What affects the Athletes' Performance in the Olympic Games?

Yilan Ye

Info640

12.14. 2019

The target audience of this paper can be people who love sports, watch the Olympic Games, care about the medal performance, or even interested in social effects on collective behaviors.

Introduction

The modern Olympic Games are the leading international sporting events held every four years. Thousands of athletes from more than 200 countries around the world participate in a variety of sporting events. Every country desires an excellent athlete performance and wins more medals in the Olympic Games to show the nation's spirit and power. As an audience, I love to watch the Olympic Games and to see how athletes break records and human limits. However, looking at the results of the national metal performance in the Olympics, I found that the individual strength may create a miracle in one competition. Still, there are more external factors that contribute to the overall athletes' performance of a country. Inspired by the research paper "Going for Gold Medals: Factors affecting Olympic Performance" by Jayantha K and Ubayachandra E. G, I would like to check out whether the social factors such as population, GDP per capita, host city advantage will influence the athletes' performance in the Olympic Games.

Hypothesis

1. The population of a country is positively correlated with the athletes' performance.
2. GDP per capita is positive affects the Olympic performance of a country.
3. The host city advantage has a positive influence on the athlete's performance.

If all of the three hypotheses are true, we can receive a solid conclusion that the Olympics performance of the athletes and the comprehensive national factors are tightly related. A country with a larger population, higher GDP per capita, and the host city right could result in a better Olympics performance.

Data Analysis

This data analysis will be the predictive data analysis, focusing on whether there is a correlation between the athletes' performance of a country in the Olympic Games and its population, GDP per capita, or the host city advantage.

Dataset

To find out the result, I tried to use the two public datasets on Kaggle: "Olympic Sports and Medals, 1896-2014" by The Guardian, and "120 years of Olympic history: athletes and results"by rgriffin. However, these two datasets had very specific personal data on every athlete, but lacked the overall statistics information for each country. Therefore, I created three datasets: "2012 London Olympic Games with Population", "2012 London Olympic Games with GDP", and "Medal Performance of Great Britain from 1896 to 2016 in Summer Olympics", based on the data resources of Olympics Glory in Proportion (www.medalspercapita.com). The site had updated its website in 2018, which contained the latest information of the PyeongChang 2018 Winter Olympic Games. In this site, the data of GDP was provided by the World Bank, and the weighting system of the medals was suggested by the New York Times. The reason I chose Great Britain as the study case was that Great Britain was one of the countries attend the Olympics since 1896, and had held three Olympics Games. In the three datasets I created, there were 14 variables: country, weighted medals, gold medals, silver medals, bronze medals, population, population per weighted medals, GDP per capita, weighted medals per GDP, year, host city, attended athletes, total weighted medals, and the ratio of the weighted medals of Great Britain over the total. The limitation of the datasets was that there was no information about the personal information of the collectors of the original data sources, which might lower certain validity of the results. To solve this problem, I compared the medals data and population data with variable public sources to make sure there were no mistakes.

Method

With this dataset, I will use exploratory data analysis and visualization to see the general athletes and medal performance of different countries during the years with either scatter points or line chart. Then using the hypothesis test based on the test statistic to find the confidence level with p-values. Also, I will try to find out whether there is a correlation between the athlete performance or the three factors and whether there exist regression lines of them. I can evaluate the model by checking out the root mean squared error or residual standard error.

In the hypothesis, the dependent variable will be the weighted medals earned by a country, which shows the athletes' performance as general. For the first hypothesis, I would choose the variables of the population and the weighted medals earned by each country in the London 2012 Summer Games. For the second hypothesis, I would also choose the variables of the weighted medals earned by each country in the London 2012 Olympic Games with the GDP per capita collected for each country. The variables for the third hypothesis will be the weighted medals earned by the British athletes during the years when the Games are held in their own country or not. I once used the total number of the medals earned by a country to measure the athletes' performance, but I found the weighted medals calculated by exponential weighting system was more accurate to be the variable. In this case, I switched to the weighted medals earned by a country to represent a country's Olympic Performance.

For the first and second hypotheses, the target is to seek the correlation between the variables and to see which is the better variable to predict the total medals earned by a country. If there exists the best-fit line with a positive correlation, it shows the hypothesis is true. In the third hypothesis, I will make a scatter points chart and draw a linear line with the total medal earned by Great Britain against years. And to predict the total medal earned by the British athletes when the host city is in Britain. Then, I can calculate the difference between the calculated ones and the actual ones. If the actual medal number is higher than the predicted one, it can imply that the hypothesis is true.

Result

To test out the hypothesis 1, I choose the Weighted Medals as the dependent variable and the Population as the independent variable. I used the exponential weighted point system (4:2:1) — gold 4 points, silver 2 points, and bronze 1 point to represent the medal performance, and choose the dataset of 2012 London Olympics Game to test whether the population of a country was positively correlated with the athletes' performance. According to the plot of the population against the weighted metals of each country, I could find a best-fitted line to show there was a positive correlation between these two variables (Fig. 1), with the adjusted R-square of 0.2001 and Pr of 1.06e-05. Therefore, hypothesis 1 was proved positive.

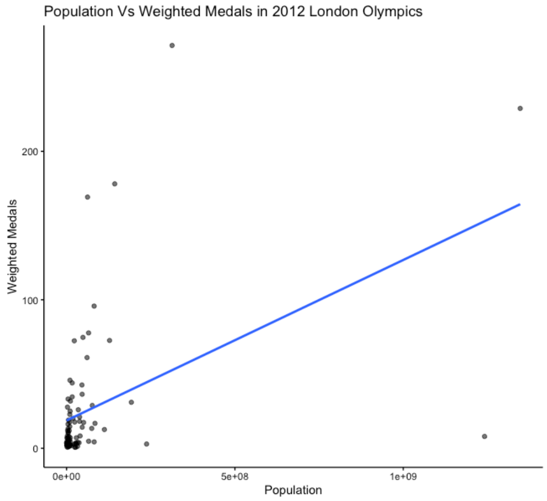

Fig.1

For hypothesis 2, I duplicated the procedures of testing hypothesis 1. This time, I choose the Weighted Medals as the dependent variable and the GDP per capita as the independent variable. Using R to draw the plot of GDP per capita against the weighted metals of each country, I got a best-fitted line to show GDP per capita is positive affects the Olympic performance of a country, which had the adjusted R-square of 0.6966 and Pr less than 2e-16 (Fig. 2). Then, hypothesis 2 was also proved to be positive.

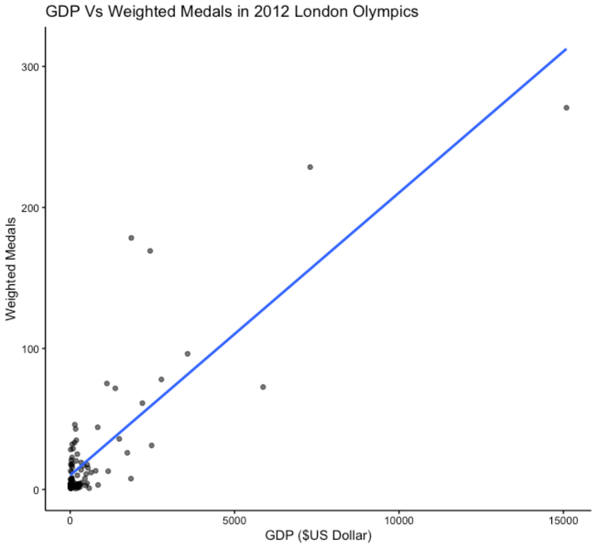

Fig.2

It would be more complex to test hypothesis 3 than the two above. In this case, I choose Great Britain as the study object, because Great Britain was one of the countries that attained all the Olympic Games since 1896 and had held three Olympic Games through the history. Besides, Great Britain had colonial effects on many counties which could also be considered as some kinds of host city advantages. I separated the whole Olympics history into three parts: 1896-1936, 1948-1992, and 1996-2016, which based on the number of countries won a medal (Fig. 3). Before 1936, there were only less than 35 countries ever won a medal in the Olympics; during 1948-1992, the number of countries won a medal was around 37 to 63; in 1996-2016, the number of countries won a medal rated to 79-88. London, the capital of Great Britain, had held one Olympics Game in each period. Since the total number of medals in different Olympics, I used the ratio of the weighted medals of Great Britain by the total weighted medals in that year to measure the athletes performance of Great Britain.

In the plot of the athletes' performance in each year, the scatter points when the Game held in London were highly above the best-fitted line. This meant that the athlete's performance of a country was better when the Game was hosted in that country. Besides, the points revealed the medal performance of Great Britain in other English Speaking districts were also much higher than the best-fitted line in general. Hence, the hypothesis 3 had proved to be correct that the host city advantage existed in the medal performance.

Medal Performance of the Great Britain

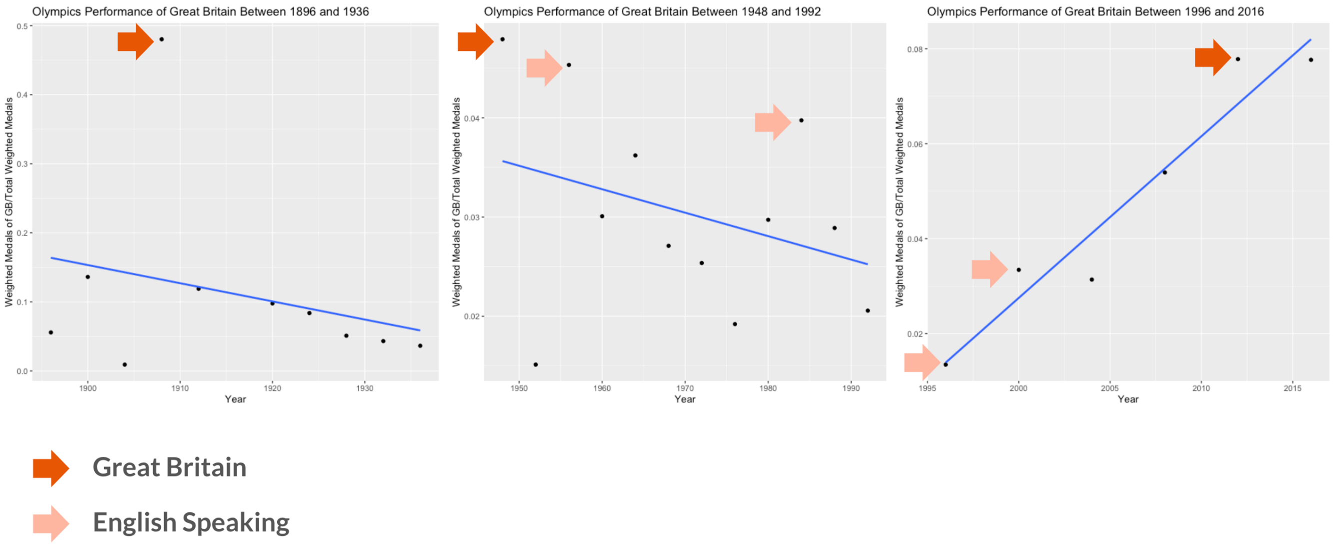

Fig.3

Reflection

During the analysis of the first two hypothesis, I found that GDP per capita as the independent variable, was better to predict the medal performance of a country than population, because the adjusted R-square of GDP per capita against weighted medal was 0.7, closer to 1, while the adjusted R-square of population against weighted medal was only 0.2.

From the analysis of hypothesis 3, I found except for the host city advantages, there were also some other elements might influence the performance of a country. One important factor was the number of athletes who attended the Olympics. According to Fig. 4 and 5, I could see when there was a larger number of athletes attended the Olympics, the correlated medal performance would be higher in general.

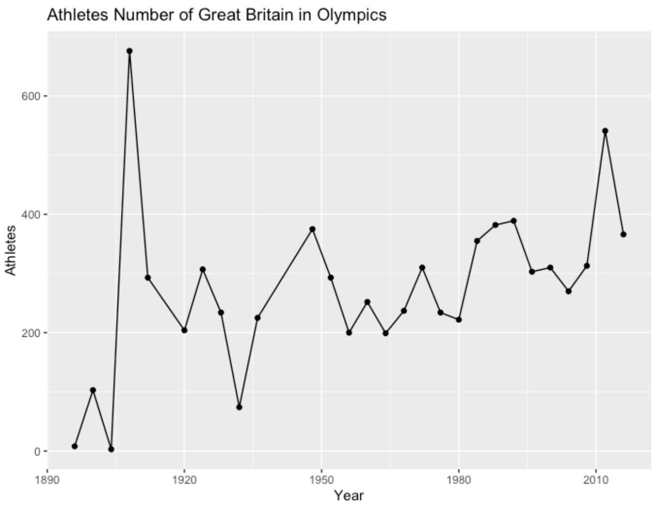

Fig.4

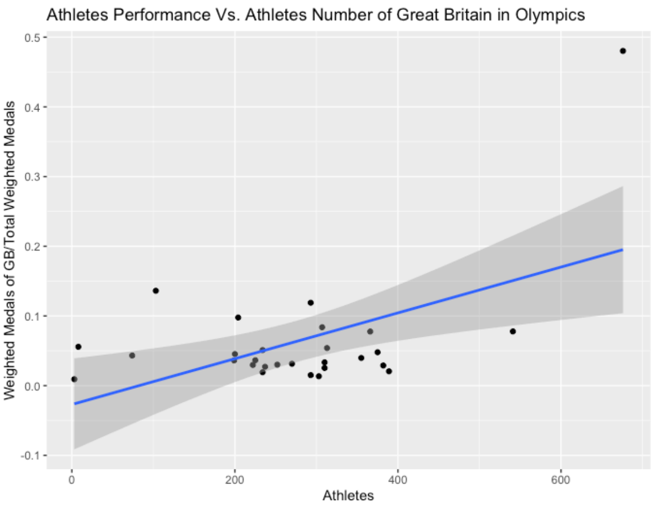

Fig.5

In this case, the reasons of increasing and decreasing the athletes' number in the Olympics should be considered. The convenience of transportation can be one vital element, because in the first period, there were two lowest points of attended athlete number of Great Britain in 1904 and 1932, which were the Games held in the USA, far from Great Britain. Another external reason could be there were more countries get involved in the Olympic Games and more events were created, especially in the second period, which resulted in the reduction of the athlete performance of Great Britain. To learn more about the reason for the upward trend of Great Britain performance in Olympics in the third period might need more research and information on the specific events performance, and other political and economical factors. In addition, more hypothesis testing on the other study cases should be taken to make a stronger claim, such as using the dataset in different years and of different countries.

Future Direction

It is really interesting to find out the triggering factors of the medal performance of athletes, and for extended research, the following topics will be intriguing to seek:

1. Find the trends of the female participants in the Olympic Games, based on the year and nation factors.
2. How anthropometric factors such as age, weight, and height will affect the athletes' performance.

Reference

1. Jayantha K, Ubayachandra E. G. Going for Gold Medals: Factors affecting Olympic, International Journal of Scientific and Research Publications, Volume 5, Issue 6, June 2015.
2. http://www.medalspercapita.com/#weighted-per-capita:2012

Appendix

```
#POPULATION
london<-read.csv(file = "Desktop/2012 Population.csv")
head(london)
summary(london)
unique(london$Country)
medaldata<-select(london,Country,Weighted_Medals, Population)
london_medal<-medaldata %>% arrange(desc(Weighted_Medals))
head(london_medal)
london_medal$Population<-as.integer(london_medal$Population)
ggplot(london_medal,aes(x=Population,y=Weighted_Medals))+geom_jitter(alpha = 0.6)+
stat_smooth(method = "lm", se=FALSE)+
  labs(title = "Population Vs Weighted Medals in 2012 London Olympics") +
scale_y_continuous("Weighted Medals")
lm_medalPOP<-lm(Weighted_Medals~Population,data=london_medal)
lm_medalPOP
summary(lm_medalPOP)
coef(lm_medalPOP)
```

```
#GDP
londonGDP<-read.csv(file = "Desktop/2012 GDP.csv")
head(londonGDP)
summary(londonGDP)

medaldata<-select(londonGDP,Country,Weighted_Medals,GDP...US.billion.)
london_medal<-medaldata %>% arrange(desc(Weighted_Medals))
head(london_medal)
ggplot(london_medal,aes(x=GDP...US.billion.,y=Weighted_Medals))+geom_jitter(alpha = 0.6)+stat_smooth(method = "lm", se=FALSE)+
  labs(title = "GDP Vs Weighted Medals in 2012 London Olympics") + scale_x_continuous("GDP ($US Dollar)")+scale_y_continuous("Weighted Medals")
lm_medalGDP<-lm(Weighted_Medals~GDP...US.billion.,data=london_medal)
lm_medalGDP
summary(lm_medalGDP)
coef(lm_medalGDP)

#Host City Advantage

GB<-read.csv(file = "Desktop/Great Britain.csv")
head(GB)
summary(GB)

GB1996<-GB%>%filter(Year>="1996")
ggplot(GB1996,aes(x=Year,y=GB.Total))+geom_point()+ stat_smooth(method = "lm", se=FALSE)+
  labs(title = "Olympics Performance of Great Britain Between 1996 and 2016") + scale_y_continuous("Weighted Medals of GB/Total Weighted Medals")
lm_GB1996<-lm(GB.Total~Year,data=GB1996)
lm_GB1996
new_GB2012 <- data.frame("Year" = 2012)
predict(lm_GB1996, newdata=new_GB2012)
summary(lm_GB1996)
coef(lm_GB1996)

GB1948<-GB%>%filter(Year>="1948")%>%filter(Year<="1992")
```

```
ggplot(GB1948,aes(x=Year,y=GB.Total))+geom_point()+ stat_smooth(method = "lm",
se=FALSE)+
  labs(title = "Olympics Performance of Great Britain Between 1948 and 1992") +
scale_y_continuous("Weighted Medals of GB/Total Weighted Medals")
lm_GB1948<-lm(GB.Total~Year,data=GB1948)
lm_GB1948
new_GB1948 <- data.frame("Year" = 1948)
predict(lm_GB1948, newdata=new_GB1948)
summary(lm_GB1948)
coef(lm_GB1948)


GB1896<-GB%>%filter(Year>="1896")%>%filter(Year<="1936")
ggplot(GB1896,aes(x=Year,y=GB.Total))+geom_point()+ stat_smooth(method = "lm",
se=FALSE)+
  labs(title = "Olympics Performance of Great Britain Between 1896 and 1936") +
scale_y_continuous("Weighted Medals of GB/Total Weighted Medals")
lm_GB1896<-lm(GB.Total~Year,data=GB1896)
lm_GB1896
new_GB1908 <- data.frame("Year" = 1908)
predict(lm_GB1896, newdata=new_GB1908)
summary(lm_GB1896)
coef(lm_GB1896)

#Others

ggplot(GB,aes(x=Year,y=Athletes))+geom_point()+ geom_line() +
  labs(title = "Athletes Number of Great Britain in Olympics")
ggplot(GB,aes(x=Year,y=GB_Weighted_Medals..Total_Weighted_Medals ))+geom_point()+
geom_line() + labs(title = "Athletes Performance of Great Britain in Olympics")
ggplot(GB,aes(x=Athletes,y=GB_Weighted_Medals..Total_Weighted_Medals ))+geom_point()+
stat_smooth(method="lm") + labs(title = "Athletes Performance Vs. Athletes Number of Great
Britain in Olympics") + scale_y_continuous("Weighted Medals of GB/Total Weighted Medals")
```