2024 Olympics Regression Analysis

Kelson Jensen
and
Layne Larson
and
Bronze Frazer

December 11, 2024

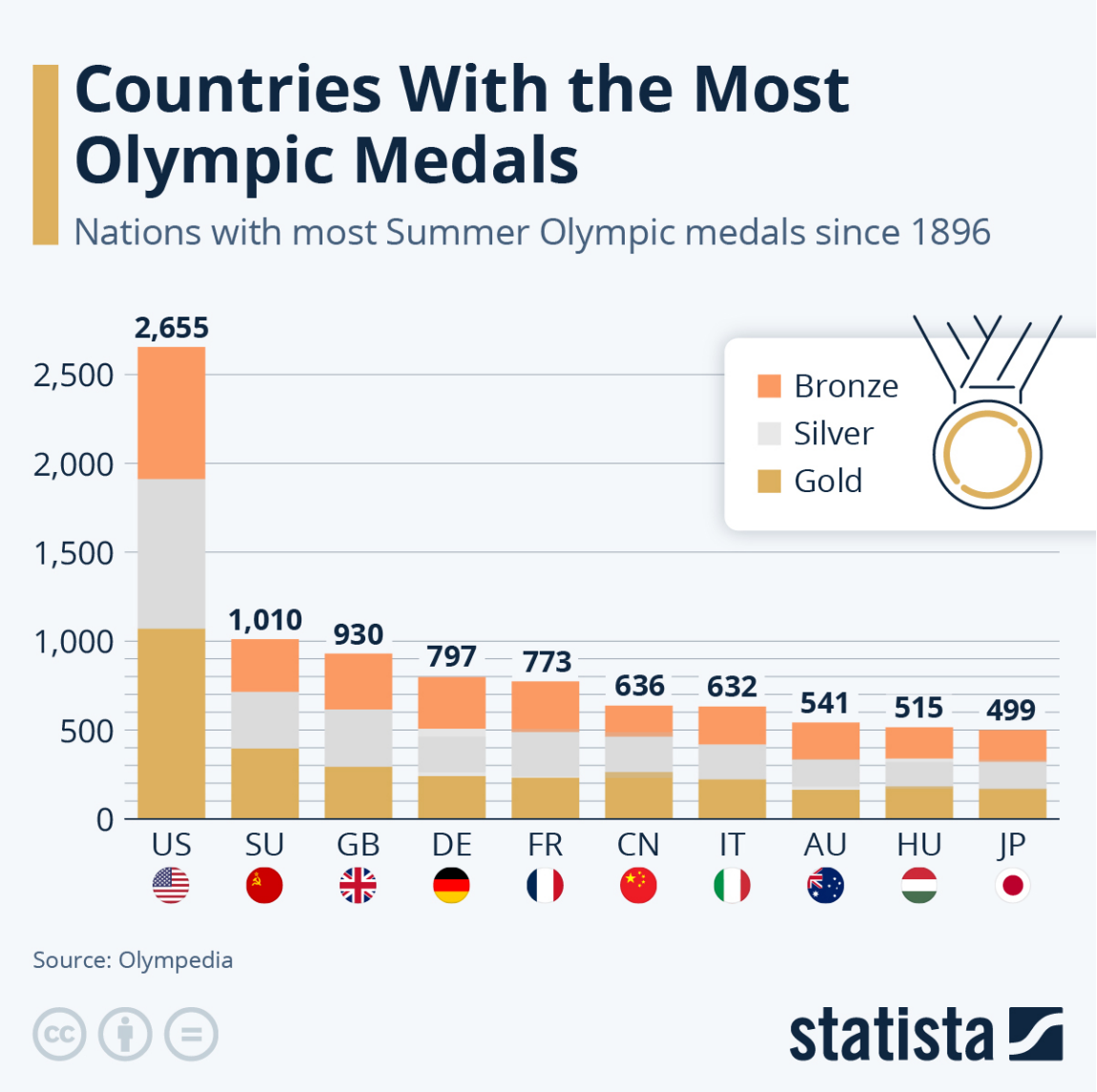
Abstract

This report explores the relationship between a country's economic and political characteristics and its medal count at the 2024 Paris Olympic Games. Using a custom dataset compiled from Kaggle and Freedom House, we analyzed variables such as GDP, Public Health Expenditure, Freedom Score, and Government Type. The key scientific question was to determine how these factors influence Olympic Medal Count. Our analysis employed variable selection methods and a linear regression model with a log-transformed response. The findings revealed Public Health Expenditure as the most significant predictor of medal count, while GDP and political variables showed little effect. Excluding outliers like the United States and China improved model fit but highlighted unique cases warranting further research. Limitations, including missing predictors like athlete count and the focus on a single Olympic Games, suggest opportunities for deeper exploration.

1 Introduction

The Olympic games are an incredible opportunity for athletes to compete at the world stage, winning medals, recognition, and prestige for their home countries. While the official reward for winning medals at the Olympics may only be bragging rights, there may be lingering effects that help boost a country's tourism and economy, as well as media coverage and acclaim for a short time. With the latest Olympics recently concluded in Paris, it became very apparent again that some few countries dominate the scene in terms of winning medals. We are interested in finding out *why* these countries win so many more medals than their competitors, and in general, what factors can help inform how many medals a country should win at the Olympic games.

In this study, we seek to answer the question- how does a country's economic and political makeup affect their ability to win medals at the Olympic games? We take into account in our study a number of different variables to help answer this question, from per capita GDP to government type.



1.1 Data Description

We gathered our country data from two main sources, and then compiled it all into one custom dataset that includes all our variables of interest. We extracted our response variable, medals, and main economic predictor variable, GDP, from a Kaggle dataset. Our other economic variables (Public Health Expenditure, Education Expenditure, and Tourism) also came from Kaggle datasets. Our Freedom Score and Government Type variables came from Freedom House and Kaggle respectively. It is worth noting that Freedom House is a nonprofit organization, best known for political advocacy surrounding issues of democracy, political freedom, and human rights (https://en.wikipedia.org/wiki/Freedom_House). They conduct a yearly survey/study that ranks each country's Freedom Score, which is a scale from 0 to 100 based on their civil liberties and political rights.

Our observational units are the countries that competed in the 2024 Paris Olympic games, and we paired all of our other variables on those countries before compiling into one dataset.

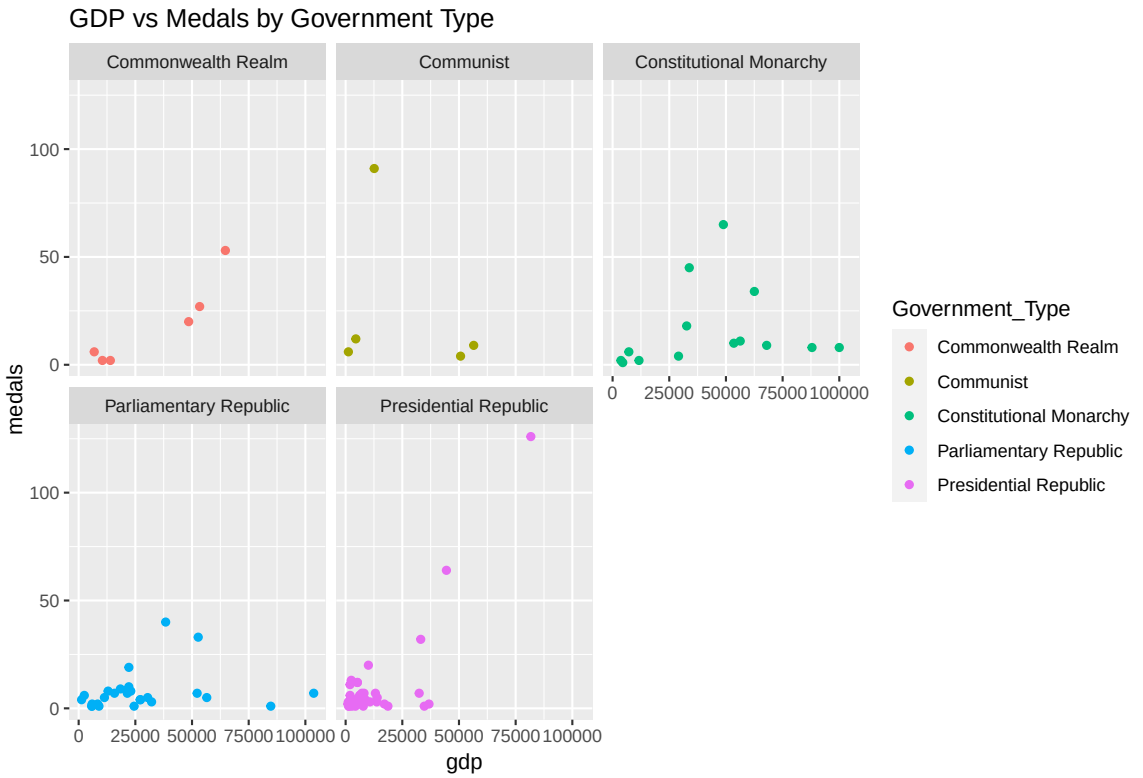


Figure 1: GDP vs Medals Facet Wrapped

Figure 1 (above) gives an exploratory analysis of the relationship between GDP and Medal Count, separated by Government Type. We can see somewhat of a positive relationship between GDP and medal count- that is to say, as GDP increases, medal count increases. But to say this trend is linear is not really true. We can see a stronger linear relationship in Figure 2 (below), showing Freedom Score vs. Medal Count. There are some outliers from this trend which we intend to look at in closer detail (United States and China).

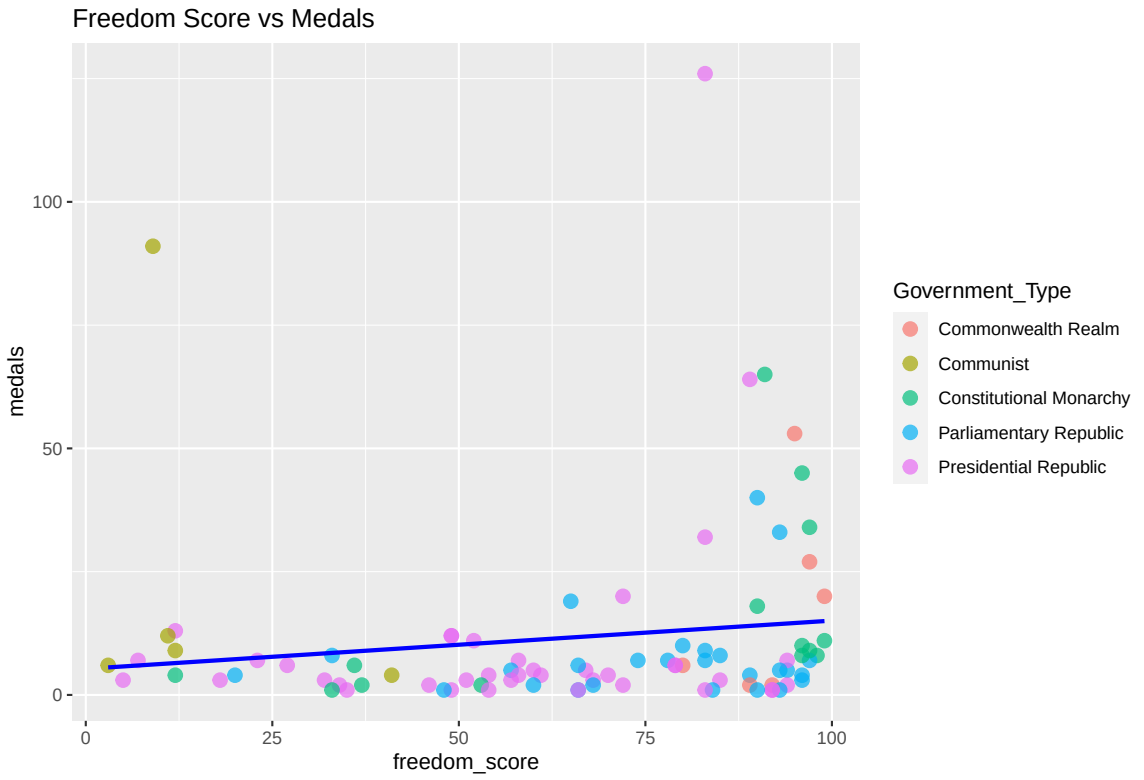


Figure 2: Freedom Score vs Medal Count

2 Model Selection and Validation

We used multiple selection methods to determine the best model to proceed with our analysis, then compared across each selection method. These methods included backwards, forwards, best subsets, and LASSO. Our results determined across all selection methods determined that we should eliminate Education Expenditure from our model. The resulting model we chose to evaluate is as follows:

$$\begin{aligned}\text{Log(Medal Count)}_i = \beta_0 &+ \beta_1 \times \text{GDP}_i \\ &+ \beta_2 \times \text{Freedom Score}_i \\ &+ \beta_3 \times \text{Public Health Expenditure}_i \\ &+ \beta_4 \times \text{I}\left(\text{Government Type}_i = \text{Constitutional Monarchy}\right) \\ &+ \beta_5 \times \text{I}\left(\text{Government Type}_i = \text{Parliamentary Republic}\right) \\ &+ \beta_6 \times \text{I}\left(\text{Government Type}_i = \text{Presidential Republic}\right) + \epsilon_i,\end{aligned} \tag{1}$$

where

$$\epsilon_i \sim N(0, \sigma^2).$$

This model includes a few key variables that help answer our initial question- how does a country's economic and political makeup affect their ability to win medals at the Olympic games? A country's economic makeup is contained in our GDP and Public Health Expenditure variables, while the social and political makeup are contained in the Freedom Score and Government Type variables. We tested different interactions between our variables,

but ultimately found none of them significant enough to include in the model.

It is worth noting that in addition to removing Education Expenditure because of our selection method, we dropped Tourism from our dataset as well. We also removed the Absolute Monarchy, Military Junta State, and Communist Government Types, because after our other variables were removed, these Government Types only had 1-4 observations, and made our model less robust. Tourism had a number of countries that were missing data points, so rather than drop entire observations we opted to remove the variable outright. Our resulting model is more clear and interpretable for this decision as well.

Along with removing variables from our analysis, we also removed two observations that were influential points (China and USA). Both of these countries had cooks distance values around 1.5, which was well above the reasonable limit for identifying influential points. We will discuss possible interpretations later on.

We ran a cross validation on our model, to assess our model fit and ability to perform on new data. We used a k-folds cross validation method, with k=10, allowing for a robust training set and using only one fold for validation. The test yielded one significant p-value for Public Health Expenditure (see Appendix: Table 2).

We also want to draw attention to our transformed response variable. We noticed in our examination of our linear assumptions that there were quite a few issues, specifically in linearity and normality of the residuals. After performing a log transformation on medal count, our linear assumptions hold much better, allowing us to proceed forward in our analysis and maintain a large amount of interpretability.

3 Analyses, Results, and Interpretation

After selecting our model, we looked at added variable plots, a summary output of our regression, and confidence intervals on our predictor variables. Together, these metrics allowed us to understand how each of the predictor variables helps explain medal count for each country.

The first and most important finding is that Public Health Expenditure has the highest impact on Medal Count for a country. It's p-value was significantly less than 0.05, suggesting it had a significant impact on the response. Compared to all our other predictor variables, this was the only variable that yielded significant results. Interestingly, the estimated coefficient for GDP was so low that it was essentially zero.

How did this happen? We identified a few reasons for these results. While we initially thought that GDP would be a much better indicator of medal count in terms of an economic factor, Public Health Expenditure was able to predict medal count much better. Perhaps this economic marker is more closely aligned with the health of a country and its citizens, leading to healthier and better performing athletes. In terms of our political markers, there were no significant effects, however the estimated coefficients were much greater than GDP and yielded actual results. While the p-values weren't low enough for significance, they are fairly low and worth some consideration. Presidential Republic specifically had the lowest p-value, suggesting a greater ability to predict medal count if a country has this government type.

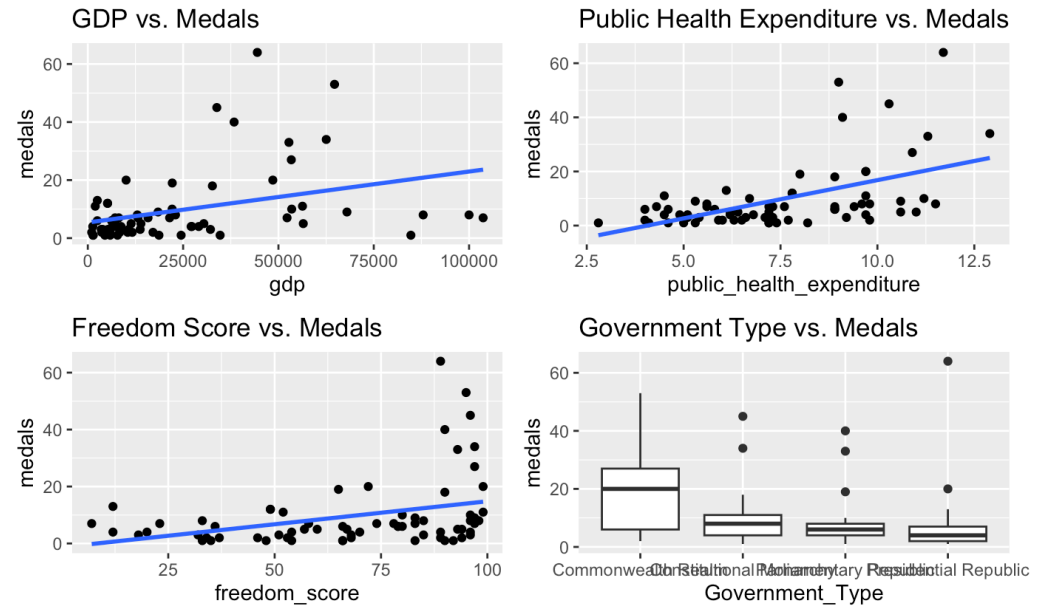


Figure 3: Illustration of model fit across selected variables

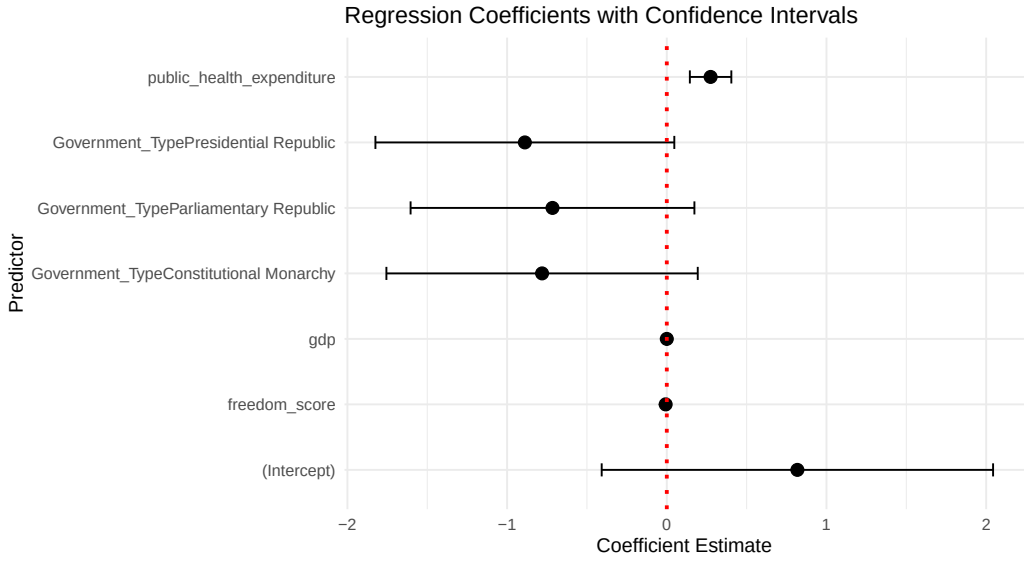


Figure 4: 95% Confidence intervals. Coefficients in for the model shown in the above equation. The dotted red line signifies the null hypothesis, that is these variables have no effect on the response. Intervals that do not contain zero suggest statistically significant effects.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.82	0.61	1.33	0.19
gdp	0.00	0.00	0.46	0.65
freedom_score	-0.01	0.01	-1.19	0.24
public_health_expenditure	0.27	0.07	4.22	0.00
Government_TypeConstitutional Monarchy	-0.78	0.49	-1.60	0.11
Government_TypeParliamentary Republic	-0.72	0.44	-1.61	0.11
Government_TypePresidential Republic	-0.89	0.47	-1.90	0.06

Table 1: Summary of the model

4 Conclusions

As stated above in our analysis, we found Public Health Expenditure to be the most significant predictor of medal count. We suggested this is because Health Expenditure could be a better indictor of population health, and subsequent athlete health and performance in the Olympic Games. However, there are some key drawbacks and weaknesses to our analysis that we want to point out here in our conclusions.

First of all, why this model? We initially ran a LASSO regression in order to perform variable selection and describe our data, but found the model to be too penalizing. It removed even more of our predictor variables till we had essentially zero, since each of the predictors were quite insignificant. While LASSO is a robust tool, we found that it performs better on more robust data that generally fits into a linear format. Our data was more difficult, and we didn't want to throw out all of our variables for the sake of discussion and interpretation, so we went with a more traditional multiple linear regression.

We also want to draw attention to the lack of significant variables. Our method for choosing which variables to include in our study at the data collection phase was not exhaustive - we selected variables that could help describe a country's economic and political makeup and that were easily attainable. One variable that we did not include but which might be quite simply a better predictor of medal count is athlete count- that is, how many athletes each country had compete at the Olympic Games. Perhaps more telling than a complex analysis of economic and political markers is simply the amount of athletes each country brings to the Games, earning them better odds at winning any specific medal. Another study including this variable, perhaps as a baseline, would be an interesting next step.

Also, as mentioned above, we removed China and The United States as influential points. These two countries earned far more medals than any other country, to a disproportionately large degree. It was hard to keep them in our analysis since they negatively impacted the fit of our model by acting as leverage points, leading to decreased predictive power. We don't have much insight as to why these two powerhouses win so many medals, and would suggest further research into these two countries specifically to find out why.

A final limitation to our study was our somewhat limited Medal Counts. We only took data from the 2024 Paris Olympic Games, and many countries had very low numbers of the response variable. Perhaps a time series analysis, analyzing medal counts over a longer time period would yield more interesting and quantifiable results.

APPENDIX

Table 2: Cross-Validation Results for Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.01	0.88	-0.01	0.99
gdp	0.00	0.00	1.67	0.10
freedom_score	-0.01	0.01	-1.08	0.29
public_health_expenditure	0.25	0.07	3.85	0.00
‘Government_TypeConstitutional Monarchy‘	0.12	0.89	0.14	0.89
‘Government_TypeParliamentary Republic‘	0.42	0.81	0.52	0.61
‘Government_TypePresidential Republic‘	0.01	0.82	0.01	1.00
‘gdp:Government_TypeConstitutional Monarchy‘	-0.00	0.00	-1.29	0.20
‘gdp:Government_TypeParliamentary Republic‘	-0.00	0.00	-1.72	0.09
‘gdp:Government_TypePresidential Republic‘	-0.00	0.00	-0.81	0.42