Medal count disparities at the Olympic Games: An econometric analysis of the determinants of national Olympic success using an economic growth framework

Annie Wan Sun

Master's thesis

Copenhagen Business School

Applied Economics and Finance (AEF)

Student number: 91766

Date of submission: 2nd of August 2020

Supervisor: Battista Severgnini

Number of characters: 181,927

Number of pages: 80

Abstract

This thesis studied the disparities between countries' Olympic medal counts and had as its main research question:

Why are some countries more successful than others at winning medals at the Olympic Games from an economic growth perspective, and how does econometrics estimate the relationship between the Olympic medal distribution and its determinants suggested by the economic growth theory framework?

Using the Herfindahl-Hirschman Index to measure the level of competitiveness at the Olympic Games in the thesis' chosen timeframe from 1996 to 2016, it was established that some countries do indeed win more medals than others, but that the competition is not dominated by any single country.

The role of technology in economic growth served as a theoretical framework to motivate the choice of regressors, representing the resources that a country has available and invests in Olympic technological progress. The proposed regressors were per capita GDP, population, team size, three host effects and tourism as a proxy for openness. Investment in sports was also suggested but could not be included in the main model due to limited data availability.

The tobit model often employed in the literature was argued to be misrepresenting the problem of modelling national Olympic success. A fixed effects regression model for panel data was chosen to estimate the relationship between medal count and the proposed regressors in consideration of preventing omitted variable bias arising from country-specific time-invariant factors.

Team size was found to be statistically significant for Olympic medals at the 5% level, with a predicted 6.3 medal increase for every 100 athlete increase in team size. Country effects were also significant for determining how many medals a country wins. Population and host effects were found to only impact medals through their effect on team size, and GDP per capita was not found to be statistically significant at all. Differences in the data used, the choice of model and the efforts to address omitted variable bias were given as possible reasons for the discrepancy between these findings and those in the literature.

A zero-inflated negative binomial model was also proposed as an alternative to the fixed effects model.

Table of Contents

1 Introduction

Every four years, people gather all over the world, some in front of their TVs and some travelling to see the spectacle in person, to experience the celebration of the Olympic Games. The Olympics are an immensely popular international sporting event and, considering the number of participating athletes and countries and the number of different sports they compete in, they are arguably the biggest sports competition in existence (Rothenbuhler, 1988, p. 64; The Olympic Museum, 2013, p. 3). The most recent Olympic Games for instance, held in 2016 in Rio, Brazil, saw the participation of over 200 countries and territories sending more than 11,000 athletes in total to represent them in competition across 306 events in a wide range of sports (The International Olympic Committee, 2017a, p. 8). Around the world, an estimated 2.6 billion people watched along on TV (The International Olympic Committee, n.d.-g, p. 5), with the average viewer consuming 9.3 hours of Olympic TV coverage (The International Olympic Committee, n.d.-g, p. 6). In other words, the scale of the Olympic Games and the reach they have are uniquely expansive.

What used to be a sports festival for amateurs (The Olympic Museum, 2013, p. 7) competing to demonstrate the utmost physical and mental strength and balance (The International Olympic Committee, 2019a, p. 11) has become increasingly commercialized over time. In 1960, the Olympic Games began to be broadcast on TV for audiences all around the world to be able to follow (The International Olympic Committee, 2020a, p. 29), and since 1984, the Olympics have been open to participation from professional athletes (The Olympic Museum, 2013, p. 7), who are important for attracting audiences (Milton-Smith, 2002, p. 134). Milton-Smith (2002) argued that the Olympic Games now operate like a business and have "been transformed by television and commercial sponsorship into global mass entertainment" (p. 134). Indeed, the Olympics bring in increasingly large revenue streams from broadcasting rights and sponsors (The International Olympic Committee, 2020a, p. 8), for whom the scale of the Olympics is an opportunity for global exposure (Milton-Smith, 2002, p. 134).

Although winning was not meant to be the focal point of Olympic competition (The Olympic Museum, 2013, p. 13), it is important to many athletes and has become the most frequent topic of discussion in the media coverage of the Olympics and its participants. Swiss tennis player Roger Federer, for example, is widely considered one of the best of the sport, but has never won Olympic gold in the men's singles (The International Olympic Committee, n.d.-n). Articles written about him in an Olympic context convey his desire to win the singles gold at the Games (BBC, n.d.; The International Olympic Committee, n.d.-n) and make note of the fact that he has not done so yet despite having had an incredibly successful career already (BBC, n.d.; The Tokyo Organising Committee of the Olympic and Paralympic Games, 2020). Thus, winning is placed at the forefront of much of the conversation surrounding the Olympics.

This contemporary view of victory as an ultimate goal not only applies to individual athletes, it extends to the way in which countries evaluate their performance at the Olympics and to the media portrayal of the Olympic Games as a competition between countries to win the most medals and be the most successful (De Bosscher et al., 2009, p. 113). Denmark, for instance, celebrated winning 15 medals at the Rio 2016 Olympics as the country's second best result ever (Daugbjerg, 2016), and it is not uncommon for countries to enter the Olympics with a medal target that they hope for their athletes to achieve collectively (BBC, 2016a; Cottrell, 2012; The Japan Times, 2018). As such, there is a certain prestige associated with winning at the Olympics that makes it desirable for participating countries (Ball, 1972, p. 186; Grix & Carmichael, 2012, pp. 81–82). The overall medal results across all sports and events at an edition of the Games are gathered into what is known as a medal table, which displays the medal distribution, i.e. the gold, silver and bronze medal counts for each country, ordered into a ranking (The Guardian, n.d.; The Washington Post, n.d.). Medal tables are commonly reported in the media during the Games, often with a focus on which country tops the ranking, thereby framing the Olympics as an inter-country competition (BBC, 2012, 2016b; Den Butter & van der Tak, 1995, pp. 27–28; I. Johnson, 2008; The Washington Post, n.d.).

In the media, especially in recent years, there seems to be an interest in trying to understand why medal tables look the way they do, i.e. which countries win Olympics medals and why some win a lot while others win none. For the answers to those questions, the media has turned to economists.[1] Amidst the reporting on Olympic competition and the medal tables, a challenge has thus emerged among these economists: by regressing the historical medal counts against different macroeconomic, social and political variables to explain the distribution of medals between countries, how well can they model and predict how many medals each country wins?

Economists' explanations for the Olympic medal distribution that have been presented in the media are rooted in the economic literature, where for decades researchers have made efforts to understand the determinants of winning Olympic medals using econometric methods. The most influential contribution has come from Bernard and Busse (2004) who popularized the idea of using GDP per capita and

[1] Examples of articles in the media exploring the topic of economists' medal count predictions include Dunbar (2016), The Economist (2012), Heuslein (2010) and Swanson (2016).

population to explain the disparity in the number of medals that different countries win. Over time, other determinants of the Olympic medal distribution have also been suggested in the literature, but per capita GDP and population have remained popular.

The purpose of this thesis is to build on these ideas and explore the Olympic medal distribution from an economic growth perspective, then apply panel data econometrics in the form of fixed effects regression to answer the main research question:

Why are some countries more successful than others at winning medals at the Olympic Games from an economic growth perspective, and how does econometrics estimate the relationship between the Olympic medal distribution and its determinants suggested by the economic growth theory framework?

In order to do so, the following secondary research questions will be addressed throughout the thesis:

- How are Olympic medals distributed between countries?
- What has existing literature suggested as the reasons for the disparities in Olympic medal counts between countries?
- How can economic growth theory be used to understand the distribution of Olympic medals, and which factors does such a framework suggest would be relevant in a regression model?
- What is the relationship between these proposed factors and the distribution of Olympic medals when using panel data econometrics in the form of a fixed effects regression model?

There are a few points to clarify in regards to the terms used above. The phrase "Olympic Games" is used in reference to the modern Olympic Games, the first of which were held in Athens, Greece in 1896 (The International Olympic Committee, n.d.-c). Outside of a short explanation of the ancient Olympic Games as the origin of the modern Games, this thesis will not concern itself with them.

Furthermore, to simplify the analysis, the scope of the thesis is limited to Summer Olympics (referred to simply as the Olympic Games). The other games in the Olympic world (e.g. Winter Olympics, Paralympics), although similar in format, are dissimilar enough in the types of sports and athletes involved that it makes better sense to differentiate between them.[2]

[2] The determinants of Olympic success could be quite different for Winter compared to Summer Olympics, e.g. an increased importance of facilities accommodating the practice of winter sports (such as ski resorts, ice rinks). In addition, countries with colder climates and mountainous regions may be at an advantage at Winter Olympics, but not the Summer ones. Similarly, the determinants of Paralympic success could be different than for Summer Olympics, perhaps in the form of an increased importance of the quality of healthcare and extent of social security available in a country.

As the research question implies, the thesis only considers Olympic medals on a country level. That is, it does not consider who exactly has won a particular medal on an individual athlete or team basis, it only addresses medal counts on the aggregate country level.

Olympic success is defined according to the total number of medals won by a country at a particular edition of the Olympic Games, such that no distinction is made between different types of medals. Gold, silver and bronze medals are weighted the same.

Lastly, for data availability reasons, only the Olympics from 1996-2016 are included in the analysis. Olympic medal data itself is readily available, with records existing of all medal distributions dating back to the first modern Olympics in 1896. However, much of the data relevant to the thesis to regress medal counts against has only been available since the 1990s.

1.1 Motivation

The motivation for studying the determinants of national Olympic success is threefold.

First of all, the extraordinary scale of the Olympic Games and the commercial interest in it as a business make the Olympics a fascinating phenomenon to study, and an interest in trying to understand what determines their outcome follows from the unique situation they provide for cross-country comparison. As previously mentioned, the Olympic Games are one of, if not the biggest international sports event (Müller, 2015; Rothenbuhler, 1988, p. 64) featuring the best athletes across many sports from all around the world on a heavily commercialized competitive scene. The most recent Olympics, with an audience in the billions (The International Olympic Committee, n.d.-g, p. 5), generated a broadcast revenue of USD 2.868 billion (The International Olympic Committee, 2020a, p. 27) and attracted sponsors for a total sponsorship revenue of over USD 1.8 billion (The International Olympic Committee, 2020a, pp. 15, 19).[3] With the Games encompassing a global audience (The International Olympic Committee, 2020a, p. 24), large corporate sponsorships for exposure and extensive media attention surrounding the medal distribution, countries have a desire to perform well and win (De Bosscher et al., 2009, p. 113; Grix &

[3] This is approximately true. Revenue from domestic sponsorships for the host country of the 2016 Olympics was USD 848 million (The International Olympic Committee, 2020a, p. 19). Revenue from The Olympic Partners worldwide sponsorships for both the 2014 Winter Olympics and 2016 Olympics combined was just over USD 1 billion (The International Olympic Committee, 2020a, p. 15). There is no figure for the 2016 Olympics alone.

Carmichael, 2012, p. 82; van Hilvoorde et al., 2010, p. 88). Thus, the scale of the Olympic Games makes it compelling to try to understand the competition between countries that inspires such media coverage and public attention to fuel sponsorships in the billions of dollars.

From a broader perspective, it could be said that although the Olympics are an important event culturally (Rothenbuhler, 1988, p. 65), the amount of money involved in the Olympic Games plays only a small part in the world economy. The revenues given above for the 2016 Olympics are relatively minor compared to the more than USD 76 trillion that constituted the world GDP in 2016 (The World Bank, n.d.-a). On the other hand, for a single event to generate revenues comparable in size to the GDP of a small country (The World Bank, n.d.-a) is not unimpressive.[4] Even more impressive perhaps is the reach and impact that the Olympics have on people. That leads to the second reason why the medal distribution of the Olympics is relevant to study.

Second of all, the determinants of success at the Olympics are interesting to explore because of the utility the Games provide to billions of people around the world. Utility is an economic concept, the meaning of which is related to the satisfaction that a consumer gets from consuming a good or service (Hirschey & Bentzen, 2014, p. 113). Olympic Games viewing in general, but also national Olympic success in particular, may be thought of as providing utility.

Rothenbuhler (1988) found that people valued and enjoyed the Olympics and made watching them a priority (p. 75, 78), stating that people "rearranged their schedules so that they could watch more" (p. 75). General sports viewing has been found to be a source of enjoyment (Gantz & Wenner, 1991, pp. 238–239; Rogers, 2018), with emotional engagement bringing viewers happiness when the team they favor wins and unhappiness when they lose (Gantz & Wenner, 1991, pp. 240–241). In an Olympic context, this could include rooting for one's home country, as Wicker et al. (2012) reported that people did in their study (p. 347). Rogers (2018) argued that even if negative emotions are experienced when one's team does not fare well, sports viewing remains meaningful (p. 379). That is, the utility to consumers of the Olympics is not necessarily dependent on the specific results, because the watching experience itself can be a source of meaning.

As for utility derived from one's country winning, according to Hallmann et al. (2013), the success of national athletes in elite sports in Germany was associated with national pride and happiness among the

[4] To give an example, Montenegro had a GDP of USD 4.377 billion in 2016 (The World Bank, n.d.-a) compared to the combined broadcast and sponsorship revenue of around USD 4.7 billion for the 2016 Olympics (with the caveat explained in footnote 3).

country's population (p. 231). Wicker et al. (2012) also found this to be the case, specifically for the Olympic Games (p. 347). Although, Kavetsos and Szymanski (2010) saw only weak evidence of increased life satisfaction when one's country performed better than expected at the Olympics (p. 164), van Hilvoorde et al. (2010) argued that the national pride felt upon national sporting success "must be preceded by a sense of belonging to a specific nation" (p. 99), and they explained that such pride may only be temporary. Nevertheless, a degree of national pride and happiness seems to be associated with watching athletes from one's own country win Olympic medals.

As such, both national success at the Olympics and the satisfaction from watching the Olympics in general may increase audience utility. It then becomes relevant to better understand what determines each country's success at winning Olympic medals and thereby influences the utility that billions of consumers of the Olympic Games get from watching and following them. Thus, the importance of the Olympic Games and of understanding their medal distribution is placed, not just on the Olympics' monetary relevance, but on the utility that the Games and their outcome provide its sizeable audience.

Lastly, having a better understanding of what determines the medal distribution at the Olympics could have policy implications. De Bosscher et al. (2009), van Hilvoorde et al. (2010) and Grix and Carmichael (2012) explained that an important reason why countries invest in elite sports is in the hopes that it will result in greater international sporting success, such as at the Olympics. If it is possible to determine to what extent, if any, an individual country's investment in sports contributes to the country's athletes winning more Olympics medals, it becomes easier to gauge whether the size of the investment is reasonable for the effect on success it is associated with. It may be that no evidence of such a relationship between sports investment and Olympic success is found in the analysis.

1.2 Structure of the thesis

The thesis is structured largely according to the research questions posed.

After this introduction, the second chapter of the thesis introduces the Olympic Games and the terminology used throughout to describe them. It then describes how medals have broadly been distributed at the 1996-2016 Olympic Games by looking at the nature of Olympic competition using the Herfindahl-Hirschman Index.

The third chapter is a literature review of studies investigating the determinants of the Olympic medal distribution, specifically those using econometric methods of regression analysis to do so.

In the fourth chapter, the theoretical framework of the thesis is established, wherein parallels are drawn between economic growth theory about the role of technology in growth and the resources required for a country to be successful at the Olympics. This leads to the choice of variables of interest, and a discussion of these follows.

Chapter five concerns the methodology of the thesis. It opens with a general discussion about the thesis' philosophical approach to knowledge and knowledge production, then discusses the choice of econometric method and the econometric theory and model assumptions that underlie the analysis in chapter six. The chapter ends with a description of the (panel) data and a discussion of how it was collected and processed.

In chapter six, the panel data is analyzed, first using descriptive statistics to get an overview of the data. Next, a fixed effects regression analysis is used to determine the effect on the number of Olympic medals won on the country level (the dependent variable) associated with the chosen regressors. A zero-inflated count model is then proposed and explained as an alternative to the fixed effects model.

Chapter seven is a discussion of the findings, including the limitations of the analysis.

Finally, chapter eight concludes the thesis and provides suggestions for further research in the field.

2 Background: The Olympic Games

Many people are likely to already be familiar with the Olympics and have at least a basic understanding of their structure and competitions. This part of the thesis will give a brief introduction to the origin and concept of the Olympic Games and highlight those of its features that are relevant for understanding later chapters. In particular, it will explain the terminology specific to the Olympics that is used throughout the thesis.

The Olympic Games, dating back as far as 776 B.C., originated in ancient Greece as recurring sports competitions held in Olympia every four years in honor of the Greek god Zeus (The Olympic Museum, n.d.). Gathering people from all over Greece and its wider sphere of influence, the ancient Olympic Games were the most important of four Panhellenic ("all Greek") Games, each celebrated at a different time in a four-year cycle. The beginning of each four-year period, known as an Olympiad, was marked by the Olympic Games. Although the other three of the Panhellenic Games were also known as such, the term "the Games" is used only in reference to the Olympics in this thesis.

In the ancient Olympic Games, only those who were men, Greek and free (i.e. not enslaved) were allowed to participate (The Olympic Museum, n.d.). Although the Games initially consisted of only a single competition, to run the length of the stadium (The International Olympic Committee, n.d.-o), over time, they grew in size to become a multi-sport event (The International Olympic Committee, n.d.-b, n.d.-s). Winning an event at the Games was regarded as a highly impressive feat achieved with the power of the gods, and those who won returned home to lifelong fame and admiration, in addition to gaining political power (The Olympic Museum, n.d.).

After centuries of Olympic competition, the ancient Olympic Games came to an end after A.D. 393 with the rise of anti-paganism (The Olympic Museum, n.d.). It was not until the late 19^{th} century that the Games were revived and the modern Olympic Games were born.

The Olympics in their current form were first held in Athens in 1896 (The International Olympic Committee, n.d.-c). Since then, they have taken place and are planned to take place every four years, except during the World Wars when the 1916, 1940 and 1944 Games were cancelled (The Olympic Museum, 2013, p. 21) and in the case of the Tokyo 2020 Olympics which, at the time of writing, have been postponed to 2021 due to the coronavirus pandemic (The International Olympic Committee, 2020c).

The modern Games bear some similarities to the ancient ones. Both are, at their core and in simple terms, culturally important organized sports events where athletes compete to win, and the victors are hailed for their abilities. The modern Olympics have also preserved the frequency with which the ancient ones were held, being celebrated once every four years. However, the Olympics today are much more formalized and strictly governed than they were in ancient Greece. Furthermore, the scope of participation in the modern Games has been greatly broadened as it has been an international event since the revival in 1896 and has included women since 1900 (The Olympic Museum, n.d.). In contrast, the ancient editions were for Greek men only. In other words, holding the modern Olympics involves many more parties and requires much more structured organization than it did in ancient Greece. Before continuing to explore the specifics of how the Olympic Games are held and unfold in the present, it is therefore necessary to establish the role of the different organizations and parties that shape and run them.

The most important organization in the world of the Olympics is the International Olympic Committee (IOC). Since its establishment in 1894 to mark the initiative to relaunch the Games, it has remained the authority on everything Olympic, and it oversees the organizing of each of the Games (The International Olympic Committee, n.d.-t, 2019a, pp. 15–17).

Secondly, the National Olympic Committees (NOCs) and the International Sports Federations (IFs) play a central role in the coordination between the IOC, the athletes that participate in the Games and the sports and events in which they compete.

The NOCs are the national institutions that represent the individual countries and territories participating in the Olympics (The International Olympic Committee, 2019a, p. 60), and Olympic athletes represent the NOC of their respective country when they compete (The International Olympic Committee, 2019a, p. 77).[5] Although some NOCs do not conform to what most people consider to be countries, e.g. Puerto Rico and Hong Kong (The International Olympic Committee, n.d.-h), the term "countries" is adopted here for simplicity and used throughout the thesis to convey the geographical areas that the NOCs represent. Broadly speaking, the NOCs serve to promote and cultivate sports and the Olympics within their country (The International Olympic Committee, 2019a, pp. 59–60), and they hold some power in regards to allocating Olympic participation spots to its national athletes (The International Olympic Committee, 2020b).

[5] This is the case for the vast majority of Olympic athletes, but there are exceptions to the general rule that athletes compete under an NOC, e.g. the Refugee Olympic Team in 2016 and Independent Olympic Athletes (The International Olympic Committee, 2016a, 2017a).

While the NOCs are concerned with the people who participate in the Olympics, the IFs are responsible for the sports these athletes compete in. The IFs are large international organizations recognized by the IOC that are well-known outside of their Olympic involvement for their role in governing each of their sports in general (The International Olympic Committee, 2019a, p. 55). Examples include FIFA in the world of football and World Athletics (previously known as the IAAF) in the world of athletics (The Association of Summer Olympic International Federations, n.d.; The International Olympic Committee, 2019a, pp. 82–83). In an Olympic setting, the IFs most notably decide the rules for their respective Olympic sports, such as setting any age limits and establishing Olympic qualification requirements (The International Olympic Committee, 2019a, pp. 56, 76, 78).

The IOC, the NOCs and the IFs are the main actors in what is called the Olympic Movement, which embodies the idea of using sport to improve the world (The International Olympic Committee, n.d.-t, 2019a, p. 15). As constituents of the Olympic Movement, all three must follow the principles and rules about their obligations and the organization of the Games that is laid out in the document known as the Olympic Charter (The International Olympic Committee, n.d.-t, 2019a, p. 9).

Having established the roles of the most important parties involved in the planning, organization and celebration of the modern Olympic Games, it becomes much easier to describe the details of how the Olympics have evolved since 1896 to who and what they encompass today and how they proceed.

As mentioned previously, the modern Games from their onset were already more inclusive than the ancient Games, encouraging participation from athletes of all nationalities and, not long after, from both men and women. Over time, this increase in diversity has only become more pronounced. At the first Games in 1896, 241 athletes from 14 countries participated (The International Olympic Committee, n.d.-c). By 1912, all five continents were represented (The Olympic Museum, n.d.). Today, the IOC oversees 206 NOCs (The International Olympic Committee, n.d.-h), and the most recent Olympic Games in 2016 saw the participation of over 11,000 athletes competing for 207 countries (The International Olympic Committee, n.d.-m, 2017a, p. 8).[6]

Female athletes, at their first Olympic appearance in 1900, made up 2.2% of the participants, a figure that has grown to reach 45% in 2016 (The International Olympic Committee, 2018, p. 4). This approach towards gender equality among Olympic participants has come as a result of the effort over the years by

[6] The number of countries is listed by the IOC as 207 (The International Olympic Committee, n.d.-m). Of those, 205 were NOCs and the remaining two were the Independent Olympic Athletes and the Refugee Olympic Team (The International Olympic Committee, 2016a, 2017a, p. 8).

the IOC to increase the number of Olympic sports women could compete in, until in 2012, all Olympic sports had female participation (The International Olympic Committee, 2018, pp. 1–2). 2012 also marked the year when there was no longer an NOC that had never had female representation (The Olympic Museum, n.d.). Thus, the Olympics today, as well as being open to participation from vastly more people than the ancient Games, have also grown increasingly international and gender equal over time.

The international dimension of the modern Olympic Games is also reflected in their geographical location. Since the ancient Games were for the Greek, they were always held in Greece. The modern Games, on the other hand, are hosted in a different city around the world from each Games to the next (The International Olympic Committee, 2019a, p. 69; The Olympic Museum, 2013, p. 6). This host city plays an important part in uniquely shaping each celebration of the Olympics. The aesthetics and atmosphere of a particular edition of the Games will be influenced by the characteristics of the hosting country, which is provided an opportunity to showcase its culture (The International Olympic Committee, 2019a, p. 75; The Olympic Museum, 2013, p. 8).

The current process to becoming a host is long, and the decision is made many years in advance. Any city interested in hosting the Olympics is selected by its respective NOC to enter into a bidding process with the IOC (The International Olympic Committee, n.d.-j, 2019a, pp. 60, 70). After formally becoming a Candidate City, it submits a proposal detailing its plans for the organization of the Games in order to compete for the designation as host (The International Olympic Committee, n.d.-j). The IOC then evaluates each candidate's submission and, after a two-year long process, elects the host.

After been elected, the host city faces the challenge of organizing the Games and ensuring that they are run successfully. An Organising Committee for the Olympic Games (OCOG) is formed by the host city's NOC as a link between the city, the NOC and the IOC (The International Olympic Committee, n.d.-k, 2019a, pp. 72–73). The fact that the host is found many years before the Games are held, e.g. Tokyo being elected in 2013 to host the 2020 Olympic Games (The International Olympic Committee, n.d.-a), gives the city time to prepare for the upcoming celebration. This includes the building of new venues, accommodation and infrastructure and improving any existing facilities that can be used (The International Olympic Committee, 2017b, pp. 81, 89). In other words, hosting the Olympics is no small undertaking for the host city given the responsibilities of organizing the Games.

In recent years, the IOC seems to have recognized this and placed an emphasis on easing the burden for the host city, e.g. by encouraging the use of existing venues where possible (The Olympic Museum, n.d.).

In addition, importance has been placed on the concept of "Olympic legacy", which was formally acknowledged in the Olympic Charter in 2003 (The International Olympic Committee, 2017c, pp. 2, 9). It describes the idea that the host city should enjoy benefits in the long-term from having hosted the Games. This includes encouraging the city to find continued use for the investments made in connection with the Games, e.g. in venues (The International Olympic Committee, n.d.-e, 2017c, p. 82). As a result, the IOC, in its evaluation of Candidate Cities, considers and greatly values sustainability and how well the city's proposal to host the Games suits the characteristics of the city and its long-term plans for development (The International Olympic Committee, n.d.-e, n.d.-j).

For the athletes, preparation for the Olympics involves qualification. In each of the Games, there is a set number of quota places to be filled for each competition/event, and athletes must earn their place by qualifying (The International Olympic Committee, 2020b). However, no more than three athletes from the same NOC may qualify and compete in each individual event at the Olympics (The International Olympic Committee, 2019a, p. 80), and the IFs may set a lower maximum in their particular sports (Hann, 2014; The International Olympic Committee, 2020b). Qualification takes place through qualification events and tournaments, where certain requirements have to be met to successfully qualify (The International Olympic Committee, 2020b). As previously mentioned, the specific criteria for qualifying for the Olympic Games are determined by the IFs of each individual sport. This includes whether quota places from a successful qualification are given to the athletes themselves with the approval of their respective NOCs, or to the NOCs which then choose their athletes.[7] The hosting country has an advantage, as it is the wish of the IOC to have the host represented in all sports. Provided that the athletes of the host country meet minimum requirements, the IFs may therefore allocate quota places in their sport to the host, allowing it to compete in the Games that it hosts without following the regular qualification process. Brazil, for example, gained such an advantage when it was automatically entered in all cycling disciplines at the Rio 2016 Olympics (Hann, 2014). As will be discussed later, this is not the only advantage of hosting the Games.

As for what the Games themselves look like and how they take place, the Olympics are multi-day events lasting no longer than 16 days (The International Olympic Committee, 2019a, p. 70) that follow the same general structure. The start of each Games is celebrated with an Opening Ceremony which comprises a

[7] The latter makes sense in team sports in particular. In handball, for example, a country may qualify for the Olympics by winning the World or European Championships held by the IF of the sport (The European Handball Federation, 2020). The Olympic quota place is then filled by the NOC of the winning country, but the individual players on the world championship winning team are not necessarily the same as those who will compete for the NOC at the Olympics.

set of Olympic traditions in addition to artistic performances from the host (The International Olympic Committee, 2019a, p. 95; The Olympic Museum, 2013, pp. 7–8). Similarly, a Closing Ceremony marks the end of each Games with its own set of rituals and features from the hosting country's culture. In the time between the two ceremonies, the competitions take place.[8]

For each edition of Olympics, the IOC determines which of the sports governed by the IFs and which specific events will be on the Olympic program (The International Olympic Committee, 2019a, p. 81). A distinction is made between the terms "sport", "discipline" and "event" in the Olympics, and this same terminology is used in this thesis. A sport carries the meaning of the broader category of organized physical activity that the IFs govern, e.g. boxing, swimming and volleyball (The International Olympic Committee, n.d.-p, 2019a, pp. 82–83). The IOC defines a discipline as "a branch of a sport" (The International Olympic Committee, n.d.-p). To give an example, track cycling and road cycling are different disciplines in the Olympic sport of cycling (Hann, 2014; Union Cycliste Internationale, n.d.). Lastly, the competition (in a sport or discipline) in which athletes are ranked and awarded medals is called an event (The International Olympic Committee, 2019a, p. 81).[9] The event specifies the terms of the specific competition, such as the format, distance and whether the competing athletes are men, women or mixed, e.g. women's doubles in badminton (The International Olympic Committee, n.d.-d) and men's 4x100m freestyle relay in swimming (The International Olympic Committee, n.d.-r).

As is customary in sports competitions, the three types of medals that Olympic athletes compete for in events are gold, silver and bronze, given to those ranked first, second and third respectively. With each athlete competing under an NOC for the country they are representing, it is possible to aggregate across all events how many medals are won on a country (NOC) level for a particular edition of the Games. That is, each NOC at some specific Games can be assigned a count for how many medals have been won in events by the athletes representing it. This practice gives rise to what is known as the medal table described in the thesis introduction. For each Olympic Games, the medal table shows a ranking of all participating countries with their gold, silver and bronze medal counts, ordered by the number of medals won by its athletes.

[8] If given permission by the IOC and the IF of the sport, a competition may start before the Opening Ceremony (The International Olympic Committee, 2019a, p. 70).

[9] The term "event" is used in the thesis with two different meanings: 1. The individual Olympic competition in a sport or discipline as defined above, and 2. The everyday interpretation of the word meaning a happening or occurrence, used for instance to describe the Olympics themselves, i.e. as a "multi-day event". Although no explicit distinction is made between the two uses of the word in the thesis, care has been taken to make the meaning of each specific use clear from the context in which it appears.

It is important to note that when the results of events are recorded, medals are counted on an event-basis and not per medal-winning athlete. In events where athletes compete individually against other individual athletes, it follows naturally that the top three athletes each win a medal that counts towards their country total. Other events are held in team sports, e.g. handball, or are team events for individual sports requiring multiple people per participating party, e.g. athletes competing in pairs in tennis mixed doubles (The International Olympic Committee, 2020b). For these, medals counting on an event-basis means that although multiple athletes from an NOC win a medal together, the medal is won for the single event and counted only once.[10]

In practice, there is not one single official medal table as different sources will rank countries differently based on the weighting they each give to the three types of medals (I. Johnson, 2008).[11] The IOC itself does not think of Olympic competitions as being between countries (The International Olympic Committee, 2019a, p. 21) nor does it produce medal tables to rank and compare the performance of different countries (The International Olympic Committee, 2019a, p. 96). Nevertheless, it has become popular in the media to regard each edition of the Olympics as a country-level competition to be the most successful of the Games (Den Butter & van der Tak, 1995, pp. 27–28; I. Johnson, 2008; Kiviaho & Mäkelä, 1978, p. 9). The BBC (2016b), for instance, stated in a headline after the 2016 Olympics that Great Britain “beat China to finish second” when ranking countries by most gold medals won. For the purposes of this thesis, the medal tables’ specific ranking of countries is not important. It is the medal counts themselves that are of interest. Thus, when mentioning the medal table in the thesis, it is this aspect of them and not any specific ranking that is being referred to. The term “medal distribution” is used throughout the thesis to refer to the medal table’s medal counts as a whole. The notions of national Olympic performance and success manifest themselves in the medal counts.

[10] The IOC’s Olympic medal count for Denmark at the 2016 Olympics and the 2016 results published by the Danish NOC provide an example to illustrate that this is the case. The IOC (n.d.-f) counts two gold medals for Denmark, one of which is in the team sport of handball (Daugbjerg, 2016) even though it was won by multiple athletes.

[11] Some medal tables treat gold, silver and bronze equally and rank countries in order of total medals won (The Washington Post, n.d.), others rank countries according gold medals won, using silver, then bronze medals, to break ties between countries (The Guardian, n.d.).

2.1 The nature of Olympic competition: Applying the Herfindahl-Hirschman Index

With a better understanding of the Olympic Games and their structure and processes, this part of the thesis explores to the question of how Olympic medals have historically been distributed between countries. Before attempting to answer the main research question of why some countries win more Olympic medals than others, it should first be established that some countries do indeed win more than others, but also that the distribution of medals is not, to begin with, an entirely predictable matter. This helps to justify why the Olympic medal distribution is interesting to study in the first place.

To do so, there are two hypothetical extremes of Olympic competition that are interesting to consider. On one hand, it is possible to imagine a scenario in which no country performs better than any other so that no country wins more medals than any other. In such a case, every event is so highly competitive and performances by athletes so evenly matched that the outcome of events is random. Thus, every country, in expectation, wins an equal number of medals. At the other extreme, there could be a hypothetical situation in which there is so little competition that one country is always dominant and wins every medal in every event.[12] Looking at how Olympic medals have historically been distributed shows that the nature of competition at the Olympic Games falls somewhere between these two extremes. This can be demonstrated numerically using the Herfindahl-Hirschman Index.

The Herfindahl-Hirschman Index (HI) is an economic measure of market concentration. It quantifies the extent to which power is concentrated in a market or industry (Pepall et al., 2014, pp. 47–49) and is calculated as the sum of squared market shares, s_i, of all N firms in an industry:

$$HI = \sum_{i=1}^{N} s_i^2 \tag{1}$$

where the market share is measured in percentage terms (Pepall et al., 2014, p. 49).

The bigger the market share of a particular firm i, the more weight is given to it, and the higher HI will be. Its highest possible value, $HI = 10{,}000$, is the case of a monopoly, where all the market power is fully concentrated to lie with one firm (Pepall et al., 2014, p. 49). The lowest possible value of HI is close

[12] In practice, this is not possible (team events only feature one team per country, for instance), but the hypothetical example is made here for illustrative purposes.

to zero and reflects a perfectly competitive market of a large number of very small firms, none of which have market power. The value of HI can thus be used to give an idea of the level of competitiveness between firms in an industry, where a high market concentration (high value of HI) indicates a low degree of competitiveness and a low market concentration (low value of HI) indicates a high degree of competitiveness.

The HI has sometimes been used to measure competitive balance in sports (Owen et al., 2007; Zimbalist, 2002, p. 112). In the context of the Olympic Games, this idea of market concentration and power can be applied to measure the competitiveness between the competing countries, treating them as if they were firms in an industry. The market share in this case is the proportion of medals won. Similarly to a monopoly, a value of $HI = 10{,}000$ is the extreme scenario in which there is no competitiveness and one country dominates to win every medal, while a low HI reflects the other extreme where all countries win the same number of medals, making the Games the most competitive imaginable.[13]

Using Olympic medal data for 1996-2016, the HI quantifies competitiveness at the Olympic Games as follows:

Year	HI
1996	443
2000	423
2004	434
2008	452
2012	433
2016	418

Table 2.1: Herfindahl-Hirschman Index, 1996-2016 Olympic Games

Appendix A contains the full medal tables of the 1996-2016 Olympics, including the medal shares and squared medal shares for every medal-winning country and the calculated HIs. As Table 2.1 indicates, the HI has remained fairly stable across the 1996-2016 period. Its value is higher than it would have been

[13] In the application to the Olympics, the lowest possible HI value (where each country's market share in percent is $\frac{1}{N} * 100$) is $\sum_{i=1}^{N} \left(\frac{1}{N} * 100\right)^2 = N * \left(\frac{100}{N}\right)^2 = \frac{10{,}000}{N}$. For approximately 200 participating countries, this would be $HI = \frac{10{,}000}{200} = 50$.

in perfect competition (see Footnote 13), implying that countries at a given Olympics have not performed equally well and have not won an equal number of medals. On the other hand, it is also far lower than the monopoly value of 10,000, indicating that the Olympic medal tables are not dominated by one country. That there is a medal count disparity between countries, but not one so large that all countries but one win no medals, helps to establish the premise of the thesis. That is, if the expected medal distribution of the Olympic Games was characterized by either extreme, it would be pointless to investigate further.

Using the Herfindahl-Hirschman Index, it has thus been established that some countries do indeed win more medals than others at the Olympics, but that the Games are also quite competitive and there is no monopoly on medals. There is merit, then, in asking the question of why some countries are more successful at the Olympics than others.

3 Literature review

The interest in why some countries win more medals at the Olympic Games than others dates back to at least the 1970s. Ball (1972), Grimes et al. (1974) and Kiviaho and Mäkelä (1978) were among the first to consider the determinants of national Olympic success, with Kiviaho and Mäkelä (1978) noting the novelty of this direction in Olympics studies (p. 5–6). Ball (1972) explored the role that population size and economic resources play in national performance at the Olympics, arguing intuitively that the more resources a country has, the more likely it should be that the country is able to foster successful Olympic athletes and win more medals (p. 189). This idea of a positive relationship between medals won and human and economic resources was echoed in the studies by Grimes et al. (1974) and Kiviaho and Mäkelä (1978).

The argument presented for population size was that the larger it is, "the larger the pool of potential winners in any event from which to select team-members" (Ball, 1972, p. 189). Grimes et al. (1974) further made explicit the assumption that innate athletic potential is "randomly distributed throughout the world's population" (p. 778), such that a country with a larger population to draw from should have a higher expected number of potential medal-winning Olympic athletes and, thus, medals won.

The case for economic status was made with the argument that countries must have the necessary nutrition, facilities, disposable income, time for training, etc. (Ball, 1972, p. 189; Grimes et al., 1974, p. 778) to realize the potential of the athletically talented population. That is, a large population in itself would not necessarily be enough to win many Olympic medals, a country would need the economic resources to cultivate its potential Olympic athletes (Kiviaho & Mäkelä, 1978, p. 5).

As Ball (1972) and Kiviaho and Mäkelä (1978) did not use regression analysis to study the determinants, their findings, although similar to the following, will not be discussed further. Grimes et al. (1974) investigated whether there was any statistically significant difference between the Olympic performance of communist and non-communist countries, controlling for population size and GNP per capita. They argued that communist countries would be able to make relatively greater efforts to nurture athletic ability and devote resources to doing so. Using two interaction terms between a communist country dummy variable and both population and per capita GNP, Grimes et al. (1974) found the coefficient estimates of the interaction terms, as well as population and per capita GNP, to be statistically significant at the 1% level. Decades later, Den Butter and van der Tak (1995) studied the relationship between the number of

Olympic medals won and national income per capita, population and a (former) socialist country dummy and compared with similar regressions on different multidimensional welfare indicators instead of national income. A country's level of welfare was found to be a determinant of national Olympic performance, but Den Butter and van der Tak (1995) noted that the "multidimensional welfare indicators do not outperform national income" (p. 34). They suggested that national income could work as a measure of welfare because it may be highly correlated with the measures included in the indicators.

The most influential contribution to the literature has come from Bernard and Busse (2004).[14] Bernard and Busse (2004) created a theoretical framework for the existing ideas that the pool of potential Olympic athletes is proportional to population size, but that economic resources are necessary to exploit it. They motivated their choice of population and GDP per capita as independent variables by framing the problem as a Cobb-Douglas production function in which the talent of a country (i.e. its "production" of potential Olympic medal winners) depends on its people (population), wealth (GDP) and some organizational ability. They then defined the medal share of a country as a function of its talent, regressing it on the logarithms of population and per capita GDP, and found them both to be statistically significant at the 1% level. In an expanded model, Bernard and Busse (2004) introduced the inclusion of a host country dummy variable. It was used to reflect advantages to the host country such as the government resources used to host the Games (e.g. invested in facilities), the reduction of participation costs to athletes and various home advantages. In addition, this expanded model included dummy variables for the former Soviet sphere of influence and other planned economies and was run both with and without those editions of the Games that were subject to boycotts.[15] Lastly, Bernard and Busse (2004) added lagged Olympic medal shares (i.e. from the previous Games) as an explanatory variable and found all coefficient estimates to be statistically significant, with a large coefficient estimate on lagged medal share. It was therefore argued that current success is closely related to past success.

In a contemporary study with that of Bernard and Busse (2004), Johnson and Ali (2004) found per capita GDP, population, host and the political system to be statistically significant at the 5% level (p. 985–986). Since then, others have built on the study by Bernard and Busse (2004) and their model, and the literature has been expanded with the introduction of new explanatory variables. Forrest et al. (2010) used the

[14] This is evidenced by the accounts of later studies, including Andreff and Andreff (2015, p. 187) and Forrest et al. (2010, p. 577), who noted the reach and influence of Bernard and Busse's (2004) study, which has also extended out of the academic world and into the media (Swanson, 2016).
[15] Throughout the years, a number of Olympic Games have been boycott, most notably in 1980 and 1984 (The Olympic Museum, 2013, p. 20).

model as a basis for theirs and included public spending on sports and an ante-host effect dummy representing the future host country of the following Olympic Games. Lagged medal share (past performance), GDP, host, the Soviet dummy and ante-host were found to be statistically significant, while public spending on sports was not at the 5% level (Forrest et al., 2010, pp. 579–580). Andreff and Andreff (2015) also based their model on Bernard and Busse's (2004) but used lagged GDP per capita and population measures, arguing that these take time to be reflected in Olympic performance (p. 188). They also added region dummy variables to capture differences in sports culture, including the role of women and country specializations in specific sports. Half of these region dummies were found to be statistically significantly different from the reference region at the 5% level, but not when past performance was added as a regressor (Andreff & Andreff, 2015, p. 194). Lagged GDP per capita, lagged population, the host effect, political structure and past performance were all statistically significant. Lui and Suen (2008) found that GDP per capita, population and the host effect were statistically significant, while life expectancy and education were not (p. 10). They also suspected significant country effects and attempted to control for them using past performance/lagged medals (Lui & Suen, 2008, p. 11). Vagenas and Vlachokyriakou (2012) introduced team size and an ex-host effect (dummy variable for the past host country of the previous Games) as determinants in a larger model and found team size to be the only statistically significant regressor (p. 213). Trivedi and Zimmer (2014) found that athlete share (relative team size) and, in some specifications, female share (relative participation from female athletes) were statistically significant for winning at least one medal and for the number of medals won given that it was at least one (p. 184, 190).

The data used by early researchers of the determinants of the Olympic medal distribution was cross-sectional data from a single edition of the Olympic Games, e.g. the 1972 Olympics (Grimes et al., 1974, p. 779), and the 1988 and 1992 Olympics (Den Butter & van der Tak, 1995, p. 30).[16] Vagenas and Vlachokyriakou (2012) also analyzed just the 2004 Olympic medal data for countries that won at least one medal (p. 212). Bernard and Busse (2004) seem to have been some of the first to use panel data, with their dataset covering medal counts from 1960 to 1996 (p. 413), along with Johnson and Ali (2004) who used data from 1952 to 2000 (p. 976). Lui and Suen's (2008) study covered the 1952-2004 Games for a simple model and 1996-2004 for their extended model. Forrest et al. (2010) used panel data for the 1992-

[16] Den Butter and van der Tak (1995) used data from two Olympics Games but considered each a separate cross-sectional dataset to analyze rather than treating them as panel data.

2004 Olympics (p. 577), Trivedi and Zimmer (2014) the 1988-2012 Olympics (p. 171) and Andreff and Andreff (2015) the 1976-2004 Olympics (p. 189).

There have been a few different approaches to the econometric regression modelling of national Olympic performance. Because Olympic medal data behaves in a way that includes many zeros for the countries that do not win any medals at the Games, many studies have used a tobit model (Andreff & Andreff, 2015; Bernard & Busse, 2004; Forrest et al., 2010; Grimes et al., 1974) or a variant thereof (Trivedi & Zimmer, 2014, p. 170). Den Butter and van der Tak (1995) stated that tobit would be the usual solution, but used a variant of Poisson regression that could accommodate their choice of a log-linear specification to avoid undefined logarithms of zero (p. 29). Lui and Suen (2008) employed a tobit, Poisson and negative binomial model for a simple model and used Poisson for their extended model. Johnson and Ali (2004) chose a fixed effects model based on OLS methods (p. 984), and although they did not make this explicit, Vagenas and Vlachokyriakou (2012) also seem to have used OLS methods for their cross-sectional study (p. 213).

There has also been a variety of ways in which studies have measured Olympic success. Grimes et al. (1974) chose the total number of medals won by a country as the dependent variable, i.e. giving equal weight to gold, silver and bronze, explaining that "all weighting schemes are arbitrary" (p. 779). Johnson and Ali (2004) also used the total number of medals (p. 984), and Den Butter and van der Tak (1995) and Vagenas and Vlachokyriakou (2012) used its logarithm. Lui and Suen (2008) assigned different weights to gold, silver and bronze medals using a points system (p. 3). Bernard and Busse (2004), Forrest et al. (2010) and Trivedi and Zimmer (2014) measured Olympic success using each country's medal share (medals won out of total medals awarded) as the dependent variable.

4 Theoretical framework: Economic growth theory

In some of the literature on the determinants of the Olympic medal distribution, there seems to have been an inclination towards justifying the choice of explanatory variables with arguments that, intuitively, make sense. Ball (1972) and Grimes et al. (1974), for instance, argued why population size and wealth could help to explain national Olympic success, but did so without reference to economic theory that might motivate using these factors as determinants. Meanwhile, Den Butter and van der Tak (1995) never seemed to address why, in the first place, there should be a relationship between the number of Olympic medals won and the welfare indicators they were interested in investigating. In fact, Bernard and Busse (2004) appear to have been among the only ones to have conducted their study on the basis of economic theory, which they did by framing the problem as one of "production technology" (p. 413) as described in the literature review.

The purpose of this chapter is to provide an alternative, but related, framework through which to view the question of what makes some countries win more Olympic medals than others. That theoretical framework is the role of technology in economic growth. The explanatory variables that are suggested when viewing the problem through this lens are ultimately not too different from those that have already appeared in the literature, but the economic theory serves to motivate these variable choices by some means other than intuition or common sense to give them more weight.

A consequence of the intuitive approach to choosing variables is that whether a variable is dropped from (or added to) a model becomes up to the researcher rather than guided by economic theory. That is, because there is no underlying theoretical reason given as to why a regressor should (or should not) be included in a model of Olympic success, there is also nothing preventing the decision to drop (or add) regressors from being made on the basis of how much of the variation in medal data they explain, e.g. represented by the regression's R^2 value. For instance, Den Butter and van der Tak (1995) used the fit of the regression (in their case measured by the log-likelihood) to argue that multidimensional welfare indicators were not more suitable than per capita national income as a determinant of winning Olympic medals (p. 34). Bernard and Busse (2004), despite having a theoretical framework, also decided that past Olympic performance was useful as a regressor on the grounds that it "further improves the fit of the model" (p. 415). Vagenas and Vlachokyriakou (2012) dropped the ex-host effect from their model with no explanation other than the R^2 remained high and a previously insignificant regressor became stastistically significant (p. 213). This does not seem to be an ideal approach to choosing explanatory variables, as the ability of the regressors to predict the dependent variable is not a testament to each regressors appropriateness in the model (Stock & Watson, 2015, p. 283).

Thus, in addition to making a more convincing argument, the introduction of a theoretical framework to guide the choice of explanatory variables is also made to prevent the inclusion of regressors in the model from being dictated solely by the fit of the regression. In other words, it serves to ensure that the analysis is driven by economic theory rather than by which regressors result in the estimated regression model having a higher R^2 value (Stock & Watson, 2015, pp. 244, 283).

As a counterargument, it is possible that this approach in the literature, where data is included based on its effect on the fit of the regression, stems from the desire to predict or forecast the results of the next Olympic Games. Especially in more recent literature, a study might be structured such that it first suggests a model and then makes an out-of-sample forecast for the next Olympics and compares the predicted to the actual realized medal distribution. Among others, this has been seen with Bernard and Busse (2004), using their regression model based on 1960-1996 data to forecast the 2000 Games (p. 416–417), Lui and Suen (2008), using 1952-2004 data to forecast the 2008 Games (p. 13), and Forrest et al. (2010), using 1992-2004 data to forecast the 2008 Games (p. 580). Likewise, the media attention that has been directed at the determinants of national Olympic success, as described in the introduction, has primarily been concerned with predicting the outcome of future Olympics. Thus, if the goal is to be able to predict the Olympics, it makes sense to base the decision to include variables on whether they add explanatory power to the model.

However, when the regression model is constructed with the purpose of forecasting rather than trying to understand the causal relationship between dependent variable and regressors, its usefulness in these regards reflects it. The ability of a model to produce forecasts is not indicative of its ability to offer an understanding of causal effect (Stock & Watson, 2015, p. 378). As Stock and Watson (2015) describe it, these two uses for regression models "are conceptually very different" (p. 378). If that is the case, it seems ambitious for models in the literature to attempt to do both. Bernard and Busse (2004), for instance, seem to have been at least somewhat guided by explanatory power in their choice of regressors, and their model managed to predict the 2008 Olympics with an R^2 of 0.96 (p. 417). Meanwhile, they also offered interpretations of the coefficient estimates when no such meaning necessarily exists for their model specification, which was intended for forecasting. To reiterate, if the purpose is to forecast the Olympic medal distribution, the chosen variables should be highly predictive of the dependent variable (Stock & Watson, 2015, p. 378), and fit of the regression becomes a useful tool. If the purpose is to understand the disparities between countries when it comes to the number of Olympic medals won, the fit of the regression should not drive the choice of regressors (Stock & Watson, 2015, p. 283). As such, this thesis does not concern itself with forecasting, and the theoretical framework, rather than the fit of the regression, is what motivates the explanatory variable choice.

4.1 Economic growth theory and the Olympic medal distribution

From the question of why some countries win more Olympic medals than others, there is a parallel to be drawn to similar questions that are central to economic growth theory, namely why some countries are richer and have higher standards of living than others, and why some have become richer (and richer faster) over time while others have not (Weil, 2016, p. 3). This pondering pertains, for example, to specific factors or country characteristics that make countries' economic status differ (Weil, 2016, p. 23), which mirrors the problem at hand of understanding the factors and country characteristics that make countries' Olympic performance differ. Thus, by likening the problem of disparities between countries' Olympic success to the problem of disparities in country prosperity, models of economic growth can become a useful theoretical framework for understanding the Olympic medal distribution.

In economic growth theory, one explanation of growth differences between countries is the role that technology and technological progress play (Weil, 2016, pp. 201, 229), with technology being defined by Weil (2016) as "the available knowledge about how inputs can be combined to produce output" (p. 31). A simple model of the relationship between technology and economic growth proposes that output growth depends positively on the size of the population and on the fraction of the population that is invested in producing new technology (Weil, 2016, p. 213). This usefulness of technological progress for economic growth is the result of it enabling improvements in productivity, i.e. the ability to produce more output with the same input (Weil, 2016, p. 201). In mathematical terms, the relationship is stated as:

$$\hat{y} = \frac{\gamma_A}{\mu} L \tag{2}$$

where $\hat{y}$ is output growth in per-worker terms, γ_A is the fraction of the population producing new technology, μ represents the cost of new technology and L is the population size (Weil, 2016, pp. 212–

213). That is, in general terms, economic growth is positively related to population size and the efforts or resources that are invested into technological improvement (Weil, 2016, p. 229), i.e. expanding available knowledge about production.

For the case of national Olympic success and thinking of technology as defined above, a country's available knowledge on how to "produce" or foster Olympic medal-winning athletes may be thought of as a technology, and the purpose that this technology works towards is winning Olympic medals rather than achieving economic growth. In that sense, the interpretation of the relationship described above becomes that winning medals at the Olympic Games depends positively on a country's population size (human resources) and the resources invested into better cultivating medal-winning Olympic athletes. The point here is not that the model for the role of technology in economic growth in its exact functional form from Equation (2) must be used prescriptively, but it serves as inspiration for understanding why population and invested resources would matter for explaining Olympic success. Broadly speaking, the economic growth framework is useful for drawing parallels between an economy needing resources to grow and a country competing in the Olympics needing resources to win medals and rise in the medal tables.

There are other aspects of economic growth theory that are also interesting to consider in an Olympic context, namely the concept of technology transfer and the role of the open economy in economic growth. That technology can be transferred stems from the fact that knowledge is non-rival, such that one country's use of technology does not prevent another country from using the same technology (Weil, 2016, p. 204). This does not mean, however, that all countries will be able to utilize the same level of technology as there may be barriers preventing less successful countries from adopting the technologies of the more successful countries (Weil, 2016, p. 224). First of all, in order for technology to be transferable, it needs to be appropriate. Say for instance that one country that is successful at the Olympics has the technology, i.e. knowledge, to construct good training facilities for its athletes. A country that performs poorly at the Olympics may hold that same knowledge without wanting to construct the same facilities because the technology is not appropriate for the country. To give an example, this could be facilities for sailing in a country with a lot of coastline compared to a landlocked country with no large bodies of water. Second of all, the technology could be tacit knowledge, e.g. unspoken experience, that is not readily transferable (Weil, 2016, p. 227). This could pertain to the quality of coaching or the day-to-day running of sports clubs in countries that are successful at the Olympics that are not as easily identified by less successful countries even though they also have coaches and sports clubs.

As for the open economy, economic growth theory argues that in terms of growth, it is beneficial to have openness to allow for trade (Weil, 2016, p. 311), which can be regarded as a transfer of technology (Weil, 2016, p. 317). Meanwhile, geographical barriers hinder this transfer by limiting openness and making trading difficult, and they are thus detrimental to economic growth (Weil, 2016, pp. 310–311). If a transfer of technology between countries is possible in an Olympic context (i.e. the technology is appropriate for the country looking to adopt it, and the technology is not tacit knowledge), then the role of openness in economic growth should imply that the more open a country is, the more trade or transfer of Olympic technology is possible, and the better the country should be at winning Olympic medals. The openness to trade can, for instance, be reflected in political openness concerning the ability to cross borders and could be exemplified by the ease with which foreign trainers, coaches and athletes can be "imported" from abroad[17] or the ease with which athletes can go abroad to train to improve their chances of winning Olympic medals. In terms of geographical barriers, a small isolated island country would find it much more difficult to have this Olympic technology exchange with other countries than a country with many neighbors in the vicinity, each with their own technology to trade. Even so, trade would become easier the better connected a country is to the rest of the world, even if it is an island (e.g. through available train and flight routes), so this accessibility aspect of geographical barriers should also be considered. Thus, a country's political openness and geographical proximity or connectedness should both be positively related to the number of Olympic medals won.

4.1.1 The proposed regressors

By providing this economic growth theory framework, country level determinants of the Olympic medal distribution can be suggested, and they fall into three general groups:

[17] It is not uncommon for athletes who are nationals of multiple countries to compete for a different NOC than they country they grew up in (Jansen & Engbersen, 2017; Vasilogambros, 2016). This transfer of athletes between countries could be regarded as a transfer of technology. Athletes who are eligible for representing more than one NOC may opt to compete for the one they believe gives them the best chance of winning. Eligibility depends on the country's openness in terms of its requirements for who can represent it as a national Olympic athlete.

1. The size of the population
2. Resources invested in technological progress
3. Openness

all of which should be positively related to the number of medals won by a country.

For the first group, the thesis proposes two determinants: population size and Olympic team size. In a population, some proportion possesses the innate technology of outstanding athletic ability that is a necessary resource to win medals. Thus, the larger the population, the more people with the abilities (technology) and potential to win at the Olympics, and the more medals a country should expect to win. In the context of the Olympics, a population may be thought of more narrowly as the national Olympic team that directly contributes towards winning Olympic medals. As mentioned in the background chapter on the Olympics, an athlete may qualify by being chosen by their NOC rather than meeting qualification requirements in competition, but it is assumed here that the choice by the NOC is fully meritocratic, i.e. it is the athlete's skills as the best Olympic candidate, and not nepotism for instance, that qualifies them. As with population, a larger Olympic team should be associated with more medals won.

For the second group, per capita GDP, investment in sports and three different host effects are suggested as determinants of the Olympic medal distribution, representing the resources invested in improving Olympic technology. The more that is invested in the cultivation of Olympics athletes, the better a country should perform at the Games. Per capita GDP, as a reflection of a country's standard of living, can be treated broadly as a measure of the available resources that can be invested into fostering Olympic medal-winning athletes. This includes investment into training facilities, recreational time for an athlete to spend on sports, the financial means for an athlete to support their sports activity and later a career in sports, government support, high levels of education and expertise from coaches, personal trainers, nutritionists, etc. Where per capita GDP is a broader measure of the potential for resources to be invested in Olympic technological progress, specific investment into sports is a narrower reflection of technological improvement effort. That is, the investment in sports, like per capita GDP, is indicative of the resources invested in fostering Olympic athletes.

The host effects described in the literature and represented by dummy variables can also be viewed in this light as the effect of invested resources in technological progress. As explained in the chapter on the Olympics, countries that host need to invest in building or improving infrastructure and sports facilities to hold the Games in, and this increased effort towards Olympic technological improvement should increase the host country's medal count. Because the investment and increased focus on the Olympics start many years in advance for the coming host, there could be, as Forrest et al. (2010) described, a pre-host effect dummy where increases in the medal count from hosting the Games can already be seen from the edition of the Games preceding the hosted ones. Similarly, a post-host effect could exist for the country that hosted the previous Games due to the increased effort from the IOC to recognize the importance of legacy, i.e. finding post-Games benefits of the Olympics to the host city. That is, training facilities may continue to be used by the host country's athletes even after the hosted Games have ended, and increased attention on lesser known Olympic sports can help fuel more interest in them (The International Olympic Committee, n.d.-e). Thus, these three host effects represented by dummy variables should be positively associated with a country's medal count.

The host effect could become even more pronounced in future Olympics as host countries are able to propose non-core sports to be included in their hosted edition of the Games. As an initiative of the so-called Olympic Agenda 2020 detailing the future direction of the Games (The International Olympic Committee, n.d.-i), host countries are given the advantage of being able to propose sports that they specialize in to be included (The International Olympic Committee, n.d.-i, p. 14). For the Tokyo 2020 Games, this has resulted in the addition of baseball/softball, skateboarding, surfing, karate and sports climbing as Olympic sports (The International Olympic Committee, 2016b), i.e. sports in which Japan is likely to possess more technology than many other countries. This should give the host country an even greater advantage for winning Olympic medals than the already increased investment in facilities.

A dummy variable for different political systems (communist, socialist countries), as has been prevalent in the literature, could also be understood as belonging to the second group of determinants, measuring systematic differences in the resources invested in Olympic technology. However, the world has looked very different in the years that are considered in this thesis (1996-2016) compared to during the Cold War when studies on the determinants of Olympic success were first conducted and Olympic success was used to legitimize ideology (Grimes et al., 1974, p. 777; Grix & Carmichael, 2012, p. 81). Because the ideological war between capitalism and communism is much less pronounced today and the effect of the dummy has been found to decrease over time (Den Butter & van der Tak, 1995, p. 30), it will not be considered as a determinant. Other sources of Olympic technology investment that differ between countries and could be considered determinants in the economic growth framework under the second group include the sports culture or tradition, which relates to variables seen in the literature pertaining to gender inequality and past Olympic performance. These cultural differences are interesting, but unobservable, and will be addressed with the choice of econometric method in the next chapter.

Lastly, for the third group, a country's openness (both political and geographical) is regarded as a determinant of Olympic success because of the possibility for technology transfer that it facilitates. Because political openness and the lack of geographical barriers are difficult to measure directly, it is necessary to find a variable that acts as a proxy for these. What this thesis proposes is to use tourism as a measure. The amount of tourism a country welcomes should be positively related both to political openness (such that a lack of political barriers to entering the country, e.g. lenient visa requirements, encourages more tourism) and to how easily accessible the country is geographically (e.g. geographical distance from neighbors, flight connections, travel time). Thus, the level of tourism in a country, as a proxy for political openness and lack of geographical barriers, should be positively associated with the number of medals won at the Olympics through its facilitation of the movement of technology (e.g. athletes, coaches) across borders.

In summary, by framing the problem using economic growth theory, this chapter has provided a theoretical justification for the explanatory variable choices, including the population and per capita GDP regressors prevalent in the literature. The following determinants have been proposed to be positively related to national Olympic success: population size, Olympic team size, per capita GDP, investment in sports, host, pre-host and post-host effects and tourism as a proxy for political openness and geographical ease of access.

5 Methodology

This chapter of the thesis comprises three parts. The first part concerns the thesis' philosophy of science approach to investigating its research question. In the second part, the choice of econometric method is presented and explained, and the model assumptions are discussed in relation to the Olympic medal distribution and its determinants suggested in the previous chapter. The third and final part describes the data used in the analysis and discusses the choices made in regards to data collection and processing.

5.1 Philosophy of science

The paradigm, or basic belief system (Guba, 1990), that most closely represents how this thesis relates to its own position as a source of knowledge creation is that of post-positivism. Building on the conventions of positivism, it answers the ontological questions of what it considers to be reality and knowledge with critical realism, i.e. the admittance that, while an objective world and truths about it exist, they cannot be fully understood because humans are incapable of perceiving them (Guba, 1990, p. 20). It becomes impossible then to remain an objective observer of the world, removed from what is studied and capable of producing knowledge without bias, as is the case in the positivist epistemological position. Instead, post-positivism recognizes that objectivity is desirable, but any study will be subject to the researcher's own bias, which should be acknowledged and made clear (Guba, 1990, p. 21).

For the thesis, this is reflected in the underlying assumption that there exists in the world some objective relationship between the number of Olympic medals won on the country level and its determinants. The truth about this relationship, however, cannot be known with perfect certainty, and the results of the analysis are influenced by the experiences and observations about the world that have shaped the direction of the thesis and the choice of theory. Care has therefore been taken to be transparent about any assumptions and decisions made and any positions taken. This inability to perfectly know the determinants of the Olympic medal distribution has an interesting parallel to the broad choice of method, the regression analysis, given its inductive nature. That is, for some true unknown population regression coefficient, for instance, statistical techniques such as hypothesis testing can make some probabilistic statement about the general true relationship inferred from the specific sample. Thus, there is inherently an element of the unknown and unknowable in the interpretation of regression results, exemplified by the tolerance for incorrect rejection of a null hypothesis implied by the significance level in hypothesis testing (Stock & Watson, 2015, p. 123). In relation to the post-positivist epistemology, it seems fitting that the results of the analysis cannot make claims about the objective world that are more certain than what is possible within the post-positivist paradigm.

To address the shortcomings of what Guba (1990) called the “overemphasis on quantitative methods” (p. 22) in the positivist paradigm, the post-positivist methodology encourages the inclusion of qualitative as well as quantitative methods. Although this has not been done for the thesis, which centers on quantitative analysis, the nature of this quantitative analysis is not the controlled experiment that is traditional for positivism (Guba, 1990, p. 20). Instead, it is better characterized by being, as Guba (1990) wrote on the post-positivist methodology, an “inquiry in more natural settings” (p. 23). With a subject matter like performance at the Olympics, it is simply not feasible to replicate the competitive conditions of the Olympics in a randomized controlled experiment. Any conclusions drawn from an attempt at such an experiment, if it were possible, would have limited generalizability to the real world, the natural setting of which should be studied instead, as it is here. Furthermore, the use of theory is less prescriptive than would be usual for positivism; the emphasis in the thesis is placed more on discovering the determinants of Olympic success by taking inspiration from theory rather than being an attempt at verifying the theory. This is more characteristic of post-positivism than positivism (Guba, 1990, pp. 22–23).

Thus, the thesis approaches the econometric study of the determinants of national Olympic success within a post-positivist paradigm, and, given this approach, the methods employed to investigate them seem appropriate.

5.2 Panel data econometrics: The fixed effects regression model

As seen in the literature review, a few different econometric models have been suggested for analyzing the relationship between the Olympic medal distribution and its determinants. The purpose here is to clarify which type of regression model is employed in the analysis of this thesis and to explain the preference of this choice over some of its alternatives.

The type of data, as has been implied in the thesis so far, is panel data, i.e. data for multiple entities observed at multiple time periods (Stock & Watson, 2015, p. 57), which in this context should be understood to be data for multiple countries in multiple Olympic Games years (1996-2016). A natural way to analyze panel data is by using linear panel data regression models, the main model of which is fixed effects regression, an extension of multiple regression using OLS (Stock & Watson, 2015, p. 396). In multiple regression for cross-sectional data, an omitted variable that is unobserved cannot be included and may lead to omitted variable bias, but regression with panel data allows for unobserved time- or entity-invariant variables to be controlled for, such that omitted variable bias may be eliminated. Time-invariant variables, also known as entity fixed effects, vary across entities but remain constant over time (Stock & Watson, 2015, p. 403), while entity-invariant variables, also known as time fixed effects, vary over time but remain constant across entities (Stock & Watson, 2015, p. 408).

Broadly speaking, there are mainly three models for regression with panel data: the pooling, the fixed effects and the random effects models (Croissant & Millo, 2008). The pooling model, as the name suggests, is a pooled OLS model treating data from different entities and time periods as if they appeared in a typical cross-sectional OLS regression (Croissant & Millo, 2008, pp. 2–3). It assumes that there are no entity or time specific effects and estimates the same regression line for all entities and time periods. In the fixed and random effects models, entity (individual) and/or time effects are incorporated such that the error term in the regression function includes an individual and/or a time component.[18] Whether a fixed or random effects model is most suitable depends on whether the individual/time component is correlated with the regressors or not. The former calls for a fixed effects model, the latter a random effects model. As is discussed later in this chapter, the individual effects in the Olympic data could include factors such as sporting culture, denoting the general tradition for a country to engage in sports, especially at an elite level. This is likely correlated with a variable such as Olympic team size (the more cultural engagement in elite sports, the more Olympic level athletes there should be), which is a regressor as suggested by the theoretical framework. A fixed effects model would therefore be more appropriate here than a random effects model. It would also be preferable to the pooling model as there are likely to be unobserved individual and time effects that would lead to omitted variable bias in the pooling model, but can be controlled for with the fixed effects model.

As Croissant and Millo (2008) explained, when the individual and/or time components of the error term are correlated with the regressors, as they are in the fixed effects model, OLS estimators for the regression

[18] Fixed and random effects models can include either an entity effect, a time effect or both (Croissant & Millo, 2008).

coefficients are inconsistent (p. 3). It is therefore conventional in fixed effects models to redefine the individual/time component of the error term as regression dummy variables to represent each entity and time period, such that OLS estimators are consistent.[19] The fixed effects model can thus be estimated with OLS methods. Because these dummy variables in practice mean that each entity and time period has a different intercept in the population regression line (Croissant & Millo, 2008, p. 3; Stock & Watson, 2015, pp. 403–404, 408), the fixed effects regression model with both entity and time fixed effects can equivalently, and more succinctly than a model with dummies, be stated as follows:

$$Y_{it} = \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} + \alpha_i + \lambda_t + u_{it} \tag{3}$$

For entity $i = 1, \ldots, n$, time period $t = 1, \ldots, T$ where Y_{it} is the dependent variable; $X_{k,it}$ is the value of the kth regressor for entity i in time period t; α_i is the entity fixed effect for entity i; λ_t is the time fixed effect for time period t and u_{it} is the error term (Stock & Watson, 2015, pp. 405, 409). This fixed effects regression model for panel data is the model that is used in the thesis' analysis of the Olympic medal distribution.

As mentioned in the literature review, a popular model choice in the literature for studying the determinants of national Olympic success has been the tobit model. While Grimes et al. (1974), Den Butter and van der Tak (1995) and Forrest et al. (2010) argued that tobit is particularly suitable for Olympic medal data because of the zero-inflation it exhibits, this sentiment seems to stem from a misunderstanding of either the nature of the medal data or tobit. Tobit models are ordinarily used for regression with a limited dependent variable, where the limitation arises from a censoring of the dependent variable above or below some cutoff, beyond which the dependent variable can no longer be observed (Amemiya, 1984, p. 6; Stock & Watson, 2015, pp. 467–468). That is, the dependent variable does exist outside the observable values, but is observed only for a limited or censored range. This, however, is not conceptually applicable to Olympic medal data. Although the number of medals is a limited dependent variable that cannot take on a value below zero, this limit of zero is not the result of censoring as seen in a tobit model. It is not the case that Olympic medal counts exist but are simply unobserved below zero, giving rise to a large number of zeros in the data. The number of Olympic medals

[19] For n entities, to avoid multicollinearity (what is known as the dummy variable trap) only $n - 1$ dummy variables are needed (Stock & Watson, 2015, p. 250). One entity (country) is captured by the intercept, and the coefficients on the entity dummies are interpreted relative to the intercept (Stock & Watson, 2015, p. 404). The same applies to time fixed effect dummies.

does not exist below zero, and a medal count of zero for a country means just that; that no medals were won.

Instead, Olympic medals as a limited dependent variable are better characterized as count data where the number of medals won by a country is a counting number (Stock & Watson, 2015, p. 469). A large number of zeros is common in count data (Zeileis et al., 2008, p. 1), which is typically analyzed using variants of the Poisson or negative binomial regression model (Stock & Watson, 2015, p. 469). Since OLS methods can also be used to model count data (Stock & Watson, 2015, p. 469), the fixed effects model remains the main choice of model for the thesis since this approach is centered around another defining characteristic of the dataset, namely that it is panel data. To, as Stock and Watson (2015) described it, "take advantage of the special structure of count data" (p. 469), which OLS cannot, the analysis also suggests a zero-inflated negative binomial model as a count model alternative to the fixed effects model.

5.2.1 Fixed effects regression assumptions

Having established the fixed effects regression model as the model choice, this part of the thesis discusses the fixed effects regression assumptions in relation to Olympic medals and the regressors suggested in the theoretical framework chapter.

The fixed effects regression assumptions mirror the least squares assumptions for regular OLS regression methods and simply extend them from cross-sectional to panel data (Stock & Watson, 2015, p. 411). The first assumption concerns the conditional mean of the error term which must be zero. In a practical sense, when this assumption holds, there is no omitted variables bias or other form of endogeneity, i.e. the error term is not correlated with the regressors (Stock & Watson, 2015, pp. 411, 471). The second assumption states that across entities (countries), the variables in the model are independently and identically distributed, which can be achieved by simple random sampling, while the variables for each country are allowed within themselves to be autocorrelated (Stock & Watson, 2015, pp. 411–412). When the third assumption holds, large outliers for the dependent and explanatory variables are unlikely. Lastly, according to the fourth assumption, there is no perfect multicollinearity. These four assumptions ensure the consistency of the fixed effects estimators and, under the central limit theorem, that they are approximately normally distributed when n, the number of countries, is large (Stock & Watson, 2015, pp. 96, 412). According to Stock and Watson (2015), $n \geq 30$ is in many cases sufficient (pp. 97–98), which the number of countries participating in the Olympics exceeds.

The first assumption requires the most discussion so is deferred until the others have been addressed. The second assumption is interesting because the data considered in this thesis is not a sample as such, but rather constitutes the whole population of countries at the Olympic Games. This is not an uncommon feature in datasets, especially in what Berk et al. (1995) referred to as "cross-national research" (p. 422) as is the case here. In relation to this second assumption, it is simply assumed, with reference to Stock and Watson's (2015) treatment of such a population dataset, that there is no sample selection problem that could induce correlation across entities and thereby induce issues with the fixed effects estimators (p. 372, 386). As for the autocorrelation that is allowed within countries, it is not difficult to imagine that it is present in the panel dataset given the time dimension (Stock & Watson, 2015, p. 412). The per capita GDP for a given country is likely to be correlated with past per capita GDP for the country, and population size is likely to be correlated with past population size. Because autocorrelation is likely to occur in the error term as well, it has implications for the standard errors that are used in the fixed effects regression model. This is addressed later in the chapter.

Regarding the third assumption, a graphical illustration using scatterplots is provided in the descriptive statistics analysis in the next chapter to support the idea that the variables are unlikely to have large outliers.

As for the absence of perfect multicollinearity required by the fourth assumption, of the suggested determinants of the Olympic medal distribution (population, Olympic team size, per capita GDP, sports investment, host effects, tourism), the only two variables whose relationship could be suspected to be perfect linear functions of each other are population and Olympic team size. The assumption that was made for population was that innate athletic talent is evenly distributed in the world, such that a fixed proportion of the population in each country has medal-winning potential. If then a fixed proportion of these talented athletes worldwide qualified for the Olympics, population and Olympic team size would exhibit perfect multicollinearity. However, in the real world, these perfectly fixed proportions are not realized, and in addition, Olympic team sizes are restricted such that no matter how many potential Olympic medal winners a country has, a maximum of only three athletes from the same NOC may compete in the same event, as described in the earlier chapter about the Olympics. Thus, none of the chosen regressors are perfectly multicollinear. Another perfect multicollinearity concern could be the potential inclusion of regressors that are constant either across countries or time periods, because these would exhibit perfect multicollinearity with the country or time fixed effect dummies.[20] Instead, it might be useful to think of these potential regressors as being incorporated in the country and time fixed effects, which will be discussed later in the chapter. For the regressors suggested by the theoretical framework, none can be said to be constant across either an entity or time dimension or both, i.e. they all vary across countries and across Olympic Games years.

To return to the first fixed effects regression assumption, there are several potential sources of omitted variable bias that are interesting to discuss and important to address to avoid biased estimators in the model. Omitted variables bias occurs when an omitted variable is correlated with a regressor and is itself a determinant of the dependent variable and therefore contained in the error term (Stock & Watson, 2015, pp. 230–231). As suggested by the theoretical framework, population, Olympic team size, per capita GDP, investment in sports, being/becoming/having been host and tourism are all determinants of Olympic success, and their mutual inclusion in many cases prevents omitted variable bias. As implied in previous discussion on the relationship between population and Olympic team size, a large population has a greater expected number of elite athletes, making the two correlated. Because of this, if population was used as a regressor while Olympic team size was not, there would be omitted variable bias.[21] Similarly, the background chapter on the Olympics suggests that the host dummy is correlated with both Olympic team size (as the NOC of the host country is given automatic qualification in some events, increasing team size) and per capita GDP (suggested by the heavy financial burden of hosting), so the omission of the host effect dummy would lead to omitted variable bias (and likewise for team size and per capita GDP). Thus, the inclusion of the determinants suggested by the theoretical framework helps to eliminate the omitted variable bias that would have occurred if only, for instance, the popular population and per capita GDP regressors had been included.

However, two concerns for omitted variable bias do remain. Firstly, because their inclusion would cause a problem with perfect multicollinearity as discussed above, no variables that are constant in either a country or time sense could be suggested as regressors. For example, tourism was chosen as a measure of political and geographical openness, but geographical accessibility may also have been represented by

[20] As noted in Footnote 19, the $n-1$ and $T-1$ (rather than n and T) country and time dummies are also made to avoid the dummy variable trap, which is a perfect multicollinearity problem (Stock & Watson, 2015, p. 250).

[21] Vagenas and Vlachokyriakou (2012), for instance, used team size as a regressor together with population among other variables and, upon finding team size to be the only statistically significant regressor, proceeded to remove team size to create their main model (p. 213–214). They did not address that this would induce omitted variable bias.

the number of land borders for any given country, if not for the fact that land borders have been practically fixed in the period 1996-2016. Including borders as a regressor would therefore have led to perfect multicollinearity with the country dummies, so it could not be used as a regressor, even though, in theory, it could be considered a determinant of Olympic success by acting as a proxy for geographical accessibility. Since the number of land borders is likely to be correlated to tourism (as they could both arguably be proxies for the same measure, and because, intuitively, a country with many neighbors likely gets more visitors than a small isolated island state), being unable to include it as a regressor could present an omitted variable bias problem. However, these types of constants within countries across time are exactly what country fixed effects can account for (and likewise for time fixed effects for constants across countries within a time period). As explained at the start of the chapter, a strength of analyzing panel data with fixed effects over a pooling model is that they provide a solution for eliminating omitted variable bias from time- and country-invariant variables (Stock & Watson, 2015, p. 366).

Secondly, because per capita GDP is used as such an all-encompassing measure of the standard of living, even if all the other proposed regressors are included in the fixed effects regression model, there are likely to still be factors contained in the error term that are correlated with per capita GDP. Those that are not captured by the country and time fixed effects would induce omitted variable bias. As a result, an instrumental variable is introduced as the solution to per capita GDP being an endogenous variable (Stock & Watson, 2015, p. 367).

How specific country and time fixed effects apply to the regression of Olympic medal counts is discussed below, followed by an explanation of the choice of instrument.

5.2.2 Country and time fixed effects

Country and time fixed effect both arguably play a role in the Olympic medal distribution. In an Olympic context, a particularly important interpretation of country fixed effects is as the effect of differences across countries in sporting culture and tradition which remain practically constant over time.

In a general sense, the level of importance placed on sports in society could be different for different countries. A cultural indifference towards sports would be associated with a lack of resources being devoted to Olympic technological progress and, for instance, a small national Olympic team. Because the cultural attitude towards and efforts to engage in sports could also be considered a resource for improvements in Olympic technology and thereby a determinant of Olympic success, not including it as a regressor would induce omitted variable bias. However, culture cannot be readily observed and measured, so including it as a country fixed effect becomes the solution. Similarly, cultural traditions for specific sports also determine how many resources a country directs towards technological progress in that sport. Because there are Olympic sports that have higher medal-winning potential than others (e.g. the multitude of swimming events compared to only two football events, one men's and one women's), a country's medal count also depends on whether it specializes in Olympic sports that award many medals across many events.

As mentioned in the literature review and briefly in the theoretical framework chapter, gender equality and the inclusion of women in sports and society in general have been suggested as a determinant in the literature and could also be seen from the economic growth perspective as an additional resource for Olympic technology and a determinant of Olympic success.[22] However, because there is no natural way to measure gender equality and because attempts to create a global index to quantify it have been relatively recent and criticized (Permanyer, 2013), a way to think about including it in the fixed effects regression model is as a country fixed effect. The level of gender equality varies across countries, but within a country, the role of women in society and in sports is likely to be an expression of a cultural attitude, which tends to be slow changing over time. The capturing of this aspect of culture in the country fixed effects eliminates any omitted variable bias arising from gender equality being a determinant of Olympic success and correlated with e.g. per capita GDP as a measure of standard of living.[23]

Similarly, a determinant that has also appeared in the literature, e.g. Bernard and Busse (2004), but may be better incorporated into the regression model as a country fixed effect, is past Olympic performance. As Andreff and Andreff (2015) noted, using past performance as a regressor induces endogeneity and omitted variable bias because a determinant of past performance would likely also be a determinant of current performance, causing the past performance regressor to be correlated with the error term (p. 195). To prevent this endogeneity, past Olympic success can be thought of as representing some country

[22] Encouraging the participation of women in sports increases the pool of athletic talent to draw Olympic athletes from and enables a country to win medals in women's events at the Olympics.
[23] Intuitively, it makes sense for this correlation to exist given that gender equality ensures a better standard of living for the female half of a population. Gender equality, relative to inequality, would also be expected to result in higher output/income for a given population size because of the contribution from women.

specific ability persistent over time (e.g. sports culture) that makes a country good at winning Olympic medals. Thus, success in past performance can be incorporated into the country fixed effects.

The time fixed effects do not address any particular concern regarding omitted variable bias, but they are included in the fixed effects model in the analysis to account for factors that change over time but are the same for all countries in a given time period. This could, for example, be the inclusion of more events over time, as Bernard and Busse (2004) also argued (p. 414). Introducing more events over time means that more medals are awarded in total at each edition of the Games, so that, on average, countries win more medals over time, not because of anything they themselves do differently, but because the number of medals available at each Olympic Games has increased for every country.

5.2.3 Instrumental variables regression

Instrumental variables regression allows for the effect of the endogenous regressor, per capita GDP, on the medal count to be isolated from the effect on the medal count that comes from per capita GDP's correlation with the error term (Stock & Watson, 2015, p. 470). The instrumental variable, or instrument, is used to isolate the part of the per capita GDP effect that is uncorrelated with the error term, so that the estimated coefficient on per capita GDP in the regression model becomes unbiased. This is done by using the two-stage least squares (TSLS) estimator, where, in the first stage, per capita GDP is regressed on the instrument and the exogenous regressors, and in the second stage, the Olympic medal distribution is regressed on the predicted values of per capita GDP from the first stage and the exogenous regressors (Stock & Watson, 2015, pp. 483–484).

In order for the instrument to be valid, it must meet two conditions (Stock & Watson, 2015, pp. 472, 485): relevance (the instrument is correlated with per capita GDP) and exogeneity (the instrument is not correlated with the error term, i.e. it is not correlated with any determinant of national Olympic success not included as a regressor in the model). Finding a valid instrument that is correlated with per capita GDP, but is simultaneously uncorrelated with other factors in the error term that determine a country's medal count is difficult. This thesis proposes the infant mortality rate, defined as the number of infants for each 1,000 live births in a year that die before reaching the age of one (The World Bank, n.d.-c).

With regards to instrument relevance, a negative relationship between per capita GDP and the infant mortality rate has been suggested (Baird et al., 2011), as there seems to be a strongly negative correlation between the two variables (Preston, 1975, p. 232; Pritchett & Summers, 1993, p. 1).[24] As Preston (1975) and Pritchett and Summers (1993) argued, infant mortality, as a measure of a country's health, should be correlated with per capita GDP, a measure of a country's standard of living, because quality of health would generally be considered an indicative aspect of the standard of living. Thus, the instrument relevance condition should be satisfied. A concern then is how relevant it is and whether it is a weak instrument that does not explain much of the variation in per capita GDP, which would make the TSLS estimator biased (Stock & Watson, 2015, pp. 489–490). To check that infant mortality is not a weak instrument, the analysis in the next chapter reports the TSLS first-stage F-statistic for the hypothesis test that the coefficient on infant mortality in the first-stage regression is zero. If the F-statistic is greater than 10 (as it is shown to be in the next chapter), infant mortality being a weak instrument is not a concern (Stock & Watson, 2015, p. 490).

As for instrument exogeneity, intuitively, it seems plausible that a country's infant mortality rate has no effect on its national Olympic success other than indirectly through its correlation with per capita GDP, i.e. the standard of living in the country. The number of medals a country wins at the Olympics most directly depends on the resources that are put into cultivating Olympic athletes and enabling the participation in elite sports. This includes the standard of living due to the general benefits (e.g. free time, funding) that a high living standard has to athletes and the facilitation of sports participation. Because infant mortality itself has very little, conceptually, to do with sports, there is no obvious reason to believe that the infant mortality rate is related to these kinds of factors that are conducive to sports and sports performance, except the standard of living. In other words, it seems likely that the infant mortality rate is correlated with per capita GDP, but not with the error term. It might be argued that since infant mortality is a measure of a country's health, other measures of health could be related to it, and if these other measures of health were included in the error term, the instrument exogeneity condition would be violated. It is not impossible to think of ways in which national health in general might affect Olympic success. For instance, being in good health is important for athletes, so good national health could help to ensure that athletes stay healthy, stay in good shape and are able to overcome injury. However, these other measures of health that may be thought of as beneficial to a country's medal count are arguably

[24] The negative correlation seems to exist between the log of per capita GDP and the log of infant mortality, as illustrated by Hague et al. (2008, p. 8).

already represented by the standard of living measured by per capita GDP. That is, are are likely not affecting Olympic success other than through per capita GDP, and they would therefore not be included in the error term. Thus, it can be reasonably assumed that the infant mortality rate is unrelated to the error term, such that the instrument exogeneity condition holds.

5.2.4 Heteroskedasticity- and autocorrelation-consistent (HAC) standard errors

As explained earlier in the chapter, there is likely to be autocorrelation in the error term of the fixed effects regression model (Stock & Watson, 2015, pp. 412–413). When that is the case, heteroskedasticity- and autocorrelation-consistent (HAC) standard errors should be used, as they are valid when the error term is both potentially (but not necessarily) heteroskedastic and autocorrelated (Stock & Watson, 2015, p. 413). The type of HAC standard errors employed in the analysis of this thesis are clustered standard errors, which allow for error term autocorrelation within, but not across, clusters (i.e. countries in this case). As Stock and Watson (2015) stated, they "allow for heteroskedasticity and autocorrelation in a way that is consistent with the second fixed effects regression assumption" (p. 413).

Thus, the chosen model for the analysis is a fixed effects regression model that can take advantage of the strengths of panel data, and it is supplemented by an example of how a zero-inflated negative binomial model that can account for the count data characteristics of the dataset might be applied to it. The fixed effects regression assumptions can reasonably be assumed to hold, such that regression coefficients are unbiased and can be consistently estimated. In particular, omitted variable bias from time- and entity-invariant variables is eliminated with country and time fixed effects, while omitted variable bias related to per capita GDP will be accounted for using infant mortality as an instrument. Clustered standard errors will be employed to ensure that standard errors are valid.

5.3 Data

This part of the thesis details the data that is used in the analysis. After a brief explanation of the intended general structure of the constructed dataset, the measures and sources of data for the dependent variable, the regressors and the instrument are described and discussed in the data collection and processing section. The chapter closes with a comment on the construction of the dataset and the choices made to prepare it for analysis.

As mentioned earlier in the chapter and evidenced by the regression model choice, the data that the thesis uses is panel data. The number of entities in the constructed dataset is $n = 206$, which is the total number of countries that have participated across the Olympics from 1996 to 2016 (see Appendix B for the list of countries). The number of time periods is $T = 6$ years for the six editions of the Olympics that are considered (1996, 2000, 2004, 2008, 2012 and 2016). That gives a total of $n * T = 206 * 6 = 1236$ observations,[25] where each observation represents a unique country-year pair. The dataset is constructed based on these observations, each of which is a row in the dataset, and each column contains a different variable whose value is given for each country-year combination. Appendix C shows the first few rows of the dataset to illustrate this.

It is important to note that not all the countries included in the dataset have participated in all six Games. For cross-sectional analysis involving a single edition of the Olympics, the dataset can be limited to only the countries that participated in those specific Games, as seen in e.g. Den Butter and van der Tak (1995, p. 28). This makes sense because only those countries have values for the dependent variable, the medal count, while a country that did not participate does not have a medal count. However, because panel data is structured such that the same entities are observed in each of the time periods, an observation is still included in the dataset for a country even if, in one year, it did not participate in the Olympic Games. Because it is a slight misrepresentation to assign this country-year observation a medal count of zero (the country's "poor performance" or lack of Olympic success occurred because it did not participate in the first place), the medal count is missing for these observations. That means the dataset is an unbalanced panel (Stock & Watson, 2015, p. 397). For some other variables, there are also missing values for some countries in some years because not all the collected data comes from balanced panels.

[25] Because the panel is unbalanced, as noted later, there are actually fewer than 1236 observations, but this description presents an easy way to understand the structure of the dataset.

5.3.1 Data collection and processing

5.3.1.1 The dependent variable: Olympic medal counts

Olympic medal data for all editions of the Games are available online on the official IOC Olympic website. However, since the Olympic website presents its medal data in a format that only allows for viewing one event in a given Olympic Games at a time (The International Olympic Committee, n.d.-l) or one country at a time (The International Olympic Committee, n.d.-f), collecting medal data from there is tedious. Fortunately, there are other sources online that gather and present data from the Olympics in a format that makes it much more easily extractable. The source chosen here is Wikipedia. Ordinarily, the use of Wikipedia in an academic context is frowned upon, but it has arguably the most current and comprehensive online collection of information about the Olympics. Criticisms of Wikipedia usually concern trustworthiness, but because Olympic medal data is factual in nature, not very open to interpretation or susceptible to biased presentation and does not require any specific expertise to report correctly, the dangers of the medal data being unreliable is minimal. Thus, the Olympic medal data has been collected from Wikipedia (n.d.-b, n.d.-d, n.d.-f, n.d.-h, n.d.-j, n.d.-l). Drug testing and doping cases mean that medals are sometimes stripped from disqualified athletes after the Olympics have ended (The International Olympic Committee, 2019b), so the medal tables on Wikipedia change over time to reflect these decisions. The Olympic medal data reported in Appendix A is current as of the 21th of July, 2020.

An important question to address with regards to the medal counts is how to measure the dependent variable. As mentioned in the literature review, Olympic performance has sometimes been defined by medal shares. Others have measured Olympic success according to a points system where gold, silver and bronze are assigned different values that weight them in order of placement. However, using the absolute total number of medals is preferred over these alternatives in the thesis, first of all because it preserves the dependent variable as count data, which will be relevant in the analysis' zero-inflated count model alternative, and second of all, because the absolute medal count has the most natural real world interpretation.[26]

From the theoretical framework, yet another way of measuring Olympic performance arises. If the economic growth model for the role of technology in Equation (2) is applied more literally to the Olympic

---

[26] Furthermore, Andreff and Andreff (2015) found no significant difference between using the absolute number of medals and the medal share as the dependent variable (p. 212), and Grimes et al. (1974) justified choosing the absolute number of medals won as the dependent variable rather than a points system by arguing that weights are arbitrary (p. 779).

context, the relevant dependent variable would be Olympic medal growth as a parallel to economic growth in the original model. This approach, however, poses two problems. Firstly, growth in a variable (when the change is small) is usually approximated in percentage terms as the difference between the logarithms of the variable in consecutive time periods (Stock & Watson, 2015, p. 317). However, when this variable is the number of Olympic medals, in the many cases of countries that win no medals it will have a value of zero, the logarithm of which is undefined.[27] Secondly, when growth (change) in the number of medals is made the dependent variable, reversion to the mean becomes a concern. After a particularly successful Olympic Games where a country wins more medals than it does on average, if the country performs worse at the following Olympics, this negative growth is not necessarily a reflection of the effects of the determinants but of reversion to the mean. This could lead the coefficients to be incorrectly estimated. For these reasons, the absolute number of medals won on the country level is used as the dependent variable in the thesis.

5.3.1.2 The regressors and the instrumental variable

GDP data is provided by the Penn World Table (PWT) version 9.1 (Feenstra et al., 2015, 2019). The measure of GDP used is the PWT's $RGDP^e$ variable, which uses prices that are constant both across countries and over time to measure real GDP (Feenstra et al., 2015, pp. 3153–3154). This allows for standard of living comparisons across countries and over time (Feenstra et al., 2015, p. 3154, 2019), which are necessary for the panel data analysis. Prices are made constant across countries by using purchasing power parity (PPP) exchange rates to express GDP in US dollars as a common currency (Feenstra et al., 2015, p. 3150). $RGDP^e$ in PWT 9.1 is expressed in 2011 US dollars (Feenstra et al., 2019).

$RGDP^e$ is not given in per capita terms, so the GDP per capita is calculated by dividing $RGDP^e$ by population. Because a natural way to think about changes in per capita GDP is in percentage terms, the per capita GDP regressor used is the (natural) logarithm of per capita GDP (Stock & Watson, 2015, p. 315). That is, the estimated linear relationship is such that a 1% change in a country's per capita GDP (regardless of the level) is associated with a fixed change in the number of medals won.

[27] This log(0) problem was also noted by Den Butter and van der Tak (1995, p. 29).

Because the effects of the standard of living, representing a country's available resources for Olympic technological progress, would not be expected to be immediately reflected in the number of medals a country wins at the Olympics, the log of GDP per capita is lagged by 4 years. This represents the resources (e.g. free time, household disposable income, training facilities, government support) that Olympic athletes have during their sports careers, particularly in the years leading up to the Olympics.[28] Thus, the 4-year lag of log of GDP per capita is used as the per capita GDP regressor.

With regards to sports investment, the IMF publishes data on government expenditures by function, one of which is expenditure on recreational and sporting services (The International Monetary Fund, 2020). Unfortunately, such data only exists for 38 countries for the 1996-2016 period, and the panel is quite unbalanced. The fixed effects regression analysis in the next chapter is therefore focused on a main model specification without the investment in sports regressor for the full dataset of 206 countries in the six Olympic Games between 1996 and 2016. A secondary regression with the investment in sports regressor included also appears in the analysis for the reduced sample of the 38 countries for the six Olympic years.[29] For the former model, there might be a concern for omitted variable bias, but if the propensity for a country to invest in sports is interpreted broadly as an expression of the cultural attitude towards and interest in sports, it is possible to regard it as being captured in the country fixed effects instead. For the latter model, the reduced sample is not ideal, but it could be useful for illustrating the effect of sports investment on the medal count for the specific sub-group of the population. It should be thought of as if, in some imagined world, it were the full population, with an interpretation of the coefficient estimates that does not apply to countries outside of it. Given that the number of countries in the reduced sample exceeds 30, large-sample normal approximation should apply (Stock & Watson, 2015, p. 98).

In the IMF dataset, the expenditure on recreations and sports is given in percent of GDP. To be able to make a fair comparison across countries (with different GDPs and population sizes) and over time, the dataset takes the calculated amount that this percentage constitutes of the $RGDP^e$ and uses this amount in per capita terms as the sports investment variable. Similarly to per capita GDP, investment in sports is lagged 4 years.

As for the remaining regressors and the instrumental variable, population data is available in the PWT (Feenstra et al., 2019). Olympic team size information, as with medal counts, is gathered from Wikipedia (n.d.-a, n.d.-c, n.d.-e, n.d.-g, n.d.-i, n.d.-k). The Olympic website has a list of host cities (The International

[28] Andreff and Andreff (2015) also used variables with a four-year lag (p. 188).
[29] See Appendix D for the list of countries in the reduced sample.

Olympic Committee, n.d.-q), and the information is included in the dataset as dummy variables taking the value of 1 for the host country and 0 otherwise. Likewise, pre- and post-host dummies take the value of 1 if the country hosts the following or hosted the previous Olympics respectively and 0 otherwise. Tourism data is collected from the World Bank (n.d.-b) with the chosen variable being the number of tourist arrivals in a country as a measure of its openness and geographical accessibility to the rest of the world. Data on the infant mortality rate, as defined earlier in the chapter, is also gathered from the World Bank (n.d.-c). As the instrument employed to account for the endogeneity of per capita GDP, infant mortality is also measured as the 4-year lag so that the year of the per capita GDP measure corresponds to the year of the infant mortality rate. Since the strong negative linear relationship (which is indicative of instrument relevance) seems to exist between the logs of the two variables (see Footnote 24), the infant mortality measure in the dataset is the 4-year lag of the log of the infant mortality rate.

5.3.2 Construction of the dataset

Details of how the full dataset was constructed from the data collected for all the individual variables are provided in Appendix E. As mentioned earlier, the first few rows of the dataset can be found in Appendix C to illustrate what it looks like.

To simplify the data processing and construction of the dataset, only current NOCs were included. That is, countries/NOCs from the earlier Olympics of the 1996-2016 period that no longer exist with the IOC were excluded (see Appendix E for more details).

6 Analysis

The analysis is divided into three parts. The first part considers descriptive statistics for the dataset in order to better understand it, particularly in relation to outliers and the data quality. The second part defines and estimates the fixed effects regression model and discusses the estimated model. The third part explains how a zero-inflated negative binomial regression could be applied as a count model alternative to the fixed effects model.

All data analysis is done using R version 4.0.2 (R Core Team, 2020), and the descriptive statistics and regression outputs are made using the R package stargazer (Hlavac, 2018). The corresponding R code for all outputs in the chapter (tables, figures, regression results) can be found across Appendix F (descriptive statistics), Appendix H and Appendix K (fixed effects regression model) and Appendix L (zero-inflated count model).

6.1 Descriptive statistics

Table 6.1 below is a summary of descriptive statistics for the variables in the dataset.[30]

Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Number of medals won	1,201	4.642	12.910	0.000	0.000	0.000	3.000	121.000
log(GDP per capita in millions 2011 USD), 1st lag (4 years)	1,056	8.896	1.250	5.408	7.897	8.955	9.959	11.888
Population (millions)	1,056	37.092	136.537	0.019	2.267	8.008	24.563	1,403.500
Team size	1,201	54.116	98.175	1.000	5.000	11.000	54.000	646.000
Number of tourists (millions)	1,079	4.705	10.588	0.001	0.210	0.895	3.877	82.682
log(infant mortality rate), 1st lag (4 years)	1,158	3.115	1.061	0.642	2.303	3.184	4.037	5.166

Table 6.1: Descriptive statistics

[30] The three host effect dummies are not included. The descriptive statistics analysis here only considers the variables appearing in the model specification for the full dataset of 206 countries, i.e. the one without the sports investment variable. The descriptive statistics and the conclusions drawn for the reduced sample of 38 countries are very similar to those for the full 206-country dataset, so it would be redundant to repeat them here. For completeness, the tables and figures pertaining to the descriptive statistics of the reduced sample are included in Appendix G.

The mean value in relation to the median and the 25th and 75th percentiles imply that some of variables are skewed, which could be an indication of outliers. For instance, the mean number of medals won is higher than the median of 0 and the 75th percentile of 3, meaning that most of the observations are low (i.e. most countries win no or few medals), but there are a few high medal counts that contribute to pulling the mean upwards. However, as has been described to be the case for Olympic medal data, it is characteristic that many countries win no medals while a few countries win many, so the skewness comes from the concentration of observations around zero, not necessarily from extreme positive outliers. This can also be seen in the scatterplot in Figure 6.1. The distribution of points along the y-axis illustrates that the largest observations for the number of medals won are not extreme, but relatively close in value to the rest of the data points.

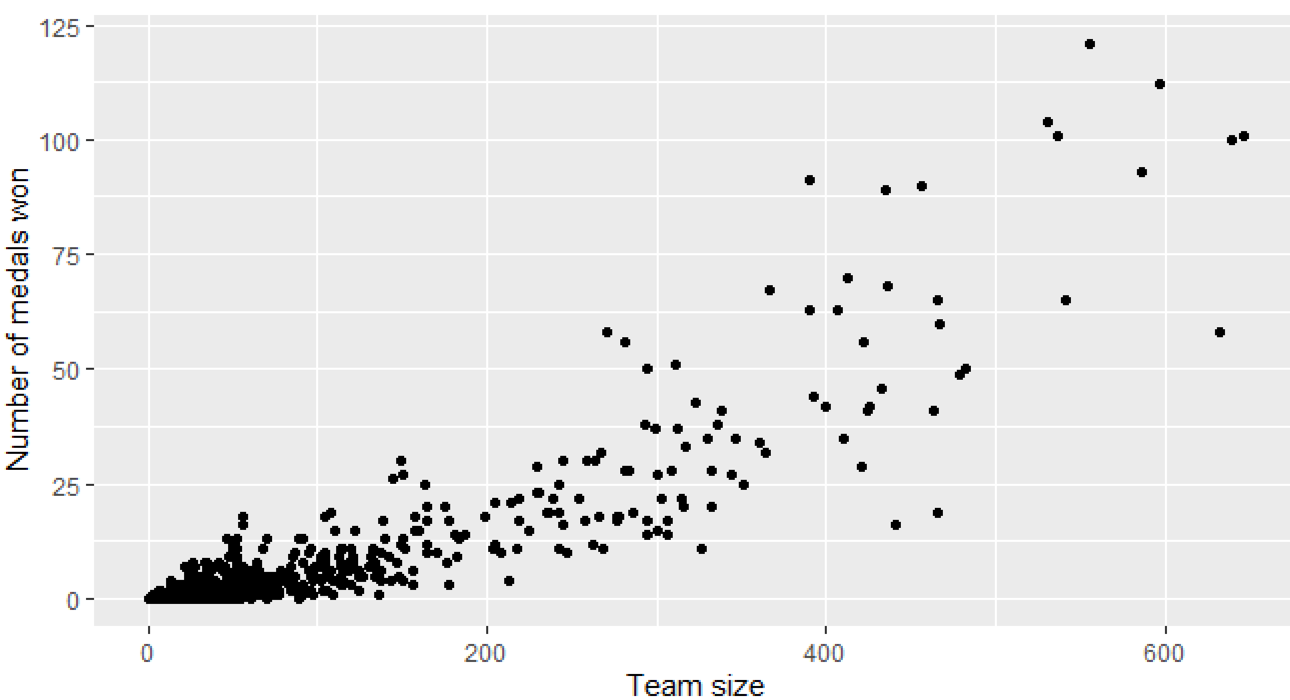

Figure 6.1: The number of medals won vs team size

From Table 6.1, similar observations can be made for population, team size and the number of tourist arrivals; the mean is higher than the median and 75th percentile, and the maximum value is much higher than the 75th percentile. That is, most observations are low in value (representing the small countries in the world with small populations, small pools of Olympic athletes and few tourists) with some high values skewing the mean, which could indicate a problem with outliers. As the distribution of points along the x-axis in Figure 6.1 above and Figure 6.2 below show, team size and the number of tourists, like the medal count, are skewed due to the presence of many small values, not by the presence of very positive outliers, as there are no extreme observations.

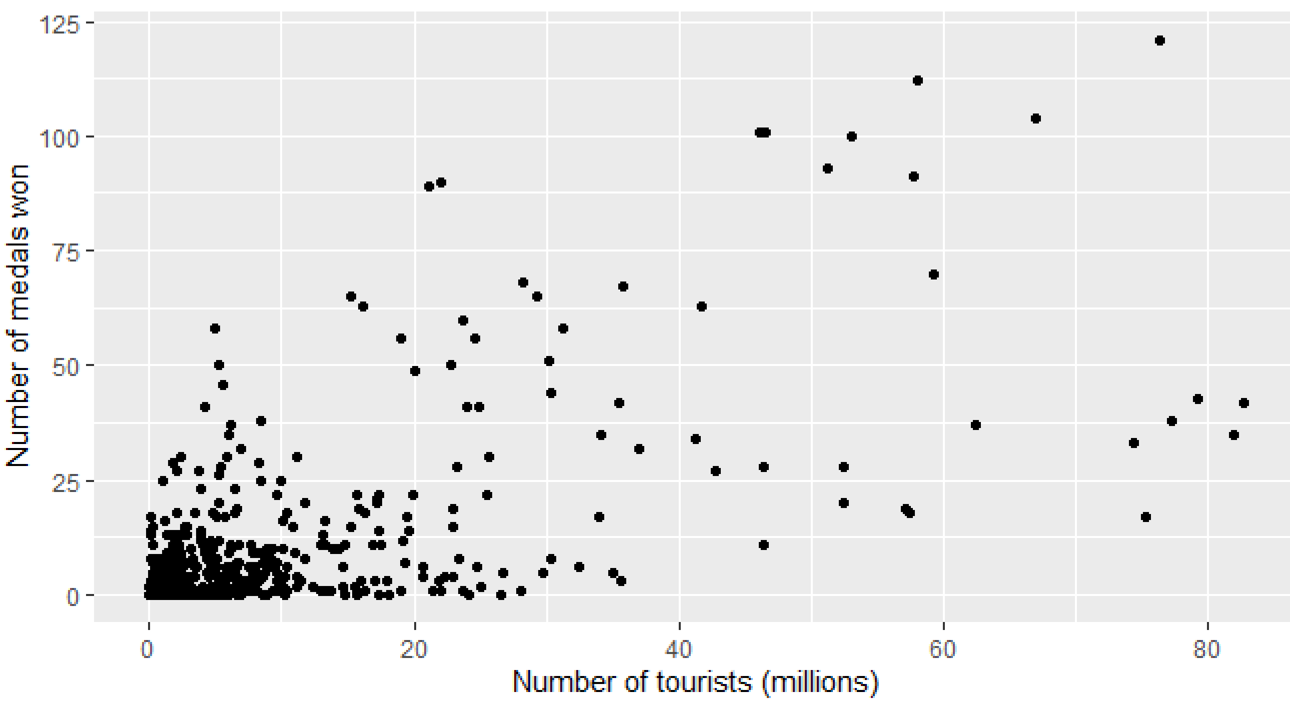

Figure 6.2: The number of medals won vs the number of tourists

For population, on the other hand, there are some positive outliers that are much larger in value than the rest of the data (see Figure 6.3). Specifically, these are the 12 observations for China and India (one each for each of the six Olympic Games from 1996 to 2016). As such, removing them would be misleading because those are the real population sizes, not outliers resulting from data being wrongly recorded or mistyped as an indication of poor data quality. Thus, even though these observations fall quite far from the remaining population sizes, they are kept in the dataset.

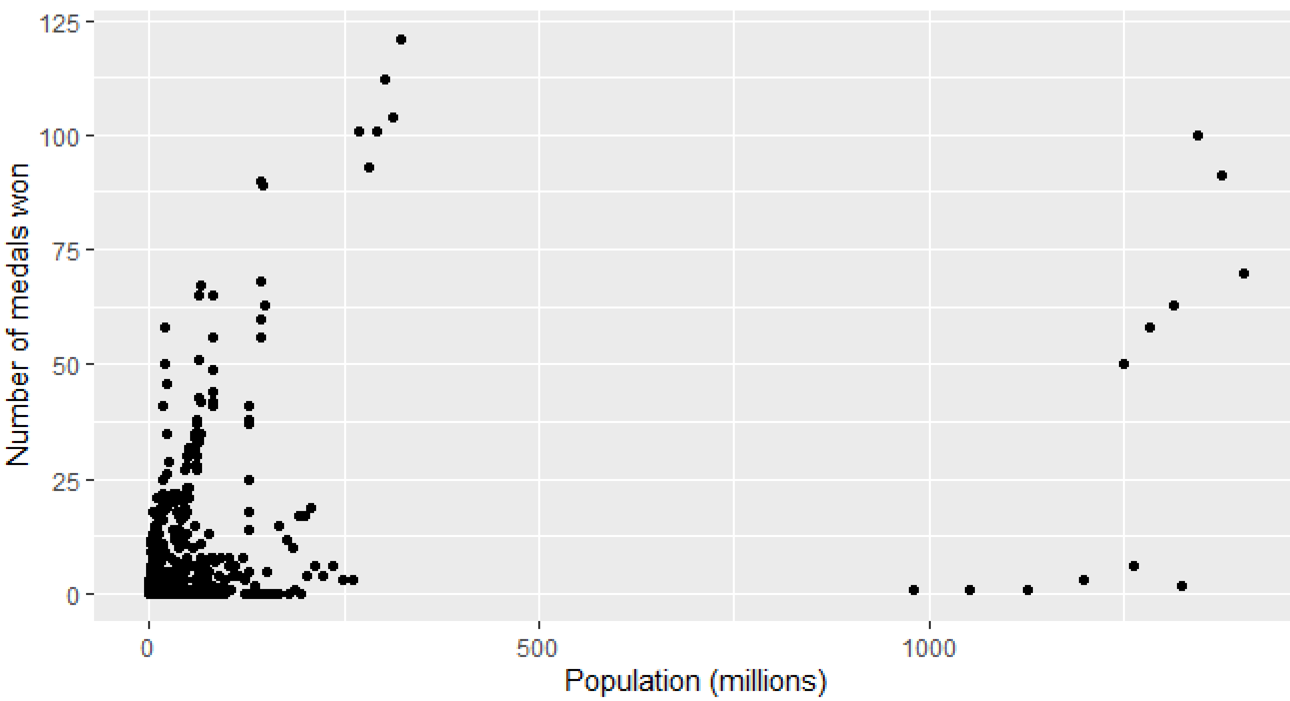

Figure 6.3: The number of medals won vs population

As for the remaining regressor (the lagged log of GDP per capita) and the instrumental variable (the lagged log of the infant mortality rate), Table 6.1 indicates that they are relatively more symmetric, i.e. have a mean closer to the median and a 25th and 75th percentile almost equidistant from the median. Figure 6.4 and Figure 6.5[31] below give no reason to believe that outliers are a concern.

Thus, the third fixed effects regression assumption discussed in the methodology chapter about large outliers being unlikely can be considered to hold.[32] There are arguably outliers in the population data, but given that they represent actual populations in the real world, it seems the best course of action is to leave them be.

[31] The lagged log of the infant mortality rate is plotted against the lagged log of GDP per capita instead of against the number of medals won (as the other variables were) because the purpose of infant mortality to the model is as an instrument to account for the endogeneity of per capita GDP. Since its validity as an instrument hinges on its lack of correlation with factors that determine the medal distribution, it would be counterintuitive to plot it against the medal count.

[32] A similar regression assumption about outliers being unlikely exists for instrumental variables regression (Stock & Watson, 2015, p. 486), which is why the instrument was also considered here.

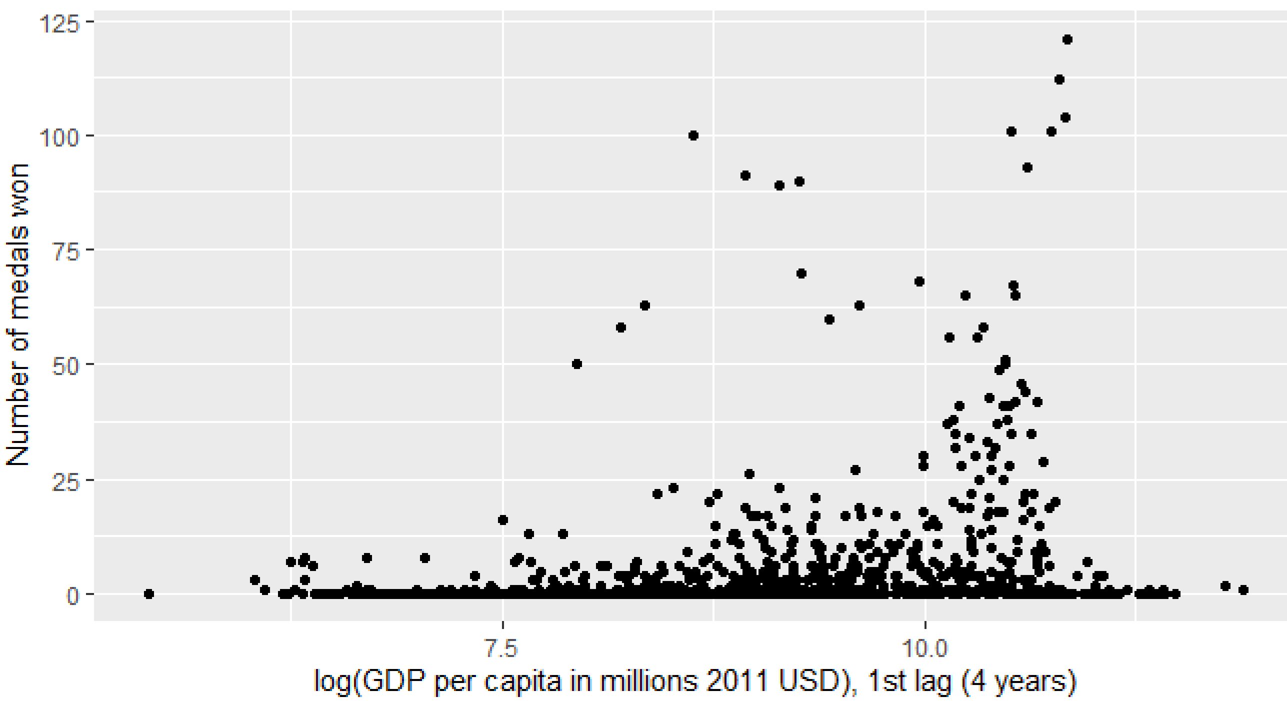

Figure 6.4: The number of medals won vs the lagged log of GDP per capita

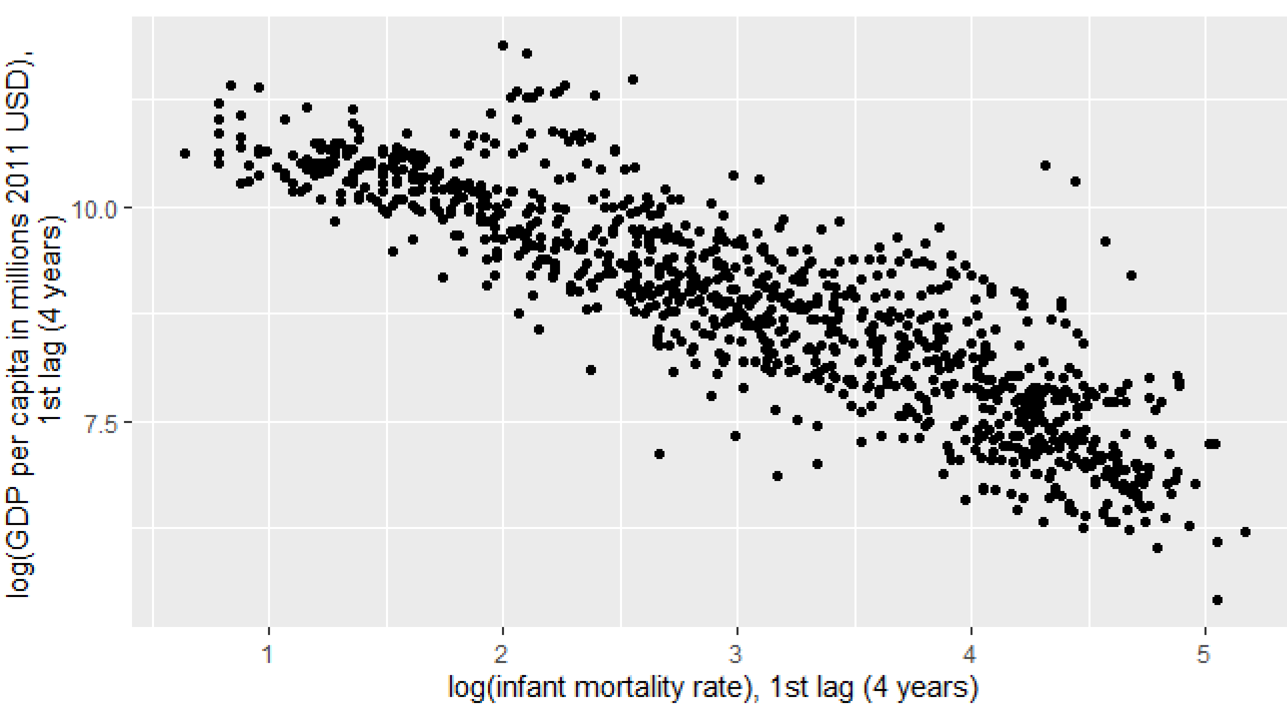

Figure 6.5: The lagged log of GDP per capita vs the lagged log of the infant mortality rate

The correlation matrix for the variables in Table 6.1 is included in Table 6.2 below.

	medals	**log(gdp), lag**	**pop**	**teamsize**	**tourism**	**log(infmort), lag**
medals	1.0000	0.2807	0.4211	0.8948	0.6935	-0.3174
log(gdp), lag	0.2807	1.0000	-0.0558	0.3917	0.3616	-0.8757
pop	0.4211	-0.0558	1.0000	0.3406	0.3362	0.0496
teamsize	0.8948	0.3917	0.3406	1.0000	0.7151	-0.4447
tourism	0.6935	0.3616	0.3362	0.7151	1.0000	-0.3811
log(infmort), lag	-0.3174	-0.8757	0.0496	-0.4447	-0.3811	1.0000

Table 6.2: Correlation matrix

Figure 6.5 can be used to visually suggest that the instrument relevance condition holds, as it illustrates that the instrument, the lagged log of the infant mortality rate, is strongly negatively correlated with the endogenous regressor, the lagged log of GDP per capita. This is also reflected in the correlation matrix in Table 6.2, which shows a strong negative correlation of -0.8757 between the two variables ("log(gdp), lag" and "log(infmort), lag"). Thus, there is support for the instrument being relevant. That infant mortality is not a weak instrument is shown later in the chapter using the F-statistic of the TSLS first-stage regression as described in the methodology chapter.

6.2 The fixed effects regression model

On the basis of the general fixed effects regression model expressed in Equation (3), the fixed effects regression model to be estimated here, using the plm package in R (Croissant & Millo, 2008), is the following:[33]

$$medals_{i,t} = \beta_1 \log(gdp)_{i,t-1} + \beta_2 pop_{i,t} + \beta_3 teamsize_{i,t} + \beta_4 host_{i,t} + \beta_5 posthost_{i,t} + \beta_6 prehost_{i,t} + \beta_7 tourism_{i,t} + \alpha_i + \lambda_t + u_{i,t} \quad (4)$$

i:	country $i = 1, \dots, 206$
t:	time period $t = 1, \dots, 6$ (representing the Olympic Games years; 1996, 2000, 2004, 2008, 2012, 2016)
$medals_{i,t}$:	the number of Olympic medals won by country i in time period t
$\log(gdp)_{i,t-1}$:	the logarithm of real GDP per capita of country i in time period $t-1$ (i.e. lagged 1 time period/four years), measured in millions of 2011 US dollars
$pop_{i,t}$:	the population size of country i in time period t, measured in millions
$teamsize_{i,t}$:	the Olympic team size of country i in time period t
$host_{i,t}$:	a dummy variable taking the value 1 if country i hosts the Olympic Games in time period t, and 0 otherwise
$posthost_{i,t}$:	a dummy variable taking the value 1 if, in time period t, country i hosted the previous Olympic Games (i.e. it was the host in time period $t-1$), and 0 otherwise
$prehost_{i,t}$:	a dummy variable taking the value 1 if, in time period t, country i hosts the next Olympic Games (i.e. it is the host in time period $t+1$), and 0 otherwise
$tourism_{i,t}$:	the number of tourist arrivals in country i in time period t, measured in millions
α_i:	country fixed effects for country i
λ_t:	time fixed effects for time period t
$u_{i,t}$:	the error term

[33] The reduced sample fixed effects model that includes the sports investment regressor is defined in Appendix I.

The lagged log of the infant mortality rate is used as an instrument and is defined as follows:

$\log(infmort)_{i,t-1}$:	the logarithm of the infant mortality rate for country i in time period $t-1$ (i.e. lagged 1 time period/four years), measured as the number of infants per 1,000 live births that die before reaching the age of one

To test the instrument relevance condition, the F-statistic of the TSLS first-stage regression is required. The first stage of TSLS regresses the lagged log of GDP per capita (the endogenous variable) on the exogenous variables, i.e. the instrument and the remaining regressors in the fixed effects regression (Stock & Watson, 2015, p. 483):

$$\log(gdp)_{i,t-1} = \pi_1 \log(infmort)_{i,t-1} + \pi_2 pop_{i,t} + \pi_3 teamsize_{i,t} + \pi_4 host_{i,t} + \pi_5 posthost_{i,t} + \pi_6 prehost_{i,t} + \pi_7 tourism_{i,t} + \alpha_i + \lambda_t + v_{i,t} \quad (5)$$

where $v_{i,t}$ is the error term. As in the fixed effects regression function, the TSLS first-stage regression above has no intercept because the country and time fixed effects, α_i and λ_t, represent sets of dummy variables where one country and one time period act as a baseline intercept.

Running the first-stage regression in R with HAC standard errors and testing the hypothesis that the coefficient on the instrument is zero gives an F-statistic of 25.615 (see Figure 6.6). As previously mentioned, the rule of thumb is that an F-statistic greater than 10 indicates that the instrument is not weak. Since the F-statistic here exceeds 10, there should be no concern that the infant mortality measure is a weak instrument.

```
Linear hypothesis test

Hypothesis:
loginfmortlag = 0

Model 1: restricted model
Model 2: loggdplag ~ loginfmortlag + pop + teamsize + host + posthost +
    prehost + tourism + ioccode + year

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    739
2    738  1 25.615 5.266e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6.6: TSLS first-stage regression F-statistic

With this, the fixed effects regression analysis can proceed.[34] Table 6.3 below shows the regression results for the estimated fixed effects model using (HAC) clustered standard errors and the lagged log of the infant mortality rate as an instrument.

[34] Usually, a Hausman test could be employed to test if the fixed or random effects model is most suitable (Croissant & Millo, 2008, p. 22). However, instrumental variables are not implemented for two-way (country and time) random effects models in the plm package in R, so it is not possible in this case. The fixed effects model should, however, be the appropriate model choice here because, as previously mentioned, the country-specific effects are likely correlated with the regressors. Furthermore, the consequences of choosing a fixed over a random effects model when the latter is more appropriate are also less severe than for the opposite case; estimating a random effects model and leaving the country effects in the error term under the incorrect assumption that they are uncorrelated with the regressors would induce endogeneity. Thus, a fixed effects model is arguably the appropriate choice over a random effects model here.

	Dependent variable:	
	medals	
	(1)	(2)
log(gdp), lag	-3.674	4.963
	(2.305)	(9.505)
pop	0.025	1.956
	(0.023)	(1.226)
teamsize	0.063***	0.045***
	(0.011)	(0.015)
host	3.188	10.826*
	(6.094)	(5.590)
posthost	8.023	8.878
	(6.723)	(9.975)
prehost	5.572	7.593
	(3.466)	(5.299)
tourism	0.199*	0.096
	(0.114)	(0.221)
sportsinv, lag		-0.007
		(0.020)
Observations	918	135
R^2	0.278	0.484
Adjusted R^2	0.103	0.186
F Statistic	314.827***	83.622***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 6.3: Fixed effects regression results using the lagged log of infant mortality as an instrument

Regression 1 in Table 6.3 estimates the full sample model, while regression 2 is for the reduced sample that includes the lag of sports investment as a regressor.

As the regression results show, all regressors except one have an estimated coefficient that is not statistically significantly different from zero at the 5% level (significance at the 5% level is indicated by at least two asterisks, as noted at the bottom of Table 6.3). In other words, for each of these explanatory variables, the null hypothesis that they have no effect on the medal count fails to be rejected. Especially with regards to the log of GDP per capita, the population and the host effect, this is in contrast with previous findings in the literature where these variables have been important factors in explaining the disparity in countries' Olympic success. An explanation for this could be omitted variables bias in previous studies stemming from unobserved country characteristics pertaining to the sports culture which is unchanging or slow-changing over time. As described in the methodology chapter, these would be captured in the model's country fixed effects. Meanwhile, the general increase in the number of medals won that is associated with more events (and medals) being introduced in the Olympics over time (i.e. which is the same for all countries but varies over time) would be incorporated in the model's time fixed effects. The significance of these country and time fixed effects in the model can be tested using either a Lagrange multiplier test or an F-test, employed with the "plmtest" and "pFtest" functions respectively in R's plm package (Croissant & Millo, 2008, pp. 21–22). Figure 6.7 and Figure 6.8 below are the respective results of these two tests for regression 1 (the full sample model).

```
        Lagrange Multiplier Test - two-ways effects (Honda) for unbalanced panels

data:  medals ~ loggdplag + pop + teamsize + host + posthost + prehost +  ...
normal = 24.105, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Figure 6.7: Lagrange multiplier test for country and time fixed effects

```
        F test for twoways effects

data:  medals ~ loggdplag + pop + teamsize + host + posthost + prehost +  ...
F = 15.341, df1 = 172, df2 = 738, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Figure 6.8: F-test for country and time fixed effects

As the figures show, the conclusions drawn from these two tests are the same; because the p-value is very small ($p < 2.2 * 10^{-16}$), the null hypothesis that there are no significant effects is rejected at the 5% significance level in favor of the alternative hypothesis that there are significant country and time fixed effects. This is also the case for the reduced sample in regression 2 (see Appendix I).

Appendix J shows the country fixed effects for regression 1. Although they are not tested for statistically significant differences between countries, the levels of the country fixed effects estimates suggest that there are indeed differences between countries. The two highest estimates belong to Russia and the USA, which have historically been among the most winning countries at the Olympics, while the lowest belongs to India, which has historically won only a few medals (see Appendix A for medal counts). Thus, the country fixed effects suggest that differences exist between countries in terms of country-specific, time-invariant factors that are associated with winning Olympic medals. Appendix J similarly shows the time fixed effects for regression 1, however, the differences in the estimates between Olympic years are relatively small, especially considering the standard errors. It is likely, then, that the significance of the two-way (i.e. both country and time) fixed effects in the model seen in Figure 6.7 and Figure 6.8 are related more to the country than the time fixed effects, although they cannot be separated out as such because the tests concern their combined effect. The suspected importance of country effects over time effects is supported by running the Lagrange multiplier test and F-test for the model as if it (hypothetically) only included time effects. This gives very large p-values (see Figure 6.9 and Figure 6.10), leading to a failure to reject the null hypothesis that there are no significant time fixed effects. In contrast, when running a hypothetical model with only country effects instead, the null hypothesis that there are no significant country fixed effects is rejected at the 5% level (see the end of Appendix H). Thus, it could be suggested that country fixed effects help explain Olympic success, while time fixed effects do not. If it is indeed the case that time effects are not significant, this implies that although the number of medals awarded at the Olympics has increased over time, there is not a tendency for countries in general to win more medals because of it. Instead, the addition of more events and medals would benefit the medal counts of only certain countries.

```
        Lagrange Multiplier Test - time effects (Honda) for unbalanced panels

data:  medals ~ loggdplag + pop + teamsize + host + posthost + prehost +  ...
normal = -1.2998, p-value = 0.9032
alternative hypothesis: significant effects
```

Figure 6.9: Lagrange multiplier test for hypothetical model with time fixed effects only

```
       F test for time effects

data:  medals ~ loggdplag + pop + teamsize + host + posthost + prehost +  ...
F = 0.095945, df1 = 5, df2 = 905, p-value = 0.9928
alternative hypothesis: significant effects
```

Figure 6.10: F-test for hypothetical model with time fixed effects only

For the estimated fixed effects regression, the only statistically significant coefficient estimate at the 5% level is for Olympic team size. According to the estimated model for the full sample in Table 6.3, the sign of the estimated coefficient is positive, meaning that a larger Olympic team size is related to a higher number of medals won, as expected from the theoretical framework chapter. This is also suggested in the descriptive statistics analysis by the scatterplot in Figure 6.1 and the highly positive correlation in Table 6.2. More specifically, an increase in team size by one athlete is associated with an expected increase in medals by 0.063 (95% confidence interval [0.041, 0.085]), or, equivalently, for an increase in a country's team size by 100 athletes, the number of medals won is expected to increase by 6.3 (95% confidence interval [4.1, 8.5]).

Similarly, for the reduced sample model, only the estimated coefficient on team size is statistically significant at the 5% level, such that an increase in team size by one athlete is associated with an increase of 0.045 medals won (95% confidence interval [0.016, 0.074]). It is also interesting to note for the reduced sample that the coefficient on the lag of sports investment is estimated to be -0.007 with a standard error of 0.020. As the 95% confidence interval of [-0.046, 0.032] also conveys, the null hypothesis that the coefficient is not statistically significantly different from zero fails to be rejected at the 5% level. That is, there is no evidence that a country's (lagged) investment in sports is associated with the number of Olympic medals it wins. As noted in the methodology chapter, because of the limited sports investment data and, as a consequence, the reduced sample, caution should be taken with the generalizability of this finding. The estimated relationship between sports investment and Olympic medal count may not hold in the wider population of countries competing at the Olympic Games. However, for the reduced sample, in an isolated sense, it suggests that for this specific subset of the NOCs, the effect of sports investment on the number of medals won is not significantly different from zero.

The idea that Olympics team size is important for explaining the Olympic medal count is interesting because understanding those two variables essentially pose similar problems. A country that succeeds at winning medals at the Olympics must necessarily have succeeded in qualifying its athletes for the Games first, which raises the question: what makes some countries more successful at qualifying for the Olympics than others? Because the qualification of athletes for the Olympics is a necessary step that precedes any medals a country wins, understanding the determinants of team size is useful for understanding the Olympic medal distribution, particularly given the regression results in Table 6.3, which quantify and support the existence of a statistically significant relationship between team size and medal count. If the earlier economic growth framework is applied differently, Olympic team size, rather than medal count, could also be thought of as the dependent variable whose determinants are the resources invested into improving a country's Olympic technology, represented by the proposed regressors. That is, the more resources that are available for Olympic technological progress in a country (per capita GDP, population size, host, post-host, pre-host effects, tourism as a proxy for openness), the more national athletes should be able to qualify for the Olympics, i.e. the bigger the national Olympic team. In turn, the bigger the national Olympic team, the more medals a country should be able to win, as suggested by the statistically significant effect of team size on the medal count.

Table 6.4 shows the estimated regression of Olympic team size against the proposed regressors for the medal count regression (except team size itself), i.e. it estimates the same model as Equation (4) except with team size as the dependent variable instead of medals and without team size as an explanatory variable. The details of the regression can be found in Appendix K. As in Table 6.3, regression 1 is the estimated model for the full sample, while regression 2 is for the reduced sample.

	Dependent variable:	
	teamsize	
	(1)	(2)
log(gdp), lag	-5.746	31.379
	(9.178)	(46.084)
pop	0.279**	5.221
	(0.112)	(5.128)
host	233.005***	248.196***
	(28.662)	(14.782)
posthost	40.916***	52.937***
	(8.138)	(16.775)
prehost	21.062**	28.243**
	(9.900)	(13.915)
tourism	0.532	-0.993
	(0.413)	(0.741)
sportsinv, lag		0.029
		(0.094)
Observations	918	135
R^2	0.574	0.593
Adjusted R^2	0.471	0.366
F Statistic	996.748***	131.516***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 6.4: Fixed effects regression results with team size as dependent variable using the lagged log of infant mortality as an instrument

As the results for regression 1 show, the coefficient estimates for the lagged log of GDP per capita and tourism remain statistically insignificantly different from zero at the 5% level. However, population and the host effects are all statistically significant at the 5% level for team size as the dependent variable, unlike in the regression of Olympic medals. The effect of population on team size is estimated such that an increase in population size of one million is associated with an increase in team size of 0.279 Olympic athletes (95% confidence interval [0.059, 0.499]). That is, the bigger the population of a country is, the more Olympic technology it will have in the form of naturally talented athletes, and, consequently, the more athletes will qualify and represent it at the Olympics. As for the host of the Olympics, the host country's team is expected to have about 233 athletes more on average than non-hosts (95% confidence interval [177, 289]). The country that hosted the previous Olympics has a team size that is approximately 41 athletes bigger in expectation (95% confidence interval [25, 57]), and the country that will host the following Olympics has a team size that is approximately 21 athletes bigger in expectation (95% confidence interval [2, 40]). Within the theoretical framework, these host effects reflect that the country that is, has been or will be host has an increased investment in Olympic technological progress in connection with hosting the Games, e.g. new sports venues and facilities. This is then associated with an increased number of athletes benefitting from these additional resources and succeeding in qualifying for the Olympics both before, during and after the hosted Games. Furthermore, in the year that the country hosts the Olympics, it has the advantage of being able to automatically qualify for a number events, which the large coefficient estimate on the host dummy should be a reflection of.

For the reduced sample, these host effects are similar, while population, the lagged log of GDP per capita and tourism are not statistically significant. It is interesting to note that investment in sports is again not statistically significantly different from zero at the 5% level. This suggests that, for this limited sample, not only does sports investment have no effect on a country's medal count, it also does not seem to have any effect on its team size. Any increased investment in sports that a country makes would not be expected to increase the number of its national athletes that successfully qualify for the Olympics.

As in the original regression with medals as the dependent variable, a Lagrange multiplier test or F-test shows that there are statistically significant country and time fixed effects (see Appendix K). Because the medal count regression analysis suggested that the significance comes from country and not time effects for the number of medals won, it could be suspected to be the case here for team size as well. Unlike before, however, running the tests for regression 1 as if there were only time fixed effects in the model shows that they are statistically significant at the 5% level (see Appendix K). Similarly, for a hypothetical model with country fixed effects only, the null hypothesis that there are no country effects is also rejected (see Appendix K). Thus, there seem to be country-specific, time-invariant factors that contribute to some countries having more athletes that qualify for the Olympics than others, as well as time effects that make Olympic team sizes bigger over time in general. As an extension of the argument made for time fixed effects in the original medal count regression, this latter effect could be explained by more events being added over time leading to more quota places needing to be filled and thereby an increasing number of athletes participating in the Olympic Games over time.

This additional regression with team size as the dependent variable is useful for better understanding the medal distribution. Although the population and the three host effects are not estimated to have a statistically significant effect directly on the number of medals a country wins, their coefficient estimates are statistically significant in the regression with team size as the dependent variable. Since team size has a statistically significant effect on a country's medal count, it could be argued then that population size and host effects are indirectly related to Olympic success through their effect on team size. The log of GDP per capita and tourism, however, remain statistically insignificant in the team size regression, as does sports investment in the reduced sample model. In other words, an improved standard of living, increased openness or additional investment in sports is not expected to increase the number of medals a country wins, neither directly nor indirectly through increasing team size.

Thus, in summary of this part of the analysis, the fixed effects model regressing medals against the chosen set of regressors shows that only the coefficient estimate for team size is statistically significant at the 5% level. Specifically, an increase in team size by 100 athletes is associated with an expected increase in medals won by 6.3. Meanwhile, it cannot be rejected that the log of GDP per capita, population, the three host dummies and tourism have no direct effect on the number of medals a country wins. This is especially surprising for the log of GDP per capita, population and the host effect given their prevalence in the literature, where they have repeatedly been found to be important determinants of Olympic success. An explanation for this could be the assumptions and model specification choices made in the thesis and the way they differ from those in the existing literature (e.g. the inclusion of country fixed effects to prevent omitted variable bias). This will be explored in more depth in the discussion in the next chapter.

To better understand what determines team size, given that it is an important determinant of the medal distribution, a fixed effects regression with team size as the dependent variable shows that population and the host effects are statistically significant. That is, although they may not be directly associated with the number of medals won, they have an effect on Olympic team size, which in turn has an effect on the medal count. The log of GDP per capita and tourism, however, seem to have no statistically significant effect on team size either.

There are significant time fixed effects determining a country's team size, however, time effects do not seem to be significant for the medal distribution. In other words, while countries in general may increase their team size over time due to the addition of more events and thus more quota places for athletes, this time effect does not seem to translate to a similar general increase in medals won over time across all countries. As for country-specific, time-invariant effects, they are statistically significant in determining both Olympic team size and medal count. As such, it implies that much of what determines national Olympic success is explained by unobserved general differences between countries and their (sports) cultures.

Lastly, regressions for the reduced sample show that investment in sports, whether as a regressor in the medal count or team size regression, has a coefficient estimate that is not statistically significant. The limited availability of sports investment data means that any conclusions drawn from the reduced sample should not be treated as representative of the effect that sports investment would have in the full sample. However, the regression results for the reduced sample tentatively suggest that investing more in sports nationally is not associated with a higher number of athletes qualifying for the Olympics or a higher number of medals won.

6.3 The zero-inflated count model

In this last part of the analysis, a zero-inflated count model is explored as an alternative to the fixed effects model to estimate a regression for the Olympic medal distribution and its determinants.

As explained in the methodology chapter, the dataset used in the thesis is panel data with a dependent variable (the number of Olympic medals won) that is characterized by being count data. The fixed effects regression model was chosen as the main model of the thesis to address the panel aspect of the dataset, while also being able to be used to analyze count data (Stock & Watson, 2015, p. 469). The purpose here is to suggest an alternative regression model in which the count data characteristics of the dataset are central. Count models employed for analyzing count data are typically Poisson or negative binomial regression models and variants thereof (Stock & Watson, 2015, p. 469; Zeileis et al., 2008, p. 2).

As a starting point for choosing a count model, the simplest and most common model, the Poisson model (Zeileis et al., 2008, p. 5), is considered. The Poisson model assumes that the mean and the variance are equal, however, this often does not hold in practice because many datasets tend to have variances greater than the mean, i.e. they are over-dispersed. By running a test for over-dispersion on a Poisson regression for the medal count and its regressors, the null hypothesis that dispersion is equal to 1 (i.e. that mean and variance are identical) is rejected at the 5% level. Dispersion is estimated to be 4.055, indicating over-dispersed data (see Figure 6.11). Instead of a Poisson model, a solution is to employ a negative binomial model, which is used in cases of over-dispersion (Zeileis et al., 2008, p. 5).

```
        Overdispersion test

data:  medals_poisson
z = 10.46, p-value < 2.2e-16
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
 4.05514
```

Figure 6.11: Over-dispersion test for Poisson model

An issue that often arises with count data is the presence of excess zeros (Zeileis et al., 2008, p. 1). As mentioned in the methodology chapter when describing the dataset, depending on how the Olympic medal data is recorded, the medal count may be purposely missing for the panel observations (country-year combinations) in which the country did not participate in those Olympic Games (as was done in the construction of the dataset for the fixed effects regression).[35] Otherwise, if it is not possible to distinguish between the absence of a country from the medal tables due to lack of success and due to non-participation, the observations for countries in years in which they did not participate in the Olympics may have a zero recorded for their medal count. In such a case, a zero-inflated count model would be appropriate for modelling the data. The zero-inflated model accounts for two different kinds of zeros and models the two different processes from which they may be generated (Zeileis et al., 2008, p. 7). One part of the model identifies that some observations always take a value of zero, i.e. they are zero, not because their count happens to be zero and they could have potentially taken on another counting value, but because it is impossible for them to be anything but zero. The other part models the count, given that it is not one of these definite zeros, using a Poisson or a negative binomial model. In the Olympic medal context, the zero-inflated model is able to distinguish between medal counts of zero that arise from non-participation (i.e. which are always zero) and those that arise from countries that do participate but are simply not successful at winning any medals. As the over-dispersion test above shows, the negative binomial model would be preferable over the Poisson model to estimate the regression for the count data that is not always zero. Thus, the zero-inflated negative binomial (ZINB) regression model could be used for modelling Olympic medals as an alternative to the fixed effects model employed in the main analysis.

The "zeroinfl" function in R's pscl package allows for the estimation of zero-inflated count models (Zeileis et al., 2008, p. 8). Unfortunately, it does not support instrumental variables, nor is there a natural way to accommodate panel data. For the instrument, it would be possible to manually estimate a TSLS first-stage regression (as was done in the fixed effects model for Equation (5) to compute the F-statistic in Figure 6.6) and use the first-stage predicted values of the endogenous variable, the lagged log of GDP per capita, as a regressor in the ZINB model instead of the observed values. However, because the ZINB model would not be able to recognize that those predicted values were estimated in a separate regression, the reported standard errors would be wrong. Similarly, it would in theory be possible to manually include country and time fixed effects as dummy variables in the specified ZINB regression. However, because

[35] This is possible when the list of participating countries at each Olympics is known and readily available in a format that makes data extraction easy.

of the large number of country dummies, it seems “zeroinfl” in R is unable to estimate so many coefficients for the data provided. Estimating the ZINB model with country and time dummies included as regressors produces an error in R, which disappears when the country dummy regressors are removed from the regression. Thus, the ZINB model estimated below has neither an instrumental variable nor fixed effects to prevent omitted variable bias, so it should be kept in mind that the coefficient estimates may be biased. This section is useful nevertheless, since the focus is on the model and its application to the data, rather than the exact estimation of the coefficients. The purpose of discussing the zero-inflated count model here is simply to suggest an alternative approach to modelling Olympic success. It is not meant to be a thorough analysis, but an explanation of how a ZINB model could be implemented for the Olympic dataset.

Table 6.5 and Table 6.6 below show the estimated regression for the ZINB model.[36]

[36] Compared to the dataset used for the fixed effects regression, in this dataset, the previously missing medal count values for countries that did not participate in a particular year have been filled with zeros. When data about participation exists, it may seem counterintuitive and backwards to edit and treat the data as if no distinction can be made between zeros stemming from non-participation and lack of success in winning. However, unless an effort is made to create a list of participating countries for each Olympic Games, the medal tables themselves do not actually allow for such a distinction to be made. Specifically, medal tables only include countries that have won at least one medal, which means that either information about non-medal-winning participating countries needs to come from elsewhere, or all non-medal-winning countries (participating or not) have to be included with medal counts of zero. For the dataset used in the fixed effects regression, this participation data was sought out, but it would have been simpler to just collect medal data on the countries that have won at least one medal and then assigning the remaining countries of the world zero medals without regards to participation. Thus, having a regression model that can naturally distinguish between the two different cases of zeros without it being specified in the data eases the data collection and processing burden.

	Dependent variable:
	medals
log(gdp), lag	0.144***
	(0.046)
pop	0.001
	(0.0003)
teamsize	0.015***
	(0.001)
host	-3.822***
	(0.569)
posthost	-1.214**
	(0.514)
prehost	-0.356
	(0.495)
tourism	-0.010
	(0.006)
Constant	-1.662***
	(0.406)
Observations	953
Log Likelihood	-1,729.300
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6.5: ZINB model regression results, count component

	Dependent variable:
	medals
participation	-20.718
	(70.410)
Constant	8.273
	(40.539)
Observations	953
Log Likelihood	-1,729.300
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6.6: ZINB model regression results, zero component

Table 6.6 shows the modelling of the zeros that do not arise from counting and cannot take any other value than zero. Since the dataset contains information about participation for each observation (i.e. whether a given country in a given Olympic Games year participated, represented by a dummy), it is already known what determines a non-count zero, so this variable is used to predict the occurrence of definite zeros. However, in cases where participation is unknown (see Footnote 36), other variables can be used to model the zeros stemming from non-participation.

Table 6.5 shows the negative binomial model estimation of the count data, which is given in a log-linear specification (UCLA, n.d.; Zeileis et al., 2008, p. 2), i.e. the coefficient estimates indicate the relationship between the regressors and the log of the expected medal count. An approximation can be made to interpret the coefficient estimates (multiplied by 100) as the percentage change in the medal count associated with a unit change in the regressor (Stock & Watson, 2015, pp. 318–319). The exception is the lagged GDP per capita which itself is a logarithm, in which case the coefficient estimate is interpreted as an elasticity (Stock & Watson, 2015, p. 320). Thus, the lagged log of GDP per capita is statistically significant at the 5% level, and a 1% increase is associated with a 0.144% increase in the predicted number of medals. The estimated coefficient on team size is also statistically significant, and a one unit increase is approximately associated with a 1.5% increase in expected medals. The coefficient estimates on the host and post-host effects are statistically significant, however, their interpretation requires calculation (exponentiation) since they are dummy variables where a one unit increase does not carry the same meaning. As aforementioned, the exact estimated effects are not what is important here, so it will simply be noted that they are both estimated to be negative, implying that being host or having been host for the previous Games is associated with a decrease in expected medals. This is surprising, but for the reasons explained earlier, these estimates should not be trusted too readily.

Thus, this part of the thesis has shown how a zero-inflated negative binomial model can be applied to the Olympic medal dataset. Due to concerns of omitted variable bias, caution should be taken with the estimated coefficients in the regression results produced here, but the ZINB model is nonetheless useful for modelling Olympic success when there should be a distinction in the dataset between zeros generated from non-participation and from a lack of Olympic medals won.

7 Discussion

As noted in the analysis, the fixed effects regression results are quite dissimilar to findings in the literature. Most notably, per capita GDP, population and the host effect were found by previous studies to be important determinants of Olympic success, but the fixed effects regression analysis in this thesis has suggested that GDP per capita has no statistically significant effect on Olympic medals won. Meanwhile, population and host effects only indirectly influence national Olympic success through team size, which has the only statistically significant coefficient estimate in the Olympic medal regression.

There are multiple reasons related to differences in data and econometric methodology that could help explain this difference in findings. First of all, the Olympic Games considered in the thesis are those from 1996 to 2016, while existing studies, owing to the time at which they were written, have used older data. It could be that the relationship between the regressors and the medal count has changed over time. Second of all, the chosen econometric model differs from most of those previously used. OLS methods for the fixed effects model have been employed here as opposed to the often used tobit model, which means that the coefficients are estimated differently. Third of all, team size was not included in the older literature but is used as a regressor in the thesis. As explained in the methodology chapter, not including team size would induce omitted variable bias because of its correlation with population and the host dummy. In other words, when team size is omitted, the coefficient estimates on population and host effect in the medal regression would capture the effect of team size on medals. Thus, when previous studies that did not include team size found that population and host were statistically significant for medals, it could be that it only appeared as if they were significant while, in fact, this was an expression of their effect on team size, which in turn has an effect on medals. That is, the omitted effect of team size was passed off as if it came from population and the host effect.

Lastly, the thesis has taken advantage of the panel characteristics of the dataset by using country fixed effects to eliminate omitted variable bias stemming from country-specific factors. In contrast, omitted variable bias may have been present in previous studies of the Olympic medal distribution that used cross-sectional data or panels with no fixed effects. As discussed in the methodology chapter, an example could be gender equality, an unobserved determinant of Olympic success correlated with per capita GDP (e.g. more people/women contributing to output/income for a given population size). Not including it or accounting for it in previous studies would have induced omitted variable bias. As a result, the statistical significance of the estimated coefficient on per capita GDP in those studies may have been a reflection of the combined effect of both GDP per capita and gender equality in society (and, by extension, sports culture) rather than the effect of GDP per capita alone. Thus, omitted variable bias in the literature from country-specific time-invariant effects could also be an explanation for the difference in findings, for instance when it comes to the significance of GDP per capita.

Apart from this discussion of the discrepancy between the findings here and in the literature, it would also be relevant in this chapter to consider the wider implications of the fixed effects regression results. Specifically, this concerns the importance of country-specific effects for determining the medal distribution and the notion suggested in the reduced sample regression that investment in sports has no statistically significant effect on the number of medals won. Because the medal count has been found to depend on country effects and team size, which in itself also depends on country effects, it could be argued that individual country performance at the Olympics is, to a great extent, determined by factors that are resistant to change. That is, an important explanation for why some countries perform better than others is found in the individual country characteristics that change very slowly over time. Consequently, this suggests that there may be little countries can do from one Olympics to the next to increase their medal count, since much of what determines Olympic success is ingrained in the national (sports) culture and the changes required to become more successful do not happen over a small number of years. A successful country is successful because of persistent country-specific characteristics that facilitate its success. Furthermore, the estimated model for the reduced sample tentatively suggests that the level of investment in sports for the countries in that sample does not seem to be associated with the number of Olympic medals that a country wins. Although this does not necessarily hold for countries in general, for the limited number of countries considered it reinforces the idea that what determines a country's Olympic success is largely outside of the country's active and immediate control. For policymakers, the implication of this could be that aiming to increase the number of Olympic medals becomes a less convincing argument for increasing spending on sports.

To conclude this chapter, it would be useful to evaluate some of the choices made in the thesis and discuss their strengths and limitations and what potential improvements, if any, could be made. The zero-inflated count model as an alternative to the fixed effects model was already suggested and discussed in the previous chapter and will not be considered further.

Firstly, the time period considered for the thesis has been 1996-2016. More so than actively choosing these years, this choice was forced by a lack of data, not for the Olympic medal counts which can be found for the entire 1896-2016 period of historical Olympic Games, but for variables of interest to serve as regressors.[37] Because global measures of different variables from before the 1990s are few, the time period was limited to 1996-2016, but that also means that potentially useful information contained in older Olympic medal data has been discarded. However, there is a tradeoff, as there are also considerable advantages to studying only the most recent Olympics. When analyzing older medal data, several data processing concerns arise. This includes the need to account for boycotts and decide how to treat medal data from successful NOCs that are no longer in existence (e.g. the Soviet Union, East and West Germany) and those that have replaced them.[38] It could also be argued that findings from more recent Games are more representative of the determinants of the Olympic medal distribution at the present time, and fixed effects are arguably more suited to prevent omitted variable bias because culture is less likely to change over the shorter term (e.g. 1996-2016 compared to 1896-2016).

Secondly, the analysis has included a regression with sports investment data for a reduced sample. Unfortunately, the findings are not readily generalizable, and it would have been preferable to obtain this data for all of the countries that participate in the Olympics and not only a subset. However, due to limited data availability, this was not possible. The reduced sample regression was nevertheless included because it has been useful for at least suggesting what the relationship between sports investment and medals may be for those limited countries. As collection of all kinds of data has only become more widespread over time, it is possible that sports investment for more countries will eventually become available.

Thirdly, for the choice of infant mortality as an instrument, there is no way to test if the instrument exogeneity condition holds because the model is exactly identified (one instrument and one endogenous regressor, GDP per capita). An improvement to the analysis would be the inclusion of more than one instrument in order to run a test of overidentifying restrictions (Stock & Watson, 2015, pp. 493–494).[39]

Lastly, the choice of the fixed effects model has allowed for the inclusion of country-specific factors, but a lot of different information is represented collectively by the country fixed effects. To find that country effects are important for determining the Olympic medal distribution is insightful, but cannot help to understand exactly which fixed effects impact the medal count and by how much (e.g. gender equality,

[37] Bernard and Busse (2004, p. 414) remarked that the lack of data made them consider only GDP and population.
[38] Andreff and Andreff (2015) noted these changes in the world's countries as the source of their panel data being unbalanced (p. 212).
[39] However, the test would need to allow for heteroskedasticity and autocorrelation. The test described by Stock and Watson (2015) only applies in cases of homoskedastic errors (p. 493).

specialization in specific sports, physical geography). The findings could therefore be considered too broad or vague, but those are the terms for using fixed effects.

8 Conclusion

The thesis has investigated the disparities between countries when it comes to the number of medals won at the Olympic Games. Its purpose has been to answer the main research question:

Why are some countries more successful than others at winning medals at the Olympic Games from an economic growth perspective, and how does econometrics estimate the relationship between the Olympic medal distribution and its determinants suggested by the economic growth theory framework?

After introducing the Olympic Games and their organization and structure, the thesis has first addressed the historical distribution of Olympic medals in its chosen timeframe 1996-2016 to establish with the Herfindahl-Hirschman Index that there are indeed countries that win more medals than others. On the other hand, Olympic competition is not simply dominated by one country either, which encourages and motivates the study of why, given that the Olympics are competitive, some countries are more successful than others.

The literature review has emphasized the prevalence of GDP per capita and population in previous studies as important determinants of the Olympic medal distribution. Other reasons suggested in the literature for disparity in countries' medal counts have included the effects of being the hosting country, a country's political structure, past performance, investment in sports, team size and gender equality. The most commonly used econometric method for studying the determinants of national Olympic success has been the tobit regression.

Parallels have then been drawn between the economic growth question of why some countries are more prosperous and have higher economic growth than others and the question here of why some countries win more Olympic medals than others. This has led to the role of technology in economic growth serving as a theoretical framework for the analysis, such that the determinants of the Olympic medal counts are framed as the resources a country has for Olympic technological progress. The suggested regressors have been GDP per capita, population, team size, three host effects and tourism as a proxy for openness. Investment in sports has also been suggested, but has not been included in the main regression model due to a lack of data. Other unobservable but country-specific time-invariant factors have also been assessed as being important for the medal distribution.

Following a discussion of why tobit may not be the most appropriate approach to modelling the number of medals won as the dependent variable, a fixed effects regression based on OLS methods has been employed in order to prevent omitted variable bias from the country-specific time-invariant effects. The regression results have shown that only team size and country fixed effects seem to be statistically significantly associated with a country's medal count at the 5% level, with an increase in team size by 100 athletes being associated with an increase in the expected number of medals won by 6.3. Using the same economic growth theory framework as for explaining the medal distribution, team size could be argued to be similarly determined. A fixed effects regression with team size as the dependent variable has shown that population, the host effects and time and country fixed effects are statistically significant. Thus, out of the regressors in the medal count regression, only team size has a statistically significant effect, while population and the host effects indirectly influence medals through team size. The country fixed effects are important for explaining disparities between countries' Olympic success, both in terms of the medal counts themselves and the team sizes that help determine them. That is, countries that succeed at winning medals are able to do so because of unobserved country characteristics that are resistant to change over time. Combined with the tentative suggestion from the reduced sample regressions that sports investment is not statistically significant for either medal count or team size, it implies that countries have limited options to actively and immediately improve their Olympic performance.

A discussion of why these regression results seem to differ from those in the literature has emphasized the difference in data, the choice of fixed effects regression rather than tobit and the thesis' efforts to address omitted variable bias, in particular with the use of country fixed effects. An explanation of how a zero-inflated negative binomial model may be applied to the Olympic medal dataset has also been provided as an alternative to the fixed effects model.

In conclusion, from an economic growth perspective, some countries are more successful than others at winning medals at the Olympic Games because they have more resources available and invested into improving the country's Olympic technology. As for the estimated relationship between the Olympic medal distribution and the determinants suggested by this framework, the fixed effects regression has shown that the resources directly relevant to national Olympics success are team size, with a predicted 6.3 additional medals for every 100 additional athletes, and resources contained in country-specific time-invariant factors. GDP per capita and tourism do not seem to have an effect on the number of medals won, and population and host dummies only have an indirect effect through team size.

Further research on the determinants of national Olympic success could pursue the use of count models for modelling medal counts. A demonstration of how a zero-inflated negative binomial model may be applied to the Olympic medal dataset has been considered in the thesis. It could also be interesting to use the approach taken here to understand the determinants of the Olympic medal distribution and apply it to the related Winter Olympics or Paralympics for comparison. Lastly, greater insight into country-specific effects, in particular the impact of specialization in specific national sports, might be able to come from cross-sectional analysis of Olympic medals on an individual athlete basis rather than on the country level. This would allow for dummy variables pertaining to the sport and event in which a medal was won or not won to be included, and interaction terms of the type of sport and event with the country of an athlete would be relevant to estimate. The data collection burden would be much heavier, but it may provide a more detailed and satisfying explanation for country differences in medal counts than just the broad understanding that some countries win more due to characteristics ingrained in their culture.

9 References

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, *24*(1–2), 3–61. https://doi.org/10.1016/0304-4076(84)90074-5

Andreff, W., & Andreff, M. (2015). Economic prediction of sport performances from the Beijing Olympics to the 2010 FIFA World Cup in South Africa: the notion of surprising sporting outcomes. In P. Rodríguez, S. Késenne, & R. Koning (Eds.), *The Economics of Competitive Sports* (pp. 185–216). Edward Elgar Publishing. https://doi.org/10.4337/9781783474769.00018

Baird, S., Friedman, J., & Schady, N. (2011). Aggregate income shocks and infant mortality in the developing world. *The Review of Economics and Statistics*, *93*(3), 847–856. https://doi.org/10.1162/REST_a_00084

Ball, D. W. (1972). Olympic Games competition: Structural correlates of national success. *International Journal of Comparative Sociology*, *13*(3–4), 186–200. https://doi.org/10.1163/156854272X00109

BBC. (n.d.). *Federer on his fading hopes of Olympic singles gold*. Retrieved July 10, 2020, from https://www.bbc.co.uk/programmes/articles/kvZ2BNMlgnWvvt845sS53P/federer-on-his-fading-hopes-of-olympic-singles-gold

BBC. (2012, August 12). London 2012: Team GB finish third in Olympic medal table. *BBC*. https://www.bbc.com/sport/olympics/19232685

BBC. (2016a, July 14). Rio 2016: Great Britain medal target set for Olympics & Paralympics. *BBC*. https://www.bbc.com/sport/olympics/36792200

BBC. (2016b, August 21). Rio Olympics 2016: Team GB beat China to finish second in medal table. *BBC*. https://www.bbc.com/sport/olympics/37085511

Berk, R. A., Western, B., & Weiss, R. E. (1995). Statistical inference for apparent populations. *Sociological Methodology*, *25*, 421–458. https://doi.org/10.2307/271073

Bernard, A. B., & Busse, M. R. (2004). Who wins the Olympic Games: Economic resources and medal totals. *The Review of Economics and Statistics*, *86*(1), 413–417. https://doi.org/10.1162/003465304774201824

Cottrell, C. (2012, August 10). Germans set to miss ambitious medals target. *Reuters*. https://uk.reuters.com/article/uk-oly-ger-medals-day14/germans-set-to-miss-ambitious-medals-target-idUKBRE8790WM20120810

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, *27*(2), 1–43. https://doi.org/10.18637/jss.v027.i02

Daugbjerg, A. F. (2016, August 21). *Historisk godt dansk OL* [Historically good Danish Olympic Games]. Danmarks Idrætsforbund. https://www.dif.dk/da/politik/nyheder/nyheder/2016/august/20160821-olsidstepresse

De Bosscher, V., De Knop, P., van Bottenburg, M., Shibli, S., & Bingham, J. (2009). Explaining international sporting success: An international comparison of elite sport systems and policies in six countries. *Sport Management Review*, *12*(3), 113–136. https://doi.org/10.1016/j.smr.2009.01.001

Den Butter, F. A. G., & van der Tak, C. M. (1995). Olympic medals as an indicator of social welfare. *Social Indicators Research*, *35*(1), 27–37. https://doi.org/10.1007/BF01079236

Dunbar, J. (2016, August 3). Predicting the Rio Olympic medal table. *BBC*. https://www.bbc.com/news/magazine-36955132

Feenstra, R. C., Inklaar, R., & Timmer, M. P. (2015). The next generation of the Penn World Table. *American Economic Review*, *105*(10), 3150–3182. https://doi.org/10.1257/aer.20130954

Feenstra, R. C., Inklaar, R., & Timmer, M. P. (2019, September 26). *PWT 9.1: Penn World Table version 9.1*. University of Groningen. https://doi.org/10.15141/S50T0R

Forrest, D., Sanz, I., & Tena, J. D. (2010). Forecasting national team medal totals at the Summer Olympic Games. *International Journal of Forecasting*, *26*(3), 576–588. https://doi.org/10.1016/j.ijforecast.2009.12.007

Gantz, W., & Wenner, L. A. (1991). Men, women, and sports: Audience experiences and effects. *Journal of Broadcasting & Electronic Media*, *35*(2), 233–243. https://doi.org/10.1080/08838159109364120

Grimes, A. R., Jr., Kelly, W. J., & Rubin, P. H. (1974). A socioeconomic model of national Olympic performance. *Social Science Quarterly, 55*(3), 777–783. https://www.jstor.org/stable/42860287

Grix, J., & Carmichael, F. (2012). Why do governments invest in elite sport? A polemic. *International Journal of Sport Policy and Politics*, *4*(1), 73–90. https://doi.org/10.1080/19406940.2011.627358

Guba, E. G. (1990). The alternative paradigm dialog. In E. G. Guba (Ed.), *The paradigm dialog* (pp. 17–27). SAGE Publications.

Hague, S., Gottschalk, R., & Martins, P. (2008). *Pro-poor growth: The evidence beyond income*. Sussex Economics Department and IDS ESRC Development Economics Conference.

http://eprints.mdx.ac.uk/5587/

Hallmann, K., Breuer, C., & Kühnreich, B. (2013). Happiness, pride and elite sporting success: What population segments gain most from national athletic achievements? *Sport Management Review*, *16*(2), 226–235. https://doi.org/10.1016/j.smr.2012.07.001

Hann, M. (2014, May 7). UCI and IOC agree qualification quotas for Rio 2016. *Reuters*. https://www.reuters.com/article/us-olympics-cycling/uci-and-ioc-agree-qualification-quotas-for-rio-2016-idUSKBN0DN0TF20140507

Heuslein, W. (2010, January 19). The man who predicts the medals. *Forbes*. https://www.forbes.com/2010/01/19/olympic-medal-predictions-business-sports-medals.html

Hirschey, M., & Bentzen, E. (2014). *Managerial economics* (13th ed.). Cengage Learning.

Hlavac, M. (2018). *stargazer: Well-formatted regression and summary statistics tables* (R package version 5.2.2). https://cran.r-project.org/package=stargazer

Jansen, J., & Engbersen, G. (2017). Have the Olympic Games become more migratory? A comparative historical perspective. *Comparative Migration Studies*, *5*, Article 11. https://doi.org/10.1186/s40878-017-0054-2

Johnson, D. K. N., & Ali, A. (2004). A tale of two seasons: Participation and medal counts at the Summer and Winter Olympic Games. *Social Science Quarterly*, *85*(4), 974–993. https://doi.org/10.1111/j.0038-4941.2004.00254.x

Johnson, I. (2008, August 13). Who’s on first in medals race? *The Wall Street Journal*. https://www.wsj.com/articles/SB121856271893833843

Kavetsos, G., & Szymanski, S. (2010). National well-being and international sports events. *Journal of Economic Psychology*, *31*(2), 158–171. https://doi.org/10.1016/j.joep.2009.11.005

Kiviaho, P., & Mäkelä, P. (1978). Olympic success: A sum of non-material and material factors. *International Review for the Sociology of Sport*, *13*(2), 5–22. https://doi.org/10.1177/101269027801300201

Lui, H.-K., & Suen, W. (2008). Men, money, and medals: An econometric analysis of the Olympic Games. *Pacific Economic Review*, *13*(1), 1–16. https://doi.org/10.1111/j.1468-0106.2007.00386.x

Milton-Smith, J. (2002). Ethics, the Olympics and the search for global values. *Journal of Business Ethics*, *35*(2), 131–142. https://doi.org/10.1023/A:1013015331517

Müller, M. (2015). What makes an event a mega-event? Definitions and sizes. *Leisure Studies*, *34*(6), 627–642. https://doi.org/10.1080/02614367.2014.993333

Owen, P. D., Ryan, M., & Weatherston, C. R. (2007). Measuring competitive balance in professional team sports using the Herfindahl-Hirschman Index. *Review of Industrial Organization*, *31*(4), 289–302. https://doi.org/10.1007/s11151-008-9157-0

Pepall, L., Richards, D., & Norman, G. (2014). *Industrial organization: Contemporary theory and empirical applications* (5th ed.). Wiley.

Permanyer, I. (2013). A critical assessment of the UNDP's Gender Inequality Index. *Feminist Economics*, *19*(2), 1–32. https://doi.org/10.1080/13545701.2013.769687

Preston, S. H. (1975). The changing relation between mortality and level of economic development. *Population Studies*, *29*(2), 231–248. https://doi.org/10.2307/2173509

Pritchett, L., & Summers, L. H. (1993). *Wealthier is healthier* (WPS 1150; Policy Research Working Papers). http://documents1.worldbank.org/curated/en/684651468741004317/pdf/multi0page.pdf

R Core Team. (2020). *R: A language and environment for statistical computing* (4.0.2). R Foundation for Statistical Computing. https://www.r-project.org/

Rogers, R. (2018). How sports entertain: Enjoyable and meaningful experiences for sports audiences. *Media Watch*, *9*(3), 372–382. https://doi.org/10.15655/mw/2018/v9i3/49486

Rothenbuhler, E. W. (1988). The living room celebration of the Olympic Games. *Journal of Communication*, *38*(4), 61–81. https://doi.org/10.1111/j.1460-2466.1988.tb02070.x

Stock, J. H., & Watson, M. W. (2015). *Introduction to econometrics* (3rd ed.). Pearson Education.

Swanson, A. (2016, July 22). How to predict Olympic results before the Games even start. *The Washington Post*. https://www.washingtonpost.com/news/wonk/wp/2016/07/22/how-to-predict-olympic-results-before-the-games-even-start/

The Association of Summer Olympic International Federations. (n.d.). *Members*. Retrieved July 5, 2020, from https://www.asoif.com/members

The Economist. (2012, July 28). Olympionomics: The dismal dash. *The Economist*. https://www.economist.com/finance-and-economics/2012/07/28/the-dismal-dash

The European Handball Federation. (2020, January 22). *How does qualification for the Tokyo 2020 Olympic Games work?* https://www.eurohandball.com/en/news/en/how-does-qualification-for-the-tokyo-2020-olympic-games-work/

The Guardian. (n.d.). *Medals and results*. Retrieved July 9, 2020, from https://www.theguardian.com/sport/ng-interactive/2016/aug/05/rio-olympics-2016-medal-table-and-results-in-full

The International Monetary Fund. (2020, July 18). *Expenditure by functions of government (COFOG).* https://data.imf.org/?sk=5804C5E1-0502-4672-BDCD-671BCDC565A9

The International Olympic Committee. (n.d.-a). *2020 host city election*. Retrieved July 7, 2020, from https://www.olympic.org/2020-host-city-election

The International Olympic Committee. (n.d.-b). *Ancient Olympic Games*. Retrieved July 2, 2020, from https://www.olympic.org/ancient-olympic-games

The International Olympic Committee. (n.d.-c). *Athens 1896*. Retrieved July 3, 2020, from https://www.olympic.org/athens-1896

The International Olympic Committee. (n.d.-d). *Badminton*. Retrieved July 8, 2020, from https://www.olympic.org/badminton

The International Olympic Committee. (n.d.-e). *Benefits of a candidature*. Retrieved July 7, 2020, from https://www.olympic.org/olympic-games-candidatures-beyond-2028

The International Olympic Committee. (n.d.-f). *Denmark*. Retrieved July 9, 2020, from https://www.olympic.org/denmark

The International Olympic Committee. (n.d.-g). *Global broadcast and audience report: Olympic Games Rio 2016*. Retrieved July 10, 2020, from https://stillmed.olympic.org/media/Document Library/OlympicOrg/Games/Summer-Games/Games-Rio-2016-Olympic-Games/Media-Guide-for-Rio-2016/Global-Broadcast-and-Audience-Report-Rio-2016.pdf

The International Olympic Committee. (n.d.-h). *National Olympic Committees*. Retrieved July 5, 2020, from https://www.olympic.org/national-olympic-committees

The International Olympic Committee. (n.d.-i). *Olympic Agenda 2020: 20+20 recommedations*. Retrieved June 26, 2020, from https://stillmed.olympic.org/media/Document Library/OlympicOrg/Documents/Olympic-Agenda-2020/Olympic-Agenda-2020-20-20-Recommendations.pdf

The International Olympic Committee. (n.d.-j). *Olympic Games candidature process*. Retrieved July 7, 2020, from https://www.olympic.org/all-about-the-candidature-process

The International Olympic Committee. (n.d.-k). *Organising Committees for the Olympic Games*. Retrieved July 7, 2020, from https://www.olympic.org/ioc-governance-organising-committees

The International Olympic Committee. (n.d.-l). *Results*. Retrieved March 12, 2019, from https://www.olympic.org/olympic-results

The International Olympic Committee. (n.d.-m). *Rio 2016*. Retrieved July 3, 2020, from https://www.olympic.org/rio-2016

The International Olympic Committee. (n.d.-n). *Roger Federer: Wizard of the Olympic tennis courts*. Retrieved July 10, 2020, from https://www.olympic.org/roger-federer

The International Olympic Committee. (n.d.-o). *"Run Leonidas Run": Meet the stars of the track at the ancient Olympic Games*. Retrieved July 2, 2020, from https://www.olympic.org/ancient-olympic-games/running

The International Olympic Committee. (n.d.-p). *Sports, programme and results*. Retrieved July 8, 2020, from https://www.olympic.org/faq/sports-programme-and-results

The International Olympic Committee. (n.d.-q). *Summer Games*. Retrieved July 21, 2020, from https://www.olympic.org/summer-games

The International Olympic Committee. (n.d.-r). *Swimming*. Retrieved July 8, 2020, from https://www.olympic.org/swimming

The International Olympic Committee. (n.d.-s). *The sports events*. Retrieved July 2, 2020, from https://www.olympic.org/ancient-olympic-games/the-sports-events

The International Olympic Committee. (n.d.-t). *Who we are*. Retrieved July 3, 2020, from https://www.olympic.org/about-ioc-olympic-movement

The International Olympic Committee. (2016a). *Rio 2016 Opening Ceremony - Flag bearers*. https://stillmed.olympic.org/media/Document Library/OlympicOrg/News/2016/08/2016-08-05-Rio-2016-Opening-Ceremony-Flag-Bearers.pdf

The International Olympic Committee. (2016b, August 3). *IOC approves five new sports for Olympic Games Tokyo 2020*. https://www.olympic.org/news/ioc-approves-five-new-sports-for-olympic-games-tokyo-2020

The International Olympic Committee. (2017a). *Factsheet: The Games of the Olympiad*. https://stillmed.olympic.org/media/Document Library/OlympicOrg/Factsheets-Reference-Documents/Games/OG/Factsheet-The-Games-of-the-Olympiad.pdf

The International Olympic Committee. (2017b). *Host city contract: Operational requirements*. https://stillmed.olympic.org/media/Document Library/OlympicOrg/Games/Summer-

Games/Games-Tokyo-2020-Olympic-Games/Host-City-Contract-Operational-Requirements-version-December-2016.pdf

The International Olympic Committee. (2017c). *Legacy strategic approach: Moving forward*. https://www.olympic.org/-/media/Document Library/OlympicOrg/Documents/Olympic-Legacy/IOC_Legacy_Strategy_Full_version.pdf

The International Olympic Committee. (2018). *Factsheet: Women in the Olympic Movement*. https://stillmed.olympic.org/media/Document Library/OlympicOrg/Factsheets-Reference-Documents/Women-in-the-Olympic-Movement/Factsheet-Women-in-the-Olympic-Movement.pdf

The International Olympic Committee. (2019a). *Olympic Charter*. https://stillmed.olympic.org/media/Document Library/OlympicOrg/General/EN-Olympic-Charter.pdf

The International Olympic Committee. (2019b, July 23). *IOC sanctions one athlete for failing anti-doping tests at London 2012*. https://www.olympic.org/news/ioc-sanctions-one-athlete-for-failing-anti-doping-tests-at-london-2012-3

The International Olympic Committee. (2020a). *Olympic marketing fact file: 2020 edition*. https://stillmed.olympic.org/media/Document Library/OlympicOrg/Documents/IOC-Marketing-and-Broadcasting-General-Files/Olympic-Marketing-Fact-File.pdf

The International Olympic Committee. (2020b). *XXXII Games of the Olympiad, Tokyo 2020: Qualification system principles*. https://stillmed.olympic.org/media/Document Library/OlympicOrg/Games/Summer-Games/Games-Tokyo-2020-Olympic-Games/Tokyo_2020_Qualification_System_Principles_Revised.pdf

The International Olympic Committee. (2020c, March 24). *Joint statement from the International Olympic Committee and the Tokyo 2020 Organising Committee*. https://www.olympic.org/news/joint-statement-from-the-international-olympic-committee-and-the-tokyo-2020-organising-committee

The Japan Times. (2018, June 5). JOC to set target of 30 gold medals for 2020 Olympics: sources. *The Japan Times*. https://www.japantimes.co.jp/sports/2018/06/05/olympics/joc-set-target-30-gold-medals-2020-olympics-sources/

The Olympic Museum. (n.d.). *Permanent exhibition* [Museum exhibition]. The Olympic Museum, Lausanne, Switzerland. Visited December 15, 2019.

The Olympic Museum. (2013). *The Modern Olympic Games*. https://stillmed.olympic.org/media/Document Library/OlympicOrg/Documents/Document-Set-Teachers-The-Main-Olympic-Topics/The-Modern-Olympic-Games.pdf

The Tokyo Organising Committee of the Olympic and Paralympic Games. (2020, June 11). *Roger Federer to miss rest of the season*. https://tokyo2020.org/en/news/roger-federer-to-miss-rest-of-the-season

The Washington Post. (n.d.). *Rio 2016 medal count*. Retrieved July 9, 2020, from https://www.washingtonpost.com/graphics/sports/olympics/rio-2016-olympic-medal-count/

The World Bank. (n.d.-a). *GDP (current US$)*. Retrieved March 5, 2020, from https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

The World Bank. (n.d.-b). *International tourism, number of arrivals*. Retrieved July 21, 2020, from https://data.worldbank.org/indicator/ST.INT.ARVL

The World Bank. (n.d.-c). *Mortality rate, infant (per 1,000 live births)*. Retrieved July 20, 2020, from https://data.worldbank.org/indicator/SP.DYN.IMRT.IN

Trivedi, P. K., & Zimmer, D. M. (2014). Success at the Summer Olympics: How much do economic factors explain? *Econometrics*, *2*(4), 169–202. https://doi.org/10.3390/econometrics2040169

UCLA. (n.d.). *Zero-inflated negative binomial regression*. UCLA Statistical Consulting, Institute for Digital Research & Education. Retrieved April 29, 2020, from https://stats.idre.ucla.edu/r/dae/zinb/

Union Cycliste Internationale. (n.d.). *Cycling at the Olympic Games*. Retrieved July 8, 2020, from https://www.uci.org/olympic-games/cycling-at-the-olympic-games

Vagenas, G., & Vlachokyriakou, E. (2012). Olympic medals and demo-economic factors: Novel predictors, the ex-host effect, the exact role of team size, and the “population-GDP” model revisited. *Sport Management Review*, *15*(2), 211–217. https://doi.org/10.1016/j.smr.2011.07.001

van Hilvoorde, I., Elling, A., & Stokvis, R. (2010). How to influence national pride? The Olympic medal index as a unifying narrative. *International Review for the Sociology of Sport*, *45*(1), 87–102. https://doi.org/10.1177/1012690209356989

Vasilogambros, M. (2016, August 5). Why some Olympians compete against their home country. *The Atlantic*. https://www.theatlantic.com/news/archive/2016/08/why-some-olympians-compete-against-their-home-country/494648/

Weil, D. N. (2016). *Economic growth* (3rd ed.). Routledge.

Wicker, P., Hallmann, K., Breuer, C., & Feiler, S. (2012). The value of Olympic success and the intangible effects of sport events - A contingent valuation approach in Germany. *European Sport Management Quarterly*, *12*(4), 337–355. https://doi.org/10.1080/16184742.2012.693117

Wikipedia. (n.d.-a). *1996 Summer Olympics*. Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/1996_Summer_Olympics

Wikipedia. (n.d.-b). *1996 Summer Olympics medal table*. Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/1996_Summer_Olympics_medal_table

Wikipedia. (n.d.-c). *2000 Summer Olympics*. Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/2000_Summer_Olympics

Wikipedia. (n.d.-d). *2000 Summer Olympics medal table*. Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/2000_Summer_Olympics_medal_table

Wikipedia. (n.d.-e). *2004 Summer Olympics*. Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/2004_Summer_Olympics

Wikipedia. (n.d.-f). *2004 Summer Olympics medal table*. Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/2004_Summer_Olympics_medal_table

Wikipedia. (n.d.-g). *2008 Summer Olympics*. Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/2008_Summer_Olympics

Wikipedia. (n.d.-h). *2008 Summer Olympics medal table*. Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/2008_Summer_Olympics_medal_table

Wikipedia. (n.d.-i). *2012 Summer Olympics*. Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/2012_Summer_Olympics

Wikipedia. (n.d.-j). *2012 Summer Olympics medal table*. Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/2012_Summer_Olympics_medal_table

Wikipedia. (n.d.-k). *2016 Summer Olympics*. Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/2016_Summer_Olympics

Wikipedia. (n.d.-l). *2016 Summer Olympics medal table*. Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/2016_Summer_Olympics_medal_table

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, *27*(8), 1–25. https://doi.org/10.18637/jss.v027.i08

Zimbalist, A. S. (2002). Competitive balance in sports leagues: An introduction. *Journal of Sports Economics*, *3*(2), 111–121. https://doi.org/10.1177/152700250200300201

10 Appendix

10.1 Appendix A: Medal tables and Herfindahl-Hirschman Index, 1996-2016

Olympic medal data was collected from Wikipedia (n.d.-b, n.d.-d, n.d.-f, n.d.-h, n.d.-j, n.d.-l).

1996:

NOC	Total medals	Medal share (%)	Medal share ^ 2
United States	101	11.9952	143.8860
Germany	65	7.7197	59.5940
Russia	63	7.4822	55.9831
China	50	5.9382	35.2627
Australia	41	4.8694	23.7107
France	37	4.3943	19.3099
Italy	35	4.1568	17.2787
South Korea	27	3.2067	10.2826
Cuba	25	2.9691	8.8157
Ukraine	23	2.7316	7.4616
Canada	22	2.6128	6.8269
Hungary	21	2.4941	6.2203
Romania	20	2.3753	5.6420
Netherlands	19	2.2565	5.0919
Poland	17	2.0190	4.0764
Spain	17	2.0190	4.0764
Bulgaria	15	1.7815	3.1736
Brazil	15	1.7815	3.1736
Great Britain	15	1.7815	3.1736
Belarus	15	1.7815	3.1736
Japan	14	1.6627	2.7646
Czech Republic	11	1.3064	1.7067
Kazakhstan	11	1.3064	1.7067
Greece	8	0.9501	0.9027
Sweden	8	0.9501	0.9027
Kenya	8	0.9501	0.9027
Switzerland	7	0.8314	0.6911

NOC	Total medals	Medal share (%)	Medal share ^ 2
Norway	7	0.8314	0.6911
Denmark	6	0.7126	0.5078
Turkey	6	0.7126	0.5078
New Zealand	6	0.7126	0.5078
Belgium	6	0.7126	0.5078
Nigeria	6	0.7126	0.5078
Jamaica	6	0.7126	0.5078
South Africa	5	0.5938	0.3526
North Korea	5	0.5938	0.3526
Ireland	4	0.4751	0.2257
Finland	4	0.4751	0.2257
Indonesia	4	0.4751	0.2257
Yugoslavia	4	0.4751	0.2257
Algeria	3	0.3563	0.1269
Ethiopia	3	0.3563	0.1269
Iran	3	0.3563	0.1269
Slovakia	3	0.3563	0.1269
Argentina	3	0.3563	0.1269
Austria	3	0.3563	0.1269
Armenia	2	0.2375	0.0564
Croatia	2	0.2375	0.0564
Portugal	2	0.2375	0.0564
Thailand	2	0.2375	0.0564
Namibia	2	0.2375	0.0564
Slovenia	2	0.2375	0.0564
Malaysia	2	0.2375	0.0564
Moldova	2	0.2375	0.0564

NOC	Total medals	Medal share (%)	Medal share ^ 2
Uzbekistan	2	0.2375	0.0564
Georgia	2	0.2375	0.0564
Morocco	2	0.2375	0.0564
Trinidad and Tobago	2	0.2375	0.0564
Burundi	1	0.1188	0.0141
Costa Rica	1	0.1188	0.0141
Ecuador	1	0.1188	0.0141
Hong Kong	1	0.1188	0.0141
Syria	1	0.1188	0.0141
Azerbaijan	1	0.1188	0.0141
Bahamas	1	0.1188	0.0141
Chinese Taipei	1	0.1188	0.0141
Latvia	1	0.1188	0.0141
Philippines	1	0.1188	0.0141
Tonga	1	0.1188	0.0141
Zambia	1	0.1188	0.0141
India	1	0.1188	0.0141
Israel	1	0.1188	0.0141
Lithuania	1	0.1188	0.0141
Mexico	1	0.1188	0.0141
Mongolia	1	0.1188	0.0141
Mozambique	1	0.1188	0.0141
Puerto Rico	1	0.1188	0.0141
Tunisia	1	0.1188	0.0141
Uganda	1	0.1188	0.0141
Total medals	**842**	**HI**	**442.8716**

Table 10.1: Herfindahl Index, 1996 Olympic Games

2000:

NOC	Total medals	Medal share (%)	Medal share ^ 2
United States	93	10.0324	100.6483
Russia	89	9.6009	92.1766
China	58	6.2567	39.1468
Australia	58	6.2567	39.1468
Germany	56	6.0410	36.4936
France	38	4.0992	16.8038
Italy	34	3.6677	13.4524
Cuba	29	3.1284	9.7867
Great Britain	28	3.0205	9.1234
South Korea	28	3.0205	9.1234
Romania	26	2.8047	7.8666
Netherlands	25	2.6969	7.2731
Ukraine	23	2.4811	6.1560
Japan	18	1.9417	3.7704
Hungary	17	1.8339	3.3631
Belarus	17	1.8339	3.3631
Poland	14	1.5102	2.2808
Canada	14	1.5102	2.2808
Bulgaria	13	1.4024	1.9667
Greece	13	1.4024	1.9667
Sweden	12	1.2945	1.6757
Brazil	12	1.2945	1.6757
Spain	11	1.1866	1.4081
Norway	10	1.0787	1.1637
Switzerland	9	0.9709	0.9426
Jamaica	9	0.9709	0.9426
Ethiopia	8	0.8630	0.7448

NOC	Total medals	Medal share (%)	Medal share ^ 2
Czech Republic	8	0.8630	0.7448
Kazakhstan	7	0.7551	0.5702
Kenya	7	0.7551	0.5702
Denmark	6	0.6472	0.4189
Indonesia	6	0.6472	0.4189
Mexico	6	0.6472	0.4189
Georgia	6	0.6472	0.4189
Turkey	5	0.5394	0.2909
Lithuania	5	0.5394	0.2909
Slovakia	5	0.5394	0.2909
Algeria	5	0.5394	0.2909
Belgium	5	0.5394	0.2909
South Africa	5	0.5394	0.2909
Chinese Taipei	5	0.5394	0.2909
Morocco	5	0.5394	0.2909
Iran	4	0.4315	0.1862
Finland	4	0.4315	0.1862
Uzbekistan	4	0.4315	0.1862
New Zealand	4	0.4315	0.1862
Argentina	4	0.4315	0.1862
North Korea	4	0.4315	0.1862
Austria	3	0.3236	0.1047
Azerbaijan	3	0.3236	0.1047
Bahamas	3	0.3236	0.1047
Nigeria	3	0.3236	0.1047
Latvia	3	0.3236	0.1047
Yugoslavia	3	0.3236	0.1047

NOC	Total medals	Medal share (%)	Medal share ^ 2
Estonia	3	0.3236	0.1047
Thailand	3	0.3236	0.1047
Slovenia	2	0.2157	0.0465
Croatia	2	0.2157	0.0465
Moldova	2	0.2157	0.0465
Saudi Arabia	2	0.2157	0.0465
Trinidad and Tobago	2	0.2157	0.0465
Costa Rica	2	0.2157	0.0465
Portugal	2	0.2157	0.0465
Cameroon	1	0.1079	0.0116
Colombia	1	0.1079	0.0116
Mozambique	1	0.1079	0.0116
Ireland	1	0.1079	0.0116
Sri Lanka	1	0.1079	0.0116
Uruguay	1	0.1079	0.0116
Vietnam	1	0.1079	0.0116
Armenia	1	0.1079	0.0116
Barbados	1	0.1079	0.0116
Chile	1	0.1079	0.0116
Iceland	1	0.1079	0.0116
India	1	0.1079	0.0116
Israel	1	0.1079	0.0116
Kuwait	1	0.1079	0.0116
Kyrgyzstan	1	0.1079	0.0116
Macedonia	1	0.1079	0.0116
Qatar	1	0.1079	0.0116
Total medals	**927**	**HI**	**423.1092**

Table 10.2: Herfindahl Index, 2000 Olympic Games

2004:

NOC	Total medals	Medal share (%)	Medal share ^ 2
United States	101	10.8954	118.7089
Russia	90	9.7087	94.2596
China	63	6.7961	46.1872
Australia	50	5.3937	29.0925
Germany	49	5.2859	27.9404
Japan	37	3.9914	15.9310
France	33	3.5599	12.6727
Italy	32	3.4520	11.9163
South Korea	30	3.2362	10.4733
Great Britain	30	3.2362	10.4733
Cuba	27	2.9126	8.4834
Ukraine	22	2.3732	5.6323
Netherlands	22	2.3732	5.6323
Spain	20	2.1575	4.6548
Romania	19	2.0496	4.2010
Hungary	17	1.8339	3.3631
Greece	16	1.7260	2.9791
Belarus	13	1.4024	1.9667
Canada	12	1.2945	1.6757
Bulgaria	12	1.2945	1.6757
Turkey	11	1.1866	1.4081
Brazil	10	1.0787	1.1637
Poland	10	1.0787	1.1637
Czech Republic	9	0.9709	0.9426
Thailand	8	0.8630	0.7448

NOC	Total medals	Medal share (%)	Medal share ^ 2
Denmark	8	0.8630	0.7448
Kazakhstan	8	0.8630	0.7448
Sweden	7	0.7551	0.5702
Austria	7	0.7551	0.5702
Ethiopia	7	0.7551	0.5702
Kenya	7	0.7551	0.5702
Norway	6	0.6472	0.4189
Iran	6	0.6472	0.4189
Slovakia	6	0.6472	0.4189
Argentina	6	0.6472	0.4189
South Africa	6	0.6472	0.4189
New Zealand	5	0.5394	0.2909
Chinese Taipei	5	0.5394	0.2909
Jamaica	5	0.5394	0.2909
Uzbekistan	5	0.5394	0.2909
Croatia	5	0.5394	0.2909
Egypt	5	0.5394	0.2909
Switzerland	5	0.5394	0.2909
Azerbaijan	5	0.5394	0.2909
North Korea	5	0.5394	0.2909
Georgia	4	0.4315	0.1862
Indonesia	4	0.4315	0.1862
Latvia	4	0.4315	0.1862
Mexico	4	0.4315	0.1862
Slovenia	4	0.4315	0.1862

NOC	Total medals	Medal share (%)	Medal share ^ 2
Morocco	3	0.3236	0.1047
Chile	3	0.3236	0.1047
Lithuania	3	0.3236	0.1047
Zimbabwe	3	0.3236	0.1047
Belgium	3	0.3236	0.1047
Portugal	3	0.3236	0.1047
Estonia	3	0.3236	0.1047
Bahamas	2	0.2157	0.0465
Israel	2	0.2157	0.0465
Finland	2	0.2157	0.0465
Serbia and Montenegro	2	0.2157	0.0465
Colombia	2	0.2157	0.0465
Nigeria	2	0.2157	0.0465
Venezuela	2	0.2157	0.0465
Cameroon	1	0.1079	0.0116
Dominican Republic	1	0.1079	0.0116
United Arab Emirates	1	0.1079	0.0116
Hong Kong	1	0.1079	0.0116
India	1	0.1079	0.0116
Paraguay	1	0.1079	0.0116
Eritrea	1	0.1079	0.0116
Mongolia	1	0.1079	0.0116
Syria	1	0.1079	0.0116
Trinidad and Tobago	1	0.1079	0.0116
Total medals	**927**	**HI**	**433.9316**

Table 10.3: Herfindahl Index, 2004 Olympic Games

2008:

NOC	Total medals	Medal share (%)	Medal share ^ 2
United States	112	11.6910	136.6800
China	100	10.4384	108.9605
Russia	60	6.2630	39.2258
Great Britain	51	5.3236	28.3406
Australia	46	4.8017	23.0560
France	43	4.4885	20.1468
Germany	41	4.2797	18.3163
South Korea	32	3.3403	11.1576
Cuba	30	3.1315	9.8064
Italy	27	2.8184	7.9432
Japan	25	2.6096	6.8100
Ukraine	22	2.2965	5.2737
Canada	20	2.0877	4.3584
Spain	19	1.9833	3.9335
Brazil	17	1.7745	3.1490
Netherlands	16	1.6701	2.7894
Kenya	16	1.6701	2.7894
Belarus	14	1.4614	2.1356
Jamaica	11	1.1482	1.3184
Poland	11	1.1482	1.3184
Hungary	10	1.0438	1.0896
Romania	9	0.9395	0.8826
Norway	9	0.9395	0.8826
New Zealand	9	0.9395	0.8826
Kazakhstan	9	0.9395	0.8826
Ethiopia	7	0.7307	0.5339
Czech Republic	7	0.7307	0.5339
Georgia	7	0.7307	0.5339
Denmark	7	0.7307	0.5339

NOC	Total medals	Medal share (%)	Medal share ^ 2
Switzerland	7	0.7307	0.5339
Slovakia	6	0.6263	0.3923
North Korea	6	0.6263	0.3923
Thailand	6	0.6263	0.3923
Argentina	6	0.6263	0.3923
Azerbaijan	6	0.6263	0.3923
Indonesia	6	0.6263	0.3923
Slovenia	5	0.5219	0.2724
Bulgaria	5	0.5219	0.2724
Turkey	5	0.5219	0.2724
Sweden	5	0.5219	0.2724
Lithuania	5	0.5219	0.2724
Nigeria	5	0.5219	0.2724
Croatia	5	0.5219	0.2724
Armenia	5	0.5219	0.2724
Mongolia	4	0.4175	0.1743
Mexico	4	0.4175	0.1743
Zimbabwe	4	0.4175	0.1743
Chinese Taipei	4	0.4175	0.1743
Finland	4	0.4175	0.1743
Uzbekistan	4	0.4175	0.1743
Latvia	3	0.3132	0.0981
India	3	0.3132	0.0981
Colombia	3	0.3132	0.0981
Greece	3	0.3132	0.0981
Austria	3	0.3132	0.0981
Ireland	3	0.3132	0.0981
Kyrgyzstan	3	0.3132	0.0981
Serbia	3	0.3132	0.0981

NOC	Total medals	Medal share (%)	Medal share ^ 2
Belgium	2	0.2088	0.0436
Dominican Republic	2	0.2088	0.0436
Estonia	2	0.2088	0.0436
Portugal	2	0.2088	0.0436
Trinidad and Tobago	2	0.2088	0.0436
Iran	2	0.2088	0.0436
Algeria	2	0.2088	0.0436
Bahamas	2	0.2088	0.0436
Morocco	2	0.2088	0.0436
Tajikistan	2	0.2088	0.0436
Egypt	2	0.2088	0.0436
Cameroon	1	0.1044	0.0109
Panama	1	0.1044	0.0109
Tunisia	1	0.1044	0.0109
Chile	1	0.1044	0.0109
Ecuador	1	0.1044	0.0109
Iceland	1	0.1044	0.0109
Malaysia	1	0.1044	0.0109
Samoa	1	0.1044	0.0109
Singapore	1	0.1044	0.0109
South Africa	1	0.1044	0.0109
Sudan	1	0.1044	0.0109
Vietnam	1	0.1044	0.0109
Afghanistan	1	0.1044	0.0109
Israel	1	0.1044	0.0109
Mauritius	1	0.1044	0.0109
Moldova	1	0.1044	0.0109
Togo	1	0.1044	0.0109
Venezuela	1	0.1044	0.0109
Total medals	**958**	**HI**	**451.8373**

Table 10.4: Herfindahl Index, 2008 Olympic Games

2012:

NOC	Total medals	Medal share (%)	Medal share ^ 2
United States	104	10.8446	117.6060
China	91	9.4891	90.0421
Russia	68	7.0907	50.2783
Great Britain	65	6.7779	45.9398
Germany	44	4.5881	21.0508
Japan	38	3.9625	15.7011
France	35	3.6496	13.3198
Australia	35	3.6496	13.3198
South Korea	30	3.1283	9.7860
Italy	28	2.9197	8.5247
Netherlands	20	2.0855	4.3493
Ukraine	19	1.9812	3.9253
Hungary	18	1.8770	3.5230
Spain	18	1.8770	3.5230
Canada	18	1.8770	3.5230
Brazil	17	1.7727	3.1424
Cuba	15	1.5641	2.4465
Iran	13	1.3556	1.8376
New Zealand	13	1.3556	1.8376
Kenya	13	1.3556	1.8376
Jamaica	12	1.2513	1.5658
Czech Republic	11	1.1470	1.3157
Poland	11	1.1470	1.3157
Kazakhstan	11	1.1470	1.3157
Belarus	10	1.0428	1.0873
Romania	9	0.9385	0.8807
Denmark	9	0.9385	0.8807
Azerbaijan	9	0.9385	0.8807
Ethiopia	8	0.8342	0.6959

NOC	Total medals	Medal share (%)	Medal share ^ 2
Sweden	8	0.8342	0.6959
Colombia	8	0.8342	0.6959
Mexico	8	0.8342	0.6959
South Africa	6	0.6257	0.3914
North Korea	6	0.6257	0.3914
Croatia	6	0.6257	0.3914
Georgia	6	0.6257	0.3914
Ireland	6	0.6257	0.3914
India	6	0.6257	0.3914
Lithuania	5	0.5214	0.2718
Mongolia	5	0.5214	0.2718
Switzerland	4	0.4171	0.1740
Norway	4	0.4171	0.1740
Argentina	4	0.4171	0.1740
Serbia	4	0.4171	0.1740
Slovenia	4	0.4171	0.1740
Trinidad and Tobago	4	0.4171	0.1740
Egypt	4	0.4171	0.1740
Thailand	4	0.4171	0.1740
Slovakia	4	0.4171	0.1740
Tunisia	3	0.3128	0.0979
Turkey	3	0.3128	0.0979
Bulgaria	3	0.3128	0.0979
Finland	3	0.3128	0.0979
Indonesia	3	0.3128	0.0979
Belgium	3	0.3128	0.0979
Uzbekistan	3	0.3128	0.0979
Dominican Republic	2	0.2086	0.0435
Chinese Taipei	2	0.2086	0.0435

NOC	Total medals	Medal share (%)	Medal share ^ 2
Latvia	2	0.2086	0.0435
Armenia	2	0.2086	0.0435
Estonia	2	0.2086	0.0435
Malaysia	2	0.2086	0.0435
Puerto Rico	2	0.2086	0.0435
Greece	2	0.2086	0.0435
Qatar	2	0.2086	0.0435
Singapore	2	0.2086	0.0435
Algeria	1	0.1043	0.0109
Bahamas	1	0.1043	0.0109
Bahrain	1	0.1043	0.0109
Grenada	1	0.1043	0.0109
Uganda	1	0.1043	0.0109
Venezuela	1	0.1043	0.0109
Botswana	1	0.1043	0.0109
Cyprus	1	0.1043	0.0109
Gabon	1	0.1043	0.0109
Guatemala	1	0.1043	0.0109
Montenegro	1	0.1043	0.0109
Portugal	1	0.1043	0.0109
Afghanistan	1	0.1043	0.0109
Cameroon	1	0.1043	0.0109
Hong Kong	1	0.1043	0.0109
Kuwait	1	0.1043	0.0109
Morocco	1	0.1043	0.0109
Saudi Arabia	1	0.1043	0.0109
Tajikistan	1	0.1043	0.0109
Vietnam	1	0.1043	0.0109
Total medals	**959**	**HI**	**433.3350**

Table 10.5: Herfindahl Index, 2012 Olympic Games

2016:

NOC	Total medals	Medal share (%)	Medal share ^ 2
United States	121	12.4358	154.6483
China	70	7.1942	51.7572
Great Britain	67	6.8859	47.4159
Russia	56	5.7554	33.1246
Germany	42	4.3165	18.6326
France	42	4.3165	18.6326
Japan	41	4.2138	17.7559
Australia	29	2.9805	8.8832
Italy	28	2.8777	8.2811
Canada	22	2.2610	5.1123
South Korea	21	2.1583	4.6581
Netherlands	19	1.9527	3.8131
Brazil	19	1.9527	3.8131
New Zealand	18	1.8499	3.4223
Kazakhstan	18	1.8499	3.4223
Azerbaijan	18	1.8499	3.4223
Spain	17	1.7472	3.0526
Hungary	15	1.5416	2.3766
Denmark	15	1.5416	2.3766
Kenya	13	1.3361	1.7851
Uzbekistan	13	1.3361	1.7851
Jamaica	11	1.1305	1.2781
Cuba	11	1.1305	1.2781
Sweden	11	1.1305	1.2781
Ukraine	11	1.1305	1.2781
Poland	11	1.1305	1.2781
Croatia	10	1.0277	1.0563
South Africa	10	1.0277	1.0563
Czech Republic	10	1.0277	1.0563

NOC	Total medals	Medal share (%)	Medal share ^ 2
Belarus	9	0.9250	0.8556
Colombia	8	0.8222	0.6760
Iran	8	0.8222	0.6760
Serbia	8	0.8222	0.6760
Turkey	8	0.8222	0.6760
Ethiopia	8	0.8222	0.6760
Switzerland	7	0.7194	0.5176
North Korea	7	0.7194	0.5176
Georgia	7	0.7194	0.5176
Greece	6	0.6166	0.3803
Belgium	6	0.6166	0.3803
Thailand	6	0.6166	0.3803
Malaysia	5	0.5139	0.2641
Mexico	5	0.5139	0.2641
Argentina	4	0.4111	0.1690
Slovakia	4	0.4111	0.1690
Armenia	4	0.4111	0.1690
Slovenia	4	0.4111	0.1690
Romania	4	0.4111	0.1690
Lithuania	4	0.4111	0.1690
Norway	4	0.4111	0.1690
Indonesia	3	0.3083	0.0951
Chinese Taipei	3	0.3083	0.0951
Venezuela	3	0.3083	0.0951
Bulgaria	3	0.3083	0.0951
Egypt	3	0.3083	0.0951
Tunisia	3	0.3083	0.0951
Bahrain	2	0.2055	0.0423
Vietnam	2	0.2055	0.0423

NOC	Total medals	Medal share (%)	Medal share ^ 2
Bahamas	2	0.2055	0.0423
Independent Olympic Athletes	2	0.2055	0.0423
Ivory Coast	2	0.2055	0.0423
Algeria	2	0.2055	0.0423
Ireland	2	0.2055	0.0423
India	2	0.2055	0.0423
Mongolia	2	0.2055	0.0423
Israel	2	0.2055	0.0423
Fiji	1	0.1028	0.0106
Jordan	1	0.1028	0.0106
Kosovo	1	0.1028	0.0106
Puerto Rico	1	0.1028	0.0106
Singapore	1	0.1028	0.0106
Tajikistan	1	0.1028	0.0106
Burundi	1	0.1028	0.0106
Grenada	1	0.1028	0.0106
Niger	1	0.1028	0.0106
Philippines	1	0.1028	0.0106
Qatar	1	0.1028	0.0106
Austria	1	0.1028	0.0106
Dominican Republic	1	0.1028	0.0106
Estonia	1	0.1028	0.0106
Finland	1	0.1028	0.0106
Morocco	1	0.1028	0.0106
Nigeria	1	0.1028	0.0106
Portugal	1	0.1028	0.0106
Trinidad and Tobago	1	0.1028	0.0106
United Arab Emirates	1	0.1028	0.0106
Total medals	**973**	**HI**	**417.5746**

Table 10.6: Herfindahl Index, 2016 Olympic Games

10.2 Appendix B: List of countries in the dataset

NOC	**IOC code**
Afghanistan	AFG
Albania	ALB
Algeria	ALG
American Samoa	ASA
Andorra	AND
Angola	ANG
Antigua and Barbuda	ANT
Argentina	ARG
Armenia	ARM
Aruba	ARU
Australia	AUS
Austria	AUT
Azerbaijan	AZE
Bahamas	BAH
Bahrain	BRN
Bangladesh	BAN
Barbados	BAR
Belarus	BLR
Belgium	BEL
Belize	BIZ
Benin	BEN
Bermuda	BER
Bhutan	BHU
Bolivia	BOL

Bosnia and Herzegovina	BIH
Botswana	BOT
Brazil	BRA
British Virgin Islands	IVB
Brunei	BRU
Bulgaria	BUL
Burkina Faso	BUR
Burundi	BDI
Cambodia	CAM
Cameroon	CMR
Canada	CAN
Cape Verde	CPV
Cayman Islands	CAY
Central African Republic	CAF
Chad	CHA
Chile	CHI
China	CHN
Chinese Taipei	TPE
Colombia	COL
Comoros	COM
Cook Islands	COK
Costa Rica	CRC
Croatia	CRO
Cuba	CUB
Cyprus	CYP
Czech Republic	CZE
Democratic Republic of the Congo	COD

Denmark	DEN
Djibouti	DJI
Dominica	DMA
Dominican Republic	DOM
East Timor	TLS
Ecuador	ECU
Egypt	EGY
El Salvador	ESA
Equatorial Guinea	GEQ
Eritrea	ERI
Estonia	EST
Ethiopia	ETH
Federated States of Micronesia	FSM
Fiji	FIJ
Finland	FIN
France	FRA
Gabon	GAB
Georgia	GEO
Germany	GER
Ghana	GHA
Great Britain	GBR
Greece	GRE
Grenada	GRN
Guam	GUM
Guatemala	GUA
Guinea	GUI
Guinea-Bissau	GBS

Guyana	GUY
Haiti	HAI
Honduras	HON
Hong Kong	HKG
Hungary	HUN
Iceland	ISL
India	IND
Indonesia	INA
Iran	IRI
Iraq	IRQ
Ireland	IRL
Israel	ISR
Italy	ITA
Ivory Coast	CIV
Jamaica	JAM
Japan	JPN
Jordan	JOR
Kazakhstan	KAZ
Kenya	KEN
Kiribati	KIR
Kosovo	KOS
Kuwait	KUW
Kyrgyzstan	KGZ
Laos	LAO
Latvia	LAT
Lebanon	LIB
Lesotho	LES

Liberia	LBR
Libya	LBA
Liechtenstein	LIE
Lithuania	LTU
Luxembourg	LUX
Macedonia	MKD
Madagascar	MAD
Malawi	MAW
Malaysia	MAS
Maldives	MDV
Mali	MLI
Malta	MLT
Marshall Islands	MHL
Mauritania	MTN
Mauritius	MRI
Mexico	MEX
Moldova	MDA
Monaco	MON
Mongolia	MGL
Montenegro	MNE
Morocco	MAR
Mozambique	MOZ
Myanmar	MYA
Namibia	NAM
Nauru	NRU
Nepal	NEP
Netherlands	NED

New Zealand	NZL
Nicaragua	NCA
Niger	NIG
Nigeria	NGR
North Korea	PRK
Norway	NOR
Oman	OMA
Pakistan	PAK
Palau	PLW
Palestine	PLE
Panama	PAN
Papua New Guinea	PNG
Paraguay	PAR
Peru	PER
Philippines	PHI
Poland	POL
Portugal	POR
Puerto Rico	PUR
Qatar	QAT
Republic of the Congo	CGO
Romania	ROU
Russia	RUS
Rwanda	RWA
Saint Kitts and Nevis	SKN
Saint Lucia	LCA
Saint Vincent and the Grenadines	VIN
Samoa	SAM

San Marino	SMR
Sao Tome and Principe	STP
Saudi Arabia	KSA
Senegal	SEN
Serbia	SRB
Seychelles	SEY
Sierra Leone	SLE
Singapore	SGP
Slovakia	SVK
Slovenia	SLO
Solomon Islands	SOL
Somalia	SOM
South Africa	RSA
South Korea	KOR
South Sudan	SSD
Spain	ESP
Sri Lanka	SRI
Sudan	SUD
Suriname	SUR
Swaziland	SWZ
Sweden	SWE
Switzerland	SUI
Syria	SYR
Tajikistan	TJK
Tanzania	TAN
Thailand	THA
The Gambia	GAM

Togo	TOG
Tonga	TGA
Trinidad and Tobago	TTO
Tunisia	TUN
Turkey	TUR
Turkmenistan	TKM
Tuvalu	TUV
Uganda	UGA
Ukraine	UKR
United Arab Emirates	UAE
United States	USA
Uruguay	URU
Uzbekistan	UZB
Vanuatu	VAN
Venezuela	VEN
Vietnam	VIE
Virgin Islands	ISV
Yemen	YEM
Zambia	ZAM
Zimbabwe	ZIM

Table 10.7: List of the 206 countries in the dataset

10.3 Appendix C: Structure of the dataset

	nocyear	noc	ioccode	year	medals	participation	loggdplag	pop	teamsize	host	posthost	prehost	tourism	loginfmortlag
1	AFG1996	Afghanistan	AFG	1996	0	1	NA	NA	1	0	0	0	NA	4.733563
2	AFG2000	Afghanistan	AFG	2000	NA	0	NA	NA	NA	NA	NA	NA	NA	4.616110
3	AFG2004	Afghanistan	AFG	2004	0	1	NA	NA	5	0	0	0	NA	4.506454
4	AFG2008	Afghanistan	AFG	2008	1	1	NA	NA	4	0	0	0	NA	4.383276
5	AFG2012	Afghanistan	AFG	2012	1	1	NA	NA	6	0	0	0	NA	4.239887
6	AFG2016	Afghanistan	AFG	2016	0	1	NA	NA	6	0	0	0	NA	4.085976
7	ALB1996	Albania	ALB	1996	0	1	7.685877	3.092228	7	0	0	0	NA	3.481240
8	ALB2000	Albania	ALB	2000	0	1	8.314810	3.121970	5	0	0	0	NA	3.328627
9	ALB2004	Albania	ALB	2004	0	1	8.406835	3.097747	7	0	0	0	NA	3.135494
10	ALB2008	Albania	ALB	2008	0	1	8.615278	2.991651	11	0	0	0	1.2470	2.906901
11	ALB2012	Albania	ALB	2012	0	1	8.955362	2.920039	12	0	0	0	3.1560	2.631889
12	ALB2016	Albania	ALB	2016	0	1	9.202155	2.926348	6	0	0	0	4.0700	2.322388

Figure 10.1: Structure of the dataset

"nocyear":	Unique identifier for each country-year combination
"noc":	NOC/country
"ioccode":	Identifier given to each NOC by the IOC
"year":	Olympic Games year
"medals":	Total number of medals won
"participation":	Whether the country participated in the Olympic in the given year
"loggdplag":	1st lag of the logarithm of per capita GDP, i.e. lagged 4 years (per capita GDP measured in millions 2011 US dollars)
"pop":	Population size (in millions)
"teamsize":	Number of Olympic athletes representing the country in the given year
"host":	Host dummy variable
"posthost":	Post-host dummy variable
"prehost":	Pre-host dummy variable
"tourism":	Number of tourist arrivals (in millions)
"loginfmortlag":	1st lag of the logarithm of the infant mortality rate, i.e. lagged 4 years

10.4 Appendix D: List of countries in the reduced sample

NOC	**IOC code**
Albania	ALB
Australia	AUS
Austria	AUT
Belgium	BEL
Bulgaria	BUL
Croatia	CRO
Cyprus	CYP
Czech Republic	CZE
Denmark	DEN
Estonia	EST
Finland	FIN
France	FRA
Germany	GER
Great Britain	GBR
Greece	GRE
Hong Kong	HKG
Hungary	HUN
Ireland	IRL
Italy	ITA
Japan	JPN
Kazakhstan	KAZ
Latvia	LAT
Lithuania	LTU
Luxembourg	LUX

Maldives	MDV
Malta	MLT
Myanmar	MYA
Netherlands	NED
Norway	NOR
Poland	POL
Portugal	POR
Romania	ROU
Slovakia	SVK
Slovenia	SLO
South Africa	RSA
Spain	ESP
Sweden	SWE
Switzerland	SUI

Table 10.8: List of the 38 countries in the reduced sample

10.5 Appendix E: Construction of the dataset

The basis for the constructed dataset was a list of participating countries in the 2012 and 2016 Olympics. Since medal tables only report medals counts for countries that win at least one medal (see for example Appendix A), these were not useful for gathering the full list of participating countries. This information instead came from the Olympic team size data, because for team size to be defined, a country must have participated.

The 2012 and 2016 Olympics were chosen so that the list would only reflect current countries. A Yugoslav team, for instance, participated in the earlier Olympics in the 1996-2016 period, but is no longer in existence as an NOC. Since including these non-current countries or entities would complicate the data processing, they were excluded from the final country list seen in Appendix B. The non-country Refugee Olympic Team and Independent Olympic Athletes were also excluded.

From the list of participating countries in Appendix B, a spreadsheet of each country-year combination for the 206 countries and 6 Olympic Games years was created. The IOC country identifier codes (see for example Appendix B) were matched to the ISO country identifier codes found in the non-Olympic data (GDP and population from the PWT, tourism and infant mortality from the World Bank, investment in sports from the IMF). Each value in the individual data across all variables could then be given a unique IOC code and year identifier for the country-year observation it represented (the "nocyear" in Appendix C) and be added to the spreadsheet to create the full dataset.

10.6 Appendix F: Descriptive statistics (R code)

R code for the descriptive statistics in Table 6.1:

```
library(stargazer)

stargazer(as.data.frame(ogmedals[c("medals",
                                   "loggdplag",
                                   "pop",
                                   "teamsize",
                                   "tourism",
                                   "loginfmortlag")]),
          type = "text",
          title = "Descriptive statistics",
          covariate.labels =
            c("Number of medals won",
              "log(GDP per capita in millions 2011 USD), 1st lag (4 years)",
              "Population (millions)",
              "Team size",
              "Number of tourists (millions)",
              "log(infant mortality rate), 1st lag (4 years)"),
          median = TRUE)
```

R code for the scatterplots in Figure 6.1-Figure 6.5:

```
library(tidyverse)

plot1 <- ggplot(ogmedals, aes(y = medals, x = teamsize)) +
  geom_point() +
  labs(x = "Team size",
       y = "Number of medals won")
print(plot1)

plot2 <- ggplot(ogmedals, aes(y = medals, x = tourism)) +
  geom_point() +
  labs(x = "Number of tourists (millions)",
       y = "Number of medals won")
print(plot2)

plot3 <- ggplot(ogmedals, aes(y = medals, x = pop)) +
  geom_point() +
  labs(x = "Population (millions)",
       y = "Number of medals won")
print(plot3)

plot4 <- ggplot(ogmedals, aes(y = medals, x = loggdplag)) +
  geom_point() +
  labs(x = "log(GDP per capita in millions 2011 USD), 1st lag (4 years)",
       y = "Number of medals won")
print(plot4)
```

```
plot5 <- ggplot(ogmedals, aes(y = loggdplag, x = loginfmortlag)) +
  geom_point() +
  labs(x = "log(infant mortality rate), 1st lag (4 years)",
       y = "log(GDP per capita in millions 2011 USD), \n1st lag (4 years)")
print(plot5)
```

R code for the correlation matrix in Table 6.2:

```
cor(ogmedals[, c("medals",
            "loggdplag",
            "pop",
            "teamsize",
            "tourism",
            "loginfmortlag")],
   use = "pairwise.complete.obs")
```

10.7 Appendix G: Descriptive statistics, reduced sample

Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Number of medals won	228	11.092	14.192	0	2	5.5	15	67
log(GDP per capita in millions 2011 USD), 1st lag (4 years)	228	9.957	0.755	6.712	9.582	10.171	10.482	11.423
Population (millions)	228	20.630	27.594	0.259	4.192	8.786	22.297	128.551
Team size	228	135.978	127.524	2	51	95.5	175.5	632
Number of tourists (millions)	220	11.569	16.196	0.202	2.198	5.996	12.994	82.682
Sports investment per capita in 2011 USD, 1st lag (4 years)	141	106.009	71.567	0.000	52.947	98.949	141.439	411.529
log(infant mortality rate), 1st lag (4 years)	222	1.928	0.796	0.788	1.367	1.723	2.238	4.359

Table 10.9: Descriptive statistics, reduced sample

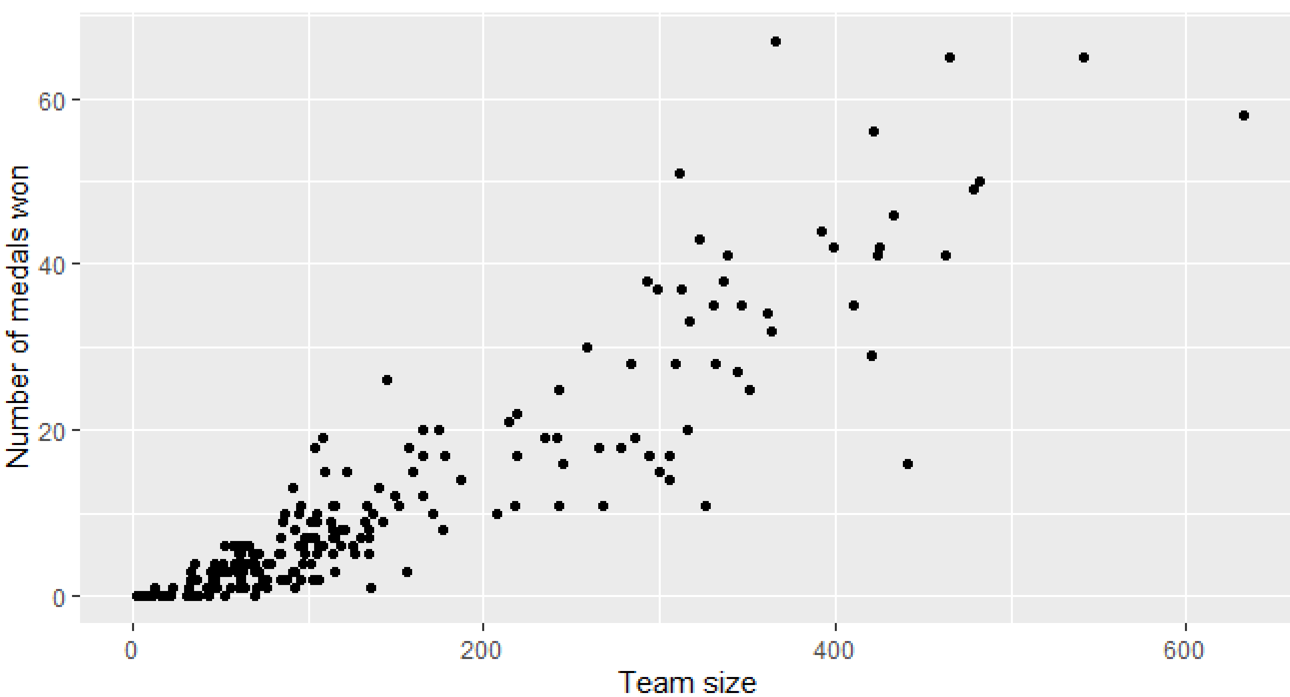

Figure 10.2: The number of medals won vs team size, reduced sample

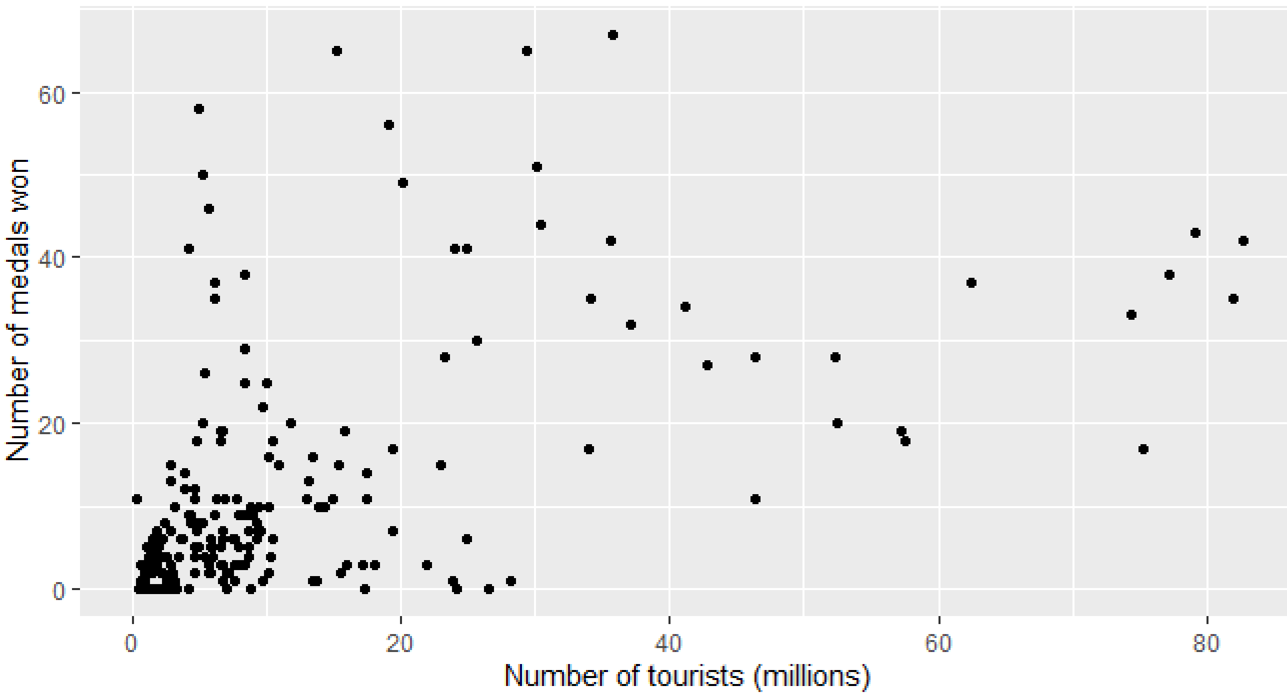

Figure 10.3: The number of medals won vs the number of tourists, reduced sample

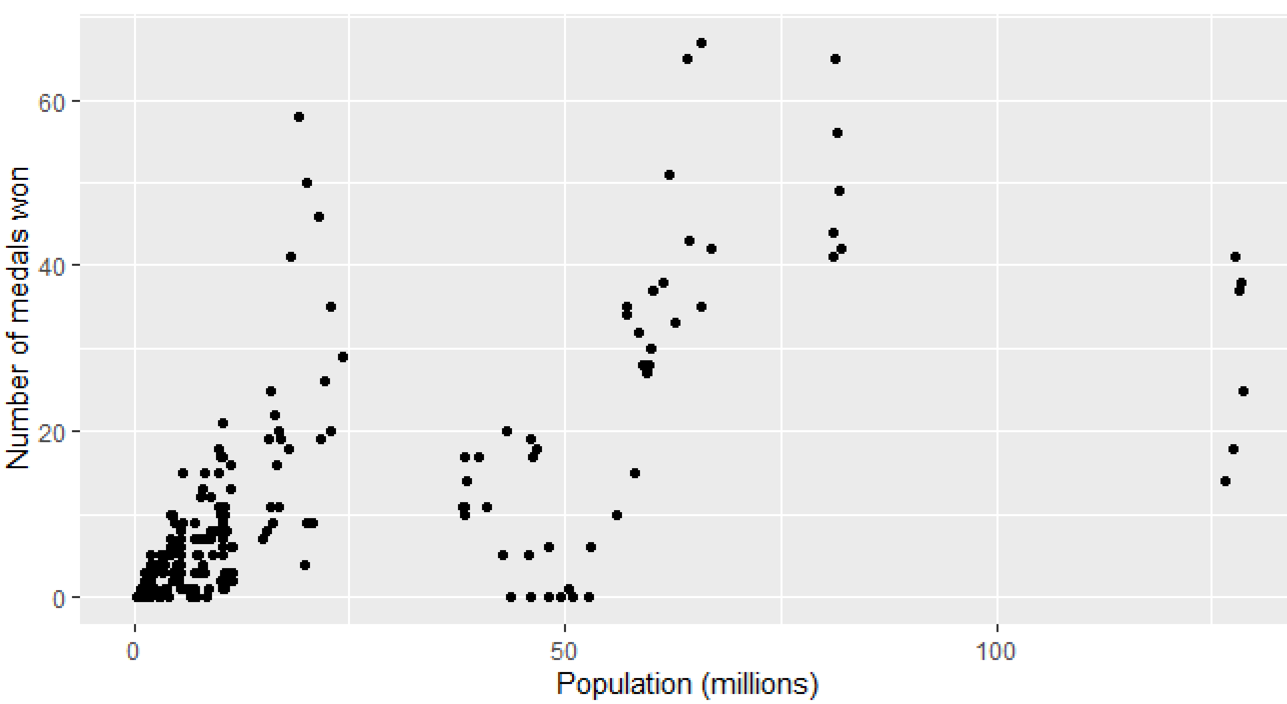

Figure 10.4: The number of medals won vs population, reduced sample

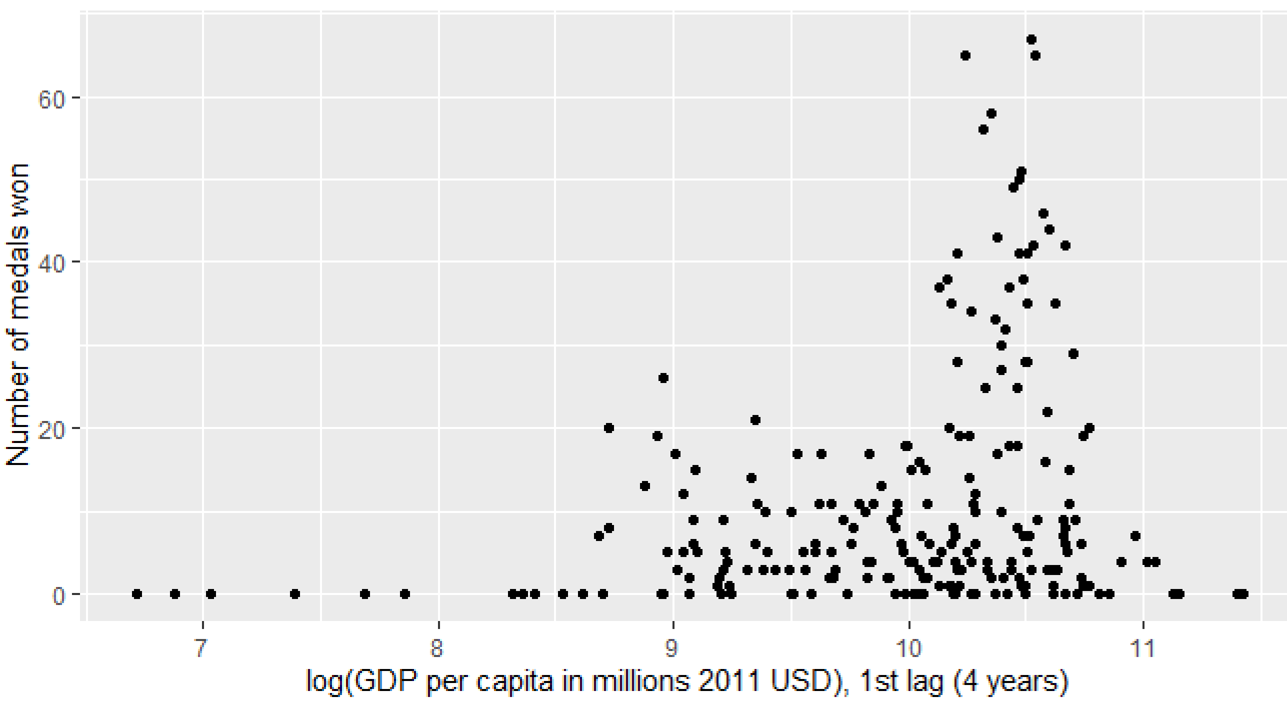

Figure 10.5: The number of medals won vs the lagged log of GDP per capita, reduced sample

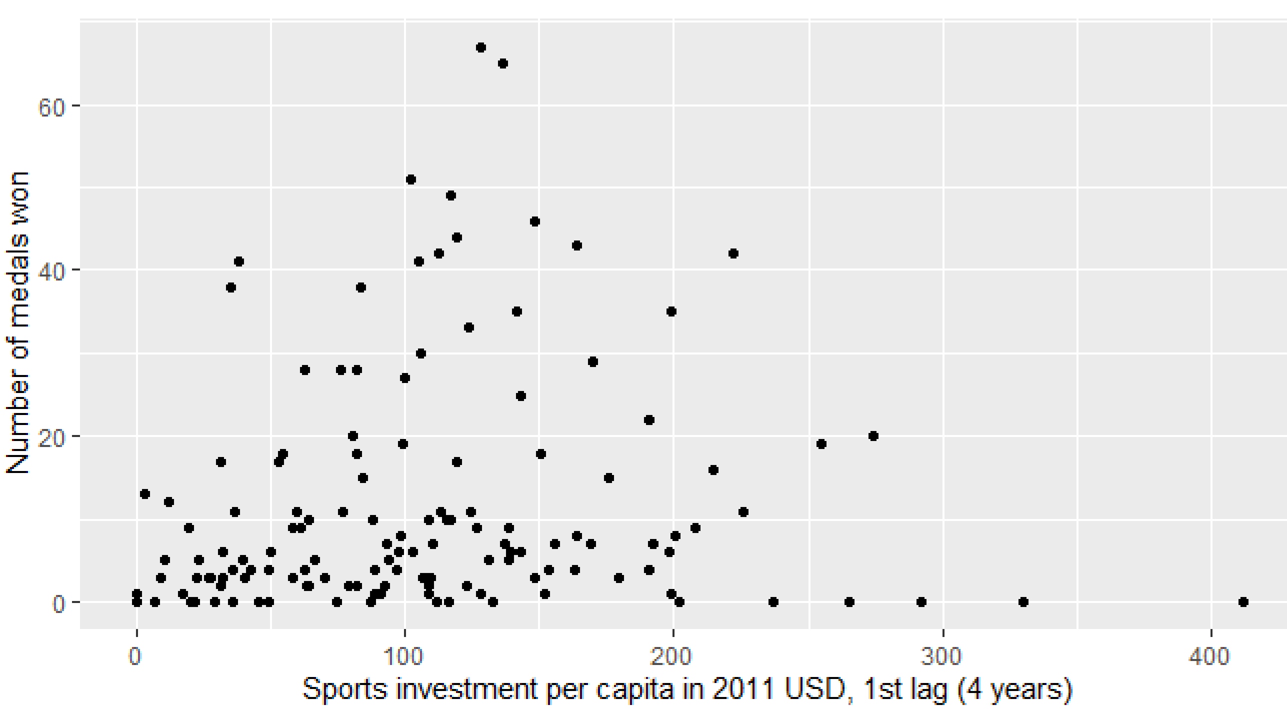

Figure 10.6: The number of medals won vs the lagged investment in sports per capita, reduced sample

Figure 10.7: The lagged log of GDP per capita vs the lagged log of the infant mortality rate, reduced sample

	medals	**log(gdp), lag**	**pop**	**teamsize**	**tourism**	**sportsinv, lag**	**log(infmort), lag**
medals	1.0000	0.2558	0.6713	0.9132	0.5416	0.1169	-0.2321
log(gdp), lag	0.2558	1.0000	0.0304	0.2972	0.2615	0.7384	-0.8892
pop	0.6713	0.0304	1.0000	0.6812	0.4998	-0.0097	-0.0359
teamsize	0.9132	0.2972	0.6812	1.0000	0.5799	0.1181	-0.2885
tourism	0.5416	0.2615	0.4998	0.5799	1.0000	0.1126	-0.2694
sportsinv, lag	0.1169	0.7384	-0.0097	0.1181	0.1126	1.0000	-0.5599
log(infmort), lag	-0.2321	-0.8892	-0.0359	-0.2885	-0.2694	-0.5599	1.0000

Table 10.10: Correlation matrix, reduced sample

R code for the descriptive statistics in Table 10.9:

```
library(stargazer)
```

```
stargazer(as.data.frame(ogmedalsreduced[c("medals",
                                          "loggdplag",
                                          "pop",
                                          "teamsize",
                                          "tourism",
                                          "sportsinvlag",
                                          "loginfmortlag")]),
          type = "text",
          title = "Descriptive statistics",
          covariate.labels =
            c("Number of medals won",
              "log(GDP per capita in millions 2011 USD), 1st lag (4 years)",
              "Population (millions)",
              "Team size",
              "Number of tourists (millions)",
              "Sports investment per capita in 2011 USD, 1st lag (4 years)",
              "log(infant mortality rate), 1st lag (4 years)"),
          median = TRUE)
```

R code for the scatterplots in Figure 10.2-Figure 10.7:

```
library(tidyverse)

plot1 <- ggplot(ogmedalsreduced, aes(y = medals, x = teamsize)) +
  geom_point() +
```

```
  labs(x = "Team size",
       y = "Number of medals won")
print(plot1)

plot2 <- ggplot(ogmedalsreduced, aes(y = medals, x = tourism)) +
  geom_point() +
  labs(x = "Number of tourists (millions)",
       y = "Number of medals won")
print(plot2)

plot3 <- ggplot(ogmedalsreduced, aes(y = medals, x = pop)) +
  geom_point() +
  labs(x = "Population (millions)",
       y = "Number of medals won")
print(plot3)

plot4 <- ggplot(ogmedalsreduced, aes(y = medals, x = loggdplag)) +
  geom_point() +
  labs(x = "log(GDP per capita in millions 2011 USD), 1st lag (4 years)",
       y = "Number of medals won")
print(plot4)

plot5 <- ggplot(ogmedalsreduced, aes(y = medals, x = sportsinvlag)) +
  geom_point() +
  labs(x = "Sports investment per capita in 2011 USD, 1st lag (4 years)",
       y = "Number of medals won")
print(plot5)
```

```
plot6 <- ggplot(ogmedalsreduced, aes(y = loggdplag, x = loginfmortlag)) +
  geom_point() +
  labs(x = "log(infant mortality rate), 1st lag (4 years)",
       y = "log(GDP per capita in millions 2011 USD), \n1st lag (4 years)")
print(plot6)
```

R code for the correlation matrix in Table 10.10:

```
cor(ogmedalsreduced[, c("medals",
                 "loggdplag",
                 "pop",
                 "teamsize",
                 "tourism",
                 "sportsinvlag",
                 "loginfmortlag")],
    use = "pairwise.complete.obs")
```

10.8 Appendix H: Fixed effects regression model (R code)

R code for the TSLS first-stage regression F-statistic using HAC standard errors in Figure 6.6:

```
library(AER)

firststage <- lm(loggdplag ~ loginfmortlag +
           pop +
           teamsize +
           host +
           posthost +
           prehost +
           tourism +
           ioccode +
           year,
          data = ogmedals)

linearHypothesis(firststage,
         "loginfmortlag = 0", vcov = vcovHAC)
```

R code for the fixed effects regression results in Table 6.3:

Clustered standard errors are employed with “vcov = function(x) vcovHC(x, method = "arellano")”, where “arellano” allows for HAC standard errors (Croissant & Millo, 2008, p. 31).

```
library(plm)

fe_model <- plm(medals ~ loggdplag +
          pop +
          teamsize +
          host +
          posthost +
          prehost +
          tourism | . - loggdplag + loginfmortlag,
        data = ogmedals,
        index = c("ioccode", "year"),
        model = "within",
        effect = "twoways")

fe_model_reduced <- plm(medals ~ loggdplag +
          pop +
          teamsize +
          host +
          posthost +
          prehost +
          tourism +
          sportsinvlag | . - loggdplag + loginfmortlag,
```

```
            data = ogmedalsreduced,
            index = c("ioccode", "year"),
            model = "within",
            effect = "twoways")

library(stargazer)

models <- list(fe_model, fe_model_reduced)

coeftest <- lapply(models, coeftest,
              vcov = function(x) vcovHC(x, method = "arellano"))

stargazer(models, type = "text", covariate.labels = c("log(gdp), lag",
                                        "pop",
                                        "teamsize",
                                        "host",
                                        "posthost",
                                        "prehost",
                                        "tourism",
                                        "sportsinv, lag"),
       se = lapply(coeftest, function(x) {
  x[, 2]
}))
```

R code for the Lagrange multiplier test in Figure 6.7:

```
plmtest(medals ~ loggdplag +
      pop +
      teamsize +
      host +
      posthost +
      prehost +
      tourism | . - loggdplag + loginfmortlag,
     data = ogmedals,
     index = c("ioccode", "year"),
     effect = "twoways")
```

R code for the F-test in Figure 6.8:

```
pFtest(medals ~ loggdplag +
      pop +
      teamsize +
      host +
      posthost +
      prehost +
      tourism | . - loggdplag + loginfmortlag,
     data = ogmedals,
     index = c("ioccode", "year"),
     effect = "twoways")
```

The R code for the hypothetical Lagrange multiplier test and F-test for time fixed effects only in Figure 6.9 and Figure 6.10 is identical to the two sections of code on the previous page, except with the argument "effect = "time"" instead of "effect = "twoways"".

Likewise, the R code for the hypothetical Lagrange multiplier test and F-test for country fixed effects only in Figure 10.8 and Figure 10.9 below is identical to the two sections of code on the previous page, except with the argument "effect = "individual"" instead of "effect = "twoways"".

```
        Lagrange Multiplier Test - (Honda) for unbalanced panels

data:  medals ~ loggdplag + pop + teamsize + host + posthost + prehost +  ...
normal = 35.39, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Figure 10.8: Lagrange multiplier test for hypothetical model with country fixed effects only

```
        F test for individual effects

data:  medals ~ loggdplag + pop + teamsize + host + posthost + prehost +  ...
F = 16.736, df1 = 167, df2 = 743, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Figure 10.9: F-test for hypothetical model with time fixed effects only

10.9 Appendix I: Fixed effects regression model, reduced sample

The fixed effects regression model to be estimated for the reduced sample of 38 countries is the following:

$$medals_{i,t} = \beta_1 \log(gdp)_{i,t-1} + \beta_2 pop_{i,t} + \beta_3 teamsize_{i,t} + \beta_4 host_{i,t} + \beta_5 posthost_{i,t} + \beta_6 prehost_{i,t} + \beta_7 tourism_{i,t} + \beta_8 sportsinv_{i,t-1} + \alpha_i + \lambda_t + u_{i,t} \quad (6)$$

Variables are defined as in the full model with these differences:

i:	country $i = 1, \dots, 38$
$sportsinv_{i,t-1}$:	the real investment in sports per capita for country i in time period $t-1$ (i.e. lagged 1 time period/four years), measured in 2011 US dollars

Tests for country and time fixed effects in the reduced sample:

```
        Lagrange Multiplier Test - two-ways effects (Honda) for unbalanced panels

data:  medals ~ loggdplag + pop + teamsize + host + posthost + prehost +  ...
normal = 3.2241, p-value = 0.0006319
alternative hypothesis: significant effects
```

Figure 10.10: Lagrange multiplier test for country and time fixed effects, reduced sample

```
        F test for twoways effects

data:  medals ~ loggdplag + pop + teamsize + host + posthost + prehost +  ...
F = 4.3731, df1 = 41, df2 = 85, p-value = 4.771e-09
alternative hypothesis: significant effects
```

Figure 10.11: F-test for country and time fixed effects, reduced sample

R code for the Lagrange multiplier test in Figure 10.10:

```
library(plm)

plmtest(medals ~ loggdplag +
          pop +
          teamsize +
          host +
          posthost +
          prehost +
          tourism +
          sportsinvlag | . - loggdplag + loginfmortlag,
        data = ogmedalsreduced,
        index = c("ioccode", "year"),
        effect = "twoways")
```

R code for the F-test in Figure 10.11:

```
library(plm)

pFtest(medals ~ loggdplag +
         pop +
         teamsize +
         host +
         posthost +
         prehost +
```

```
  tourism +
  sportsinvlag | . - loggdplag + loginfmortlag,
data = ogmedalsreduced,
index = c("ioccode", "year"),
effect = "twoways")
```

10.10 Appendix J: Fixed effect regression model, country and time fixed effects

	Estimate	Std. Error	t-value	Pr(>|t|)	
ALB	31.5450	14.6362	2.1553	0.031464	*
ALG	31.4971	14.9291	2.1098	0.035213	*
ANG	27.5181	13.2663	2.0743	0.038398	*
ANT	34.9056	15.7544	2.2156	0.027022	*
ARG	27.3009	15.3934	1.7735	0.076551	.
ARM	31.4519	13.7831	2.2819	0.022778	*
AUS	48.0343	17.5129	2.7428	0.006240	**
AUT	31.9389	17.1551	1.8618	0.063032	.
AZE	38.2116	14.3454	2.6637	0.007898	**
BAH	37.6629	16.8651	2.2332	0.025835	*
BAN	23.1960	11.5145	2.0145	0.044319	*
BAR	35.6551	16.1384	2.2093	0.027458	*
BDI	23.6575	10.6569	2.2199	0.026728	*
BEL	35.5430	17.0440	2.0854	0.037379	*
BEN	26.4023	11.9779	2.2043	0.027815	*
BHU	30.6335	13.7052	2.2352	0.025704	*
BIH	31.4248	14.2822	2.2003	0.028097	*
BIZ	32.2989	14.4822	2.2303	0.026031	*
BLR	36.9016	15.3171	2.4092	0.016232	*
BOL	29.2518	13.3501	2.1911	0.028754	*
BOT	33.1479	15.0596	2.2011	0.028036	*
BRA	24.0331	14.3965	1.6694	0.095468	.
BRN	35.7805	16.6200	2.1529	0.031653	*
BRU	40.7924	18.2569	2.2344	0.025759	*
BUL	35.5056	15.1426	2.3447	0.019304	*
BUR	25.1194	11.4457	2.1946	0.028500	*
CAF	24.5937	11.1327	2.2091	0.027471	*
CAM	26.1502	12.0648	2.1675	0.030518	*
CAN	33.5813	17.3015	1.9410	0.052644	.
CGO	29.6288	13.5088	2.1933	0.028597	*
CHA	26.5167	12.0463	2.2012	0.028029	*
CHI	32.2530	15.4228	2.0913	0.036847	*
CHN	32.9957	14.7917	2.2307	0.026001	*

CIV	27.6548	12.6886	2.1795	0.029610	*
CMR	27.1396	12.9160	2.1012	0.035959	*
COD	21.7063	10.2513	2.1174	0.034558	*
COL	30.5773	14.5772	2.0976	0.036279	*
COM	28.8832	12.9624	2.2282	0.026165	*
CPV	29.9174	13.4399	2.2260	0.026315	*
CRC	33.2546	15.0730	2.2062	0.027675	*
CRO	32.5148	15.6348	2.0796	0.037903	*
CYP	36.1250	16.7090	2.1620	0.030938	*
CZE	36.8817	16.4510	2.2419	0.025264	*
DEN	38.9630	17.1295	2.2746	0.023215	*
DJI	28.5820	12.8804	2.2190	0.026788	*
DMA	33.1664	14.8490	2.2336	0.025810	*
DOM	31.2962	14.5927	2.1446	0.032307	*
ECU	30.6407	14.3503	2.1352	0.033074	*
EGY	24.4516	13.6489	1.7915	0.073627	.
ESA	29.9793	13.7218	2.1848	0.029217	*
ESP	22.3751	16.7687	1.3343	0.182505	
EST	34.1078	15.7008	2.1724	0.030146	*
ETH	26.8935	10.3073	2.6092	0.009260	**
FIJ	31.0956	14.2916	2.1758	0.029888	*
FIN	36.5735	16.9513	2.1576	0.031284	*
FRA	38.2989	17.1231	2.2367	0.025605	*
GAB	33.0987	14.9759	2.2101	0.027402	*
GAM	28.4972	12.8244	2.2221	0.026578	*
GBR	48.8372	16.9210	2.8862	0.004013	**
GBS	26.2154	11.8869	2.2054	0.027734	*
GEO	36.2698	14.6936	2.4684	0.013797	*
GER	53.4425	17.1997	3.1072	0.001961	**
GHA	27.4662	13.0267	2.1085	0.035328	*
GRE	27.8229	16.4711	1.6892	0.091604	.
GRN	32.9246	14.6989	2.2399	0.025392	*
GUA	30.4638	14.1870	2.1473	0.032094	*
GUI	27.5646	12.4847	2.2079	0.027560	*

HAI	26.0718	11.9052	2.1899	0.028840	*
HON	28.2913	13.2284	2.1387	0.032789	*
HUN	38.2095	16.4020	2.3296	0.020098	*
INA	26.2576	12.9655	2.0252	0.043207	*
IND	-4.2951	13.4109	-0.3203	0.748856	
IRI	35.0888	14.8297	2.3661	0.018233	*
IRL	35.3046	17.0465	2.0711	0.038698	*
IRQ	29.0150	13.3054	2.1807	0.029520	*
ISL	37.6942	17.2145	2.1897	0.028860	*
ISR	35.9376	16.7931	2.1400	0.032681	*
ITA	37.9365	17.0417	2.2261	0.026309	*
JAM	37.4267	14.2777	2.6213	0.008939	**
JOR	30.0854	13.8239	2.1763	0.029847	*
JPN	41.6854	16.7479	2.4890	0.013030	*
KAZ	36.3010	15.1329	2.3988	0.016696	*
KEN	33.9446	12.3690	2.7443	0.006211	**
KGZ	28.8785	13.1693	2.1929	0.028629	*
KOR	45.8332	16.4651	2.7837	0.005512	**
KSA	33.2888	16.6428	2.0002	0.045846	*
KUW	37.6819	17.4631	2.1578	0.031265	*
LAO	27.5227	12.5365	2.1954	0.028445	*
LAT	33.9753	15.4909	2.1932	0.028601	*
LCA	33.1897	14.9260	2.2236	0.026476	*
LES	27.4297	12.4442	2.2042	0.027817	*
LIB	32.6031	14.8178	2.2003	0.028098	*
LTU	34.5444	15.6304	2.2101	0.027406	*
LUX	40.0802	18.1473	2.2086	0.027509	*
MAD	25.0941	11.5592	2.1709	0.030256	*
MAR	28.2191	13.8280	2.0407	0.041634	*
MAS	30.6985	15.5889	1.9693	0.049297	*
MAW	24.6434	11.2262	2.1952	0.028462	*
MDA	27.5851	13.0190	2.1188	0.034439	*
MDV	32.9084	14.8311	2.2189	0.026799	*
MEX	25.7498	15.0765	1.7079	0.088068	.

MGL	31.3618	13.7581	2.2795	0.022921	*
MKD	32.2815	14.6027	2.2107	0.027366	*
MLI	25.1145	11.5541	2.1736	0.030049	*
MLT	35.9971	16.3121	2.2068	0.027637	*
MNE	32.2554	15.3047	2.1075	0.035407	*
MOZ	23.6092	11.0020	2.1459	0.032207	*
MRI	34.1402	15.6184	2.1859	0.029136	*
MTN	28.0408	12.8970	2.1742	0.030007	*
MYA	25.3241	11.8482	2.1374	0.032896	*
NAM	31.6798	14.3076	2.2142	0.027120	*
NCA	29.0867	13.3167	2.1842	0.029260	*
NED	42.1846	17.3313	2.4340	0.015169	*
NEP	25.3866	11.7110	2.1678	0.030495	*
NGR	21.3045	11.3710	1.8736	0.061383	.
NIG	23.7704	10.7724	2.2066	0.027649	*
NOR	40.1988	17.4567	2.3028	0.021569	*
NZL	36.1214	16.9068	2.1365	0.032967	*
OMA	38.1105	17.2785	2.2057	0.027715	*
PAK	23.8322	12.4144	1.9197	0.055278	.
PAN	33.4313	15.0870	2.2159	0.027002	*
PAR	30.3956	13.8187	2.1996	0.028145	*
PER	29.6483	14.0938	2.1036	0.035747	*
PHI	27.4630	13.3683	2.0543	0.040295	*
PLE	29.5914	13.3334	2.2193	0.026767	*
POL	29.8642	15.6341	1.9102	0.056496	.
POR	31.6024	16.3446	1.9335	0.053556	.
QAT	42.8325	19.4035	2.2075	0.027589	*
ROU	38.9102	15.0750	2.5811	0.010040	*
RSA	28.9410	14.8349	1.9509	0.051450	.
RUS	72.3643	15.4823	4.6740	3.511e-06	***
RWA	24.7327	11.3585	2.1775	0.029763	*
SEN	26.9143	12.8500	2.0945	0.036556	*
SEY	35.6050	16.0904	2.2128	0.027216	*
SGP	36.8013	17.3856	2.1168	0.034613	*

SKN	35.0727	15.7593	2.2255	0.026347	*
SLE	25.4497	11.5249	2.2082	0.027535	*
SLO	36.0945	16.4640	2.1923	0.028667	*
SRB	32.2326	15.2588	2.1124	0.034989	*
SRI	30.2712	13.8767	2.1814	0.029466	*
STP	27.6038	12.3862	2.2286	0.026141	*
SUD	26.9853	12.3386	2.1871	0.029051	*
SUI	37.9939	17.5482	2.1651	0.030698	*
SUR	33.0224	14.8511	2.2236	0.026480	*
SVK	35.9426	15.9906	2.2477	0.024888	*
SWE	36.8542	17.1678	2.1467	0.032142	*
SWZ	32.1546	14.4837	2.2200	0.026719	*
SYR	25.9802	11.9718	2.1701	0.030316	*
TAN	24.6858	11.5443	2.1384	0.032816	*
THA	30.5574	14.6125	2.0912	0.036854	*
TJK	28.0938	12.5117	2.2454	0.025038	*
TKM	32.0085	14.5641	2.1978	0.028276	*
TOG	26.0840	11.7314	2.2234	0.026489	*
TTO	36.1353	15.8974	2.2730	0.023311	*
TUN	29.6701	14.7424	2.0126	0.044523	*
TUR	30.4707	15.2381	1.9996	0.045905	*
UAE	39.9959	18.4764	2.1647	0.030730	*
UGA	24.5208	11.4569	2.1403	0.032661	*
UKR	33.9370	14.4865	2.3427	0.019411	*
URU	33.1650	15.3768	2.1568	0.031342	*
USA	87.9096	16.8905	5.2047	2.521e-07	***
UZB	31.2155	13.7256	2.2743	0.023236	*
VEN	29.8985	14.9316	2.0024	0.045612	*
VIE	25.5750	12.4095	2.0609	0.039660	*
VIN	32.7364	14.6877	2.2288	0.026126	*
YEM	25.5617	11.6873	2.1871	0.029046	*
ZAM	26.3583	12.0232	2.1923	0.028671	*
ZIM	27.9167	12.6133	2.2133	0.027185	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 10.11: Country fixed effects

```
     Estimate Std. Error t-value Pr(>|t|)
1996   31.264     13.914  2.2470  0.02494 *
2000   31.605     14.002  2.2572  0.02429 *
2004   31.810     14.168  2.2453  0.02504 *
2008   31.727     14.223  2.2306  0.02601 *
2012   32.783     14.629  2.2409  0.02533 *
2016   32.914     14.859  2.2150  0.02706 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 10.12: Time fixed effects

R code for the country fixed effects in Table 10.11 and the time fixed effects in Table 10.12:

```
library(plm)

fe_model <- plm(medals ~ loggdplag +
          pop +
          teamsize +
          host +
          posthost +
          prehost +
          tourism | . - loggdplag + loginfmortlag,
        data = ogmedals,
        index = c("ioccode", "year"),
        model = "within",
        effect = "twoways")
summary(fixef(fe_model))
summary(fixef(fe_model, effect = "time"))
```

10.11 Appendix K: Fixed effects regression model, team size as dependent variable

The fixed effects regression model with team size as the dependent variable is defined as follows:

$$teamsize_{i,t} = \beta_1 \log(gdp)_{i,t-1} + \beta_2 pop_{i,t} + \beta_3 host_{i,t} + \beta_4 posthost_{i,t} + \beta_5 prehost_{i,t} + \beta_6 tourism_{i,t} + \alpha_i + \lambda_t + u_{i,t} \quad (7)$$

As in the original regression of Olympic medals, the (lagged) logarithm of the infant mortality rate is used as an instrument because the (lagged) logarithm of GDP per capita is an endogenous variable. Running the TSLS first-stage regression defined in (8 below gives an F-statistic of 25.64 (see Figure 10.12), which exceeds 10. Thus, the logarithm of the infant mortality rate is not a weak instrument.

$$\log(gdp)_{i,t-1} = \pi_1 \log(infmort)_{i,t-1} + \pi_2 pop_{i,t} + \pi_3 host_{i,t} + \pi_4 posthost_{i,t} + \pi_5 prehost_{i,t} + \pi_6 tourism_{i,t} + \alpha_i + \lambda_t + v_{i,t} \quad (8)$$

```
Linear hypothesis test

Hypothesis:
loginfmortlag = 0

Model 1: restricted model
Model 2: loggdplag ~ loginfmortlag + pop + host + posthost + prehost +
    tourism + ioccode + year

Note: Coefficient covariance matrix supplied.

  Res.Df Df     F    Pr(>F)
1    740
2    739  1 25.64 5.201e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10.12: TSLS first-stage regression F-statistic without team size as an exogenous regressor

Tests for two-way fixed effects indicate that there are significant country and time fixed effects (see Figure 10.13-Figure 10.16).

```
        Lagrange Multiplier Test - two-ways effects (Honda) for unbalanced panels

data:  teamsize ~ loggdplag + pop + host + posthost + prehost + tourism |  ...
normal = 31.399, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Figure 10.13: Lagrange multiplier test for country and time fixed effects with team size as dependent variable

```
        F test for twoways effects

data:  teamsize ~ loggdplag + pop + host + posthost + prehost + tourism |  ...
F = 78.524, df1 = 172, df2 = 739, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Figure 10.14: F-test for country and time fixed effects with team size as dependent variable

```
        Lagrange Multiplier Test - two-ways effects (Honda) for unbalanced panels

data:  teamsize ~ loggdplag + pop + host + posthost + prehost + tourism +  ...
normal = 6.3109, p-value = 1.387e-10
alternative hypothesis: significant effects
```

Figure 10.15: Lagrange multiplier test for country and time fixed effects with team size as dependent variable, reduced sample

```
        F test for twoways effects

data:  teamsize ~ loggdplag + pop + host + posthost + prehost + tourism +  ...
F = 32.418, df1 = 41, df2 = 86, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Figure 10.16: F-test for country and time fixed effects with team size as dependent variable, reduced sample

Running the Lagrange multiplier test and F-test for a hypothetical model in which only time fixed effects are included shows that there are significant time fixed effects (see Figure 10.17 and Figure 10.18).

```
        Lagrange Multiplier Test - time effects (Honda) for unbalanced panels

data:  teamsize ~ loggdplag + pop + host + posthost + prehost + tourism |  ...
normal = 4.9962, p-value = 2.924e-07
alternative hypothesis: significant effects
```

Figure 10.17: Lagrange multiplier test for hypothetical model with time fixed effects only with team size as dependent variable

```
        F test for time effects

data:  teamsize ~ loggdplag + pop + host + posthost + prehost + tourism |  ...
F = 3.8178, df1 = 5, df2 = 906, p-value = 0.001987
alternative hypothesis: significant effects
```

Figure 10.18: F-test for hypothetical model with time fixed effects only with team size as dependent variable

Running the Lagrange multiplier test and F-test for a hypothetical model in which only country fixed effects are included shows that there are significant country fixed effects (see Figure 10.19 and Figure 10.20).

```
        Lagrange Multiplier Test - (Honda) for unbalanced panels

data:  teamsize ~ loggdplag + pop + host + posthost + prehost + tourism |  ...
normal = 39.408, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Figure 10.19: Lagrange multiplier test for hypothetical model with country fixed effects only with team size as dependent variable

```
        F test for individual effects

data:  teamsize ~ loggdplag + pop + host + posthost + prehost + tourism |  ...
F = 81.007, df1 = 167, df2 = 744, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Figure 10.20: F-test for hypothetical model with country fixed effects only with team size as dependent variable

R code for the TSLS first-stage regression using HAC standard errors in Figure 10.12:

```
library(AER)

firststage1_ts <- lm(loggdplag ~ loginfmortlag +
                     pop +
                     host +
                     posthost +
                     prehost +
                     tourism +
                     ioccode +
                     year,
                    data = ogmedals)

linearHypothesis(firststage1_ts,
              "loginfmortlag = 0", vcov = vcovHAC)
```

R code for the Lagrange multiplier test and F-test in Figure 10.13 and Figure 10.14:

```
plmtest(teamsize ~ loggdplag +
        pop +
        host +
        posthost +
        prehost +
        tourism | . - loggdplag + loginfmortlag,
       data = ogmedals,
```

```
        index = c("ioccode", "year"),
        effect = "twoways")

pFtest(teamsize ~ loggdplag +
         pop +
         host +
         posthost +
         prehost +
         tourism | . - loggdplag + loginfmortlag,
       data = ogmedals,
       index = c("ioccode", "year"),
       effect = "twoways")
```

The R code for the hypothetical Lagrange multiplier tests and F-tests for time fixed effects only in Figure 10.17 and Figure 10.18 and for country fixed effects only in Figure 10.19 and Figure 10.20 is identical to the two sections of code above, except with the arguments “effect = “time”” and “effect = “individual”” respectively instead of “effect = “twoways””.

R code for the Lagrange multiplier test and F-test for the reduced sample in Figure 10.15 and Figure 10.16:

```
plmtest(teamsize ~ loggdplag +
          pop +
          host +
          posthost +
          prehost +
          tourism +
```

```
  sportsinvlag | . - loggdplag + loginfmortlag,
 data = ogmedalsreduced,
 index = c("ioccode", "year"),
 effect = "twoways")

pFtest(teamsize ~ loggdplag +
  pop +
  host +
  posthost +
  prehost +
  tourism +
  sportsinvlag | . - loggdplag + loginfmortlag,
 data = ogmedalsreduced,
 index = c("ioccode", "year"),
 effect = "twoways")
```

R code for the fixed effects regression results with team size as the dependent variable using clustered standard errors in Table 6.4:

```
library(plm)

fe_model_ts <- plm(teamsize ~ loggdplag +
          pop +
          host +
          posthost +
          prehost +
```

```
    tourism | . - loggdplag + loginfmortlag,
  data = ogmedals,
  index = c("ioccode", "year"),
  model = "within",
  effect = "twoways")

fe_model_reduced_ts <- plm(teamsize ~ loggdplag +
          pop +
          host +
          posthost +
          prehost +
          tourism +
          sportsinvlag | . - loggdplag + loginfmortlag,
        data = ogmedalsreduced,
        index = c("ioccode", "year"),
        model = "within",
        effect = "twoways")

library(stargazer)

models_ts <- list(fe_model_ts, fe_model_reduced_ts)

coeftest_ts <- lapply(models_ts, coeftest,
        vcov = function(x) vcovHC(x, method = "arellano"))

stargazer(models_ts, type = "text", covariate.labels = c("log(gdp), lag",
                     "pop",
```

```
                      "host",
                      "posthost",
                      "prehost",
                      "tourism",
                      "sportsinv, lag"),
se = lapply(coeftest_ts, function(x) {
  x[, 2]
}))
```

10.12 Appendix L: Zero-inflated count model (R code)

R code for the over-dispersion test in Figure 6.11:

```
library(AER)

medals_poisson <- glm(medals ~ loggdplag +
                      pop +
                      teamsize +
                      host +
                      posthost +
                      prehost +
                      tourism, data = ogmedalszi, family = poisson)

dispersiontest(medals_poisson)
```

R code for the ZINB regression results in Table 6.5 and Table 6.6:

```
library(pscl)

medals_zinb <- zeroinfl(medals ~ loggdplag +
                        pop +
                        teamsize +
                        host +
                        posthost +
                        prehost +
```

```
              tourism | participation,
             data = ogmedalszi,
             dist = "negbin")

library(stargazer)

stargazer(medals_zinb, type = "text", zero.component = FALSE,
      covariate.labels = c("log(gdp), lag",
                  "pop",
                  "teamsize",
                  "host",
                  "posthost",
                  "prehost",
                  "tourism"))

stargazer(medals_zinb, type = "text", zero.component = TRUE,
      covariate.labels = c("participation"))
```