Olympic medals: Does the past predict the future?

How many medals will each country win in Rio this August? Recent history provides some strong indications, but a more sophisticated approach might produce better predictions. **Julia Bredtmann**, **Carsten J. Crede** and **Sebastian Otten** investigate

The 31st Olympic Games gets underway on 5 August, with athletes from more than 200 nations travelling to Brazil to compete for glory – and medals, of course. Regular Olympics watchers will know that the USA, China and Russia stand a good chance of topping the medals table based on past performance. Indeed, a closer look at the rankings reveals that these three nations have tended to win the most medals at all recent Olympic Games.

Why should this be? The simple answer is that the USA, China and Russia have better athletes than other countries, and more of them to boot. Of course, it helps that these countries have a huge pool of potential talent to draw from. Provided that world-class athletic ability is uniformly distributed across the world's population, larger countries should generally produce more top athletes. Indeed, Figure 1a shows a clear correlation between population size and total Olympic medals. But there may be other factors that determine the success of these nations.

Academic studies have found a number of socio-economic variables that are reliable predictors of how well a country will do in the Olympic Games. Like population size, a country's gross domestic product (GDP) is strongly correlated with medal wins (see Figure 1b).[1] Of course, GDP by itself has no direct impact on an athlete's performance, but it is a proxy variable for other things that do have an impact: in a wealthy country, the population can dedicate more time to leisure activities and can afford to support a class of professional athletes – both of which lead to investments in better sports infrastructure which can deliver more effective training.

Julia Bredtmann is a postdoctoral researcher at Rheinisch-Westfälisches Institut für Wirtschaftsforschung. Her research interests cover labour economics, migration economics and education economics

Carsten J. Crede is a PhD student at the Centre for Competition Policy and the School of Economics at the University of East Anglia. His research interests include industrial organisation and applied econometrics

Sebastian Otten is a postdoctoral researcher at Rheinisch-Westfälisches Institut für Wirtschaftsforschung. His research interests cover labour and migration economics

Main image: Alexandru Vlad/iStock/Thinkstock

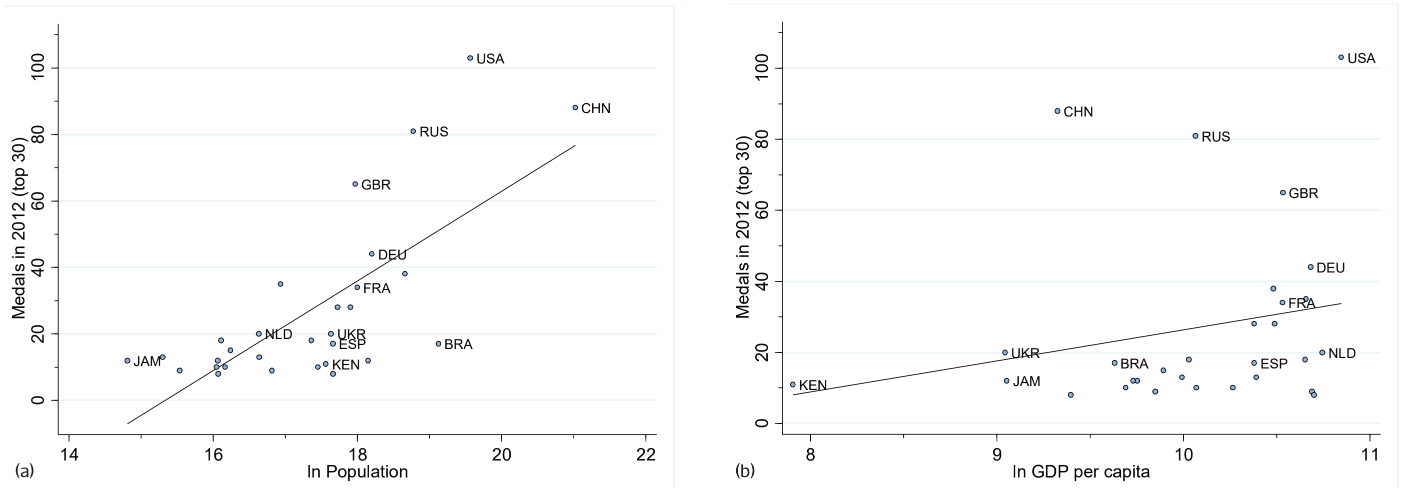

FIGURE 1 Correlation between (a) the natural logarithm of population and total medals, and (b) the natural logarithm of GDP per capita and total medals for the 30 top-scoring countries. As can be seen from the linear regression line, larger and wealthier countries generally win more medals. However, there are some outliers. The USA, China, and Russia, for example, won more medals at the 2012 Games than one would expect from their GDP per capita alone

Other predictors include a country's past or present political system (autocratic nations and planned economies tend to make greater investments in athletes to obtain prestige, and countries formerly featuring such political systems continue to profit from past efforts to develop sports), whether a country is the host, or the next to host, the Olympics (these nations tend to increase investments in athletes, while the host country itself benefits from a certain amount of "home advantage"), and whether women are equally likely to participate in sports and are able to train to become athletes.[2,3]

Knowing all this, we set out to make some predictions of our own of what the medals table will look like at the end of the 2016 Games by exploiting the correlation between socio-economic variables and Olympic success, as well as the persistence in Olympic success over time.

Predicting Olympic medals

For this exercise, we used regression analysis to conduct out-of-sample predictions (see box on page 24). Our initial goal was to predict the number of medals won by each country at the London 2012 Olympics, using data from Olympic Games held between 1996 and 2008. We did not use data from Games that took place before 1992 (either as current or lagged observations) because of the fundamental social and political changes that took place around the world in the early 1990s – in particular, the collapse of the Soviet Union, which led to many more countries participating in the Games. The data we did have provided 676 country–year observations, which we then used to produce out-of-sample predictions for 2012.

We derived predictions from two different models, and these predictions were then compared to the actual results for 2012 to see which model performed best. The first model we considered, which we call the *naive* model, primarily described a country's medals as a function of the number of medals won at the previous Olympic Games – that is, a nation's medals

The estimated models

Both models are estimated using ordinary least squares. The naive model is

$$\mathit{Medals}_{it} = \beta_1 + \beta_2 \mathit{Lag\ Medals}_{it} + \beta_3 \mathit{Year}_t + \varepsilon_{it}$$

where *Medals* denotes the total number of medals (gold, silver and bronze) won by a country at the corresponding Olympic Games, *Lag Medals* contains a country's medals at the preceding Games, and *Year* denotes the year of the Olympic Games. The latter variable is included to capture the steady increase in the total number of medals awarded at the Olympic Games over time.

The sophisticated model includes additional explanatory variables that capture country-specific characteristics:

$$\begin{aligned}\mathit{Medals}_{it} = {} & \gamma_1 + \gamma_2 \mathit{Lag\ Medals}_{it} + \gamma_3 \ln \mathit{GDP}_{it} + \gamma_4 \ln \mathit{Pop}_{it} + \gamma_5 \mathit{Host}_{it} + \gamma_6 \mathit{Next\ Host}_{it} \\ & + \gamma_7 \mathit{Planned}_{it} + \gamma_8 \mathit{Muslim}_{it} + \gamma_9 \mathit{Year}_t + \zeta_{it}\end{aligned}$$

where ln *GDP* denotes the natural logarithm of a country's GDP per capita and ln *Pop* it is the natural logarithm of a country's population. Both variables are included in logs to acknowledge that the positive effects of GDP per capita and population size on Olympic medals diminish with increasing values of these variables. *Host* and *Next Host* are indicator variables for the current host and the upcoming host country. *Planned* is an indicator variable denoting whether a country has or had a fully centralised planned economy (such as former members of the Soviet Union, China and Cuba) and controls for the higher expenditure on sports in such countries to promote national prestige. In case of a country's switch to another economic system, it measures the extent to which it still profits from previous investments in sports infrastructure. Finally, *Muslim* is an indicator for countries with a majority Muslim population, which tend to send fewer female athletes,[3] and tend to have a lower share of the female population active in professional sports. For all explanatory variables considered, we do not assume them to have a causal impact on a country's Olympic success, but are aware that they might capture both direct effects and indirect effects of other, unobserved country-specific factors.

TABLE 1 Actual and predicted medals for the 2012 Olympic Games

Country	Actual results		Naive model		Sophisticated model		
	2012 rank	Medals	Predicted medals for 2012	Diff.	Predicted medals for 2012	Diff.	Predicted 2012 rank
USA	1	103	110	7	105	2	1
China	2	88	100	12	96	8	2
Russia	3	81	73	8	70	11	3
UK	4	65	47	18	60	5	4
Germany	5	44	41	3	39	5	6
Japan	6	38	25	13	24	14	12
Australia	7	35	46	11	44	9	5
France	8	34	41	7	39	5	6
Italy	9	28	27	1	26	2	10
South Korea	9	28	31	3	30	2	8
Netherlands	11	20	16	4	16	4	17
Ukraine	11	20	27	7	27	7	9
Canada	13	18	19	1	19	1	14
Hungary	13	18	10	8	10	8	21
Brazil	15	17	15	2	26	9	10
Spain	15	17	18	1	18	1	15
Sum		654	646	106	649	93	
MAE				6.6		5.8	
MFE			3.4		3.2		

Note: 16 countries are included as Brazil and Spain were ranked joint 15th

Out-of-sample predictions

Consider a simple regression model,

$$Y_{it} = \beta_1 + \beta_2 X_{it} + \varepsilon_{it}$$

where Y is the outcome variable, X is the explanatory variable, ε is the error term, and indices i and t denote countries and time, respectively. Assume that we have three time periods: two in the past and one in the future. To predict Y_{it} in the future period, we can estimate the above model using only observations from the two previous periods, i.e. $t = \{1, 2\}$, and obtain the corresponding estimated coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$. Using the estimated past relationship between X and Y in periods 1 and 2, and assuming that this relationship holds for the future period, we can predict the values of the outcome variable in period 3 by plugging in values of X in period 3 into

$$\hat{Y}_{i3} = \hat{\beta}_1 + \hat{\beta}_2 X_{i3}$$

to obtain predictions of the outcome variable for each country i.

at the 2008 Games are predicted primarily based on how many medals it won in 2004. This captures the persistence in performance of countries over the years. The only other explanatory variable in this model was a linear time trend capturing the years of the different Olympic Games; this was done to take account of the fact that the overall number of medals increases over time because of an increasing number of events over time.

The second model, which we label the *sophisticated* model, expands the naive model with additional socio-economic variables. These include a country's GDP per capita and population size, whether the country is or was a planned economy, whether the country is the host (for 2012 this was the UK) or upcoming host (in 2012 this was Brazil), and whether it has a predominantly Muslim population (as such countries tend to send fewer female athletes and win fewer medals in women's events). For more details, see the box on page 23.

To assess whether the sophisticated model outperforms the naive model, we can calculate measures of fit and compare them between the two models. For this purpose, we calculate the mean absolute error (MAE) and the mean forecast error (MFE) for both the naive and the sophisticated model (see box on page 25).

For the full sample, including 181 countries, the MAE is 1.43 for the naive model, and 1.41 for the sophisticated model. In other words, on average, in the naive model predictions are off by 1.43 medals from the true number of medals won by each country, and by 1.41 in the sophisticated model. But if we only consider those countries ranked in the top 15, the difference in MAE is more pronounced, with a reduction from 6.6 to 5.8 when switching from the naive to the sophisticated model. At the same time, the predictions in the sophisticated model are subject to a lower uncertainty, as the MFE is 3.2 in the sophisticated model compared to 3.4 in the naive model. Therefore, these results suggest that including socio-economic variables in the model slightly increases the precision of the predictions.

Table 1 compares the actual results for the top-ranked countries at the 2012 Olympics with the predictions based on the two different models. For the naive and the sophisticated model, it shows both the number of predicted medals as well as the absolute difference between the predicted and the actual medals won by each country. For the sophisticated model, it further shows the predicted rank.

We see that the naive model already predicts Olympic success fairly well. Due to the persistence in Olympic success, one can determine how many medals a country wins quite reliably by merely looking at the results of the last Olympic Games. In fact, due to the nearly constant number of awarded medals at the 2008 and 2012 Olympic Games, the naive model delivers exactly the same results as an "ultra-naive model", which obtains the number of medals won at the 2012 Games by simply multiplying the total number of medals to be awarded in 2012 by the proportion of medals won by each nation in the 2008 Games. In total, the naive model wrongly predicts 106 medals of the 654 medals won by these countries. The most notable outlier is the UK, the 2012 host country, which won 18 more medals than predicted by the naive model as that ignores the host country effect.

Turning to the results of the sophisticated model, it can be seen that it performs slightly better than the naive model. This is in line with the finding of smaller MAE and MFE values indicating a higher precision of the sophisticated model. The absolute difference between predicted and actual medals reduces from 106 to 93 medals. The increase in precision of predictions primarily arises from smaller differences for the UK, as 2012 host, and the USA. For the UK, the difference goes down from 18 to only 5 medals. This is a result of controlling for the host country effect outlined above. Yet, for Brazil the prediction error increases from 2 to 9 medals. This follows from the fact that, compared to other previous upcoming host nations, Brazil underperformed at the 2012 Games. With only 17 medals won, it was 9 medals from where it was expected to be, according to the sophisticated model's assumption of an 'upcoming host' advantage.

The 2016 predictions

Given that the sophisticated model performs slightly better in predicting the 2012 Olympics, we use it to predict the Olympic medals for the top 15 nations at the 2016 Games. The predictions can be found in Table 2, while predictions for all countries are at bit.ly/1Xi20Bx.

The model predicts that the USA, China, Russia, and the UK will retain their top positions in the medals ranking (though this rests on the presumption that no country is excluded from participating at any events due to doping concerns), but Brazil and Japan are expected to make the biggest gains.

As the host country, Brazil will send more athletes than it did to previous Games (431 in 27 sports, compared to 258 in 24 sports in 2012) and, in preparation for hosting, it will have invested more in the development of its national talent pool. Furthermore, Brazilian athletes will likely be the stadium visitors' favourites, receiving the most support from the audience. Meanwhile, Japan will host the 2020 Summer Olympics. Therefore, the investment it makes in its athletes in preparation for the 2020 Games is expected to pay off this year.

How much should we trust these predictions? There are only indirect ways to assess their quality. One possibility is to look at the MFE for the 2016 predictions, which goes down to 3.1 for the 2016 Games from 3.2 for the 2012 Games. Therefore, the model predictions feature a similar level of uncertainty as those for the 2012 Olympic Games. Thus, it is likely that the accuracy of the 2016 Olympic medals prediction will be in the same range as those for the 2012 Games (though readers should note here that Table 1 and Table 2 feature 16 and 15 countries respectively, so MFE comparisons between the two tables are somewhat artificial).

TABLE 2 Predicted medals for the 2016 Olympic Games

Country	Predicted rank for 2016	Predicted medals for 2016	Rank in 2012	Medals in 2012	Diff. in ranks	Diff. in medals to 2012
USA	1	98	1	103	0	–5
China	2	84	2	88	0	–4
Russia	3	77	3	81	0	–4
UK	4	62	4	65	0	–3
Japan	5	46	6	38	1	8
Germany	6	42	5	44	–1	–2
Australia	7	33	7	35	0	–2
Brazil	7	33	15	17	8	16
France	7	33	8	34	1	–1
Italy	10	27	9	28	–1	–1
South Korea	10	27	9	28	–1	–1
Ukraine	12	20	11	20	–1	0
Netherlands	13	19	11	20	–2	–1
Canada	14	18	13	18	–1	0
Hungary	14	18	13	18	–1	0
MFE		3.1				

Mean absolute error and mean forecast error

The mean absolute error (MAE) is a measure of the average inaccuracy associated with a set of model-produced estimates. It compares the predicted values $\hat{y}$ for the outcome variable with the true values y across all individual estimates i, and calculates a measure of the absolute value of differences:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|$$

The lower the MAE, therefore, the better the fit of the model.

A second measure of prediction inaccuracy is the mean forecast error (MFE). The forecast error is the standard error of the point prediction and thus expresses the uncertainty in estimating the unknown value of y for an individual observation with known X values. A forecast error can be obtained for each country's predicted medals. For the purpose of model comparison, we calculate the mean value of these errors for the countries ranked in the top 15.

However, supporters of "underdog" nations need not give up hope entirely. Yes, past success is a good predictor of future success in the Olympic Games, and the bigger, better-off countries are almost guaranteed to do well given the socio-economic advantages they enjoy, but it should be remembered that a certain level of unpredictability remains in any sporting competition. The history of the Olympic Games is full of surprising performances by individual athletes, and we should expect to see more of them in Rio in August. ■

References

1. Bernard, A. B. and Busse, M. R. (2004) Who wins the Olympic games: Economic resources and medal totals. *Review of Economics and Statistics*, **86**(1), 413–417.

2. Johnson, D. K. and 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.

3. Bredtmann, J., Crede, C. J. and Otten, S. (2016) Participation and success at the Olympic Games – the role of gender equality. Mimeo.