Robert Shepherd
2431907S

Predicting Olympic medal counts

APM Project Assignment 2018-19

Executive summary

This report investigates the capability of predicting Olympic medal counts using historic data. Data relating to Olympic events between 2000 and 2012 was used to build a series of models to predict the number of medals won by each country in the 2016 Olympic games.

Exploratory analysis found a strong correlation between medals won at the previous Olympics event, the number of athletes, GDP and population. Continent information (e.g. Asia, Europe) was added to the dataset and this was found to be influential in improving model accuracy. Other factors, such as whether the nation was hosting the event, were found to be influential but did not significantly improve our ability to predict and tended to overfit the data.

Model building took an iterative approach with a focus on identifying models with a high predictive accuracy. A benchmark model was created by assuming the number of medals won by each country in 2016 would be consistent with 2012. This generally performed quite well and exceeded the performance of other models with one exception.

The linear model outperformed the benchmark model by a significant margin and improved our ability to predict medals won in 2016. The final model was a separate lines model (for each continent) based on previous medals won, number of athletes, GDP and population as predictors. While the model was highly predictive, the residuals suggested that not all of the non-random structure of the data was captured meaning there may be scope to improve the model further.

The report ends by identifying potential future developments to the model that may improve predictive accuracy, such as the inclusion of data from other (more recent) events to help bridge the four-year gap between the Olympic games.

Introduction

The challenge

This report will aim to develop a model to predict the number of medals won in the 2016 Olympics using data from previous events between 2000 to 2012. In developing the model, the report aims to answer the below:

1. Which variables are associated with the total number of medals won in the 2012 Olympics?
2. How well does a model based on data up to and including 2012 predict Olympic performance in the 2016 Games?
3. What improvements might be made to the model/data collected in order to better predict Olympic medal counts for future Games?

The report uses the 'rioolympics.csv' dataset available below (as of July 2019):

http://www.stats.gla.ac.uk/~tereza/rp/rioolympics.csv

Assumptions

Two top level assumptions have been made about the dataset:

1. Variables below were consistent over all events

- BMI
- Soviet
- Communist
- Muslim
- Oneparty
- Altitude

2. Variable info for the below was available prior to each event:
    - Total medals available
    - Number of athletes entered
    - Population
    - GDP
    - Host

Data description and methodology

Data preparation

Additional data sources

Continent and region information for each country was appended from (accessed June 2019):

https://cloford.com/resources/codes/index.htm

Some values were missing and were added manually for the below countries:

- Hong Kong: Asia / East Asia
- Kosovo: Europe / South East Europe
- Montenegro: Europe / South East Europe
- Romania: Europe / South East Europe
- Serbia: Europe / South East Europe

The data was added to test the hypothesis that countries in the same region will have comparable performance.

The **host** variable has been transformed so that it changes for each event so that only the country that is hosting is flagged as below:

- 2000: Australia
- 2004: Greece
- 2008: China
- 2012: United Kingdom
- 2016: Brazil

Reformatting the data

The data was reformatted in to a 'long' format so that one record was present for each country and year (previously there was only one record for each country). This was done to:

- Increase the number of records available to build the model
- Avoid the risk of a skewed model by focusing only on predicting one year (e.g. UK did particularly well in 2012 when it was hosting the Olympics)

- Ensure the host variable was relevant by making it a flag for each year (instead of just one flag over all the events)

In addition to reformatting the data, the below fields were calculated:

- prev_gold: Number of gold medals a country won in the previous event
- prev_medals: Number of total medals a country won in the previous event
- prev_gold_total: Total number of gold medals given in the previous event
- prev_medals_total: Total number of medals given in the previous event
- year: Year of the event
- gold_share: Share of gold medals (gold / total_gold)
- total_share: Share of total medals (total / total_medals)
- prev_gold_share: Share of previous gold medals (prev_gold / prev_gold_total)
- prev_total_share: Share of previous total medals (prev_medals / prev_medals_total)

As the previous year's figures were used as a predictor variable, data for 2000 was not used as no 'previous' figures were available (the exception to this is when developing the linear mixed model which tested including 2000 data separately).

Note: The final dataset used is available to view in the appendix.

Missing/erroneous values

Some values were corrected as below:

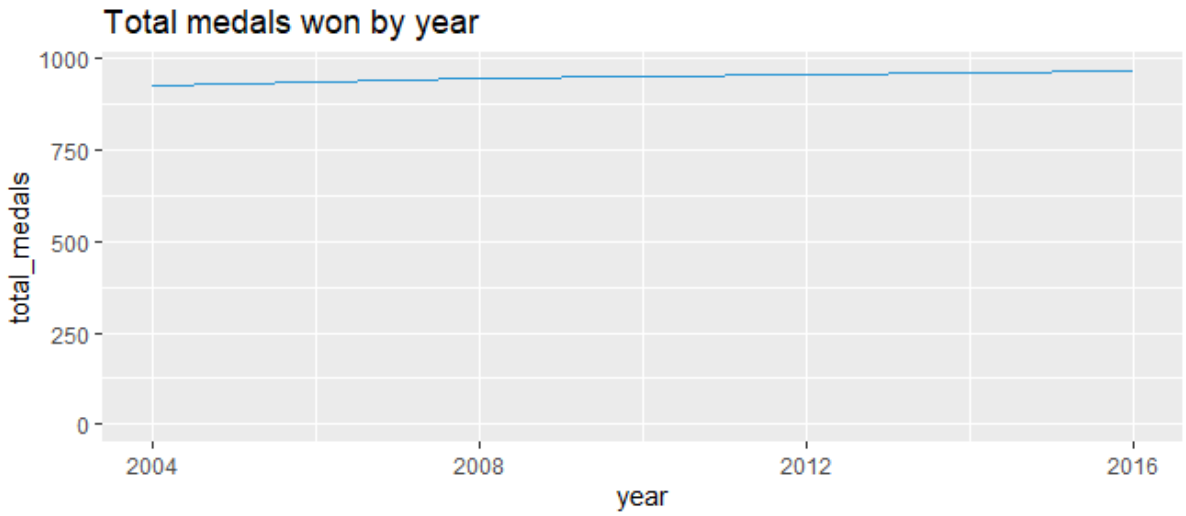
- GDP contained 2 missing values in 2016 and 1 in 2000
  - Data for 2016 was simulated using 2012 data
  - Missing data for 2000 was ignored as data for this year was not used
- BMI contained 27 missing values across all countries
  - Mean BMI calculated for populated countries to simulate missing values
- Total gold medals and total medals values for 2016 were smaller than the sum of the medals awarded to each country
  - Values were updated to match the sum of medals awarded to the individual countries

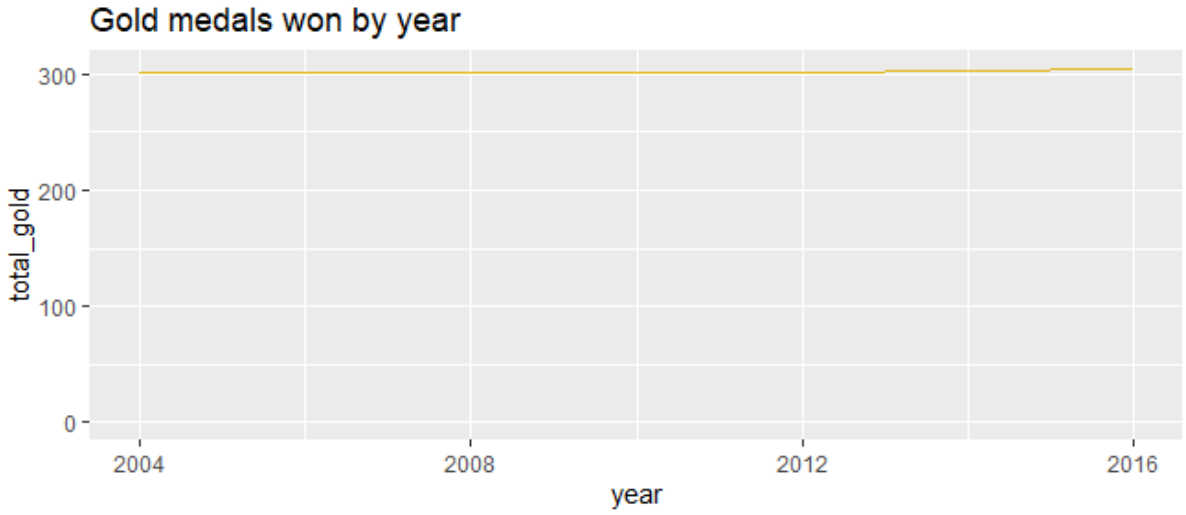
Exploratory analysis

Total medals or medal share?

If the total amount of medals changes significantly from year to year it would be more valuable to focus on the share of medals won.

Looking at the total amount of medals:

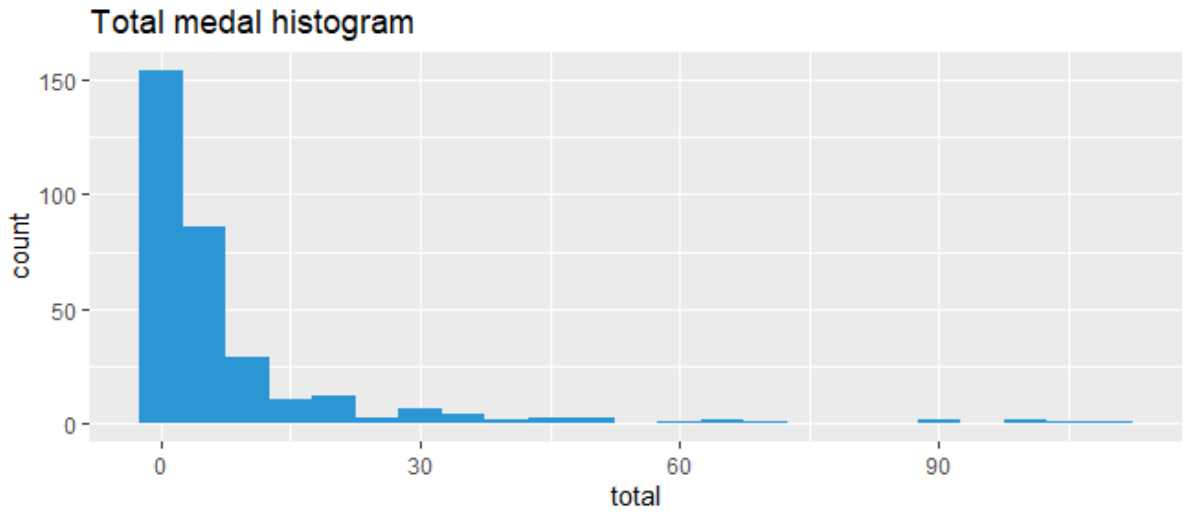


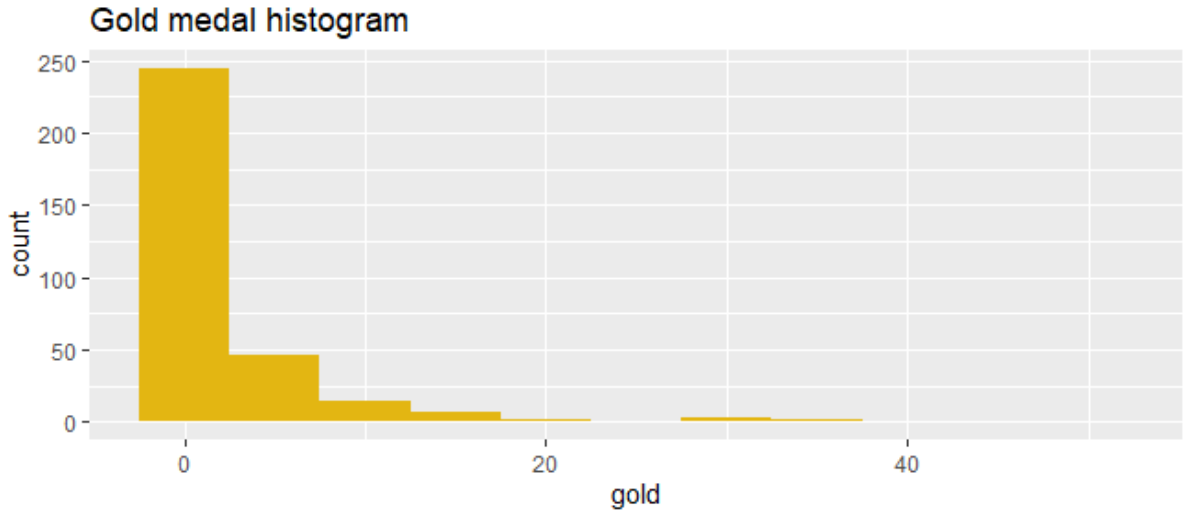


The total volume of medals does not vary much from year to year so the models will focus on predicting the total number of medals to keep the modelling process as simple and interpretable as possible.

Medal distribution and outlier detection

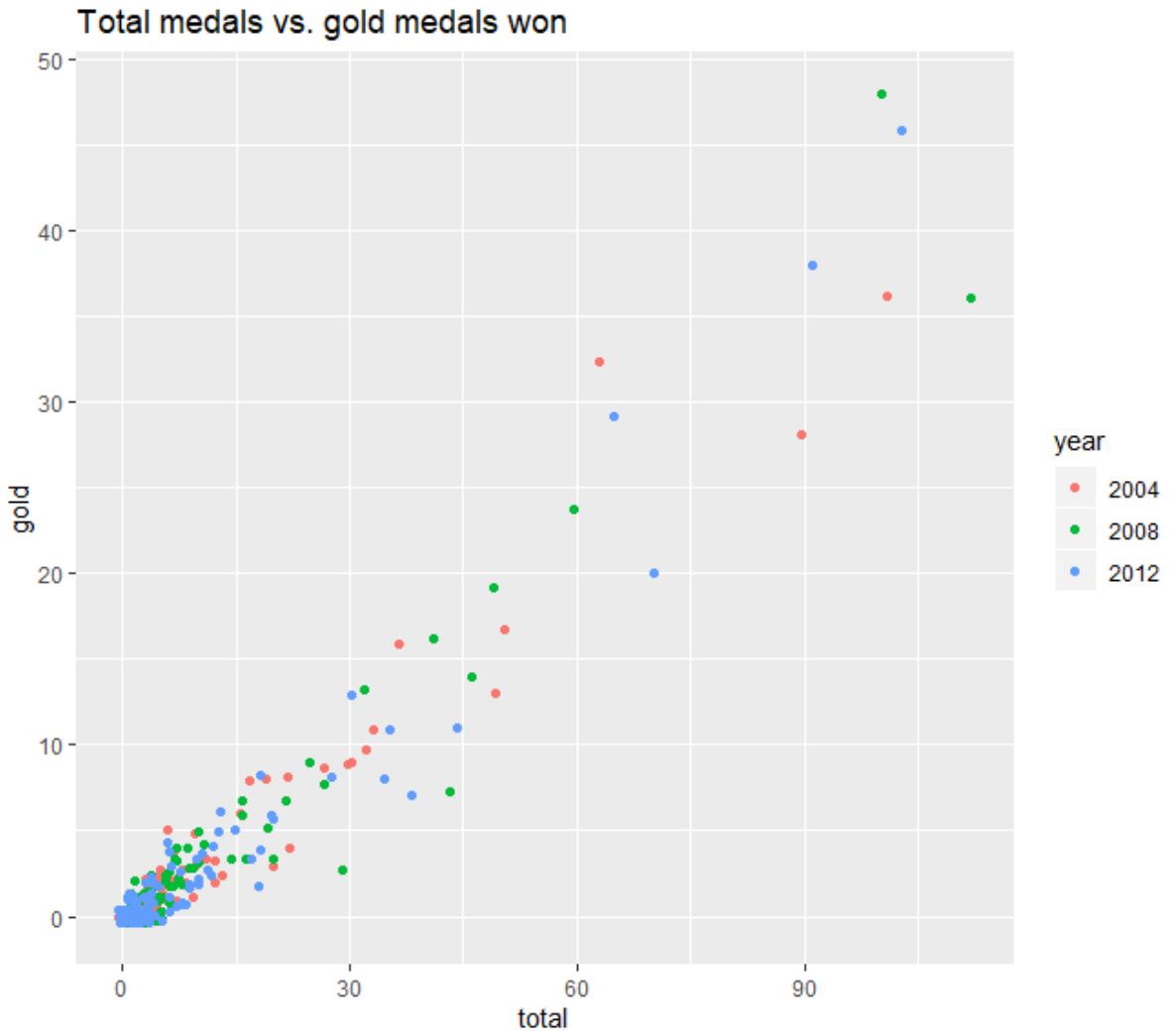
Examining the distribution of medals won helps inform what models are likely to fit the data well:





Both distributions are similar, following a zero-inflated Poisson distribution. Both distributions also have a long tail of high values suggesting a transformation may also help with fitting the data.

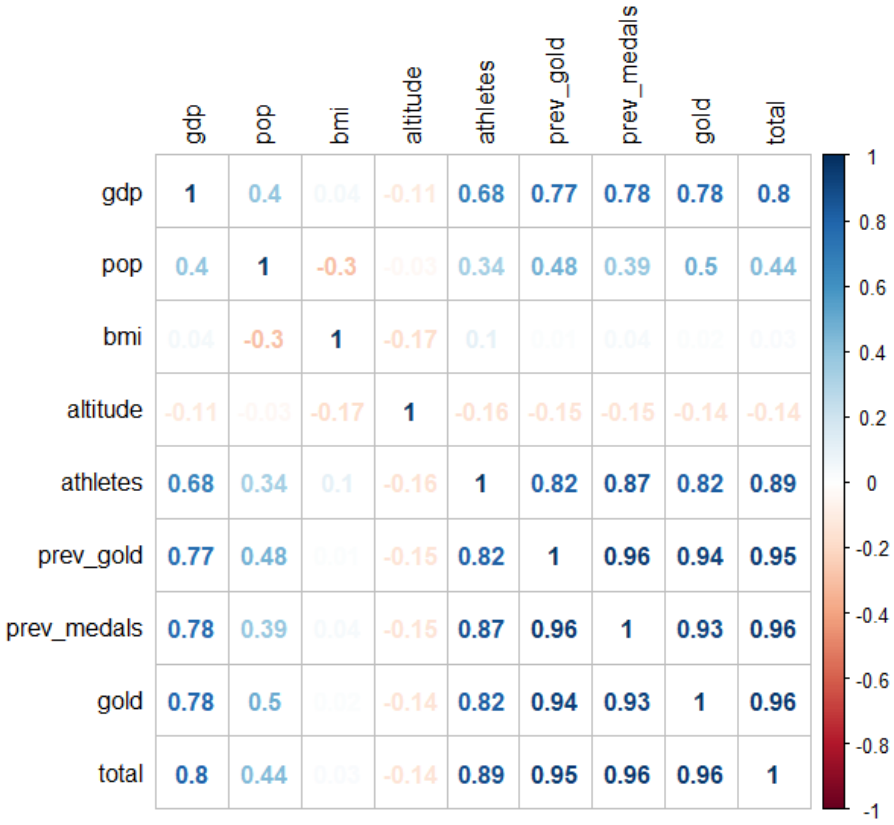
Gold medals appear to have a shorter distribution with a higher concentration of 0 values.



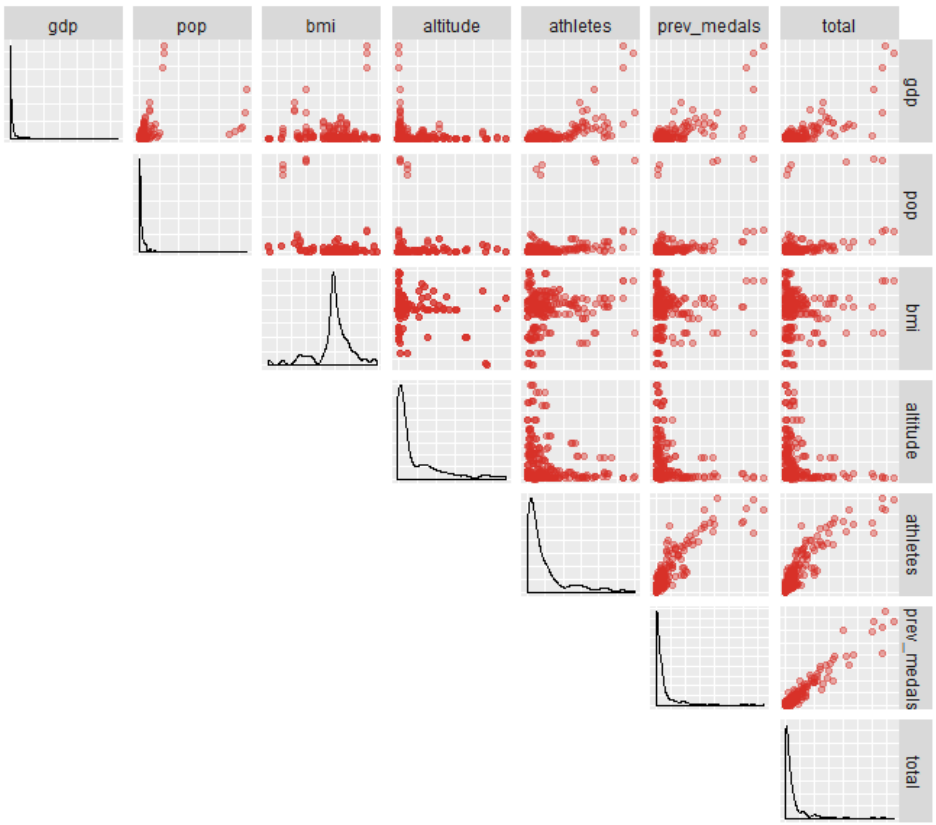
Mapping total medals vs. gold medals shows a linear correlation. No extreme values are observed so no outliers have been removed.

Examining correlations

Correlation plot



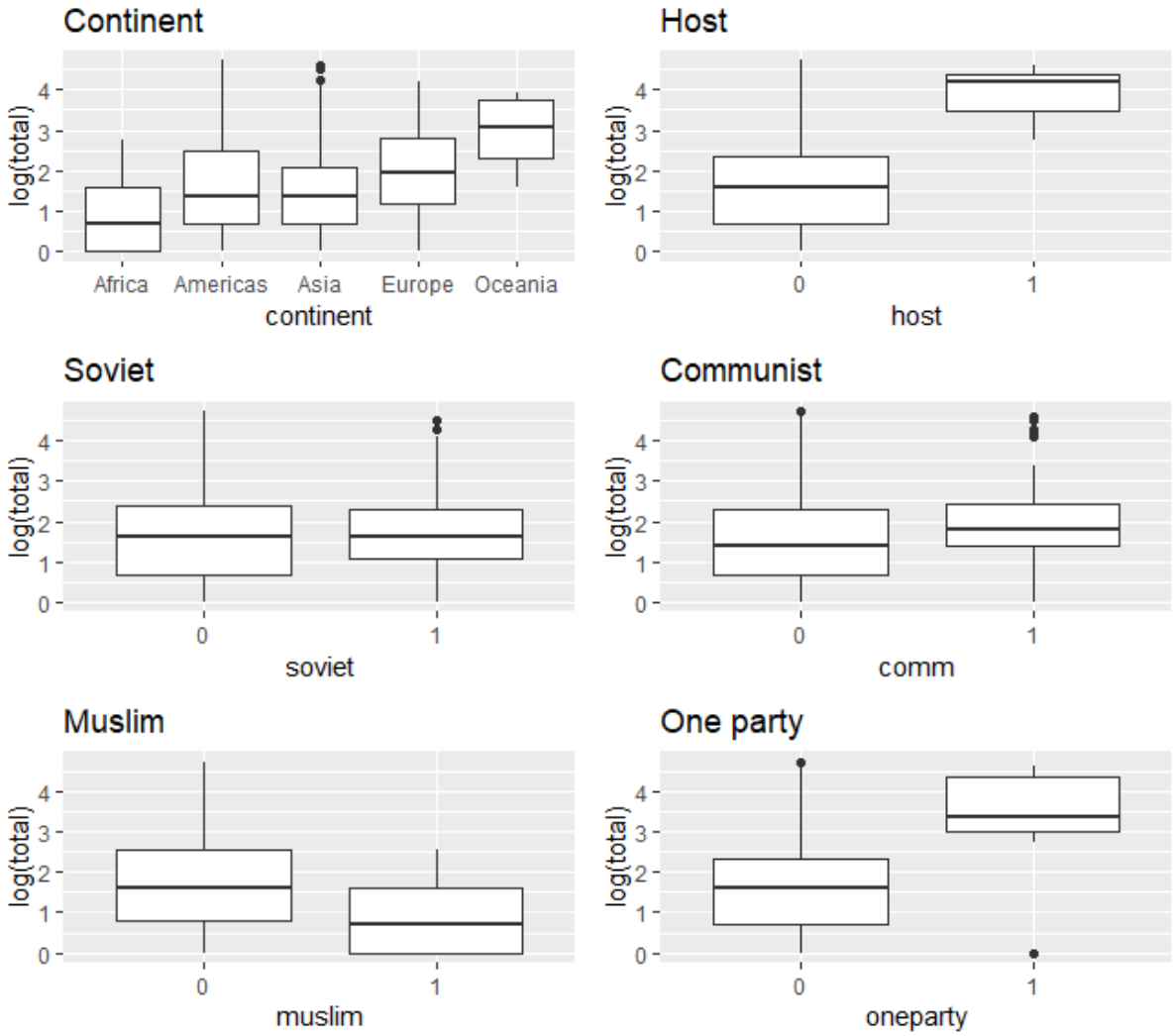
	gdp	pop	bmi	altitude	athletes	prev_gold	prev_medals	gold	total
gdp	1	0.4	0.04	-0.11	0.68	0.77	0.78	0.78	0.8
pop	0.4	1	-0.3	-0.03	0.34	0.48	0.39	0.5	0.44
bmi	0.04	-0.3	1	-0.17	0.1	0.01	0.04	0.02	0.03
altitude	-0.11	-0.03	-0.17	1	-0.16	-0.15	-0.15	-0.14	-0.14
athletes	0.68	0.34	0.1	-0.16	1	0.82	0.87	0.82	0.89
prev_gold	0.77	0.48	0.01	-0.15	0.82	1	0.96	0.94	0.95
prev_medals	0.78	0.39	0.04	-0.15	0.87	0.96	1	0.93	0.96
gold	0.78	0.5	0.02	-0.14	0.82	0.94	0.93	1	0.96
total	0.8	0.44	0.03	-0.14	0.89	0.95	0.96	0.96	1



The correlation plot and pairs plot highlight some important considerations:

- Total and gold medals have a very high correlation
- Previous medals, athletes, population and GDP have a positive correlation with total/gold medals won and should be included as predictors
- The high level of correlation between predictors means including all variables may bias the models
- BMI and altitude have little correlation with total/gold medals won and are unlikely to add predictive value to models

Exploring factors



Looking at how the distribution of medals (log transformed excluding 0 medal wins) shows which factors are likely to improve predictive power:

- 'Region' (see appendix), 'Continent', 'Host', 'Muslim' and 'One party' all show differences in distribution
- 'Soviet' and 'Communist' do not show significant differences in the distribution

Methodology

The data will be split in to two datasets:

Training: Data for events between 2004 and 2012 (with 2000 used to create the 'previous medal' fields for 2004)

Test: Data for the 2016 event

A series of models will be developed using the training data as below:

- Linear models
- Generalised linear models
- Linear mixed models (including generalised linear)

The models will aim to **predict the total number of medals won**. Model performance will be assessed based on how well they fit the data and, most importantly, how accurately they predict medals won using the test data.

Predictive accuracy will be assessed based on the root-mean-squared-error (RMSE) and the mean absolute error (MAE). For each model type, a best performing (lowest error) model will be carried forward for a final comparison.

To add a frame of reference for our model performance, a benchmark model will be created using non-statistical methods to understand whether our models improve upon our 'best guess'.

Analysis

Benchmark model

The benchmark model was created by assuming the medals won by each country in 2016 would be the same as 2012.

This produces the below errors:

- RMSE on test data: **4.17**
- MAE on test data: **2.33**

Linear model

A series of linear models were built and tested using the below process:

1. A 'full' linear model was built using 'GDP', 'Population', 'Athletes', 'Previous medals', 'BMI' and 'Altitude'
    - Some variables were not included as they showed a high correlation with other predictors in the exploratory analysis
2. 'BMI' and 'Altitude' were then removed using a stepwise regression function (both directions) to retain as much information as possible with fewer variables
    - This was consistent with the variables showing little correlation with total medals in the exploratory analysis
3. Analysis of co-variance (ANCOVA) was used to decide whether to include factors within the model and whether they should be included as a parallel or separate lines model
4. Data was transformed using a square root transformation and compared to the performance of the untransformed model

5. Fitted values below 0 were changed to 0 (as it is not possible to get less than 0 medals)

Two models stood out as having a high performance, significantly beating the benchmark model:

Model 1

- No transformation
- Continuous predictors: GDP, population, athletes, previous medals won
- Categorical predictors: Continent (separate lines model)
- Diagnostic charts show residuals are non-normal and residuals are skewed suggesting model **does not** capture all of the non-random structure of the data (see appendix)
- RMSE on test data: **3.51**
- MAE on test data: **2.14**

Model 2

- Square root transformation applied to total medals and all continuous predictors
- Continuous predictors: GDP, population, athletes, previous medals won
- Categorical predictors: Continent (separate lines model)
- Diagnostic charts show residuals are relatively normal and residuals are more spread (though some pattern remains) suggesting the model **does** capture most of the non-random structure of the data (see appendix)
- RMSE on test data: **3.68**
- MAE on test data: **2.18**

Of the two, 'Model 1' (untransformed) was selected to represent a linear modelling approach as it had the better predictive accuracy (determined by RMSE/MAE). The non-normal and patterned residuals suggest this could be improved further, potentially by fitting a model that fits a different distribution.

Generalised linear model

A series of generalised linear models were built and tested using the below process:

1. As per the linear model, a 'full' generalised linear model was built using 'GDP', 'Population', 'Athletes', 'Previous medals', 'BMI' and 'Altitude' by fitting a Poisson distribution
2. The stepwise regression function was applied (both directions) but this did not remove any variables
3. Visualising the residuals showed evidence of overdispersion (see appendix)
4. A dispersion parameter was estimated and used to select which variables should be included
    - GDP and altitude were subsequently dropped from the model
5. A Quasi-Poisson model was built but this failed to significantly improve the predictive accuracy of the model
6. A negative binomial model was built and refined using stepwise regression however the prediction errors were large and the model was subsequently discarded
7. As the exploratory analysis identified a zero inflated distribution of medals and residuals showed poor performance over lower medal estimates, both a hurdle model and a zero inflated Poisson model were built
    - Both models performed similarly and improved upon other GLM models

Of the two high performing models (hurdle and zero-inflated Poisson), the hurdle model returned slightly lower predictive errors as below:

- RMSE on test data: **4.82**
- MAE on test data: **3.04**

This failed to improve upon the benchmark model.

Linear mixed model

A series of linear mixed models were built and tested using the below process:

1. A linear mixed model with a random effect on the intercept was created by country
2. Data from 2000 was added to the training but this decreased predictive accuracy (determined by RMSE and MAE) and so was subsequently removed
3. A random effect on the slope was introduced by country/year however this also decreased predictive accuracy
4. The random intercept model was repeated as a hierarchical model with continent added as a factor but this, again, decreased predictive accuracy
5. A generalised linear mixed model (GLMM) was created (with the same random effect on the intercept) was created by fitting a Poisson distribution but the predictive accuracy was poor

Based on this, the original model (a linear mixed model with a random effect by country) was carried forward which returned:

- RMSE on test data: **4.91**
- MAE on test data: **2.88**

This failed to improve upon the benchmark model.

Conclusions and discussion

Which variables are associated with the total number of medals won in the 2012 Olympics?

The exploratory analysis showed a positive correlation between the below continuous variables and total number of medals won:

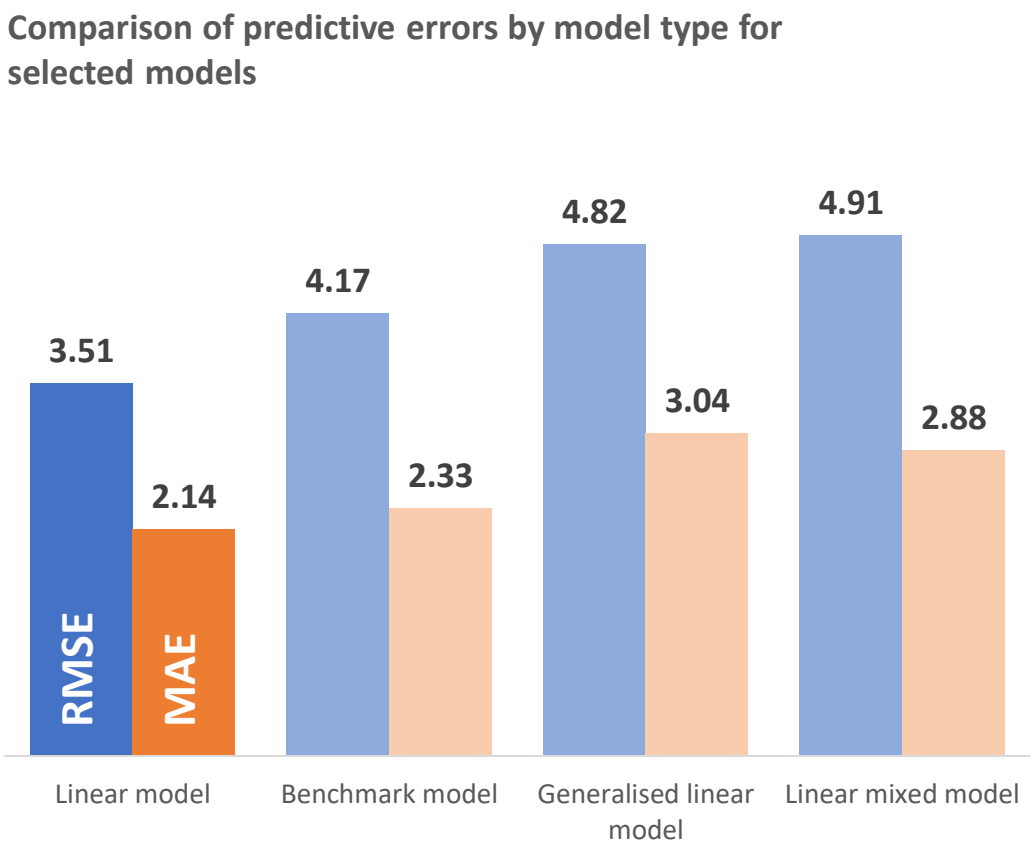
- Previous medals won
- Number of athletes
- GDP
- Population

In addition, the below factors also showed a difference in the distribution of medals won:

- Continent
- Host

One party/Muslim factors also showed some difference in the distribution of medals won but the factors were discarded during the model build process as they did not significantly improve the predictive accuracy of models.

How well does a model based on data up to and including 2012 predict Olympic performance in the 2016 games?



The linear model significantly improved our ability to predict the total number of models won when comparing against the benchmark model returning low predictive errors. Other model types struggled to achieve the same predictive accuracy and performed worse than the benchmark model.

The final model selected was therefore a linear regression model with the below variables:

- Previous medals
- Number of athletes
- Population
- GDP
- Continent (separate lines model)

What improvements might be made to the model/data collected in order to better predict Olympic medal counts for future games?

The model could potentially be developed further by breaking down the medals by event type and building individual models that aggregate up to total medals. This is on the hypothesis that certain countries have a higher propensity to win certain types of competitions. Account for this may improve predictive accuracy by taking in to account trends across different event types.

As the Olympics is only every four years, there may also be value in understanding performance over other competitions, such as the Commonwealth games. This would test the hypothesis that success in more recent competitions may translate to success in the Olympics.

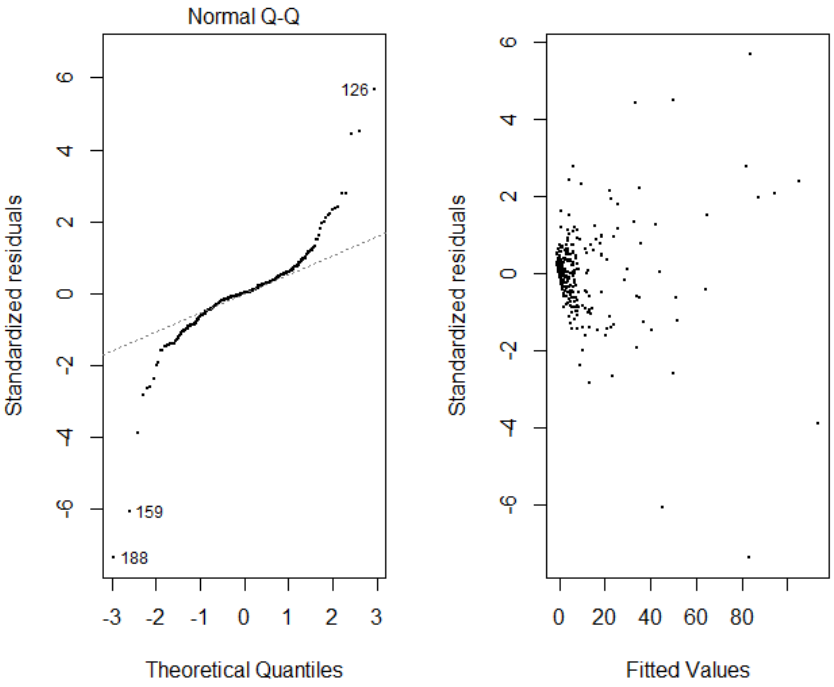
Finally, the models may be potentially be improved by including dimension reduction techniques (such as principle components analysis) or clustering techniques before creating the models. This is evidenced by the high amount of correlation between predictor variables observed in the exploratory analysis. It was also found that simpler models tended to perform better suggesting some element of overfitting that could be avoided by generating more distinct predictors.

Appendix

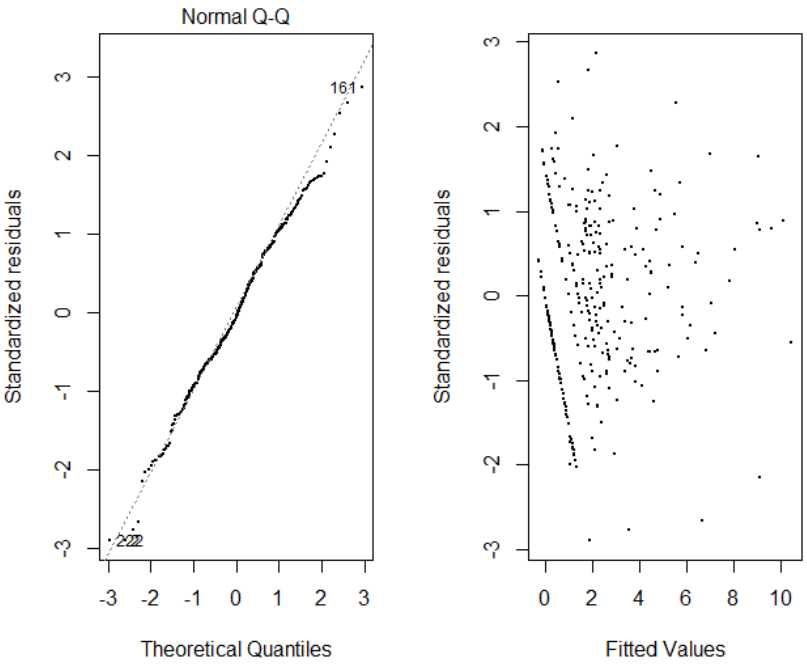
Dataset for linear and generalised linear models

Variable	Example	Role
country	United Kingdom	Reference
country.code	GBR	Predictor
continent	Europe	Predictor
region	Western Europe	Predictor
gdp	2398555	Predictor
pop	59988	Predictor
soviet	0	Predictor
comm	0	Predictor
muslim	0	Predictor
oneparty	0	Predictor
gold	9	Outcome
total	30	Outcome
total_gold	301	Reference
total_medals	924	Reference
bmi	27.25	Predictor
altitude	14	Predictor
athletes	264	Predictor
host	0	Predictor
prev_gold	11	Predictor
prev_medals	28	Predictor
prev_gold_total	298	Predictor
prev_medals_total	915	Predictor
year	2004	Reference
gold_share	0.030	Outcome
total_share	0.032	Outcome
prev_gold_share	0.037	Predictor
prev_total_share	0.031	Predictor

Linear model 1 - diagnostics



Linear model 2 - diagnostics



Poisson GLM residuals showing evidence of overdispersion