APM Assignment 1 2020-21

Predicting Olympic medal counts

The aim of this assignment is to develop models for predicting the number of medals won by each country at the Rio Olympics in 2016 using information that was available prior to the Rio Games.

Data

Data are available on the number of medals (total and gold) won by each country for 108 countries participating in the Rio 2016 Olympics, along with information on previous Olympic performance (from the 2000, 2004, 2008 and 2012 Games) and other variables.

The dataset `rioolympics.csv` has 108 observations and the following variables:

- `country`: the country's name,
- `country.code`: the country's three-letter code,
- `gdpYY`: the country's GPD in millions of US dollars during year YY,
- `popYY`: the country's population in thousands in year YY,
- `soviet`: 1 if the country was part of the former Soviet Union, 0 otherwise,
- `comm`: 1 if the country is a former/current communist state, 0 otherwise,
- `muslim`: 1 if the country is a Muslim majority country, 0 otherwise,
- `oneparty`: 1 if the country is a one-party state, 0 otherwise,
- `goldYY`: number of gold medals won in the YY Olympics,
- `totYY`: total number of medals won in the YY Olympics,
- `totgoldYY`: overall total number of gold medals awarded in the YY Olympics,
- `totmedalsYY`: overall total number of all medals awarded in the YY Olympics,
- `altitude`: altitude of the country's capital city,
- `athletesYY`: number of athletes representing the country in the YY Olympics,
- `host`: 1 if the country has hosted/is hosting/will be hosting the Olympics, 0 otherwise.

The first observation, corresponding to Afghanistan, is shown below.

```
oldat <- read.csv(url("http://www.stats.gla.ac.uk/~tereza/rp/rioolympics.csv"))
oldat$gdp16 <- as.numeric(oldat$gdp16)
```

```
Warning: NAs introduced by coercion
```

```
head(oldat,1)
```

```
      country country.code gdp00 gdp04 gdp08 gdp12 gdp16 pop00 pop04 pop08
1 Afghanistan          AFG  #N/A  5285 10191 20537 19469 20094 24119 27294
  pop12 pop16 soviet comm muslim oneparty gold00 gold04 gold08 gold12 gold16
1 30697 34656      0    0      1        0      0      0      0      0      0
  tot00 tot04 tot08 tot12 tot16 totgold00 totgold04 totgold08 totgold12
1     0     0     1     1     0       298       301       301       301
  totgold16 totmedals00 totmedals04 totmedals08 totmedals12 totmedals16
1       298         915         924         949         956         949
  altitude athletes00 athletes04 athletes08 athletes12 athletes16 host
1     1790          0          5          4          6          3    0
```

Note: The dataset was put together from various sources, and as such it has some missing values etc. Please document how you deal with these data issues in your report.

Tasks

1. Explore which variables are associated with the number of medals (total/gold or both) won in the 2012 Olympics.
2. Develop appropriate models for the training data up to (and including) 2012 and use these models to predict Olympic performance in the 2016 Games (test data). Evaluate the predictive performance of these models for the test data.
3. Describe your findings and comment on what improvements might be made to the model/data collected in order to better predict Olympic medal counts for future Games.

Guidelines

- You should submit both a .pdf and an R code (or Rmarkdown) file. The report should not be more than ten pages long, so please consider carefully what output and figures to include. Check the marking scheme to make sure you have included everything that's expected.
- Introduce the problem, discuss any data cleaning/pre-processing decisions you made, and provide some exploratory analysis. This doesn't need to be long, just enough to give an idea of what you are trying to do to someone not familiar with the assignment.
- When fitting GLMs try at least **two** different distributions for the response. These can include the normal linear model.

- The emphasis is on prediction. Interpretation of regression coefficients is not that important.
- As the objective is prediction, you need to have a measure of out-of-sample predictive performance by which to compare models, such as the root mean squared error. Make sure you state what measure(s) you are using with appropriate definition(s) provided.
- Train your models on previous years' data and use 2016 as a test set on which to evaluate predictive performance using the above measure(s). Summarise your findings.
- For the final part of your report you should discuss the limitations of your approach(es) and any further work you would do if you had more time to improve your predictive model.