Predicting The Performance At The 2016 Olympics Games

The aim of this project is to build two models to predict the number of medals won by each country at the Rio Olympics 2016. This will be achieved by using information that was available prior to the event.

The data set contains information regarding the performance of 108 participating countries throughout the following years: 2000, 2004, 2008, 2012 and 2016.

Data Overview and Management

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 initial data set has been split into three smaller datasets as follows:

olympics12: contains data from the 2012 games. It was used for variable selection and model building.
training.data: contains data for the years 2000, 2004, 2008 and 2012. The year references in the variable names have been removed. This data set was used for training the models.
olympics16: contains the data from year 2016. This data set was used for testing and comparison of the two models.

The three subsets contain the following variables:

• country,
• gdpYY: renamed to gdp
• popYY: renamed to population,
• GDPpercapita: calculated as gdp / population,
• soviet,
• comm,
• muslim,
• oneparty,
• totYY: renamed to medals,
• altitude,
• athletesYY: renamed to athletes,
• host.

Missing Data

The data contains missing values in the gdpYY variable for the following countries:

• Afghanistan
• Cuba
• Syrian Arab Republic

The missing values have been replaced by the mean gross domestic product for that particular year.

Exploratory Analysis

In order to be able to view the relationships between variables clearly, the exploratory analysis was performed on the olympics12 data set.

The variables soviet, comm, muslim, oneparty and host have been re-coded as factors with two levels.

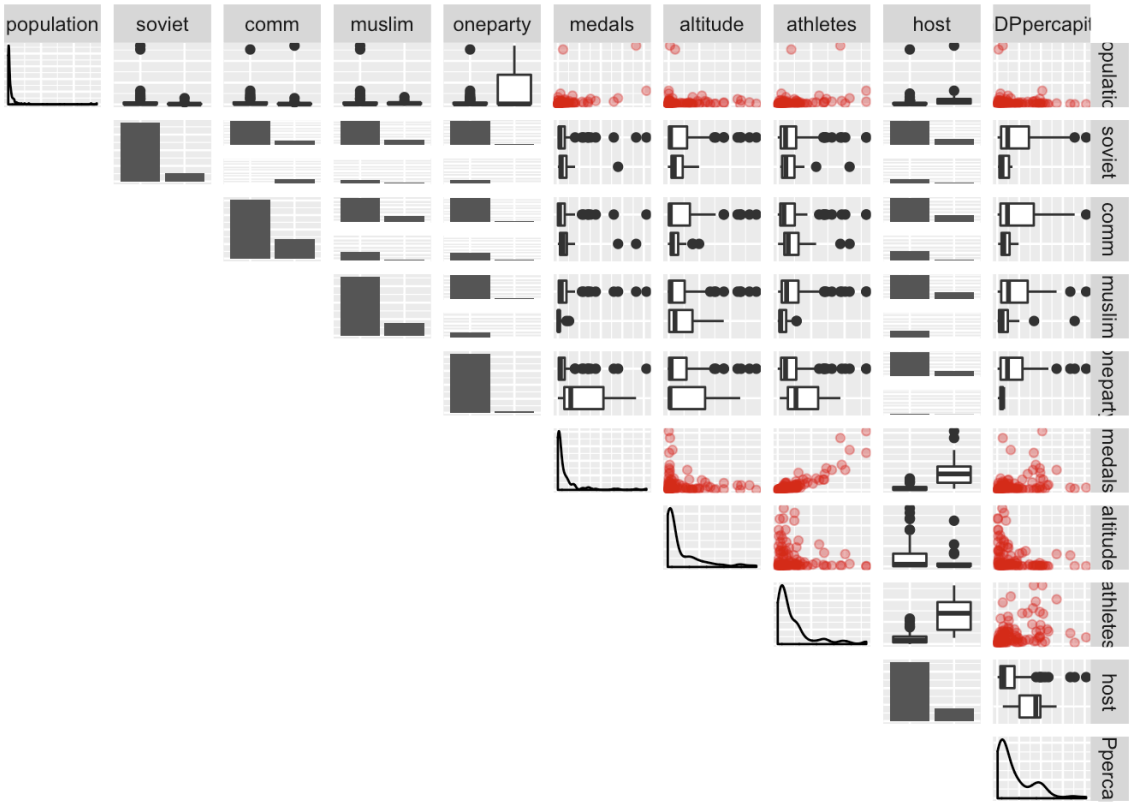
```
summary(olympics12)
```

```
##         country          gdp              population      soviet comm   muslim
##  Afghanistan:  1   Min.   :     800   Min.   :    105   0:94   0:81   0:89
##  Algeria    :  1   1st Qu.:   27850   1st Qu.:   4523   1:14   1:27   1:19
##  Argentina  :  1   Median :  172868   Median :  11086
##  Armenia    :  1   Mean   :  685062   Mean   :  55944
##  Australia  :  1   3rd Qu.:  496410   3rd Qu.:  39072
##  Austria    :  1   Max.   :16155255   Max.   :1350695
##  (Other)    :102
##  oneparty     medals           altitude          athletes      host
##  0:105    Min.   :  0.000   Min.   : -28.00   Min.   :  0.00   0:89
##  1:  3    1st Qu.:  1.000   1st Qu.:  11.62   1st Qu.: 19.75   1:19
##           Median :  3.000   Median :  78.00   Median : 47.00
##           Mean   :  8.852   Mean   : 395.71   Mean   : 90.15
##           3rd Qu.:  9.000   3rd Qu.: 586.25   3rd Qu.:111.25
##           Max.   :103.000   Max.   :2850.00   Max.   :530.00
##
##   GDPpercapita
##  Min.   :  0.3915
##  1st Qu.:  3.8556
##  Median : 10.7476
##  Mean   : 19.5311
##  3rd Qu.: 28.6279
##  Max.   :101.6595
##
```

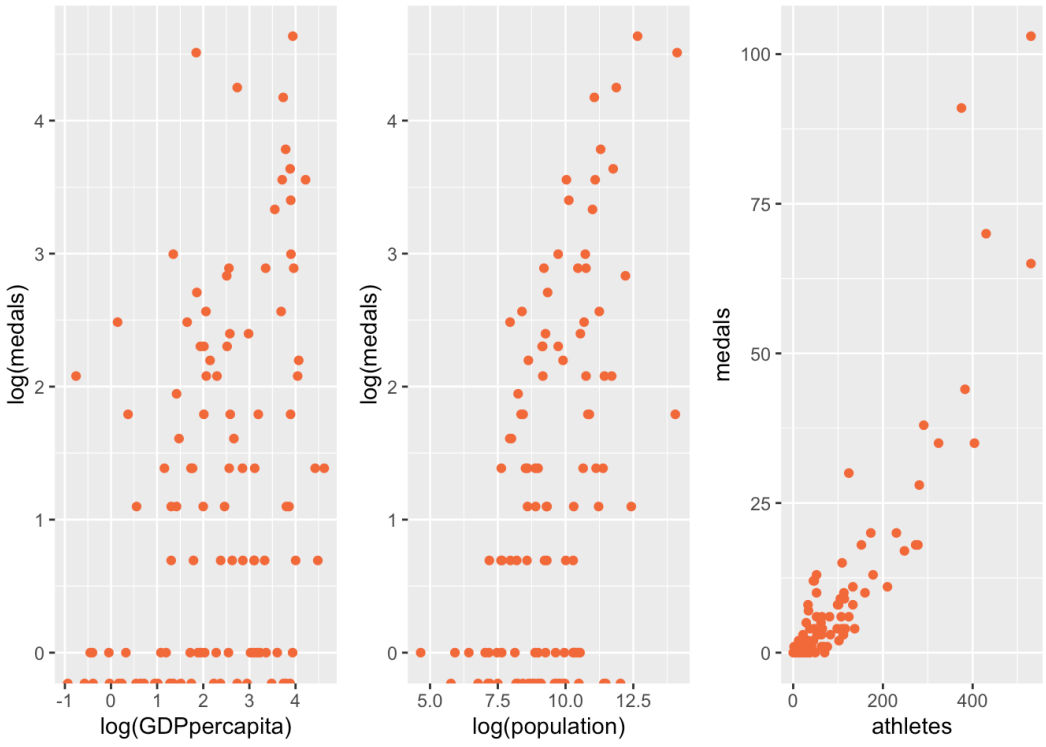
As we can see from the summary there is a wide range of values for variables gdp, population and GDPpercapita, therefore the log transformed versions were used moving forward.

From the plots below we see that there seems to be a positive relationship between athletes and medals, host and medals, oneparty and medals, GDPpercapita and medals and possibly between population and medals.

There also seems to be a positive linear relationship between athletes and GDPpercapita, athletes and host, athletes and oneparty and host and GDPpercapita; also between oneparty and population. There is a negative relationship between GDPpercapita and soviet, GDPpercapita and muslim, GDP per capita and comm.



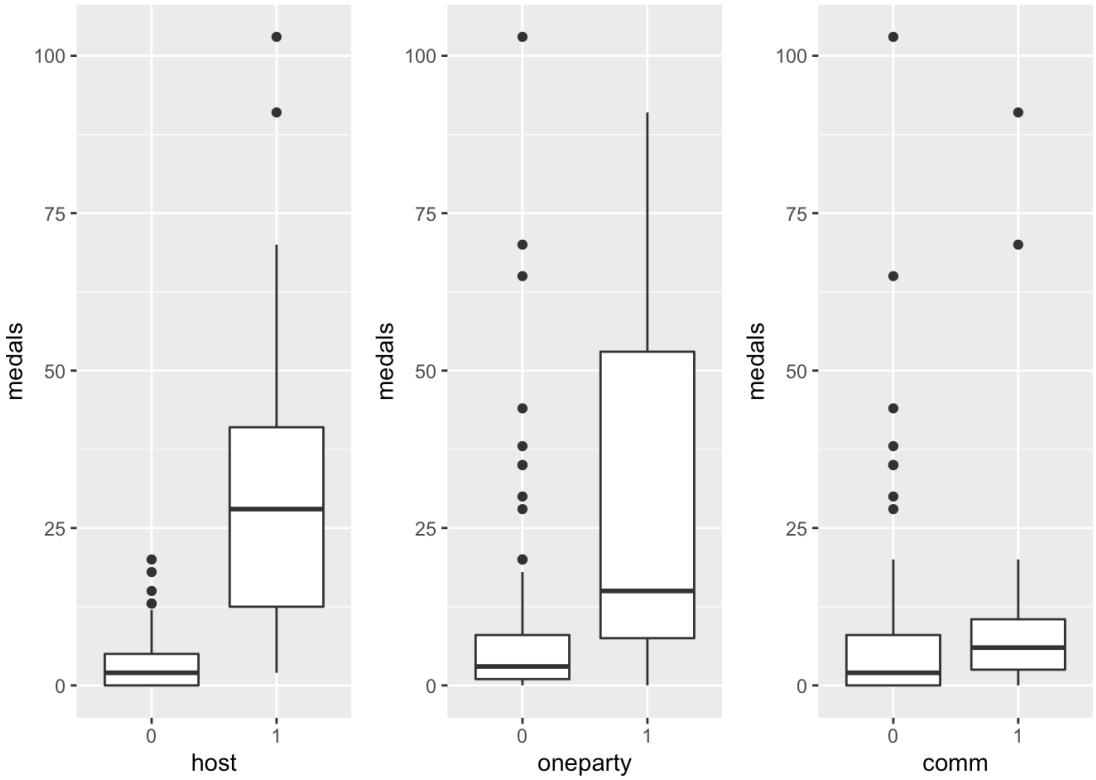
Below, a few pairs of variables have been chosen and looked at individually. We can see from the plots that there is a positive association between log(medals) and log(GDPpercapita); log(medals) and log(population) and medals and athletes. There are also a few potential outliers present.



We can see from the first plot below that the median number of medals won by countries who didn't host the Olympics lies below the median number of medals won by the countries who are/were/will be hosts.

We can see from the second plot that the median number of medals won by countries which are not one-party states lies below the median number of medals won by countries ruled by a one-part system. However, one thing to note is that there are only 3 countries with this type of political system, one of them being China, which has a very high population and has supplied a very high number of athletes.

The third plot shows us that the median number of medals won by countries who are/were communist lies above the median number of medals won by countries who never had this type of political regime.



GLM1

The first model we will fit will be a linear regression model with the response variable following the normal distribution. First, we fit a full model containing all the variables in our olympics12 dataset.

At the first glance, only two of the variables have p-values lower than 0.05: athletes and oneparty.

An outlier test has been performed on the full model using the function outlierTest from the *car* library. This has indicated that two of the observations, China and the USA are outliers. However, since both of the observations are genuine and they represent the natural variability of the data, the decision has been made not to remove them from further analysis.

For variable selection I have used a combination of forward selection and backward selection based on the AIC criteria.

From the first variable selection attempt, athletes and oneparty appeared as the most relevant variables. However, we have seen previously in the case of oneparty that there are only 3 countries with this type of political system, with China - an outlier - being one of them. Therefore, oneparty was removed from the model and the variable selection process was re-run.

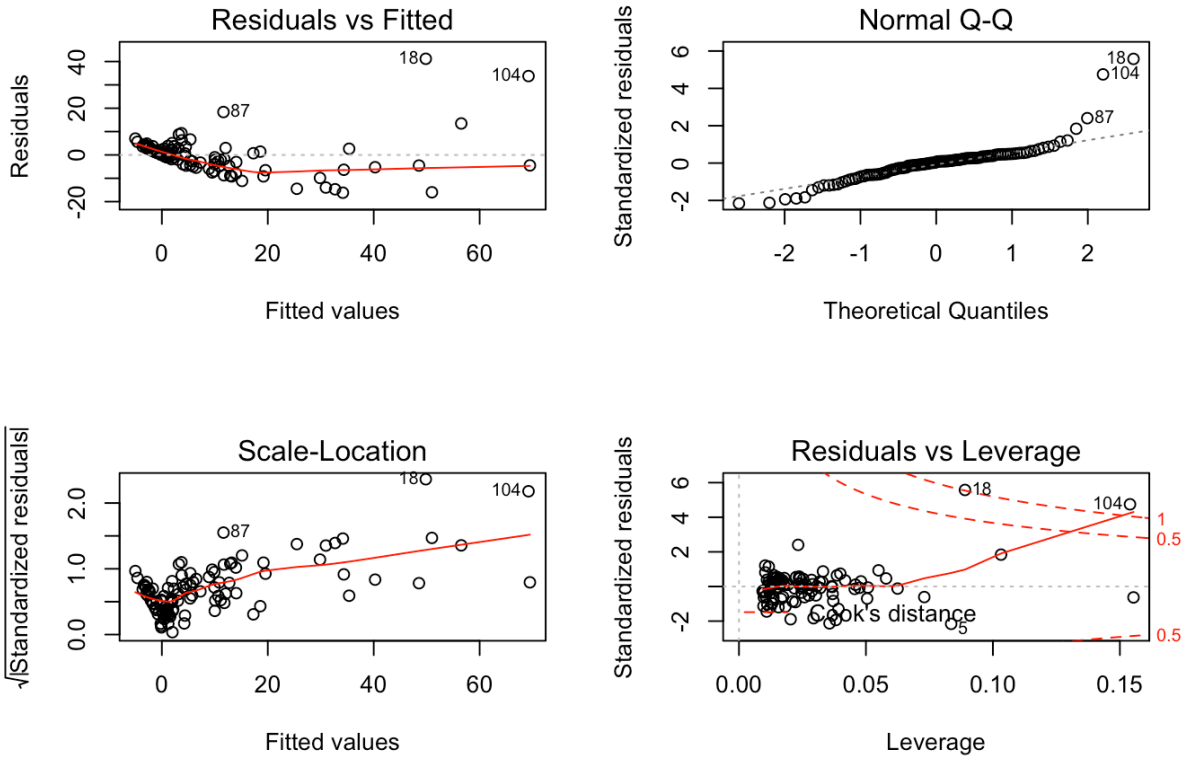
On the second attempt, log(GDPpercapita) and athletes were selected as the most significant variables for winning Olympic medals. Therefore, I fit a linear regression model with them as the explanatory variables.

The first GLM is:

E(medals_i|log(GDPpercapita)_i, athletes_i) = alpha + beta1log(GDPpercapita)_i + beta2athletes_i

Since the p-values for athletes and log(GDPpercapita) are lower than 0.05, we conclude that there is a statistically significant relationship between the two variables and the response variable, medals.The adjusted R-squared for the model is 0.7966, which means that log(GDPpercapita) and athletes explain almost 80% of the variability in our data.

Upon producing the residuals vs fitted values plot it has been noted that the mean of errors is not zero, meaning the model doesn't accurately capture deterministic part of the data. Also, the residuals appear clustered, and we have potential heteroscedascity. They also clearly point towards the presence of outliers, so it is a possibility that the plots would change for the better if those observations were to be removed. However, with the current circumstances, the suitability of the linear model is questionable.



The next step was to train the model on the training.data dataset. The summary can be viewed below:

```
##
## Call:
## lm(formula = medals ~ log(GDPpercapita) + athletes, data = training.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.432  -1.903   0.287   2.159  35.662
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.721036   0.975437  -1.764   0.0806 .
## log(GDPpercapita) -1.093912   0.477193  -2.292   0.0239 *
## athletes          0.128012   0.006349  20.164   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.286 on 105 degrees of freedom
## Multiple R-squared:  0.805,  Adjusted R-squared:  0.8013
## F-statistic: 216.7 on 2 and 105 DF,  p-value: < 2.2e-16
```

First, we notice that the intercept is negative. We can interpret the regression coefficients of trained model in the following way:

The coefficient for log(GDPpercapita) is negative meaning for every one unit increase in log(GDPpercapita), the number of medals decreases by -1.093 on average. On the other hand, for every unit increase in athletes, the number of medals won increases by 0.128.

GLM2

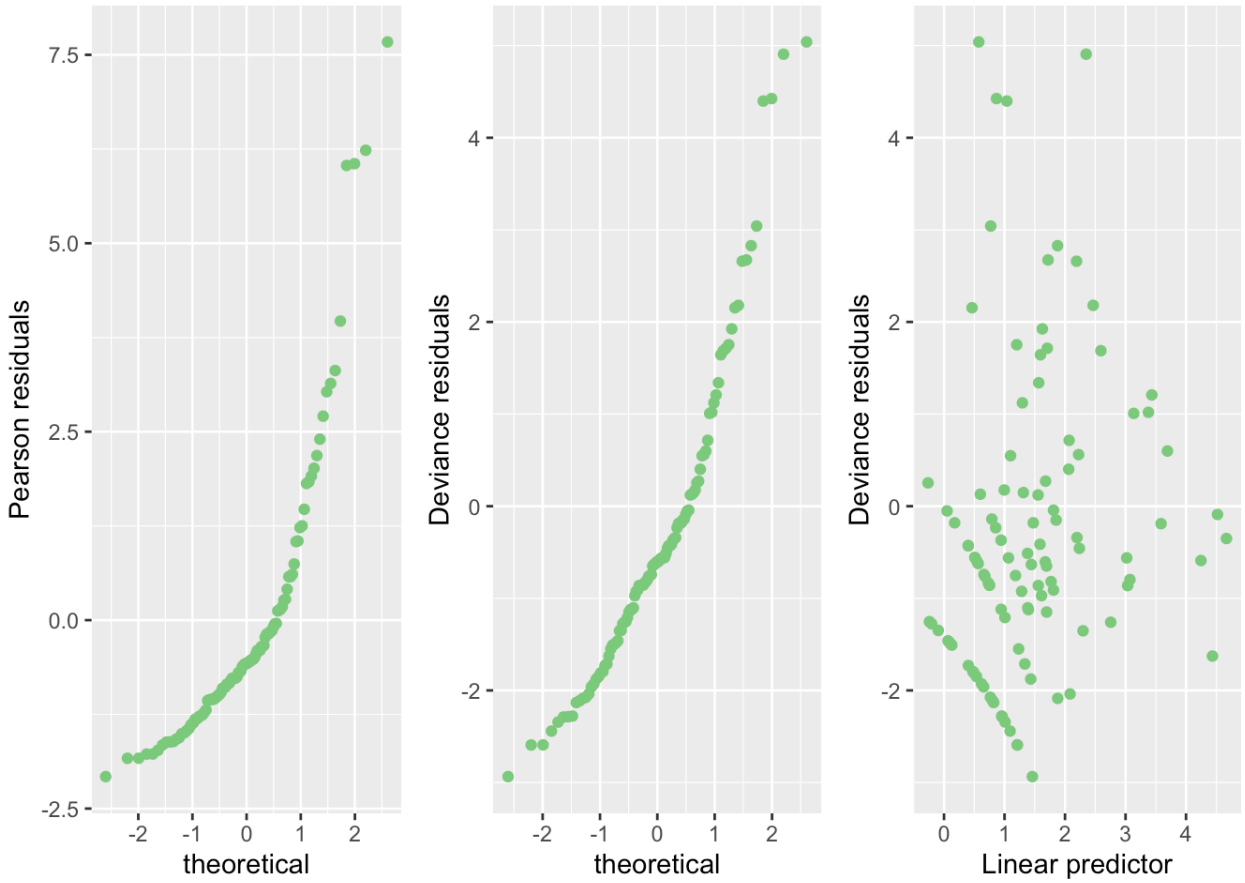
For the second model, we will fit a general linear model with the response variable following a Poisson distribution. Again, we first fit a full model, using all the explanatory variables in the olympics12 dataset.

At first glance, the residual deviance appears to be very large compared to the Chi-squared value.

For variable selection I have used a combination of forward selection and backward selection based on the AIC criteria.

Upon performing the variable selection process, the variables soviet, oneparty and altitude were flagged up as being not significant, and they were removed from the model. Unfortunately the residual deviance had increased further. This could be due to the presence of outliers in the data, the excess zeros or potentially overdispersion.

The Pearson residuals and deviance residuals plot were examined in the search for further insight.



The normal probability plots show some large positive residuals, and a few points deviating from the line. The deviance residuals plotted against the linear predictor shows a pattern in the data, pointing towards nonlinearity.

Next, I checked for overdispersion by computing the value of the dispersion parameter. Its value is 3.39, which is higher than 1, indicating we do have a slight overdispersion issue. In order to address this issue, I fit Quasi-Poisson model to the olympics12 dataset.

In order to determine the significance of the regression coefficients in the Quasi-Poisson model, I have used an F test. The reason for doing so, is because due the use of the dispersion parameter Wald tests are not reliable. Following the F test, muslim did not appear significant and was removed from the model.

The residual deviance from the Quasi-Poisson has decreased in comparison to our previous Poisson model, but it is still much larger than the Chi-squared value. Possible reasons include the excess zeros in the response variable, the presence of outliers or lack of fit of the model.

The second GLM is therefore the following:

log(E(medals_i|log(GDPpercapita)_i, log(population)_i, athletes_i, comm_i, host_i)) = alpha + beta1log(GDPpercapita)_i + beta2log(population)_i + beta3athletes_i + beta4**I**comm1_i + beta5**I**host1_i

The model was then trained on the training.data dataset. The summary can be viewed below:

```
##
## Call:
## glm(formula = medals ~ log(GDPpercapita) + log(population) +
##     athletes + comm + host, family = quasipoisson, data = training.data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.821  -1.648  -0.905   0.723   4.734
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.5132177  0.5596468  -2.704 0.008031 **
## log(GDPpercapita)  0.2043031  0.0729671   2.800 0.006114 **
## log(population)    0.1925184  0.0537785   3.580 0.000528 ***
## athletes           0.0040420  0.0005047   8.008 1.96e-12 ***
## comm1              1.2047803  0.1899336   6.343 6.25e-09 ***
## host1              0.6952732  0.2311166   3.008 0.003309 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 3.479668)
##
##     Null deviance: 2021.32  on 107  degrees of freedom
## Residual deviance:  340.33  on 102  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

For the second GLM, log(GDPpercapita), log(population), athletes, comm and host have positive coefficients, suggesting that the medal winning rate increases with increased log(GDPpercapita), log(population) and number of athletes. In the case of the binary variables, the presence of communism and the host factor is associated with a higher rate of winning medals.

The coefficients can be interpreted in the following way:

For every unit increase in log(GDPpercapita) the rate (for winning medals) increases by a factor of exp(0.2073458) = 1.230.

Likewise, exp(0.1971309) = 1.217 is the rate ratio associated with every unit increase in log(population).

For every unit increase in athletes, the rate increases by a factor of exp(0.0039341) = 1.004.

The comm coefficient indicates that the presence vs absence rate ratio is exp(1.1856268) = 3.272.

Similarly, in the case of host, the presence vs absence rate ratio is exp(0.7522124) = 2.121.

Predictions and findings

In the next step, the trained models were used to predict the number of medals won by each country at the 2016 Olympic games.

In the case of the first GLM, the linear model predicts values smaller than 0 for some countries. However if we set the negative values to 0, we can see that most of the predicted smaller values are not very far off from the observed values, the most significant exception to this being Argentina, for which our model over predicts by a very large margin.

The model seems to perform relatively well on mid-range values, with a few exceptions. However, the predictions for large values, such as the USA, seem to be completely off track.

In the case of the second GLM, which assumes a Poisson distribution, the model performs quite well at predicting mid-range values (between 20 and 50), but it constantly over predicts values in the smaller ranges (<10). Also for higher range values, the performance is not consistent, for example the predictions for USA are close to the number of medals won, but that is not the case for China.

In order to formally assess and compare the predictive performance of the two models we will use the root mean square error. The root mean square error (RMSE) is used as a measure for the difference between values predicted by a model and the values observed. The RMSE can be obtained by taking the square root of the mean of squared differences between predicted values and observed values.

In other words, RMSE is a measure of accuracy used to compare forecasting errors of different models for a particular dataset. (Hyndman et al., 2006)

RMSE is always non-negative and a value of 0 indicates a perfect fit to the data. In practice, this is almost never achieved.

Next, the RMSE was used to asses the performance of the two models.

The result for GLM1:

```
## [1] 9.360309
```

The result for GLM2:

```
## [1] 11.21795
```

From the results of the RMSE, GLM1 performs better in predicting the results for the 2016 Olympic Games than GLM2.

In summary, despite the diagnostic plots being far from ideal, the linear regression model provides the best prediction.

The Poisson model, which should be the most suitable model for these data, are not performing as well as the linear regression.

Discussion and Limitations

The project's aim has been to predict the number of medals won by each country during the 2016 Olympic Games. This has been achieved by using the olympics12 dataset for variable selection and model diagnostics and the dataset training.data has been utilised for training the model. The dataset olympics16 has been used to test the prediction performance of the models.

During the exploratory analysis, two outliers have been identified, and the observations have not been removed from the data. Also, it has been highlighted that several continuous variables such as log(GDPpercapita), athletes and log(population) appeared to have a linear relationship with the response variable. Several binary variables such as host, comm or oneparty were shown to affect the number of won medals.

Following this, two models were fitted to the data. For the linear regression model, the variables were selected using the AIC criteria. The diagnostic plots showed a poor fit to the data. In the case of the second model, following a Poisson distribution, the AIC criteria was used again in order to identify the most significant variables. Due to the over dispersion in the data, the final model used was a Quasi-Poisson model.

The models were then trained and the test dataset was used to generate predictions.

Finally, the root mean square error was used to determine and compare the predictive performance of the two models. Doing so has revealed that despite the apparent lack of fit, GLM1 is better at predicting the number of medals than GLM2.

In order to improve the results of the analysis, there are a few possible steps that could be taken. First of all, I think the removal of the outliers could improve the fit of the two models. This is an option that could be looked at and tested.

Secondly, exploring the use of additional variables would be helpful. For example, previous years' performance could be used as a predictor. I believe this could significantly improve the prediction capability for two models, since present performance is likely to be conditioned on past performance in this case.

Lastly, considering the excess zeros in the response variables I believe it would be worth fitting a zero inflated Poisson model.

References

Hyndman, R.& Koehler, A. (2006). *"Another look at measures of forecast accuracy"*. International Journal of Forecasting. **22** (4): 679–688.