------------------------------------------------------------------------
F test of absorbed indicators: F(105, 203) = 4.499            Prob > F = 0.000


------------------------------------------------------------------------------
          Model |          N      ll(null)  ll(model)      df        AIC        BIC
----------------+-------------------------------------------------------------
              . |        323     -143.3047  -123.0111      15   276.0222    332.687
------------------------------------------------------------------------------
```

TABLE 3: LHS Semi-Log Regression Model (as panel data)

The LHS semi-log regression model shown in TABLE 3 has an adjusted R^2 of 0.871, indicating that 87% of the variance in the dependent variable, $MEDALCOUNT_{it}$, is explained by the independent variables, including the absorbed country effects. This adjusted R^2 is lower compared to that of TABLE 2. Additionally, this model has just three statistically significant variables with

one being statistically significant at the 1% level. Therefore, this is a worse model when compared to the one in TABLE 2. Here, the coefficient of $GNIPPP_{it}$ means that if the number of athletes representing $country_i$ at the Olympics increase by 1, their medal count would increase by roughly 0.35%, ceteris paribus.

It is important to note, however, that there are 323 observations in the model rather than the 716 observations from the dataset. Using $LnMEDALCOUNT_{it}$, the log form of the dependent variable $MEDALCOUNT_{it}$, excludes 393 observations with a medal count of zero (0), as log 0 is undefined and not a real number. The relationships estimated by the model are now biased because the model now only reflects cases where medals are won. Although the AIC and BIC values in TABLE 3 are much lower than those for TABLE 2, they are non-comparable because the dependent variable in not in the same functional form–one is in log, while the other is not.

Lastly, the regression model estimated below appeared to be the best predictor of medal counts at the Olympics:

$$MEDALCOUNT_{it} = \beta_0 + \beta_1 LnGNIPPP_{it} + \beta_2 POPULATION_{it} + \beta_3 URBPOPPCT_{it} + \beta_4 URBPOPPCT^2_{it} + \beta_5 HEALTHEXPENDPPP_{it} + \beta_6 HOSTNOW_{it} + \beta_7 HOSTINFOUR_{it} + \beta_8 HOSTLASTFOUR_{it} + \beta_9 TIMESHOST_{it} + \beta_{10} FAILEDBID_{it} + \beta_{11} NUMATHLETES_{it} + \beta_{12} MEDALLASTFOUR_{it} + \alpha_1 CNTRY_1 + \ldots + \alpha_{184} CNTRY_{184} + \rho_1 YR_1 + \ldots + \rho_3 YR_3 + \varepsilon_{it}$$

```
Linear regression, absorbing indicators                     Number of obs     =        716
Absorbed variable: Country                                  No. of categories =        185
                                                            F(15, 516)        =      15.72
                                                            Prob > F          =     0.0000
                                                            R-squared         =     0.9825
                                                            Adj R-squared     =     0.9758
                                                            Root MSE          =     2.2242

------------------------------------------------------------------------------
  medalcount | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    lngnippp |  -.4056893   .5881478    -0.69   0.491    -1.561148    .7497694
  population |  -.0017872     .01061    -0.17   0.866    -.0226313    .0190568
   urbpoppct |  -.1764805   .1261443    -1.40   0.162    -.4243001    .0713392
  urbpoppct2 |   .1984389   .1021162     1.94*  0.053    -.0021757    .3990536
healthexpendppp |   .0600674   .0237498     2.53***  0.012     .0134092    .1067257
     hostnow |   7.277349   2.288057     3.18***  0.002     2.782296     11.7724
  hostinfour |   4.791971   1.579029     3.03***  0.003     1.689854    7.894088
hostlastfour |   2.783943   1.564166     1.78**   0.076    -.2889742     5.85686
   timeshost |   4.819905   2.038806     2.36***  0.018     .8145223    8.825287
   failedbid |   .3539503   .5142593     0.69   0.492    -.6563492     1.36425
 numathletes |   .0362403     .00671     5.40***  0.000     .0230579    .0494226
medallastfour |   .0572526   .0424241     1.35*  0.178    -.0260926    .1405979
         Yr1 |  -.0866798   .5400361    -0.16   0.873     -1.14762    .9742599
         Yr2 |  -.0151693   .3993777    -0.04   0.970    -.7997755     .769437
         Yr3 |  -.2239757   .3243986    -0.69   0.490    -.8612801    .4133286
       _cons |   5.018036   4.025408     1.25   0.213    -2.890167    12.92624
------------------------------------------------------------------------------
F test of absorbed indicators: F(184, 516) = 3.605            Prob > F = 0.000


------------------------------------------------------------------------------
       Model |          N      ll(null)  ll(model)      df        AIC        BIC
-------------+----------------------------------------------------------------
           . |        716     -1605.784  -1471.054      16   2974.108   3047.287
------------------------------------------------------------------------------
```

TABLE 4: Regression model with modifications to $GNIPPP_{it}$ and $URBPOPPCT_{it}$ (as panel data)

The regression model shown in TABLE 4 has an adjusted R^2 of 0.9758, indicating that 97.6% of the variance in the dependent variable, $MEDALCOUNT_{it}$, is explained by the independent variables, including the absorbed country effects. This is 0.02% better than the regression model in TABLE 2 and has slightly lower AIC and BIC values, showing that the regression model in TABLE 4 is a superior model, even though the difference is minimal.

TABLE 4 adds modifications to the variables $GNIPPP_{it}$ and $URBPOPPCT_{it}$. GNI per capita data is often right skewed (or positively skewed), with few countries having significantly higher values when compared to the rest of the world. Using the log form reduces this skewness.

Additionally, $URBPOPPCT^2_{it}$, the polynomial form of the variables $URBPOPPCT_{it}$ is added. From the variable's definitions, a nation with a high urban population percentage often has more sports infrastructure, training centers, and coaching resources which can lead to more medals at the Summer Olympics. However, a very high urban population percentage might cause overcrowding problems that limit access to quality training for some and might also mean the nation prioritizes academics or work over sports engagements. This hints at an inverted U-shaped curve when graphed. However, the coefficients of $URBPOPPCT_{it}$ and $URBPOPPCT^2_{it}$ are negative and positive, respectively, meaning a U-shaped curve which is the opposite of what was expected.

Here, the coefficient of $LnGNIPPP_i$ means that if the GNI per capita of $country_i$ increases by 1%, their medal count decreases by 0.00041, ceteris paribus and the the slope can be calculated thus:

$$\frac{\Delta MEDALCOUNT}{\Delta URBPOPPCT_1} = \beta_1 + 2\beta_1 URBPOPPCT_1$$

$$\frac{\Delta MEDALCOUNT}{\Delta URBPOPPCT_1} = -0.1764805 + 2(0.1984389)(URBPOPPCT_1) = 0.445$$

This means that as the urban population percentage increases by 1%, the medal count for a country would change by 0.445, ceteris paribus.

Even though the relationship isn't as expected, this still suggests a potential non-linear relationship between urban population percentage and medal count even though the variable is statistically insignificant after running a t-test.

VI. Evaluation

When running a multiple regression analysis, many things could go wrong. This is how I mitigate these possible problems.

1. Omitted Variable

   This is the omission of a relevant independent variable. I mitigated this by conducting a literature review and including relevant variables that other researchers considered. I decided to add variables like $FAILEDBID_{it}$ because I thought a country bidding to host the Olympics could indicated the resources and sports investments which could potentially influence medal count. I also added the variable $TIMESHOST_{it}$ to provide a

holistic analysis of the "hosting effect" alongside the variables $HOSTNOW_{it}$, $HOSTINFOUR_{it}$, and $HOSTLASTFOUR_{it}$.

2. Irrelevant Variable

   This is the inclusion of a variable that does not belong in the equation, thereby reducing the accuracy of the standard errors and affecting the t-scores and confidence intervals. During my analysis, I did not remove any variables because I found none to be irrelevant.

3. Incorrect Functional Form

   Incorrect functional forms result in biased estimates, poor fit, and difficult interpretation. An example of this is the model that uses LnMEDALCOUNT. This model excluded 300+ observations and introduced biased. I did not proceed with that model.

4. Multicollinearity

   This is when some of the independent variables are imperfectly or perfectly correlated, causing the estimates of standard errors and t-scores to be unreliable. Through testing, there was multicollinearity found between the variables NUMATHLETES and MEDALLASTFOUR. This was no concern are the latter served as a lagged variable of the dependent variable.

5. Serial Correlation

   This is when the observations of the error term are correlated. As this only affects time series data, it is not a problem for this study

6. Heteroskedasticity

   This is when the variance of the error term is not constant for all observations. As this study used panel data, heteroskedasticity is not a problem.

VIII. Conclusion

This study gives a comprehensive analysis of the factors influencing national performance at the Summer Olympic Games, with complex relationships being evident between economic, demographic, and historical variables and medal counts. Through multiple regression analysis', the research shows the importance of the "hosting effect", number of athletes representing a nation, and past Olympic performance significantly to a nation's medal count. Notably, the most refined model explained approximately 97.6% of the variance in medal counts, highlighting the intricate dynamics that contribute to Olympic success.

The results challenge some conventional assumptions, such as the direct relationship between gross national income (or gross domestic product) and medal counts. The research also found interesting relationships, with variables like urban population percentage showing a potential non-linear relationship with medal count. By using panel data analysis and carefully addressing potential statistical problems, the study offers valuable insights for policymakers, sports administrators, and nations seeking to improve on their medal tally. The research underscores that Olympic success is not merely about economic resources, but a complex interaction of national investments, infrastructure, athletic development, and strategic planning.

Bibliography

Bernard, Andrew B., and Meghan R. Busse. "Who wins the Olympic Games: Economic resources and medal totals." *Review of economics and statistics* 86, no. 1 (2004): 413-417.

Forrest, David, Ismael Sanz, and J.D. Tena. “Forecasting National Team Medal Totals at the Summer Olympic Games.” *International Journal of Forecasting* 26, no. 3 (July 2010): 576–88. https://doi.org/10.1016/j.ijforecast.2009.12.007.

Hosein, Roger, Jeetendra Khadan, and Nicholas Paul. "An Assessment of the Factors Determining Medal Outcomes at the Beijing Olympics and Implications for CARICOM Economies." *Social and Economic Studies* (2013): 177-199.

Johnson, Daniel K., and Ayfer Ali. “A Tale of Two Seasons: Participation and Medal Counts at the Summer and Winter Olympic Games*.” *Social Science Quarterly* 85, no. 4 (October 29, 2004): 974–93. https://doi.org/10.1111/j.0038-4941.2004.00254.x.

Lui, Hon-Kwong, and Wing Suen. “Men, Money, and Medals: An Econometric Analysis of the Olympic Games.” *Pacific Economic Review* 13, no. 1 (January 9, 2008): 1–16. https://doi.org/10.1111/j.1468-0106.2007.00386.x.

Moosa, Imad A., and Lee Smith. “Economic Development Indicators as Determinants of Medal Winning at the Sydney Olympics: An Extreme Bounds Analysis.” *Australian Economic Papers* 43, no. 3 (September 2004): 288–301. https://doi.org/10.1111/j.1467-8454.2004.00231.x.

Sun, Ang, Rui Wang, and Zhaoguo Zhan. "A medal share model for Olympic performance." *Economics Bulletin* 35, no. 2 (2015): 1065-1070.