Assessed Practical: Predicting the Olympic Games

Ahmed Ghafouri

10 Mar 2021

Introrduction

The olympic games are one of, if not the most famous sporting event in history. Each country can compete to get the most number of medals as nation. However, the medal counting system has come under scrutiny for not reflecting the success of countries with less GDP and/or lower population compared to other nations. Many nations believe that GDP and population have a strong effect on how many medals are won at the olympics and results in poor reflection of some countries athletic prowess.

This report will use of linear regression and multiple linear regression to assess how much affect GDP and population have on the number of medals won. This report will also use regression in assessing how well the number of medals won at the olympics can be predicted using just GDP and population.

Task 1:

A multiple linear regression model will be built with GDP and population variables as inputs within the generalized linear model (glm) function for 2008 and 2012 medal count.

The glm function will calculate the coefficients of the linear regression for each year using numerical methods. The coefficients for y intercept, population and GDP are labeled, beta 0 beta1 and beta 2 respectively in equation 1.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (eq\ 1)$$

The family input for the glm function was not specified resulting in a guassian base function being used to build the glm. The code the glm can be seen for 2008 and 2012 bellow.

```
medal_pop_gdp_data_statlearn=read.csv2('C:/Users/Ahmed/Desktop/medal_pop_gdp_data_statlearn.csv' ,heade

model2008 = glm(Medal2008 ~Population + GDP ,family = gaussian, data = medal_pop_gdp_data_statlearn)

model2012 = glm(Medal2012 ~Population + GDP ,family = gaussian, data = medal_pop_gdp_data_statlearn)
```

The estimate for the coefficients for intercept (β_0), the coefficient for population variable (β_1) and coefficient for GDP variable (β_2) can be seen under the column labeled estimate bellow for 2008 and 2012 respectively

```
print(summary(model2008)$coefficients[, 1:3])
```

```
##                 Estimate   Std. Error   t value
## (Intercept) 5.613267e+00 1.505558e+00  3.728364
## Population  8.434824e-09 7.219509e-09  1.168338
## GDP         7.613135e-03 7.352773e-04 10.354100
```

```r
print(summary(model2012)$coefficients[, 1:3])
```

```
##                 Estimate   Std. Error    t value
## (Intercept) 6.076092e+00 1.499954e+00  4.0508527
## Population  5.246750e-09 7.192637e-09  0.7294612
## GDP         7.564081e-03 7.325406e-04 10.3258183
```

Task 2

The coefficients act as a sort of weight on each variable and indicate how much of an effect each variable has on the output variable. Th higher the coefficient, the more affect that variable has on the number of medals won that year.

As can be seen by the estimate of the coefficients above, the coefficient for population is 8.43 x 10ˆ-9 and decreases to 5.24 x 10ˆ-9. This suggests that population is having less of an effect on number of medals won, at least between 2008 and 2012. The estimate for the coefficient of GDP is 7.61 x 10ˆ-3 and decreases to 7.56 x 10ˆ-3 suggesting GDP is also having less of an effect on number of medals won, at least between 2008 and 2012 as well.

Task 3

Once the regression models have been built, they can be used to make predictions on unseen data. The model built using the medal count in 2012 can be used to make predictions about 2016 medal count for each country. This can be done by pulling out the coefficients of the model built using the 2012 medal count and applying these coefficients to population and GDP data in 2016. However, there is only one value for GDP and population for each country in the data set. As there is no value for GDP and population for each country for 2008, 2012 and 2016, the GDP and population used for all 3 years will be fed into the model built with the 2012 medal count. By feeding the GDP and population values into the 2012 regression model(with the 2012 GDP and population coefficients) a prediction for the 2016 medal count can be made that wont be the most accurate prediction due to the GDP and population data not varying with the years in the data set.

The code bellow shows the values predicted for 2016 medal count by the regression model built using the 2012 model count in the medal2016_pred line of code. The medal2016_pred being the number of medals predicted for each country.

```r
x1=medal_pop_gdp_data_statlearn$GDP
x2=medal_pop_gdp_data_statlearn$Population

model2012 = glm(Medal2012 ~Population + GDP ,family = gaussian, data = medal_pop_gdp_data_statlearn,)
cgdp=summary(model2012)$coefficients[3,1]

cpop=summary(model2012)$coefficients[2,1]

inter=summary(model2012)$coefficients[1,1]
medal2016_pred<-inter+cgdp*x1 + cpop*x2
medal2016_actual<-medal_pop_gdp_data_statlearn$Medal2016
```

Task 4

The predicted number of medals won and actual number of medals won by each country will be compared by inputting both data sets for predicted and actual medals won by each country into a correlation function as seen below. A value of 0.88 for the correlation between the predicted and actual medal count is given. This correlation suggests there is a strong positive relationship between the medal predicted and actual values which suggests the predictions are fairly good. However, the more accurate the prediction is, the closer each data point is to the y=x line when plotting the predicted values against the actual values.

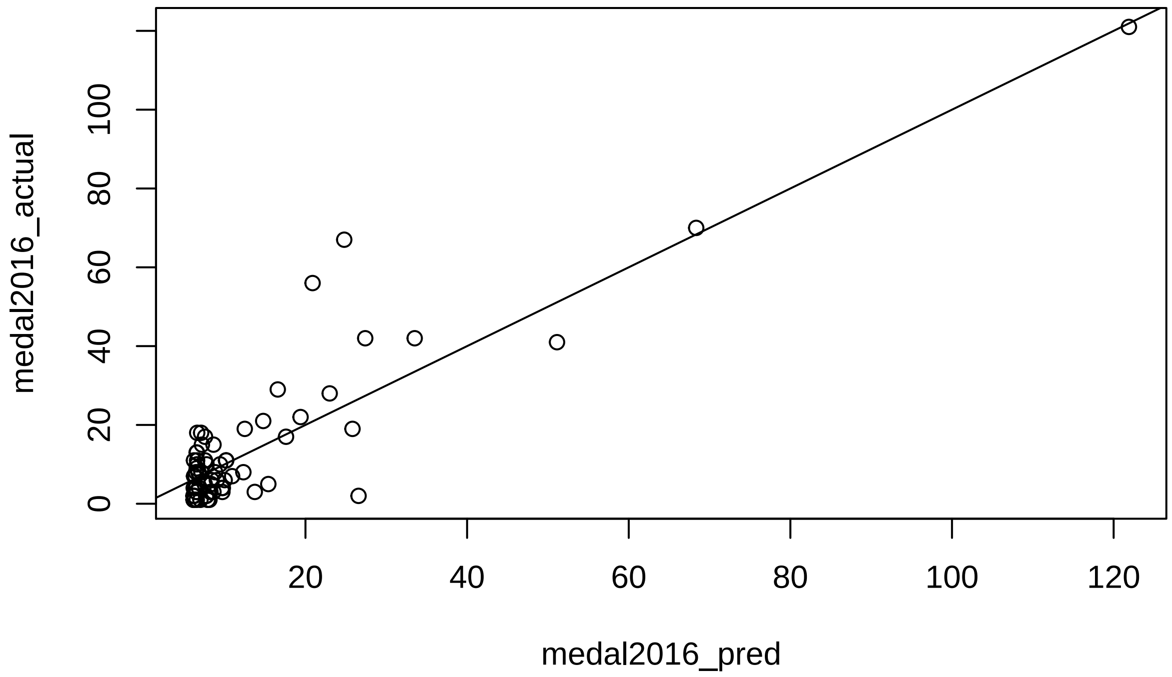
The plot for the predicted and actual medal count can be seen with the y=x line in the figures bellow. The only difference between the 2 plots is the 2nd plot is on the log scale for both predicted and actual values. The 2nd plot is in the log scale to make outliers more apparent.

For a country to be considered an outlier, the residual of each data point would have to be found. Any residual higher than a certain threshold could be considered an outlier. The residuals of the data points from the y=x line were not calculated in this report.

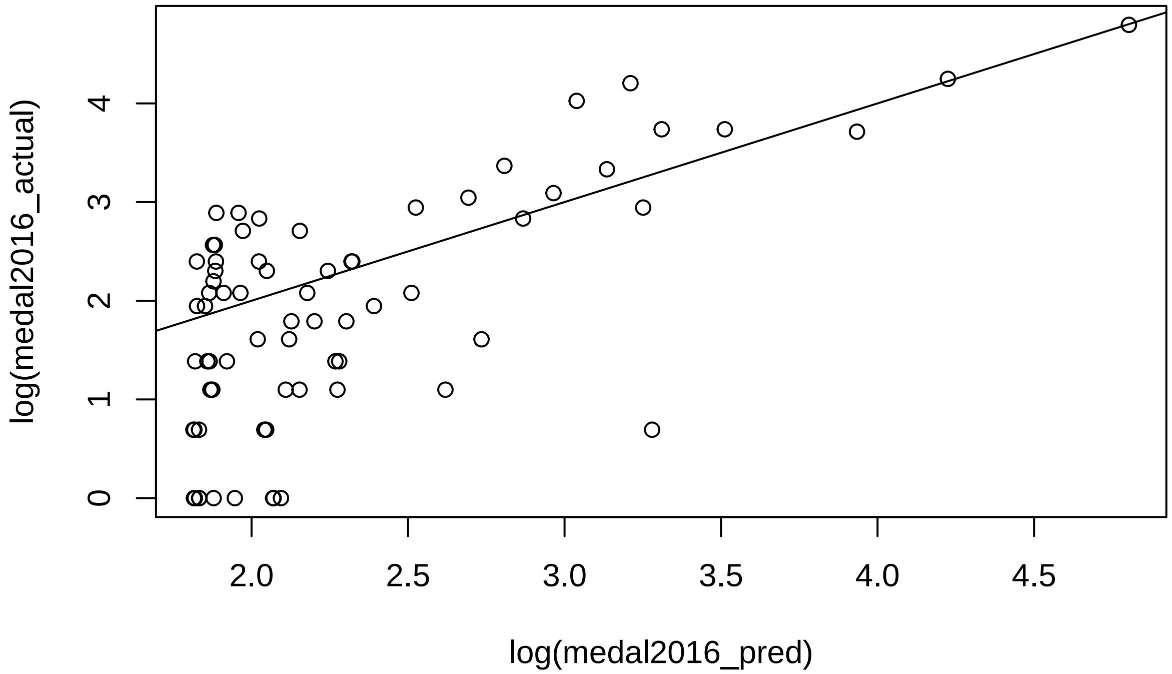
```
cor(medal2016_pred,medal2016_actual)
```

```
## [1] 0.8830621
```

```
plot(medal2016_pred, medal2016_actual)
abline(a=0, b=1)
```



```
plot(log(medal2016_pred),log(medal2016_actual))

abline(a=0, b=1)
```



Section 2: Model selection

Task 1

Three regression models will be built using the medal count for 2012 using inputs: (i) Population alone; (ii) GDP alone; (iii) Population and GDP.

.

```
pop_model <- glm(Medal2012 ~ Population, data = medal_pop_gdp_data_statlearn)
GDP_model <- glm(Medal2012 ~ GDP, data = medal_pop_gdp_data_statlearn)
pop_GDP_model <- glm(Medal2012 ~ GDP+Population, data = medal_pop_gdp_data_statlearn)
```

Pop model AIC: 618.15 GDP model AIC: 551.74 GDP and Pop AIC: 553.19

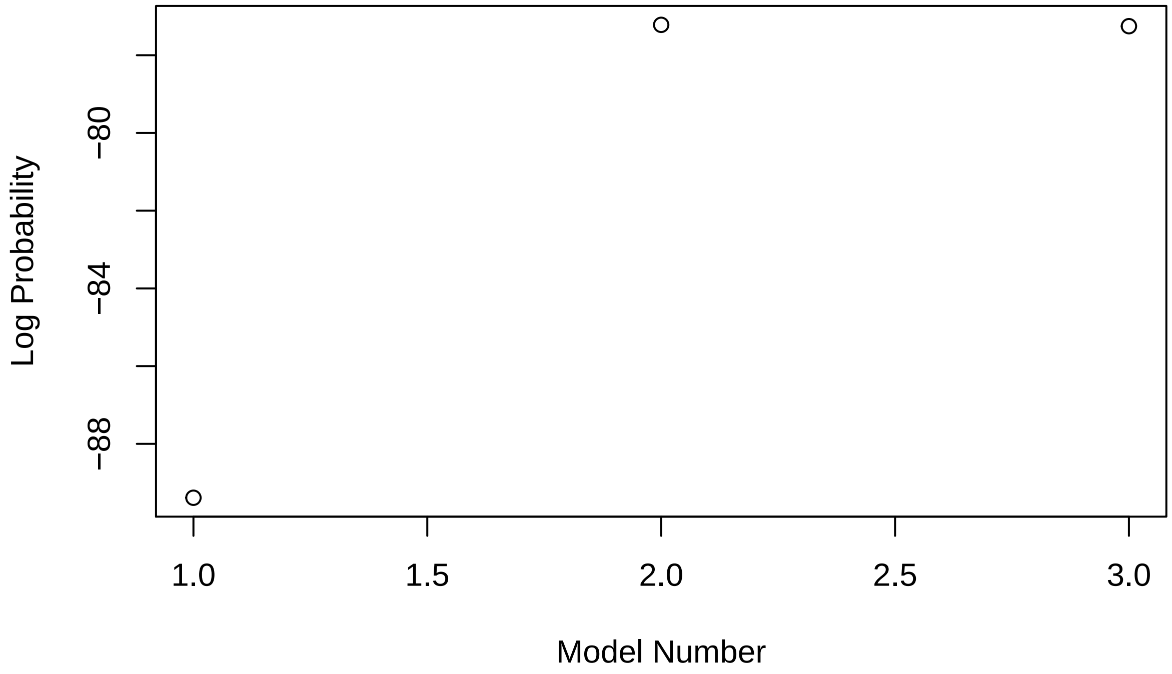
The AIC values above are used for model selection where the lower the AIC value, the better the model. The AIC is used as it penalizes too many parameters being used in the model. Too many unnecessary parameters is something to be avoided as too many parameters could lead to overfitting of the regression model. If the model overfits, it will struggle to generalize well when using unseen data to make predictions

From the AIC results of each model, the model using just GDP has the lowest AIC and therefore performs the best

Task 2

Cross validation can also be used for model selection. Cross validation will be used to assess the same 3 models built in the previous section where the inputs for each model are (i) Population alone; (ii) GDP alone; (iii) Population and GDP.

```r
X1=medal_pop_gdp_data_statlearn$GDP
X2=medal_pop_gdp_data_statlearn$Population
Y=medal_pop_gdp_data_statlearn$Medal2012
mydata = medal_pop_gdp_data_statlearn

idx = sample(1:71, 50) #sample 100 points in 1...200 without replacement
train_data = medal_pop_gdp_data_statlearn[idx, ]; test_data = medal_pop_gdp_data_statlearn[-idx, ]

formulas = c("Medal2012 ~ Population", "Medal2012 ~ GDP", "Medal2012 ~ GDP+Population")


predictive_log_likelihood = rep(NA, length(formulas))
for (i in 1:length(formulas)){

  current_model = glm(formula = formulas[i], data = medal_pop_gdp_data_statlearn)

  sigma = sqrt(summary(current_model)$dispersion)

  ypredict_mean = predict(current_model, test_data)

  predictive_log_likelihood[i] = sum(dnorm(test_data$Medal2012,
                                           ypredict_mean, sigma, log=TRUE))

}

plot(1:length(formulas), predictive_log_likelihood,
     xlab="Model Number", ylab="Log Probability")
```



The figure above shows that model 2 which was model with just GDP as an input had the highest log probability which suggests it is the best model. The way the data was split into training and test data affects the outcome of each model. Due to this, each model would have to be tested at different splits and the frequency of the oh which model came out with the highest log probability would have to be calculated in order to assess which model was best. From trial and error of many values, there is a strong suggestion that the model using just GDP as an input is the best model. The cross validation also suggests the GDP model performs the best

Task 3

The three fitted models from Model Selection Task 1 will be used to predict the results of Rio 2016 as seen by the code bellow.

```
x1=medal_pop_gdp_data_statlearn$GDP
x2=medal_pop_gdp_data_statlearn$Population

pop_model <- glm(Medal2012 ~ Population, data = medal_pop_gdp_data_statlearn)
GDP_model <- glm(Medal2012 ~ GDP, data = medal_pop_gdp_data_statlearn)
pop_GDP_model <- glm(Medal2012 ~ GDP+Population, data = medal_pop_gdp_data_statlearn)
#################################################

Pop_intercept=summary(pop_model )$coefficients[1,1]
Pop_grad=summary(pop_model )$coefficients[2,1]

pop_pred<-Pop_intercept+Pop_grad*x2
```

```r
pop_actual<-medal_pop_gdp_data_statlearn$Medal2016
cor(pop_pred,pop_actual)
```

```
## [1] 0.3575079
```

```r
#############################################
GDP_intercept=summary(GDP_model )$coefficients[1,1]
GDP_grad=summary(GDP_model )$coefficients[2,1]

GDP_pred<-GDP_intercept+GDP_grad*x1
GDP_actual<-medal_pop_gdp_data_statlearn$Medal2016
cor(GDP_pred,GDP_actual)
```

```
## [1] 0.8888765
```

```r
##########################################

Pop_GDP_intercept=summary(pop_GDP_model)$coefficients[1,1]
Pop_GDP_grad1=summary(pop_GDP_model)$coefficients[2,1]
Pop_GDP_grad2=summary(pop_GDP_model)$coefficients[3,1]

Pop_GDP_pred<-GDP_intercept+Pop_GDP_grad1*x1+Pop_GDP_grad2*x2
Pop_GDP_actual<-medal_pop_gdp_data_statlearn$Medal2016

cor(Pop_GDP_pred,Pop_GDP_actual)
```

```
## [1] 0.8830621
```

Correlations

Population: 0.3575079 GDP: 0.8888765 GDP and Population:0.8830621

The correlation function will be used to assess how well each model does in predicting the medal count of 2016 for each country. The correlation for the model using GDP alone had the highest correlation suggesting it had the best predictive capabilities. Though the model using GDP and population did almost as well, the model using GDP alone has a lower AIC.

Conclusion

From assessing models using the following inputs: (i) Population alone; (ii) GDP alone; (iii) Population and GDP, it was found that the model with only GDP as the input did the best at predicting medal counts. This would make the most sense as a country with a high population would not mean that the country can afford to do many sports. Many sports in the olympics not only require expensive training equipment but also training facilities and the ability to earn a salary whilst playing their competitive sport. Many countries have high populations but not enough people with the sort of money to play the sports that are in the olympics. Indonesia is a good example of this. Indonesia is the 4th most populated country in the world but no where near the top of the olympic charts.

Appendix

(*)

Introrduction

The olympic games are one of, if not the most famous sporting event in history. Each country can compete to get the most number of medals as nation. However, the medal counting system has come under scrutiny for not reflecting the success of countries with less GDP and/or lower population compared to other nations. Many nations believe that GDP and population have a strong effect on how many medals are won at the olympics and results in poor reflection of some countries athletic prowess.

This report will use of linear regression and multiple linear regression to assess how much affect GDP and population have on the number of medals won. This report will also use regression in assessing how well the number of medals won at the olympics can be predicted using just GDP and population.

Task 1:

A multiple linear regression model will be built with GDP and population variables as inputs within the generalized linear model (glm) function for 2008 and 2012 medal count.

The glm function will calculate the coefficients of the linear regression for each year using numerical methods. The coefficients for y intercept, population and GDP are labeled, beta 0 beta1 and beta 2 respectively in equation 1.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (eq\ 1)$$

The family input for the glm function was not specified resulting in a guassian base function being used to build the glm. The code the glm can be seen for 2008 and 2012 bellow.

medal_pop_gdp_data_statlearn=read.csv2(‘C:/Users/Ahmed/Desktop/medal_pop_gdp_data_statlearn.csv’,header = TRUE,sep = “,”, quote = “"”,dec = “.”)

model2008 = glm(Medal2008 ~Population + GDP ,family = gaussian, data = medal_pop_gdp_data_statlearn)

model2012 = glm(Medal2012 ~Population + GDP ,family = gaussian, data = medal_pop_gdp_data_statlearn)

The estimate for the coefficients for intercept ($\beta_0$), the coefficient for population variable (

```
print(summary(model2008)$coefficients[, 1:3])
```

print(summary(model2012)$coefficients[, 1:3])

## Task 2

2. How consistent are the effects of Population and GDP over time?

The coefficients act as a sort of weight on each variable and indicate how much of an effect each varia

As can be seen by the estimate of the coefficients above, the coefficient for population is 8.43 x 10^-

## Task 3

Once the regression models have been built, they can be used to make predictions on unseen data. The mo

```
The code bellow shows the values predicted for 2016 medal count by the regression model built using the
```

```
x1=medal_pop_gdp_data_statlearn$GDP
x2=medal_pop_gdp_data_statlearn$Population

model2012 = glm(Medal2012 ~Population + GDP ,family = gaussian, data = medal_pop_gdp_data_statlearn,)
cgdp=summary(model2012)$coefficients[3,1]

cpop=summary(model2012)$coefficients[2,1]

inter=summary(model2012)$coefficients[1,1]
medal2016_pred<-inter+cgdp*x1 + cpop*x2
medal2016_actual<-medal_pop_gdp_data_statlearn$Medal2016
```

Task 4

The predicted number of medals won and actual number of medals won by each country will be compared by inputting both data sets for predicted and actual medals won by each country into a correlation function as seen below. A value of 0.88 for the correlation between the predicted and actual medal count is given. This correlation suggests there is a strong positive relationship between the medal predicted and actual values which suggests the predictions are fairly good. However, the more accurate the prediction is, the closer each data point is to the y=x line when plotting the predicted values against the actual values.

The plot for the predicted and actual medal count can be seen with the y=x line in the figures bellow. The only difference between the 2 plots is the 2nd plot is on the log scale for both predicted and actual values. The 2nd plot is in the log scale to make outliers more apparent.

For a country to be considered an outlier, the residual of each data point would have to be found. Any residual higher than a certain threshold could be considered an outlier. The residuals of the data points from the y=x line were not calculated in this report.

cor(medal2016_pred,medal2016_actual)

plot(medal2016_pred, medal2016_actual) abline(a=0, b=1)

plot(log(medal2016_pred),log(medal2016_actual))

abline(a=0, b=1)

which countries are outlines

```
## Section 2: Model selection

## Task 1
Three regression models will be built using the medal count for 2012 using inputs: (i) Population alone



.
```

```
pop_model <- glm(Medal2012 ~ Population, data = medal_pop_gdp_data_statlearn)
GDP_model <- glm(Medal2012 ~ GDP, data = medal_pop_gdp_data_statlearn)
pop_GDP_model <- glm(Medal2012 ~ GDP+Population, data = medal_pop_gdp_data_statlearn)
```

Pop model AIC: 618.15 GDP model AIC: 551.74 GDP and Pop AIC: 553.19

The AIC values above are used for model selection where the lower the AIC value, the better the model. The AIC is used as it penalizes too many parameters being used in the model. Too many unnecessary parameters is something to be avoided as too many parameters could lead to overfitting of the regression model. If the model overfits, it will struggle to generalize well when using unseen data to make predictions

From the AIC results of each model, the model using just GDP has the lowest AIC and therefore performs the best

Task 2

Cross validation can also be used for model selection. Cross validation will be used to assess the same 3 models built in the previous section where the inputs for each model are (i) Population alone; (ii) GDP alone; (iii) Population and GDP.

X1=medal_pop_gdp_data_statlearn$GDPX2 = medal_{p}op_{g}dp_{d}ata_{s}tatlearn$Population Y=medal_pop_gdp_data_statlearn mydata = medal_pop_gdp_data_statlearn

idx = sample(1:71, 50) #sample 100 points in 1…200 without replacement train_data = medal_pop_gdp_data_statlearn[idx]; test_data = medal_pop_gdp_data_statlearn[-idx,]

formulas = c("Medal2012 ~ Population", "Medal2012 ~ GDP", "Medal2012 ~ GDP+Population")

predictive_log_likelihood = rep(NA, length(formulas)) for (i in 1:length(formulas)){

current_model = glm(formula = formulas[i], data = medal_pop_gdp_data_statlearn)

sigma = sqrt(summary(current_model)$dispersion)

ypredict_mean = predict(current_model, test_data)

predictive_log_likelihood[i] = sum(dnorm(test_data$Medal2012, ypredict_mean, sigma, log=TRUE))

}

plot(1:length(formulas), predictive_log_likelihood, xlab="Model Number", ylab="Log Probability")

```
The figure above shows that model 2 which was model with just GDP as an input had the highest log probal
The cross validation also suggests the GDP model performs the best

## Task 3
The three fitted models from Model Selection Task 1 will be used to predict the results of Rio 2016 as s
```

```
x1=medal_pop_gdp_data_statlearn$GDP
x2=medal_pop_gdp_data_statlearn$Population

pop_model <- glm(Medal2012 ~ Population, data = medal_pop_gdp_data_statlearn)
GDP_model <- glm(Medal2012 ~ GDP, data = medal_pop_gdp_data_statlearn)
pop_GDP_model <- glm(Medal2012 ~ GDP+Population, data = medal_pop_gdp_data_statlearn)
##################################################

Pop_intercept=summary(pop_model )$coefficients[1,1]
Pop_grad=summary(pop_model )$coefficients[2,1]
```

```
pop_pred<-Pop_intercept+Pop_grad*x2
pop_actual<-medal_pop_gdp_data_statlearn$Medal2016
cor(pop_pred,pop_actual)
#############################################
GDP_intercept=summary(GDP_model )$coefficients[1,1]
GDP_grad=summary(GDP_model )$coefficients[2,1]

GDP_pred<-GDP_intercept+GDP_grad*x1
GDP_actual<-medal_pop_gdp_data_statlearn$Medal2016
cor(GDP_pred,GDP_actual)
##########################################

Pop_GDP_intercept=summary(pop_GDP_model)$coefficients[1,1]
Pop_GDP_grad1=summary(pop_GDP_model)$coefficients[2,1]
Pop_GDP_grad2=summary(pop_GDP_model)$coefficients[3,1]

Pop_GDP_pred<-GDP_intercept+Pop_GDP_grad1*x1+Pop_GDP_grad2*x2
Pop_GDP_actual<-medal_pop_gdp_data_statlearn$Medal2016

cor(Pop_GDP_pred,Pop_GDP_actual)
```

Correlations

Population: 0.3575079 GDP: 0.8888765 GDP and Population:0.8830621

The correlation function will be used to assess how well each model does in predicting the medal count of 2016 for each country. The correlation for the model using GDP alone had the highest correlation suggesting it had the best predictive capabilities. Though the model using GDP and population did almost as well, the model using GDP alone has a lower AIC.

Conclsuion

From assessing models using the following inputs: (i) Population alone; (ii) GDP alone; (iii) Population and GDP, it was found that the model with only GDP as the input did the best at predicting medal counts. This would make the most sense as a country with a high population would not mean that the country can afford to do many sports. Many sports in the olympics not only require expensive training equipment but also training facilities and the ability to earn a salary whilst playing their competitive sport. Many countries have high populations but not enough people with the sort of money to play the sports that are in the olympics. Indonesia is a good example of this. Indonesia is the 4th most populated country in the world but no where near the top of the olympic charts.)