VU BI2 - Exercise 1: Olympics Data

Asil Cetin / 01100130

23/11/2017

Changelog

23.11.2017

Changes made in this version are:

- Labels of countries on scatterplots are added.
- Correlation functions for observed parameters are added.
- Multiple regression analysis is added.
- Confidence intervals for the model coefficents are added.
- Borda points method is questioned.
- Some text changes and minor improvements.

06.11.2017

Changes made in this version are:

- Initial version with loading and cleaning data.
- Descriptive analysis.
- Linear regression analysis of selected parameters.
- Presentation data and text are created.

Data Description

In this exercise we are investigating a dat set from London 2012 Olympics. The data set gives the names of the 203 participating countries as well as the number of gold, silver and bronze medals won by country, the total number of medals won by country, the Borda points by country, income per capita (in $1.000), population size (in 1.000.000), gross domestic product (GDP= income per capita multiplied by population size) and the polynomial variables of income per capita squared, population size squared, income per capita cubed, population size cubed, gross domestic product squared, gross domestic product cubed, natural log of income per capita, natural log of population size, and natural log of GDP.

Analysis Questions

We are mainly interested in the corelation between the overall success in London 2012 Olympics - which is represented in the parameter “BordaPoints”, since it ranks the countries weighted on the value of different medals - and the population and income levels of a given country.

Thus our first analysis question can be stated as:

Do parameters of population size and income have any effect on success in the London 2012 Olympics?

After investigating the possible correlations between olympic success and parameters of population size and income, we would want to know to what degree these parameters have an effect. Thus our second question would be:

If population size and income has an effect on olympic success, what are the factoring weights of these parameters?

Correlation of Observed Parameters

Total of Borda points, income per person (ln) and the total population of a country (ln) are the relevant selected parameters for us. To have a first glance at these parameters and their relationships we can make use of a correlation analysis.

First let's look at the correlation between Borda points and income per person (ln):

```
## 
##  Pearson's product-moment correlation
## 
## data:  Ln(Income) and BordaPoints
## t = 3.7178, df = 201, p-value = 0.0002605
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1201471 0.3781601
## sample estimates:
##     cor 
## 0.25366
```

Let's observe the correlation between Borda points and population size (ln) as well:

```
## 
##  Pearson's product-moment correlation
## 
## data:  Ln(PopnSize) and BordaPoints
## t = 5.8417, df = 201, p-value = 2.055e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2567258 0.4928220
## sample estimates:
##      cor 
## 0.380967
```

As we'll do in the later steps, we want to observe the countries with at least 5 Borda points separately from the whole data (this analysis will be elaborated in the next sections). Here is the correlation between Borda points and income per person (ln) of those countries with at least 5 Borda points:

```
## 
##  Pearson's product-moment correlation
## 
## data:  FBC$`Ln(Income)` and FBC$BordaPoints
## t = 1.8366, df = 50, p-value = 0.07222
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02309386  0.49063153
## sample estimates:
##       cor 
## 0.2513908
```

Let's observe the correlation between Borda points and population size (ln) of those countries with at least 5 Borda points as well:

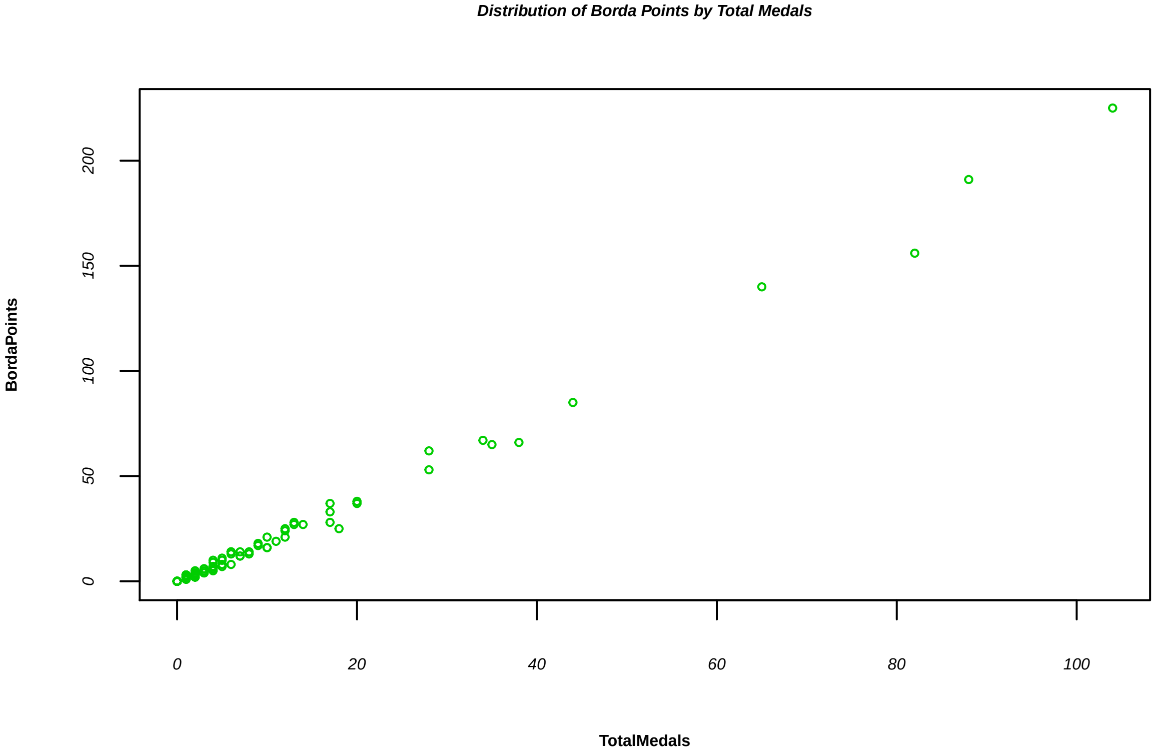
```
##
##  Pearson's product-moment correlation
##
## data:  FBC$`Ln(PopnSize)` and FBC$BordaPoints
## t = 4.6751, df = 50, p-value = 2.259e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3279813 0.7165676
## sample estimates:
##       cor
## 0.5515174
```

Borda Points, an Accurate Parameter for Success?

One might as if Borda points method deliver an accurate representation for the success in olympics. First it may be useful to explain the method of calculation for Borda points:

For each medal in the olympics a value for Borda points is defined to have different weights for different medals. Thus for gold medals three (3) Borda points are given, for silver two (2) and for bronze one (1) Borda points are added to the sums of countries.

Let's look at the relationship between total medals gained in olympics and their respective Borda point calculations:



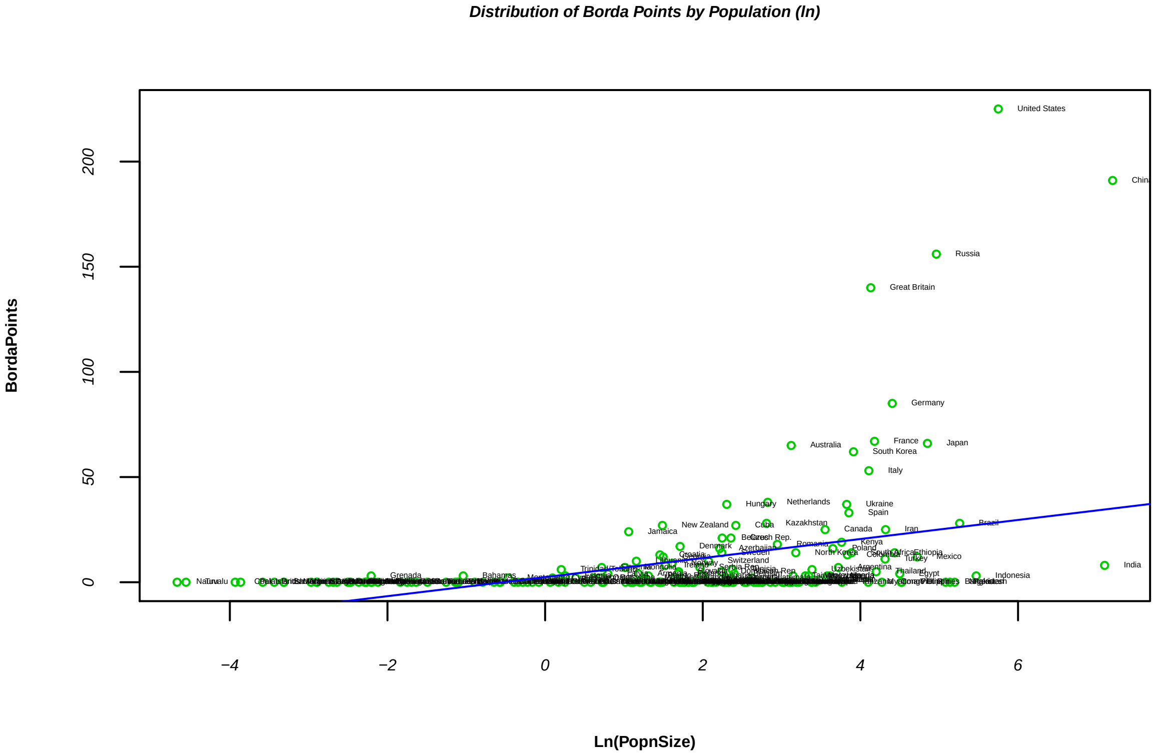
```
##
##  Pearson's product-moment correlation
##
## data:  TotalMedals and BordaPoints
## t = 168.2, df = 201, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9953406 0.9973208
## sample estimates:
##       cor
## 0.9964665
```

As we can observe, Borda points method is a direct representation of total medals awarded in olympics. The only difference this method brings to the table is the different weights of medals so that the the success of countries with different amount of different medals can be compared in a more accurate way.

Population and Success

First parameter we want to analyse is the population. Here we will use the natural logarithm of population size (in 1.000.000) since it's preferred to naturalize the enourmous population size differences between some contries.

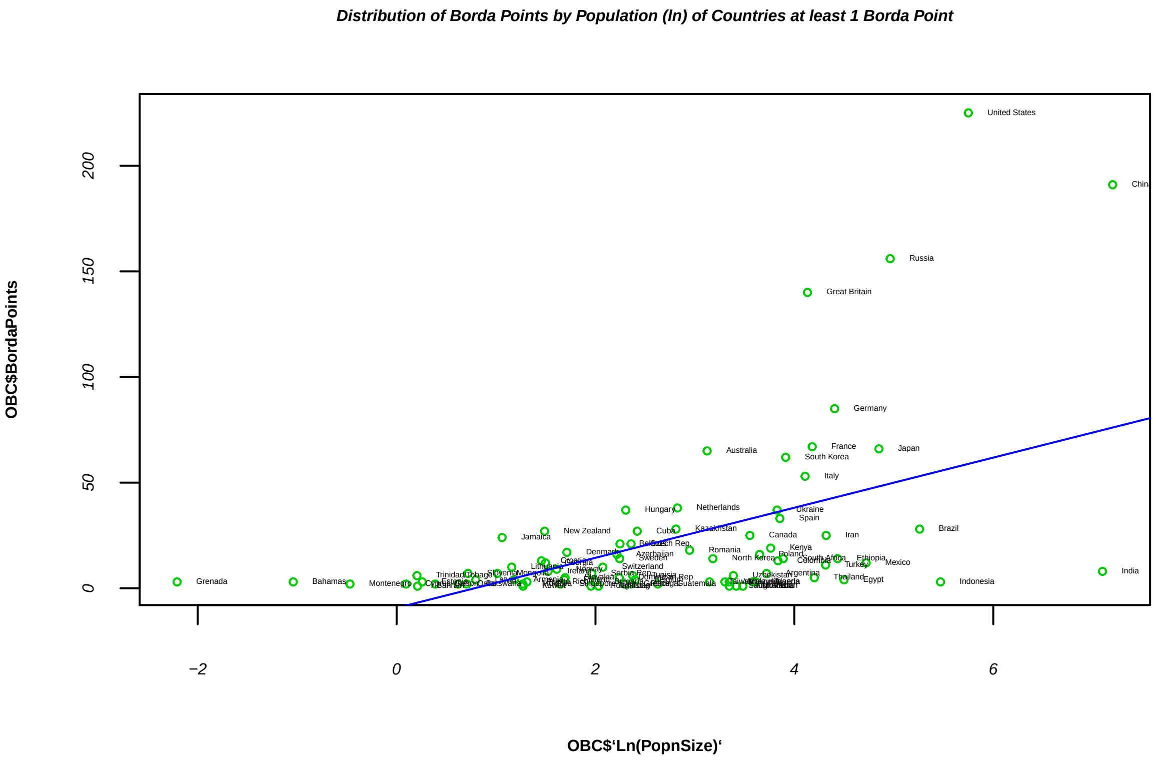
The plot of all participating countries in the 2012 London Olympics looks like as follows:



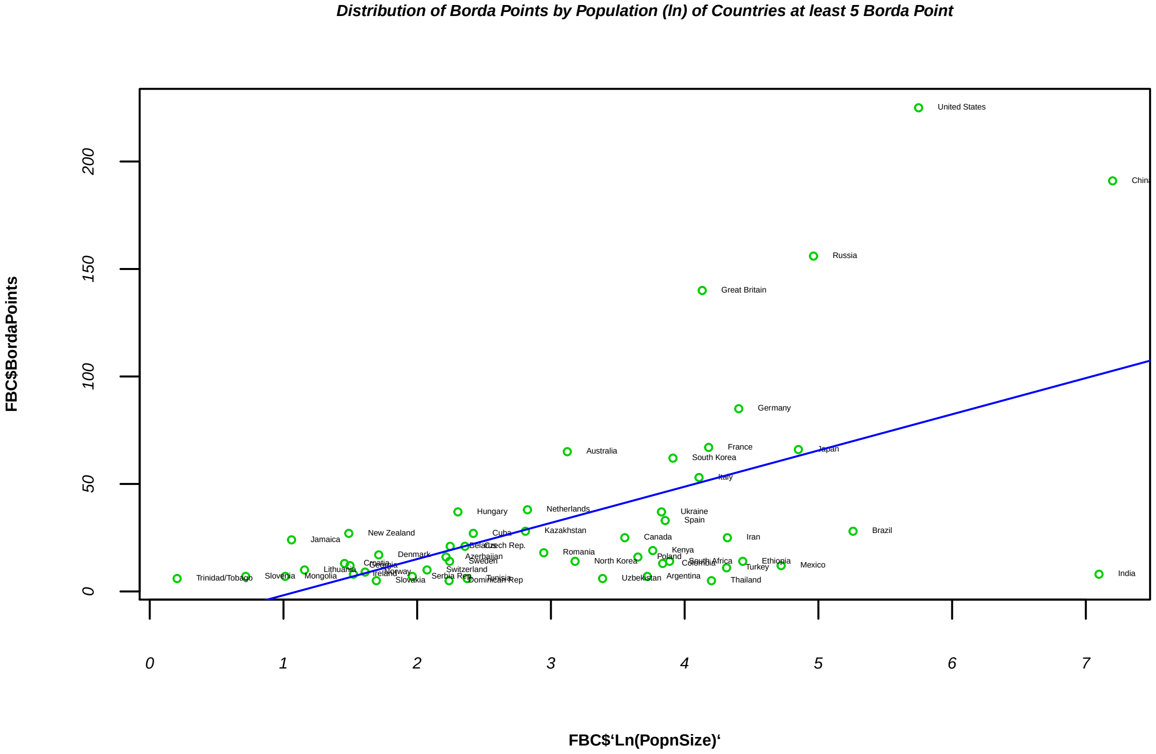
Population and success at the olympics don't show a strict correlation at the first glance.

However one should not forget that we we observe many countries - with big and small populations - having no success (0 Borda Points), which may hinder us coming to effective conclusions.

How would the same graphic look, if we only take countries with at least 1 Borda Point into account? This may be a better comparision since those are the countries which showed at least a minimal level of competitive participance in the events. The following plot depicts those countries with at least 1 Borda Point:



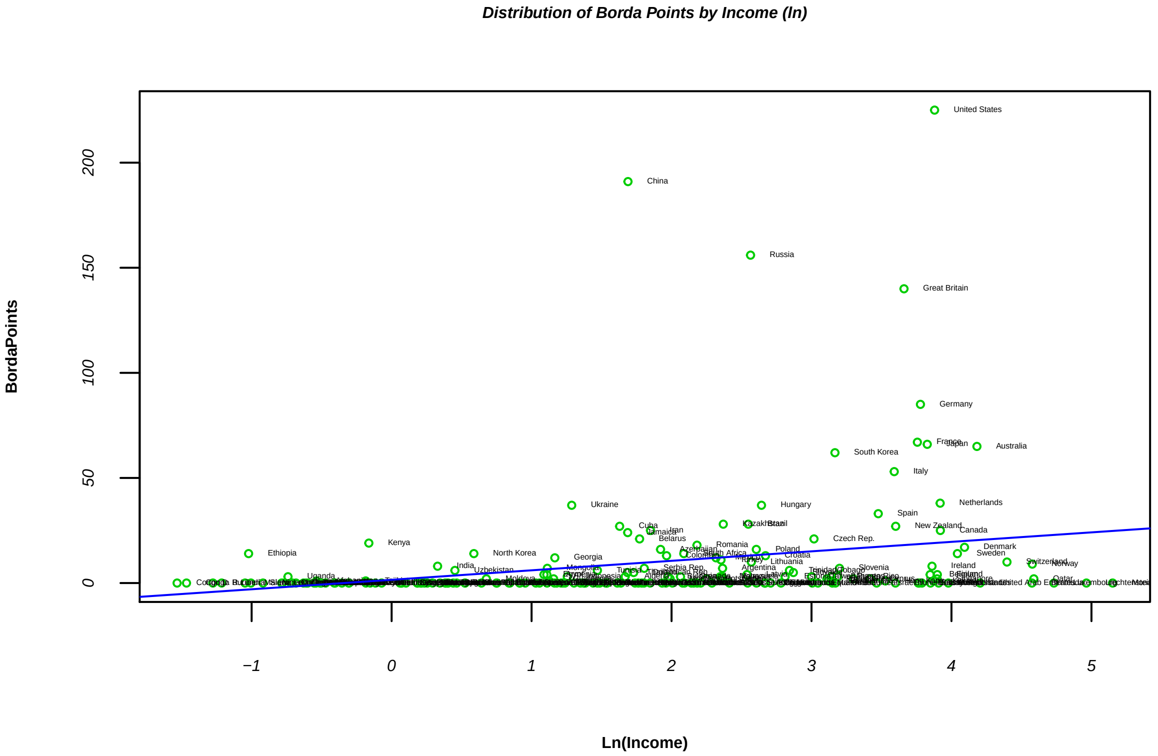
If we would like to further increase our minimal requirement of Borda Points from 1 upto 5, it might be argued that this would lead to a more precise analysis, since getting one Borda Point is done by only one bronze medal, which would be too small of a difference between 0 and 1 Borda Point countries. The new plot would look like the following:



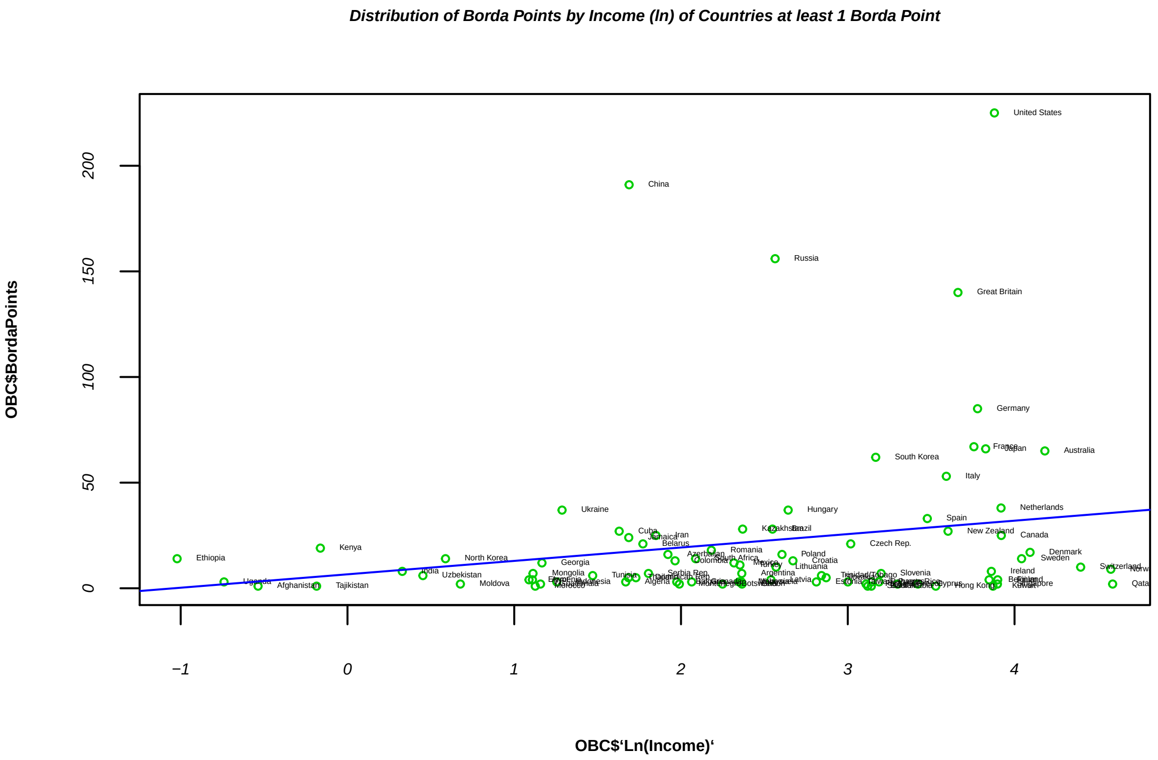
Income and Success

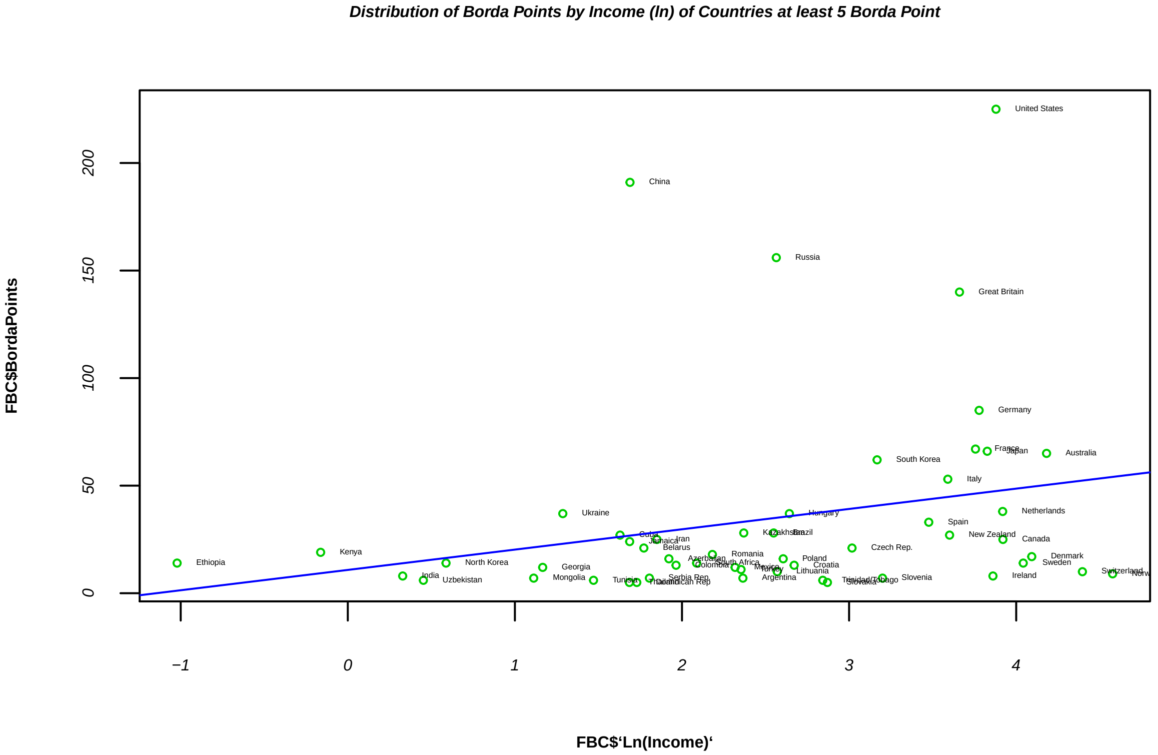
Second parameter we want to analyse is the level of income Here we will use the natural logarithm of income per capita (in $1.000) since it's preferred to naturalize the enourmous income size differences between some contries. Moreover income per capita paramater is preferred over GDP of a country, because we're interested in individuals' financial oppurtunities rather than a countries total production.

The plot of all participating countries in the 2012 London Olympics looks like as follows:



Similar to the first plot of population size vs. success, this plot doesn't represent a direct correlation due to a high number of countries with 0 Borda Points. Let's increase our Borda Point requirement to at least one and than to five points:


Distribution of Borda Points by Income (ln) of Countries at least 1 Borda Point


Distribution of Borda Points by Income (ln) of Countries at least 5 Borda Point

Our first descriptive analysis on both income and population have shown a certain degree of correlation with success, especially after restricting our data to countries which have achieved at least minimal level of medals in the competitions.

Top 50 and Bottom 50 Countries and Averages

In this section we'd like to compare the summaries of (1) all participating countries, (2) top-50 countries by Borda Points and (3) bottom-50 countries by Borda Points.

First parameter to be analysed is the population size (in in 1.000.000):

Summary of all countries' population size:

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
##    0.0094    1.2306    6.4980   34.0245   22.6124 1340.0865
```

Summary of top-50 countries' population size:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##    1.227    8.256   23.381   92.089   61.901 1340.086
```

Summary of bottom-50 countries' population size:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##   0.0196   0.3520   3.4213  10.2194   9.9751 161.0137
```

Second parameter we like to analyse in the same matter is the income per capita (in $1.000):

Summary of all countries' income per capita:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.216   1.526   5.638  15.728  17.401 172.676
```

Summary of top-50 countries' income per capita:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.360   5.699  12.890  22.677  38.330  97.254
```

Summary of bottom-50 countries' income per capita:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.2160  0.8662  3.9100 12.0196 14.1977 97.0000
```

Under- and Overperformers and Outliers

On our plots we see some countries as obvious exceptions which are over- or underperforming in comparison to ther population or income level parameters. Some examples for those countries are:

Averages of all participating countries as a reference:

```
##      Income           Popsize        
##  Min.   :  0.216   Min.   :   0.0094  
##  1st Qu.:  1.526   1st Qu.:   1.2306  
##  Median :  5.638   Median :   6.4980  
##  Mean   : 15.728   Mean   :  34.0245  
##  3rd Qu.: 17.401   3rd Qu.:  22.6124  
##  Max.   :172.676   Max.   :1340.0865
```

Pakistan, 7th most populus country in the world:

```
##          BordaPoints  Popsize Income
## Pakistan           0 180.4564  1.201
```

Nigeria, 8th most populus country in the world:

```
##         BordaPoints  Popsize Income
## Nigeria           0 170.1175   1.49
```

Monaco, 1st highest income country in the world:

```
##        BordaPoints Popsize  Income
## Monaco           0  0.0364 172.676
```

Liechtenstein, 2nd highest income country in the world:

```
##               BordaPoints Popsize  Income
## Liechtenstein           0  0.0363 143.151
```

Jamaica, 138th most populus country in the world:

```
##         BordaPoints  Popsize Income
## Jamaica          24 2.887784  5.402
```

Ethiopia, 6th lowest income country in the world:

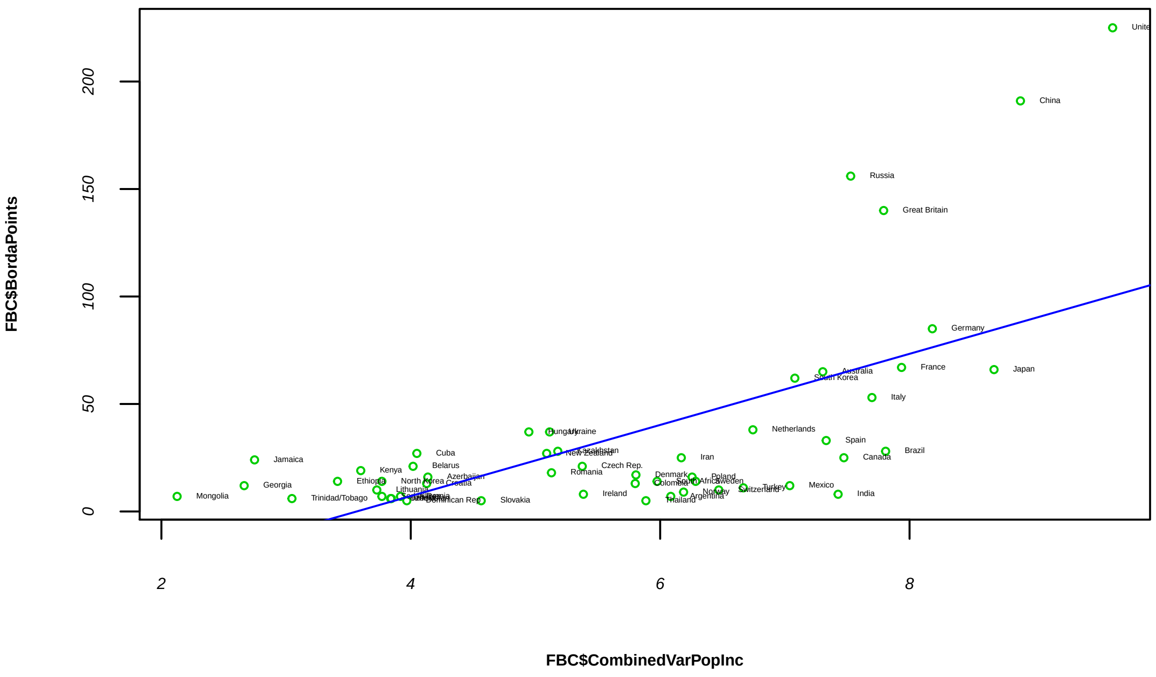
```
##          BordaPoints  Popsize Income
## Ethiopia          14 84.30907   0.36
```

Combining parameters

Answering the second question of our analysis may help us to combine population size and income parameters with correct weights and coming up with a fitting regression model can help us explain these inconsistencies.

An experimental plot to combine these to parameters and create a better correlation would be:

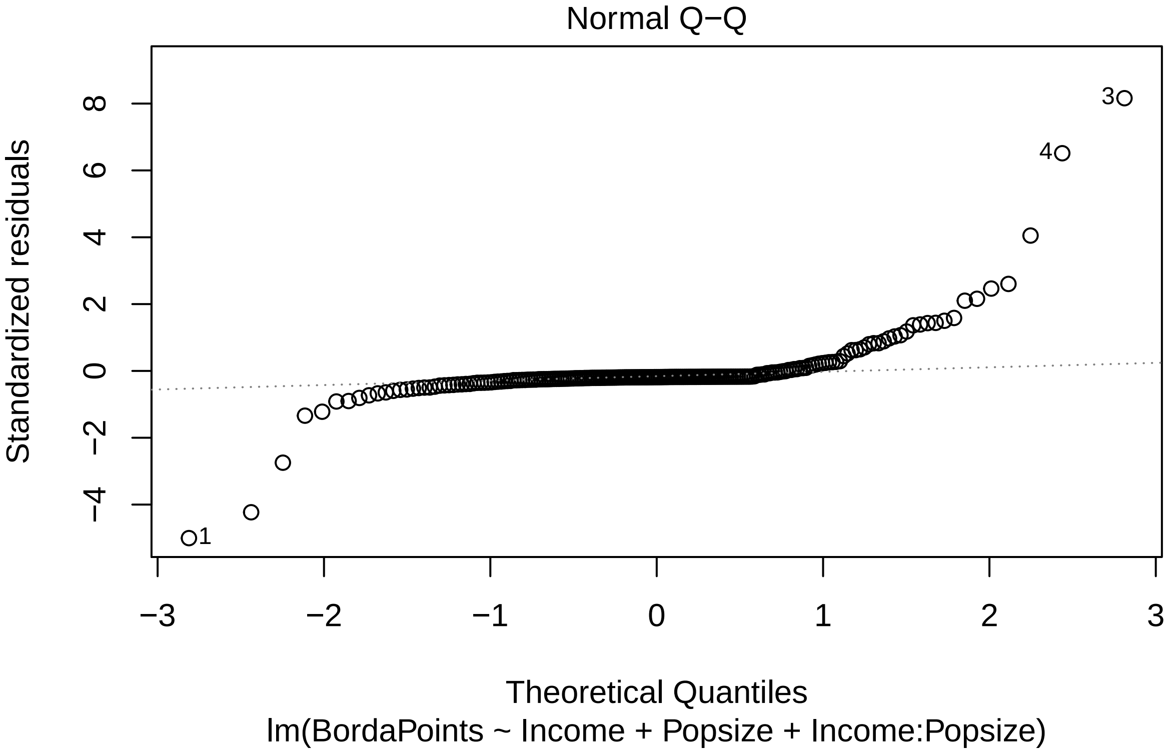
Distribution of Borda Points by Ln(PopSize)+(Ln(Income) where a country has >=5 Borda Points



Multiple Regression Analysis

In our multiple regression analysis we'll again look at four different variations of parameters.

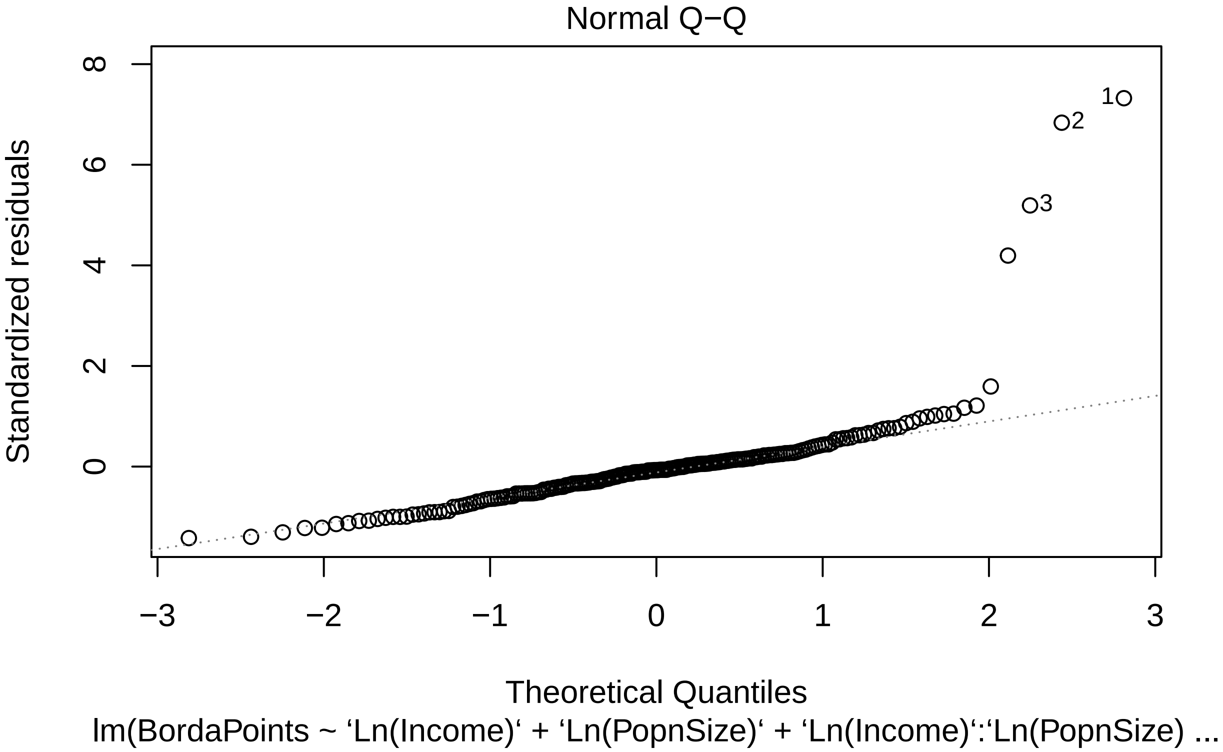
First let's observe the multiple linear regression model for the raw data in terms of income per person and population size:



```
##                        2.5 %      97.5 %
## (Intercept)    -0.0270928811 4.94600330
## Income         -0.0639039667 0.10528735
## Popsize        -0.0001255451 0.03636172
## Income:Popsize  0.0149627629 0.01862703
```

```
##
## Call:
## lm(formula = BordaPoints ~ Income + Popsize + Income:Popsize)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.644  -3.621  -2.735  -0.993 119.454
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.4594552  1.2609548   1.950   0.0525 .
## Income         0.0206917  0.0428994   0.482   0.6301
## Popsize        0.0181181  0.0092515   1.958   0.0516 .
## Income:Popsize 0.0167949  0.0009291  18.077   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.73 on 199 degrees of freedom
## Multiple R-squared:  0.723,  Adjusted R-squared:  0.7188
## F-statistic: 173.2 on 3 and 199 DF,  p-value: < 2.2e-16
```

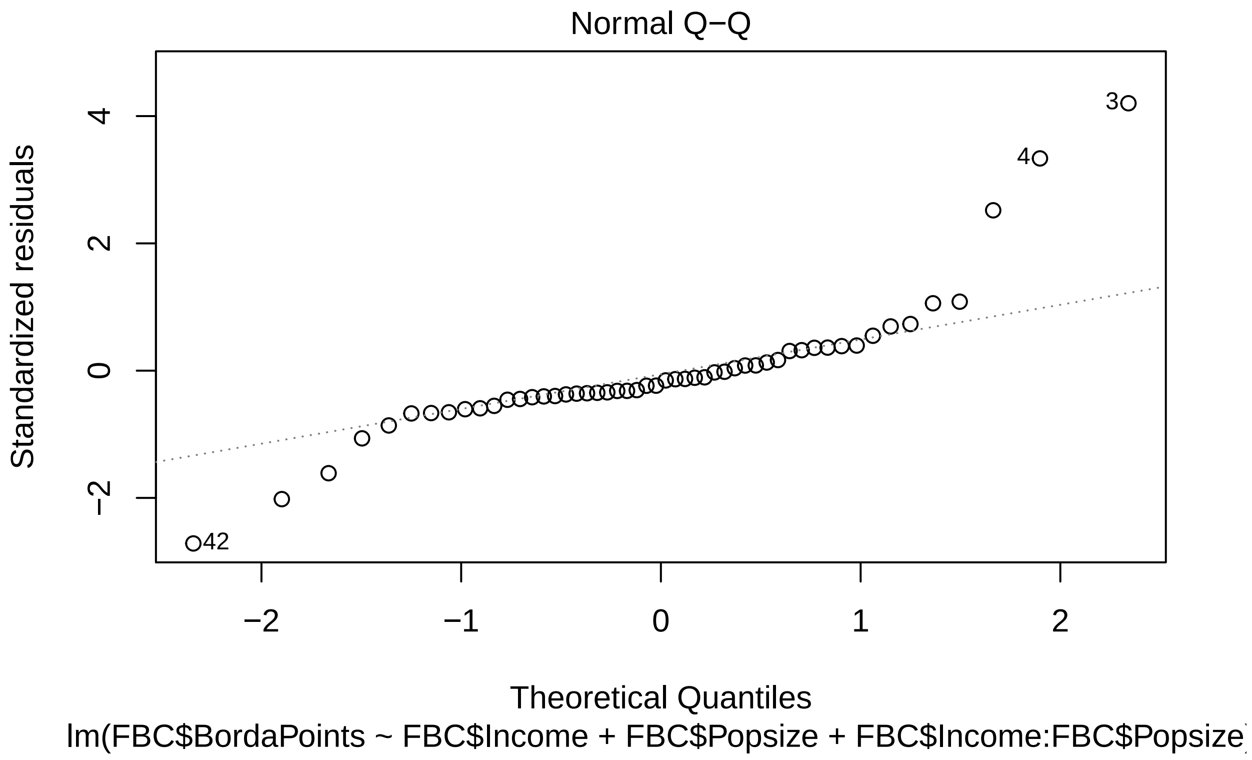
Now we model the the raw data of income per person and population size using values on natural logaritm:



```
##                                   2.5 %   97.5 %
## (Intercept)                  -7.6287282 6.064672
## `Ln(Income)`                  0.1127035 5.399378
## `Ln(PopnSize)`               -1.2967795 3.535823
## `Ln(Income)`:`Ln(PopnSize)`   1.1379222 3.026594
```

```
##
## Call:
## lm(formula = BordaPoints ~ `Ln(Income)` + `Ln(PopnSize)` + `Ln(Income)`:`Ln(PopnSize)`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.292 -10.429  -1.592   5.119 162.212
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.7820     3.4720  -0.225   0.8220
## `Ln(Income)`                  2.7560     1.3405   2.056   0.0411 *
## `Ln(PopnSize)`                1.1195     1.2253   0.914   0.3620
## `Ln(Income)`:`Ln(PopnSize)`   2.0823     0.4789   4.348 2.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.92 on 199 degrees of freedom
## Multiple R-squared:  0.3292, Adjusted R-squared:  0.3191
## F-statistic: 32.55 on 3 and 199 DF,  p-value: < 2.2e-16
```

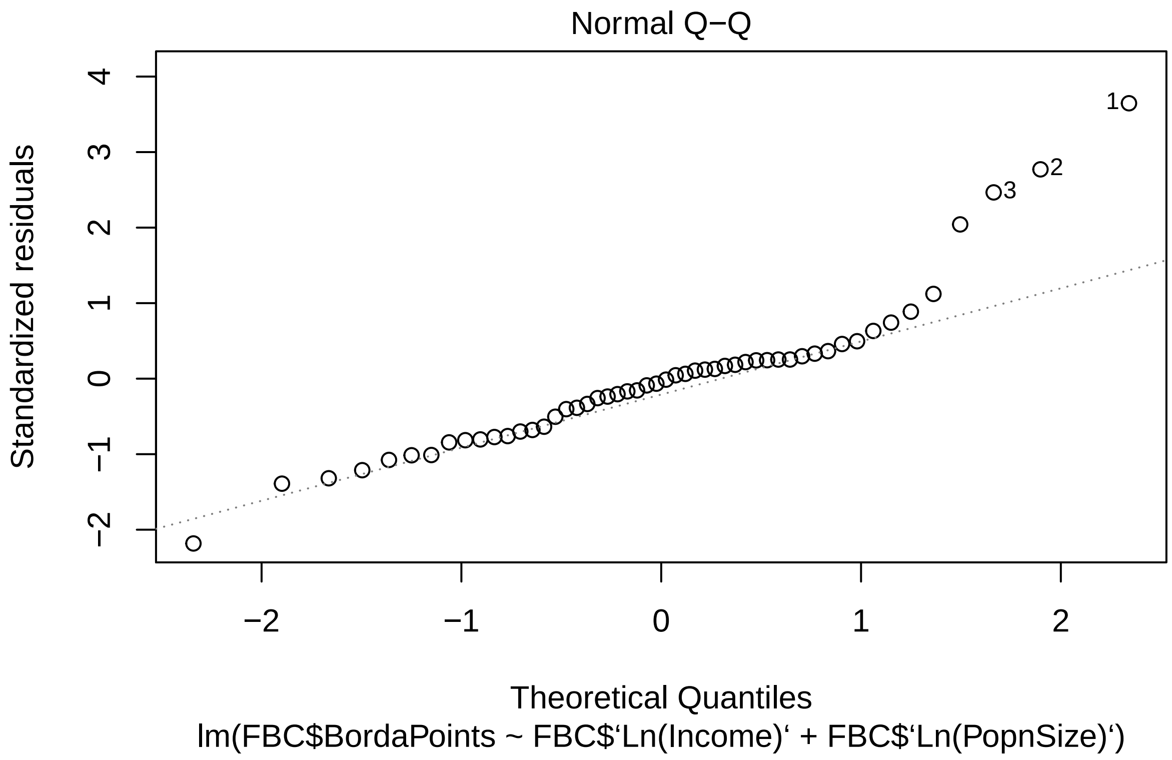
As we did in the other sections, let’s separate us from the countries which have less than 5 Borda points, firstly using the raw income per person and population values:



```
##                              2.5 %      97.5 %
## (Intercept)             4.67232260 26.87211300
## FBC$Income             -0.37588882  0.33931769
## FBC$Popsize            -0.02216773  0.04960009
## FBC$Income:FBC$Popsize  0.01159707  0.01898564
```

```
##
## Call:
## lm(formula = FBC$BordaPoints ~ FBC$Income + FBC$Popsize + FBC$Income:FBC$Popsize)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.048 -11.046  -5.090   8.127 110.073
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            15.772218   5.520593   2.857   0.0063 **
## FBC$Income             -0.018286   0.177856  -0.103   0.9185
## FBC$Popsize             0.013716   0.017847   0.769   0.4459
## FBC$Income:FBC$Popsize  0.015291   0.001837   8.322  7.2e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.54 on 48 degrees of freedom
## Multiple R-squared:  0.6959, Adjusted R-squared:  0.6769
## F-statistic: 36.62 on 3 and 48 DF,  p-value: 1.842e-12
```

Now we observe the the data for countries more that have more than 5 Borda points, using values on natural logaritm:



```
##                         2.5 %    97.5 %
## (Intercept)        -90.579501 -23.46128
## FBC$`Ln(Income)`     4.875141  21.47884
## FBC$`Ln(PopnSize)`  11.839530  25.32663
```

```
##
## Call:
## lm(formula = FBC$BordaPoints ~ FBC$`Ln(Income)` + FBC$`Ln(PopnSize)`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.221 -23.993  -1.415   9.275 124.061
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -57.020     16.700  -3.414  0.00129 **
## FBC$`Ln(Income)`     13.177      4.131   3.190  0.00248 **
## FBC$`Ln(PopnSize)`   18.583      3.356   5.538  1.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.16 on 49 degrees of freedom
## Multiple R-squared:  0.4238, Adjusted R-squared:  0.4003
## F-statistic: 18.02 on 2 and 49 DF,  p-value: 1.361e-06
```

Summary

After our analysis it can be argued that the population size and income per capita parameters of a country have an obvious effect on the countries' success in olympics. However as in many socioeconomic topics only two parameters cannot directly explain an outcome in every case. It’s obvious that there are many exceptions or extreme cases where one of the parameters or both of them show a contrary relationship with the success in olympics. Thus it can be stated that there is still enough room for further exploration in this analysis.

Further Investigation

Our investigation reveals that the data set may possibly be expanded with further parameters about the countries for us to arrive at better conclusions. In this case our further analysis question could be:

If population size and income are not fully enough to predict the success in olympics, what other measurable metrics can be added to the data to explain the olympics success more precisely?

Possible metrics to expand the dataset may be:

- Historical olympics success data of countries
- Goverment investment into sports infrastructures of countries
- Other development/welfare metrics of countries, such as HDI etc.
- Population survey data about interest in sports activities.
- Health and genetic metrics per country
- Climate and geographical metrics of countries.