MATH5743M: Statistical Learning - Assessed Practicals

Dr Seppo Virtanen, School of Mathematics, University of Leeds

Semester 2: 2022

Assessed Practical I: Predicting the Olympic Games

Assignment due: 11:59pm Friday March 18

The Olympic Games have always mixed pure sporting spectacle with national competition. During the Cold War the USA and the Soviet Union competed fiercely to win the most medals in each games. On a somewhat milder level, in Britain we often compare our medal count with that of Australia, one of our traditional sporting rivals. If you were in the UK during the summers of 2012 and 2016 you cannot have missed the excitement caused by the UK's success relative to previous years.

This competition is usually expressed in terms of the number of medals won by each country's athletes ((Figure 1 top panel). However, many interested watchers, especially those from smaller countries, have pointed out that the medal table is hardly a fair reflection of a country's sporting prowess. Some countries have a strong tradition of sporting excellence, but are simply too small to make an impact in terms of total medals. These commentators would rather look at the per capita medal count (Figure 1 bottom panel).

Looking at the per capita map above though, we see that large areas of the world are still very under-represented. Specifically, poorer countries do not win many medals per head of population. There are many reasons for this, including a lack of investment in sport and facilities, and fewer individuals who are wealthy enough to devote their life to training. As such, it has been suggested that we should compensate for wealth when measuring a country's Olympic performance.

In this practical you will investigate how the number of medals a country wins can be predicted from national population and GDP, and how consistent these relationships are, using the glm function in R.

Please check and follow the instructions given in Practicals 1 on documenting your work. You need to return your report using Turnitin in Minerva by the deadline.

Begin by downloading the data file medal_pop_gdp_data_statlearn.csv from Minerva.

This data file contains the following information for 71 countries (those that won at least one gold medal in each of the last three games):

- Country name (as recognised by the IOC)
- Population
- GDP (in billions of US dollars)
- Medals won in Beijing 2008, London 2012 and Rio 2016

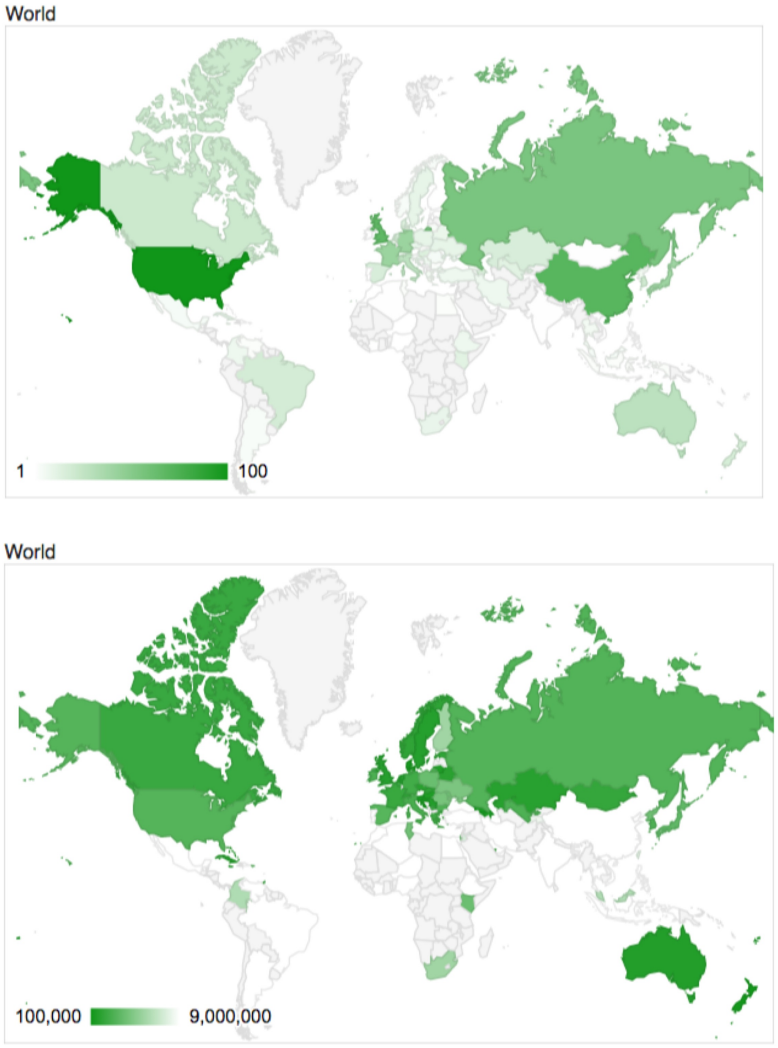

Figure 1: Total medals per country (top) and medals per capita (bottom) in the Rio 2016 Olympic Games (credit: http://www.medalspercapita.com/)

Tasks

1. Perform linear regression for the medal count in 2012 from Population and GDP, and make a prediction for the results of 2016. Explain the model. Report results and comment on your findings. Plot your predictions against the actual results of 2016. If the results are hard to see, use a transformation of the axes to improve clarity. Comment on your findings. How good are the predictions? Which countries are outliers from the trend? (Hint: Assume that the GBP and Population for each country do not significantly change from 2012 to 2016.)
2. Repeat the task 1 for linear regression for log-transformed medal counts. In addition, discuss differences and potential benefits of using the log-transformation compared to modelling raw medal counts. (Hint: For the predictions remember to transform the predicted counts suitably so that you can compare them to the actual medal counts.)
3. Repeat the task 1 for Poisson regression. (Hint: Remember to specify suitably the family argument for

the glm function. For the predict function remember to specify type="response".)

4. Repeat the task 1 for Negative Binomial regression. (Hint: You need to use MASS package and specify the family argument for the glm function using family=negative.binomial(theta = thetaVal). For the predict function remember to specify type="response".). In addition, try different values for theta ranging from 0.001 to 1000 and comment on how the results change. Find an optimal value for theta that gives the best predictive performance.

5. Perform model selection for the four different models, tasks 1 to 4, respectively, and report your results. For Negative Binomial regression, use the model with your optimal value for theta. Which model would you choose to accurately predict the medal count? Justify your reasoning carefully.