Department of Data Analytics
University of Houston-Downtown
Applied Regression Analysis

APPLICATION OF MULTIPLE LINEAR REGRESSION ANALYSIS TO EXPLAIN AND PREDICT GROSS DOMESTIC PRODUCT PER CAPITA OF A COUNTRY

Armando Bendayan
armandbendayan@gmail.com
832-870-4620

Contents

APPLICATION OF MULTIPLE LINEAR REGRESSION ANALYSIS TO EXPLAIN AND PREDICT GROSS DOMESTIC PRODUCT PER CAPITA OF A COUNTRY 1

INTRODUCTION 3

OBJECTIVE 5

METHODS OF ANALYSIS 5

DATA SOURCES 5

DATA STRUCTURE 5

EXPLORATORY DATA ANALYSIS 6

EXPLORATORY PLOTS 7

MULTIPLE LINEAR REGRESSION MODELS 9

MODEL #1 (USING ALL VARIABLES) 9

MODEL #2 (WITHOUT RELIGION FREEDOM) 11

BACKWARD AND FORWARD ELIMINATION 12

FORWARD ELIMINATION METHOD 12

BACKWARD ELIMINATION METHOD 13

COMPARISON TABLE 14

MULTICOLLINEARITY 15

RESIDUAL PLOT ANALYSIS 16

DATA ANALYSIS 18

MODEL #3 WITH TRANSFORMATION 19

CONFIDENCE INTERVALS 21

PREDICTING VALUES FROM THE MODEL #3 21

PREDICTING VALUES FROM A TRAINING AND TEST SET 22

ASIDE Q 22

CONCLUSION 23

REFERENCES 24

INTRODUCTION

The purpose of this project is to make a prediction of a country's Gross Domestic Product (GDP) by doing a Multiple Linear Regression Analysis.

Gross Domestic Product (GDP), is one of the primary indicators used to measure the economic health of a country. In simple words, it is often explained as the size of the economy of a country. It represents the total dollar value over a specific period of time (In this paper dollars will be the currency used). Annual GDP figures are often considered the benchmark for the size of the economy; for example, if the GDP of a country is up by 5% over a year, the economy of that country has grown by 5% in that year.

Calculating GDP is not an easy task; there are two ways of doing it, either by adding up the earnings of all the population of that country in a year, or by adding what everybody spent. In this paper the calculation will not be done, but the calculations are going to be taken from *The World Bank* Database.

Another aspect that this paper will review is Intelligence Quotient (IQ). In science, the term intelligence typically refers to what we could call academic or cognitive intelligence. In their book on intelligence, professors Resing and Drenth answer the question 'What is intelligence?' using the following definition: *"The whole of cognitive or intellectual abilities required to obtain knowledge, and to use that knowledge in a good way to solve problems that have a well described goal and structure"* (Resing, W. and P. Drenth, 2007). In short, it is supposed to gauge how well someone can use information and logic to answer questions or make predictions.

It has been argued that human intelligence is a measurable variant as stated in the book *The Bell Curve* by psychologist Richard J. Herrnstein (2010). A series of IQ studies carried out from 2002 to 2006, were summarized in their book *IQ and the Wealth of Nations* and *IQ and Global Inequality.* Based on their analysis, they showed that these IQs are highly correlated with per capita income and rates of economic development. They argued that the income could be predicted, since intelligence is correlated with earnings among individuals. Nations are aggregates of individuals, so the same correlation would be expected across nations. They claimed to have shown that this is indeed the case and that the correlations between per capita income and rates of economic development are around 0.7

Intelligence appears to be a major component of human development. There is substantial relationship at the level of individuals of IQ with school grades and tests (Dansen, Lee, & Detternan, 2003). In several studies in several countries, the correlation of IQ with these educational outcomes has typically been found to be around 0.5 to 0.7 (Mackintosh, 1998) (Jensen, 1998) (Jencks, 1972), and sometimes as high as 0.80.

Low intelligence is a significant determinant of unemployment, poverty, welfare dependency, single motherhood, mortality, and crime (Brand, 1987; Herrnstein and Murray, 1994). The advantages of having a high IQ are greatest when dealing with complex tasks such as those involved in professional and managerial occupations. A high IQ is less advantageous for dealing with routine tasks in semi-skilled and unskilled work, but even in these types of work a high IQ confers some advantage (Lynn R. M., 2007).

Besides IQ, there are other variables that possibly have correlation with the GDP; for example, the religious freedom may affect the productivity of the nation. Religious freedom correlates closely with civic, political, and economic freedom, making it an integral part and a building block of what we consider a liberal democracy (Alon & Chase, 2016).

"Lack of religious freedom can lead to religious fractionalization and conflict. Disenfranchised religious minorities may form militias to protect their rights, properties, or national identity. Destabilizing terrorist groups may emerge from repressed religious minorities" (Alon & Chase, 2016)

On the other hand, there is the freedom of economy, which is *"the fundamental right of every human to control his own labor and property. In an economically free society, individuals are free to work, produce, consume, and invest in any way, also, in economically free societies, governments allow labor, capital and goods to move freely, and refrain from constraint of liberty beyond the extent necessary to protect and maintain liberty itself"* (Miller, Kim, & Roberts, 2018).

Today's successful economies are not necessarily geographically large or richly blessed with natural resources. Many economies have managed to expand opportunities for their citizens by enhancing their economic dynamism. The Economic Freedom Index results have shown that such economic dynamism can be sustained when governments adopt economic policies that empower individuals and firms with more choices, thereby encouraging greater entrepreneurship. In other words, economic freedom is closely related to openness to entrepreneurial activity (Miller, Kim, & Roberts, 2018).

OBJECTIVE

The objective of the present paper is to address the following question:

- **Are correlations between Gross Domestic Product per capita produced by their dependence with Economic Freedom, Religion Freedom, and Intellectual Quotient (IQ)?**

METHODS OF ANALYSIS

Correlation analysis is best suited to measure the strength of the linear relationship between national IQs and alternative measures of GDP per capita, also using education level by country and religiosity level by country and economic freedom index as categorical variables. Correlations should be clearly positive and relatively strong. Negative correlations would falsify the hypothesis that national IQs are a significant determinant of gross domestic product.

In this paper a multiple linear regression will be performed using data from different well-accredited sources for each of the variables and then they will be merge together into a single data frame in order to be analyzed in the most efficient way.

The dependent variable will be GDP per capita and IQ will be the independent variable; also, two independent categorical variables will be used. Economic freedom index and religious freedom, taking that into consideration a multiple linear regression model will be created and analyzed to understand the relationship and correlation between the variables. Using the coefficient of determination, it will be possible to explain the proportion of the variance in the dependent variable that is predicted from the independent variables.

DATA SOURCES

- GDP per capita data: The World Bank
- IQ data: Richard Lynn's and Tatu Vanhanen's "Intelligence and the Wealth and Poverty of Nations"
- Index of Economic Freedom data: Wall Street Journal and the Heritage Foundation
- Religion Freedom data: Pew Research Center Religion & Public Life

DATA STRUCTURE

- GDP per capita– Numbers ($ in billions)
- IQ – Numbers
- Index of Economic Freedom – Free, Mostly Free, Moderately Free, Mostly Unfree, Repressed
- Religion Freedom – Very High, High, Moderate, Low

EXPLORATORY DATA ANALYSIS

First, we are going to look at the data, clean it, and merge it together so the data has consistency and is fit to use in a multiple linear regression.

This is how the data looks:

```
head(Data2)
```

```
##   Country Name GDP per Capita (PPP)       IQ EconfreeCategory  RelFree
## 1  Afghanistan             1918.599 82.11964   Mostly Unfree     High
## 2      Albania            11840.228 81.74895   Moderately Free   Low
## 3      Algeria            15026.461 75.99773   Repressed         Very High
## 4       Angola             6844.433 75.10000   Repressed         High
## 5    Argentina            20047.489 86.62904   Mostly Unfree     Moderate
## 6      Armenia             8620.975 88.81870   Moderately Free   Moderate
```

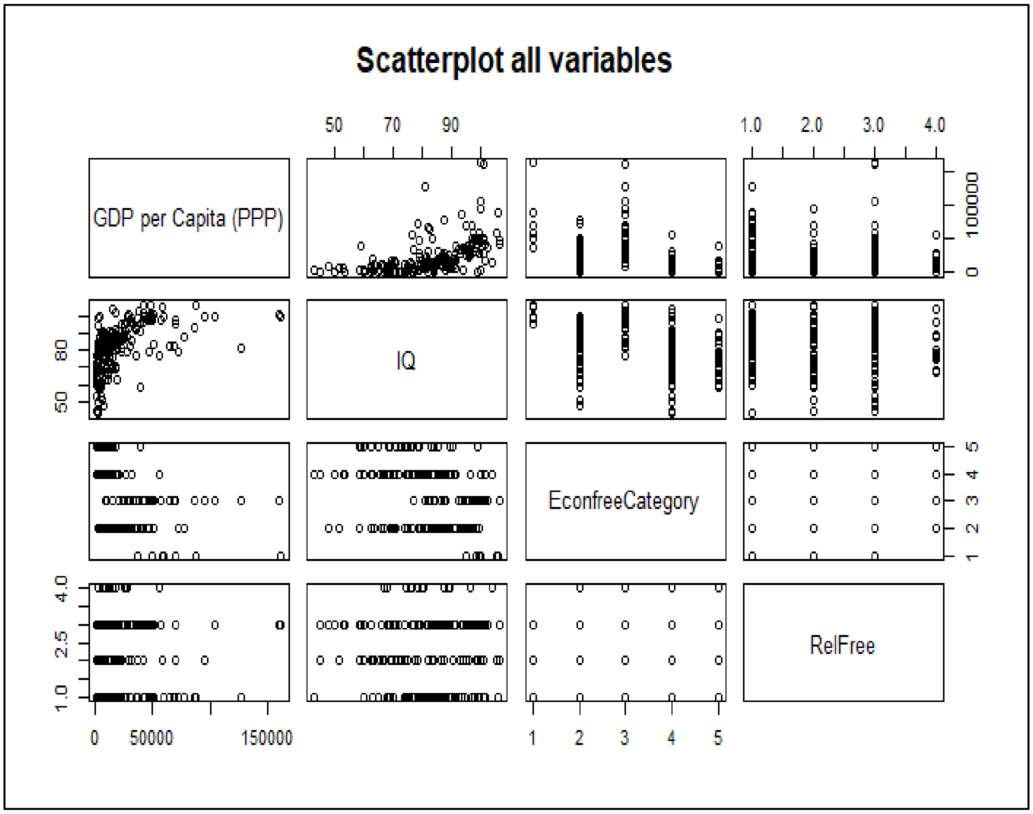
And as we can see in the table below, EconFreeCategory and RelFree are set as factors or will be categorical variables, with 5 and 4 levels in that order.

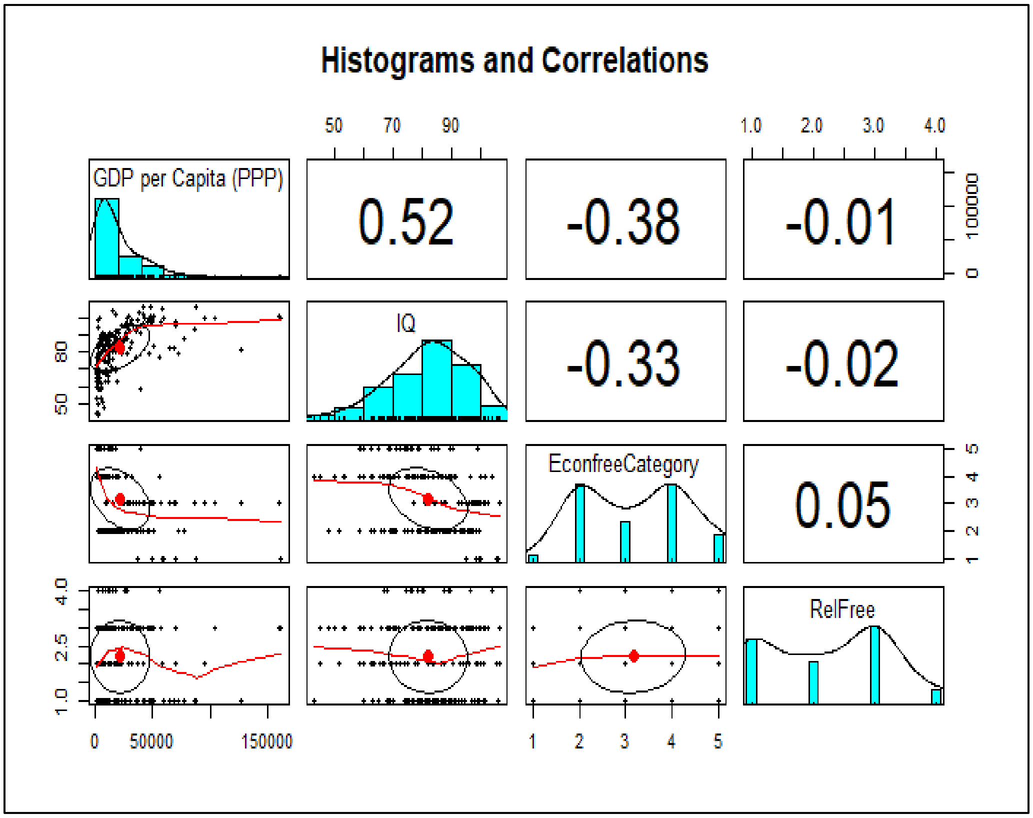
```
str(Data2)
```

```
## 'data.frame':    166 obs. of  4 variables:
##  $ GDP per Capita (PPP): num  1919 11840 15026 6844 20047 ...
##  $ IQ                  : num  82.1 81.7 76 75.1 86.6 ...
##  $ EconfreeCategory    : Factor w/ 5 levels
##  $ RelFree             : Factor w/ 4 levels
```

Before continuing with the EDA, it is important to look for missing values, which were fixed. After making sure there are no missing values or NA's, a few histograms and scatterplots are important to see, so it is possible to understand the type of data that we have, and to also see if there is correlation and if we are working with a normal distribution.

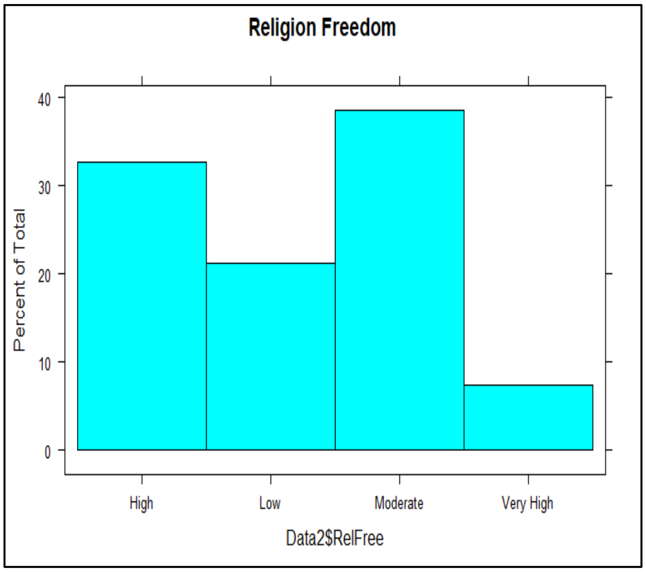
EXPLORATORY PLOTS

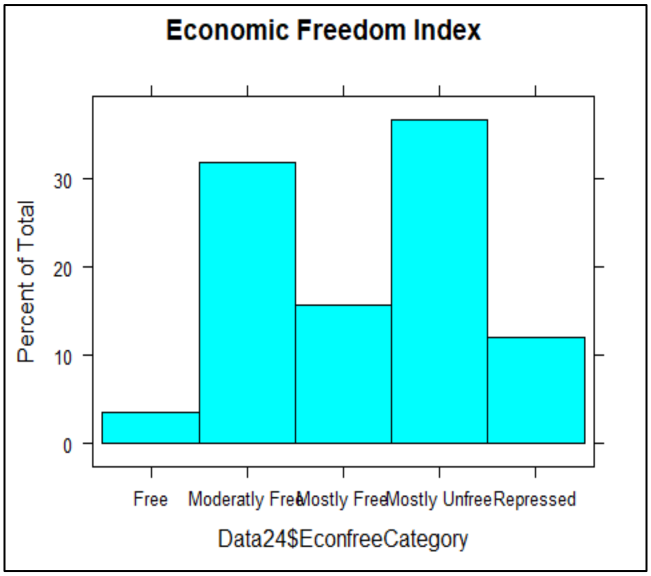




By looking at these two plots we can see several things:

1. Making scatterplots of categorical variables does not make much sense.
2. There is a pattern when looking at GDP vs. IQ, when looking at the correlation we can see that those variables have positive correlation and is expected to be between 0.4 and 0.7.
3. IQ data have a normal distribution and GDP is skewed.
4. For the categorical variables, it is better to create histograms to see the frequency of each variable.





By looking at these two graphs, there are some things that are very impressive. First, we need to have a better and more compassionate world, with more freedom to believe in the religion that each individual chooses and also to perform economic activities freely. If we pay attention to the Economic Freedom Index, it is possible to see that less than 5 countries in the world are completely free, and there are many more repressed countries than free countries, but most of the countries are between “Moderately Free” and “Moderately Unfree”.

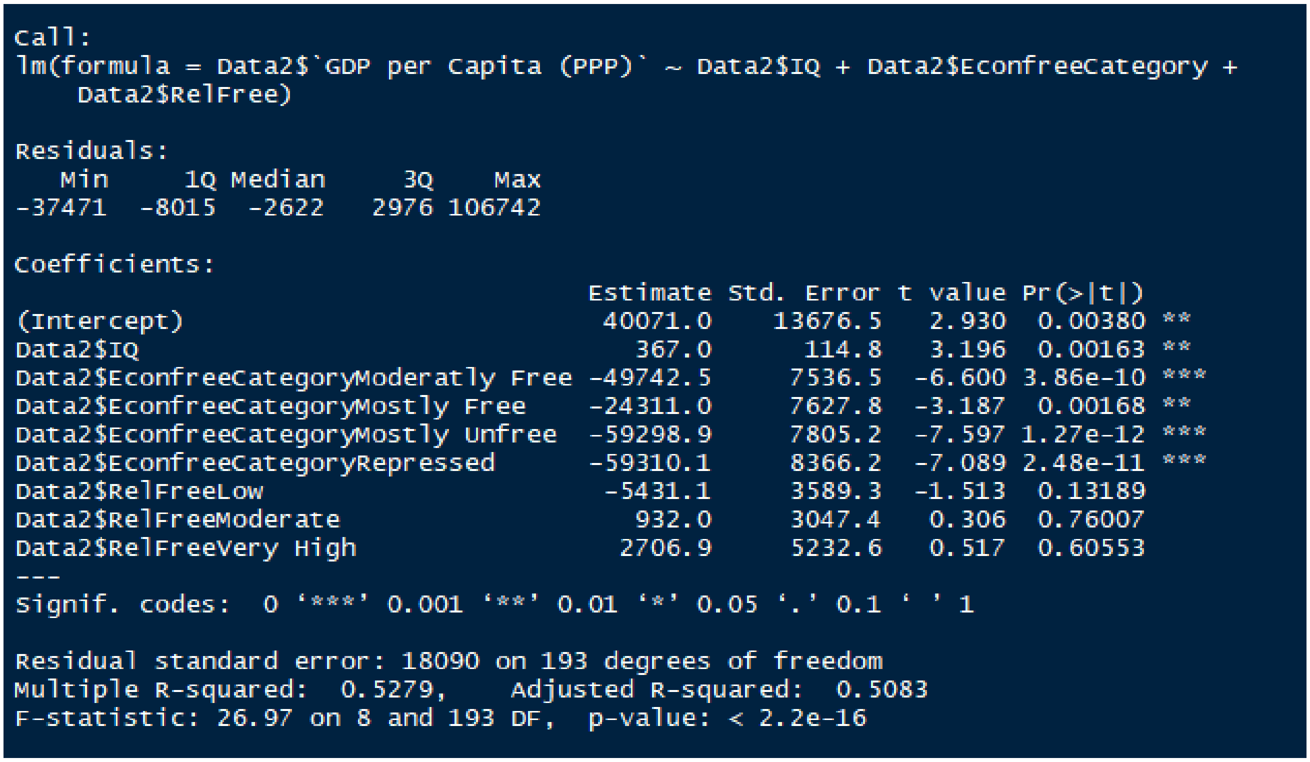
On the other hand, Religion Freedom reveals a similar scenery; most of countries are divided between “Moderate” and “High” but there are more countries with “Low” religion freedom than countries with “Very High” religion freedom.

MULTIPLE LINEAR REGRESSION MODELS

After fitting the data for the regression model and performing an exploratory analysis, it is time to create the multiple linear regression models.

The first model will use all the variables that we have in order to understand better if there is a relationship between the independent variables and the dependent one. Our dependent variable is GDP per capita and the independent variables are IQ, Economic Freedom Index and Religious Freedom.

MODEL #1 (USING ALL VARIABLES)



```
Call:
lm(formula = Data2$`GDP per Capita (PPP)` ~ Data2$IQ + Data2$EconfreeCategory +
    Data2$RelFree)

Residuals:
   Min     1Q Median     3Q    Max
-37471  -8015  -2622   2976 106742

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                           40071.0    13676.5   2.930  0.00380 **
Data2$IQ                                367.0      114.8   3.196  0.00163 **
Data2$EconfreeCategoryModeratly Free -49742.5     7536.5  -6.600 3.86e-10 ***
Data2$EconfreeCategoryMostly Free    -24311.0     7627.8  -3.187  0.00168 **
Data2$EconfreeCategoryMostly Unfree  -59298.9     7805.2  -7.597 1.27e-12 ***
Data2$EconfreeCategoryRepressed      -59310.1     8366.2  -7.089 2.48e-11 ***
Data2$RelFreeLow                      -5431.1     3589.3  -1.513  0.13189
Data2$RelFreeModerate                   932.0     3047.4   0.306  0.76007
Data2$RelFreeVery High                 2706.9     5232.6   0.517  0.60553
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18090 on 193 degrees of freedom
Multiple R-squared:  0.5279,    Adjusted R-squared:  0.5083
F-statistic: 26.97 on 8 and 193 DF,  p-value: < 2.2e-16
```

From the summary output it is important to observe the significance of each variable with the independent variable It is also important to look at the multiple R-Squared and Adjusted R-Square; and the p-value of the model and each variable.

As we can see, the Religious Freedom does not have much significance to the model. Having p-values higher than 0.05, it is possible to say that a new model without those variables is recommended to perform.

By looking at the R-Squared and Adjusted R-Squared, we can see that the value is over 0.5, meaning that the dependent variable is more than 50% explained by the independent variables, meaning there is correlation between the variables.

On the other hand, there are a lot of things to say about the Beta values. First let's create the equation for the model.

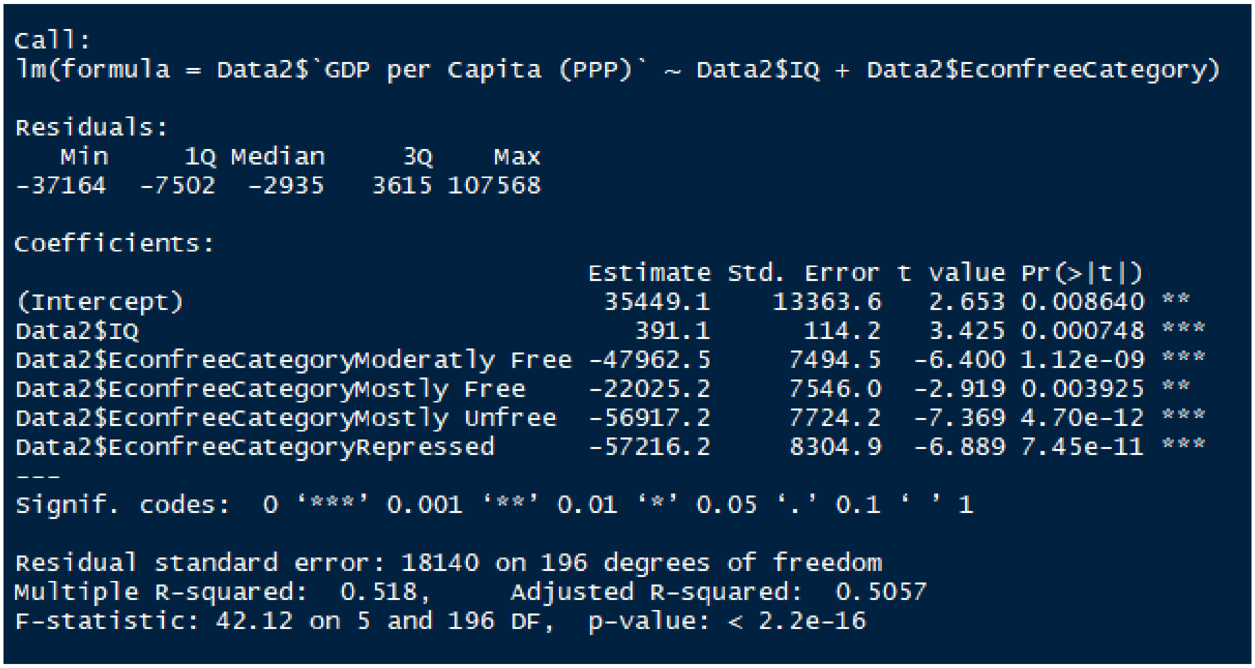
$$\boldsymbol{GDP = 40071 + 367 * (IQ) - 49742.5 * (EconFree\$ModeratlyFree) - 24311 * (EconFree\$MostlyFree) - 59298.9 * (EconFree\$MostlyUnfree) - 59310.1 * (EconFree\$Repressed) - 5431.1 * (Religion\$Low) + 932 * (Religion\$Moderate) + 2706.9 * (Religion\$VeryHigh)}$$

We can see how the independent variables contribute to the value of the dependent variable, and the intercept gives a positive value and the IQ. The other variables can give negative or positive values to the GDP depending on the which one is it; for example, a economically repressed country with low religion freedom will subtract value from the GDP, or an economically free country with very high religion freedom will add value to the GDP per capita of the country.

This is important to notice and understand because this implies that having high IQ and living in an economically free country with high religious freedom will add to the GDP per capita, or a country with very high IQ but with an economically repressed freedom with low religious freedom, no matter what, the GDP per capita will be low.

Is also important to look for outliers and grouping of countries that are not fulfilling this "rule", but this will be performed for the second model.

MODEL #2 (WITHOUT RELIGION FREEDOM)



```
Call:
lm(formula = Data2$`GDP per Capita (PPP)` ~ Data2$IQ + Data2$EconfreeCategory)

Residuals:
   Min     1Q Median     3Q    Max 
-37164  -7502  -2935   3615 107568 

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           35449.1    13363.6   2.653 0.008640 ** 
Data2$IQ                                391.1      114.2   3.425 0.000748 ***
Data2$EconfreeCategoryModeratly Free -47962.5     7494.5  -6.400 1.12e-09 ***
Data2$EconfreeCategoryMostly Free    -22025.2     7546.0  -2.919 0.003925 ** 
Data2$EconfreeCategoryMostly Unfree  -56917.2     7724.2  -7.369 4.70e-12 ***
Data2$EconfreeCategoryRepressed      -57216.2     8304.9  -6.889 7.45e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18140 on 196 degrees of freedom
Multiple R-squared:  0.518,	Adjusted R-squared:  0.5057 
F-statistic: 42.12 on 5 and 196 DF,  p-value: < 2.2e-16
```

On the second model, the variable religious freedom was not included, and now it is possible to see that almost all variables have a very small p-value, meaning that those variables have better significance. On the other hand, the Multiple R-Squared decreased by a very small factor, and the Adjusted R-Squared also decreased by a small factor. The dependent variables explain the independent variable by more than 50% (Same as the Model #1), and the overall p-value stayed the same (2.2e-16), which also means that is a reliable model.

It is important into take in consideration that the model did not improve significantly and the cost for that “improvement” was taking out the religious freedom variable.

Now let’s take a look at the equation of the model:

$$\begin{aligned} \boldsymbol{GDP} &= \mathbf{35449.1} + \mathbf{391.1} * (\boldsymbol{IQ}) - \mathbf{47962.5} * (\boldsymbol{EconFree\$ModeratlyFree}) - \mathbf{22025.2} \\ &\quad * (\boldsymbol{EconFree\$MostlyFree}) - \mathbf{56917.2} * (\boldsymbol{EconFree\$MostlyUnfree}) \\ &\quad - \mathbf{57216.2} * (\boldsymbol{EconFree\$Repressed}) \end{aligned}$$

Because the Model #2 has one less variable, the other values for the variables are adjusted to create a better model. In Model #2 the variable IQ now has greater positive impact on the GDP per capita, and for the other variable, the less economic freedom the country has, the less GDP per capita the Country will have.

BACKWARD AND FORWARD ELIMINATION

On this segment backward and forward elimination were performed to see if there is consistency with both methods and compare the selected models by these method with the two models created before.

FORWARD ELIMINATION METHOD

```
step(nullmodel, scope=list(lower=nullmodel, upper=fullmodel), direction="forward")

## Start:  AIC=4104.86
## Data2$`GDP per Capita (PPP)` ~ 1
## 
##                      Df  Sum of Sq        RSS    AIC
## + EconfreeCategory  4 6.5435e+10 6.8348e+10 3977.2
## + IQ                1 3.6690e+10 9.7092e+10 4042.1
## <none>                          1.3378e+11 4104.9
## + RelFree           3 2.4451e+09 1.3134e+11 4107.1
## 
## Step:  AIC=3977.2
## Data2$`GDP per Capita (PPP)` ~ EconfreeCategory
## 
##            Df  Sum of Sq        RSS    AIC
## + IQ       1 3859922329 6.4488e+10 3967.5
## <none>                 6.8348e+10 3977.2
## + RelFree  3 1842184570 6.6506e+10 3977.7
## 
## Step:  AIC=3967.46
## Data2$`GDP per Capita (PPP)` ~ EconfreeCategory + IQ
## 
##            Df  Sum of Sq        RSS    AIC
## <none>                 6.4488e+10 3967.5
## + RelFree  3 1324570969 6.3163e+10 3969.3

## 
## Call:
## lm(formula = Data2$`GDP per Capita (PPP)` ~ EconfreeCategory + 
##     IQ, data = Data2)
## 
## Coefficients:
##                    (Intercept)  EconfreeCategoryModeratly Free  
##                        35449.1                        -47962.5  
##    EconfreeCategoryMostly Free   EconfreeCategoryMostly Unfree  
##                       -22025.2                        -56917.2  
##      EconfreeCategoryRepressed                              IQ  
##                       -57216.2                           391.1
```

BACKWARD ELIMINATION METHOD

```
step(fullmodel, data=Data2, direction="backward")

## Start:  AIC=3969.27
## Data2$`GDP per Capita (PPP)` ~ IQ + EconfreeCategory + RelFree
## 
##                    Df  Sum of Sq        RSS    AIC
## - RelFree           3 1.3246e+09 6.4488e+10 3967.5
## <none>                           6.3163e+10 3969.3
## - IQ                1 3.3423e+09 6.6506e+10 3977.7
## - EconfreeCategory  4 3.2851e+10 9.6014e+10 4045.9
## 
## Step:  AIC=3967.46
## Data2$`GDP per Capita (PPP)` ~ IQ + EconfreeCategory
## 
##                    Df  Sum of Sq        RSS    AIC
## <none>                           6.4488e+10 3967.5
## - IQ                1 3.8599e+09 6.8348e+10 3977.2
## - EconfreeCategory  4 3.2604e+10 9.7092e+10 4042.1

## 
## Call:
## lm(formula = Data2$`GDP per Capita (PPP)` ~ IQ + EconfreeCategory, 
##     data = Data2)
## 
## Coefficients:
##                    (Intercept)                              IQ  
##                        35449.1                           391.1  
## EconfreeCategoryModeratly Free     EconfreeCategoryMostly Free  
##                       -47962.5                        -22025.2  
##  EconfreeCategoryMostly Unfree       EconfreeCategoryRepressed  
##                       -56917.2                        -57216.2
```

Both methods ended up with the same model, which is the same model as Model #2. In the Model, the dependent variable is GDP per capita, and the independent variables are IQ and Economic Freedom Index.

These results make sense because the final Adjusted R-Squared is the greatest that can be achieved with these variables. Also these two methods use the Akaike Information Criterion (AIC) value to give a better prediction of the model. The lower the AIC value the better the model.

There are other values that can help to decide which model is better; calculating and comparing the AICc, AIC, Bayesian information criterion (BIC), Adj.R-Squared and p-value in a table could help to decide and provide a more in-depth investigation of the models and variables. Also it is crucial to check multicollinearity between the variables.

The multicollinearity explains some mistakes or errors that can happen between the variables- for example, when using categorical variables or dummy variables in a wrong way, including a variable that is computed from another variable from the same dataset, repetition of the same variable, or when the correlation between the variables is too high.

COMPARISON TABLE

As mentioned before, a table with the AIC, AICc, BIC, Adj.R-Squared and p-value was created to compare the models.

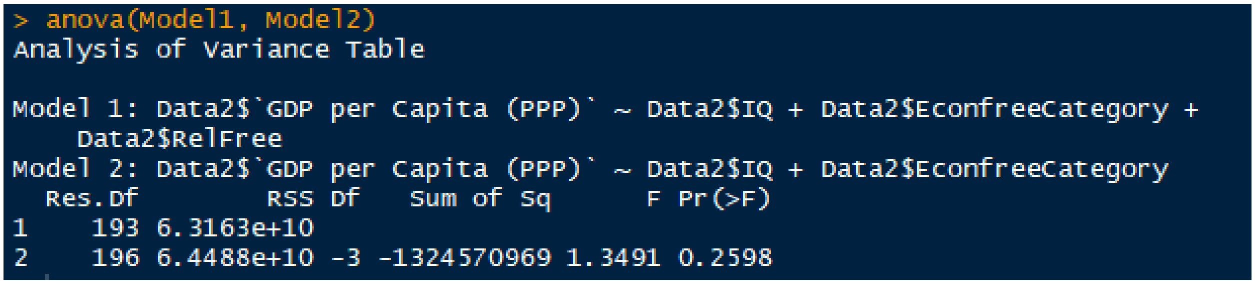
The Hypothesis is:

H_0: Reduced model is adequate

H_a: Full model is better

```
ModelComparison

##                            Model 1             Model 2
## Adjusted R^2            5.082952e-01        5.056678e-01
## AIC                     4.544517e+03        4.542709e+03
## AICc                    4.545668e+03        4.543286e+03
## BIC                     4.577599e+03        4.565867e+03
## P-Value                 6.006467e-21        6.341687e-21
```

From this table, it is possible to say that the Model #2 is a better model, the Adj.R-Squared is higher, the AIC, AICc and BIC have lower values than the ones in Model #1, and also by looking at the p-value which is very small, it is clear to say that Model #2 is the best model of the two.



```
> anova(Model1, Model2)
Analysis of Variance Table

Model 1: Data2$`GDP per Capita (PPP)` ~ Data2$IQ + Data2$EconfreeCategory +
    Data2$RelFree
Model 2: Data2$`GDP per Capita (PPP)` ~ Data2$IQ + Data2$EconfreeCategory
  Res.Df        RSS Df    Sum of Sq      F Pr(>F)
1    193 6.3163e+10
2    196 6.4488e+10 -3 -1324570969 1.3491 0.2598
```

By using the Anova function, we can deduce that the reduced model is better.

MULTICOLLINEARITY

To check multicollinearity the Variable Inflation Factor (VIF) was performed. If the value is higher than 5, it is recommended to check the variables and see which ones are causing problems. If the value is 10 or more, the collinearity is clearly present, and if the value is below 4, there is no multicollinearity between the variables.

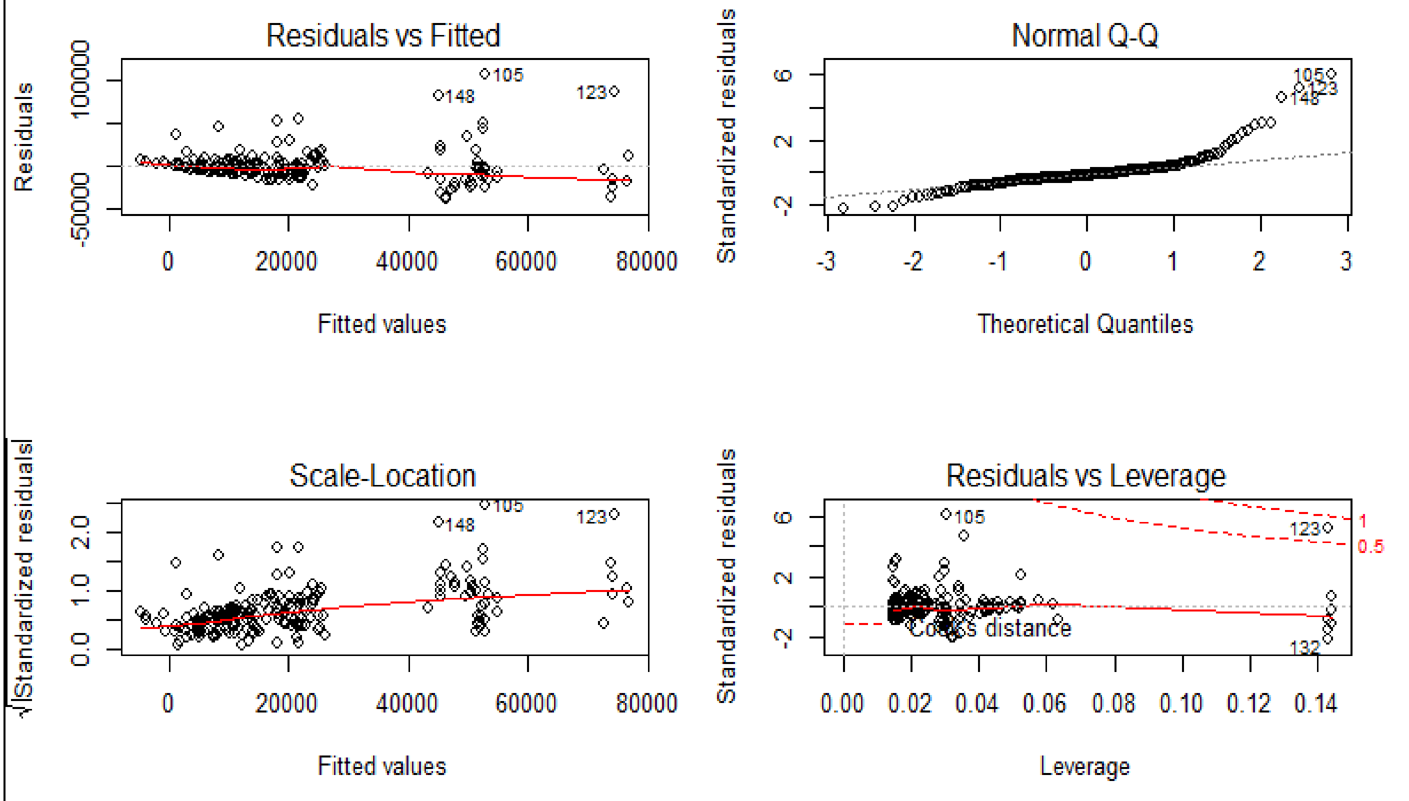
```
vif(Model1)

##                            GVIF Df GVIF^(1/(2*Df))
## Data2$IQ               1.460116  1        1.208353
## Data2$EconfreeCategory 1.540893  4        1.055532
## Data2$RelFree          1.083643  3        1.013478

vif(Model2)

##                            GVIF Df GVIF^(1/(2*Df))
## Data2$IQ               1.436311  1        1.198462
## Data2$EconfreeCategory 1.436311  4        1.046300

eigen(cor(Data2$`GDP per Capita (PPP)`, Data2$IQ))$values

## [1] 0.5236902
```

The values of VIF on both models are around 1 and 1.5, meaning that there is no multicollinearity between the variables used. The eigen distance and correlation was performed between the variables that have numeric value (GDP per capita and IQ), and we can see that those two variables have a value of 0.5419, when the value is around 0.5 and below 1. It is possible to say that there is no multicollinearity between them.

RESIDUAL PLOT ANALYSIS

In this section we are going to perform an analysis of the residual plots of Model #2 to see if there are violations of the assumptions and to understand the behavior of the residual in the regression model.

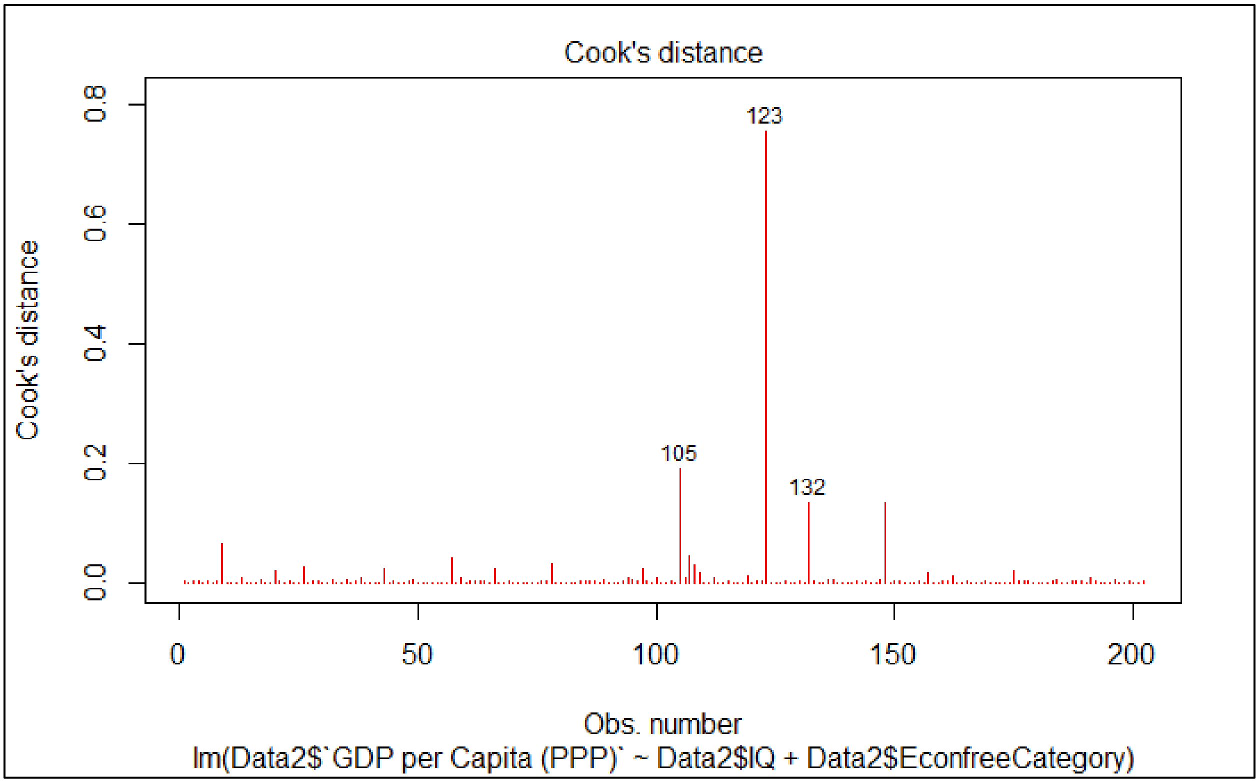
Model #2 Residual Plot



By looking at the residual plots we can see violations of the assumptions:

- Residuals vs. Fitted: Variance seem non-constant, but it has megaphone shape meaning that it has an increase pattern.
- Residuals vs. Fitted: The red line does not maintain a straight line around "0", but it is close.
- Normal Q-Q: The data does not seem that normal, on -2 and 1, the residual values move away from the line, meaning that normality is in question.
- Scale-Location: The residuals also has increasing pattern.
- Residuals vs. Leverage: We can notice several points that are influential for the regression line, getting out the accepted range.

Let’s take a closer look to the Cook’s Distance:

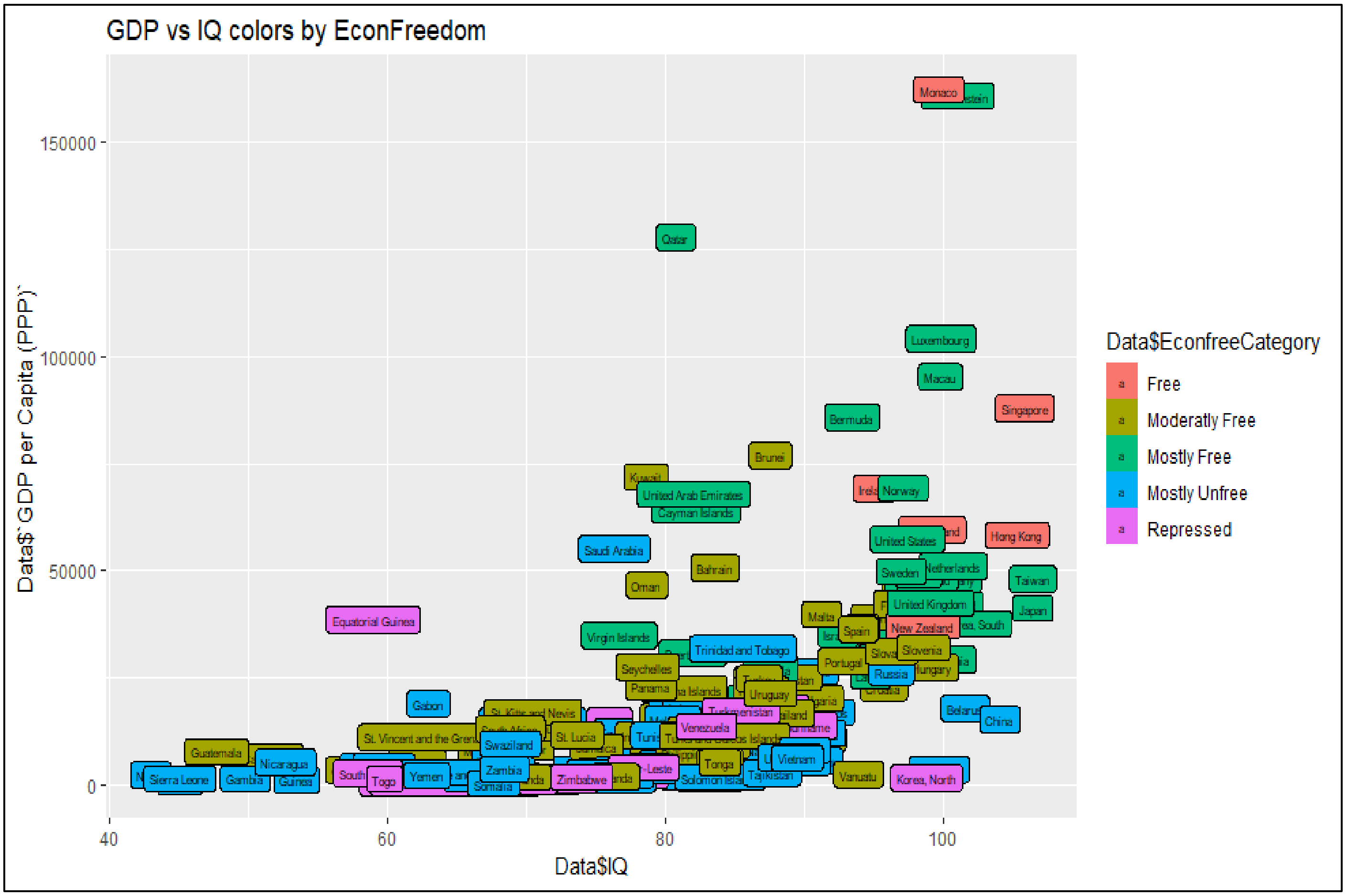


We can observe that the data point #123 is way outside the range, being an Influential Point with leverage. It might be worth it to remove this outlier; this data point is for Liechtenstein, a rich Country with 37,810 people.

Another suggestion is to perform a logarithmic or square root transformation to fix the megaphone or V-shape pattern in the residuals.

DATA ANALYSIS

It is also important to see if there are any other types of patterns or explanations for the result the model is giving. In order to do this, we are going to create a scatterplot for GDP per capita vs. IQ but also using Economic Freedom Index with a color indicator.

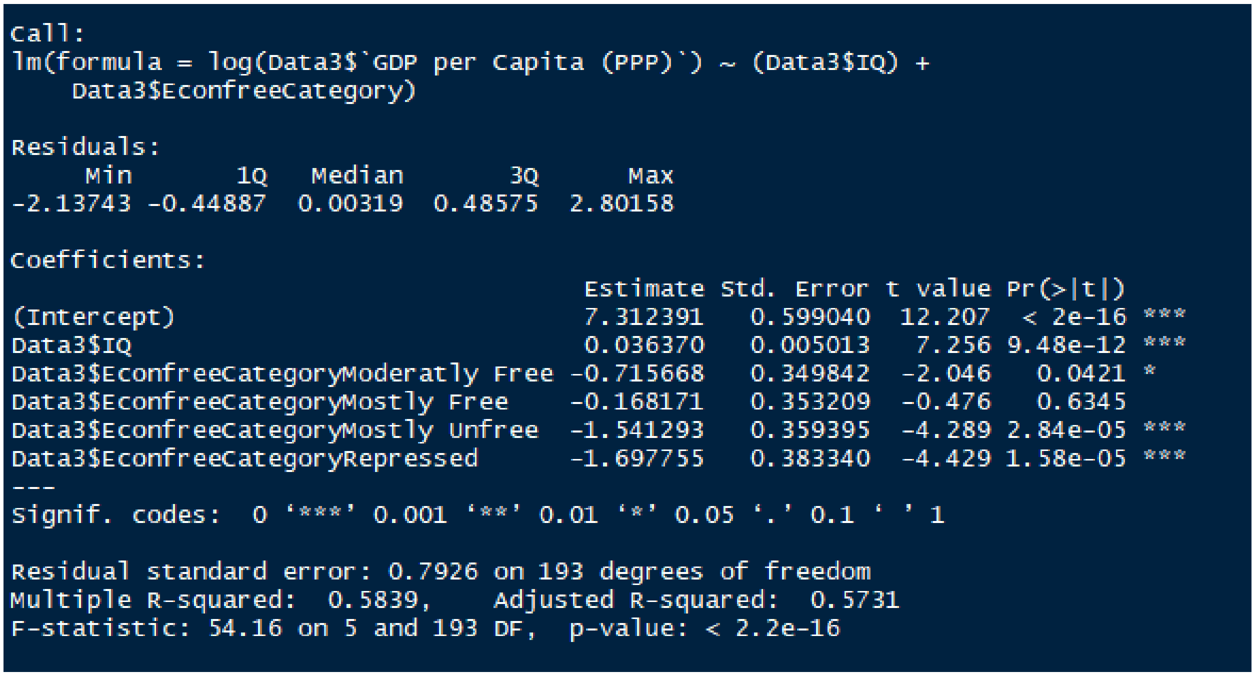


By looking at the graph above, we can see that we have some interesting findings. Monaco and Liechtenstein are outliers that we can take out because they are wealthy countries with small population.

On the other hand, we can see a group of countries with high GDP (above 5,0000) but with IQ in the range of 80-85. Those countries are big oil exporters, and it might be possible that the IQ and Economic Freedom have less impact on the GDP per capita because their economy is based on oil production.

MODEL #3 WITH TRANSFORMATION

We performed a logarithmic transformation of the GDP per capita and took care of the outliers that had high leverage and residuals, and the result improved by a good margin.



```
Call:
lm(formula = log(Data3$`GDP per Capita (PPP)`) ~ (Data3$IQ) +
    Data3$EconfreeCategory)

Residuals:
     Min       1Q   Median       3Q      Max
-2.13743 -0.44887  0.00319  0.48575  2.80158

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                          7.312391   0.599040  12.207  < 2e-16 ***
Data3$IQ                             0.036370   0.005013   7.256 9.48e-12 ***
Data3$EconfreeCategoryModeratly Free -0.715668   0.349842  -2.046   0.0421 *
Data3$EconfreeCategoryMostly Free    -0.168171   0.353209  -0.476   0.6345
Data3$EconfreeCategoryMostly Unfree  -1.541293   0.359395  -4.289 2.84e-05 ***
Data3$EconfreeCategoryRepressed      -1.697755   0.383340  -4.429 1.58e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7926 on 193 degrees of freedom
Multiple R-squared:  0.5839,    Adjusted R-squared:  0.5731
F-statistic: 54.16 on 5 and 193 DF,  p-value: < 2.2e-16
```

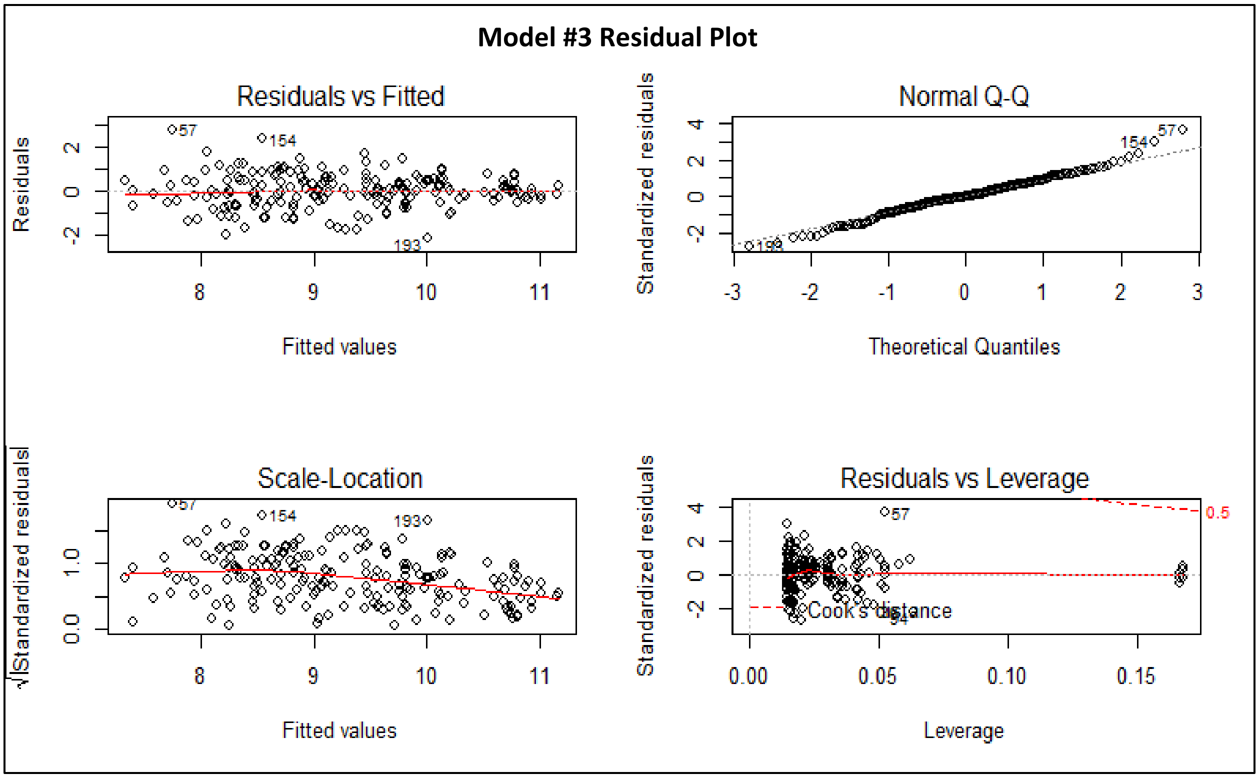
By looking at the summary of the regression model, we can see that now the most significant variable is IQ. Also the Multiple R-squared and Adjusted R-squared increased the value, so now the value is around 0.57, which means that the independent variables explain the dependent variable by 57%. Also the p-value also decreased for all variables.

The equation for this model is:

$$\boldsymbol{Ln(GDP) = 7.312 + 0.03637 * (IQ) - 0.71565 * (EconFree\$ModeratlyFree) \\ - 0.1681 * (EconFree\$MostlyFree) - 1.5412 \\ * (EconFree\$MostlyUnfree) - 1.6977 * (EconFree\$Repressed)}$$

After using the equation it is important to take out the transformation to have normal data again by taking the log of the equation using exponential.

Now let’s have a look of the residual plot to see if the assumptions are met.

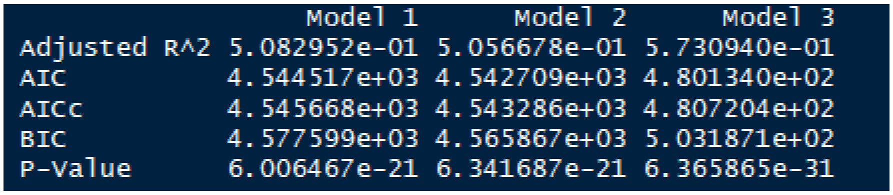


Model #3 Residual Plot

By looking at the residual plots we can see that now the assumptions are met:

- Residuals vs. Fitted: Variance seem non-constant and they are randomly spread, there is a little pattern and it looks that the residuals are decreasing.
- Residuals vs. Fitted: The red line does maintain a straight line around "0".
- Normal Q-Q: The data seem normal, there are some residuals that are not in the line but is acceptable.
- Scale-Location: The residuals seems with a decreasing pattern but not a big one.
- Residuals vs. Leverage: We can notice several points that might be influential, but they are on the range.

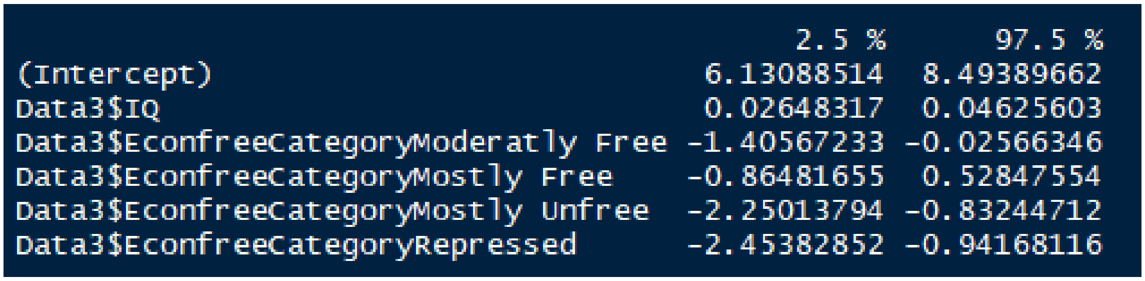
Now that the assumptions are met, and the summary of the regression model gives better result than the previous 2 models, we can say that the Model #3 is the best model. Even though when comparing the AIC, BIC, AICc, the Model #2 can look better, but the Model #2 violates some of the assumptions.



```
                  Model 1      Model 2      Model 3
Adjusted R^2 5.082952e-01 5.056678e-01 5.730940e-01
AIC          4.544517e+03 4.542709e+03 4.801340e+02
AICc         4.545668e+03 4.543286e+03 4.807204e+02
BIC          4.577599e+03 4.565867e+03 5.031871e+02
P-Value      6.006467e-21 6.341687e-21 6.365865e-31
```

CONFIDENCE INTERVALS

After selecting the best model, let's calculate de confidence interval for this model:

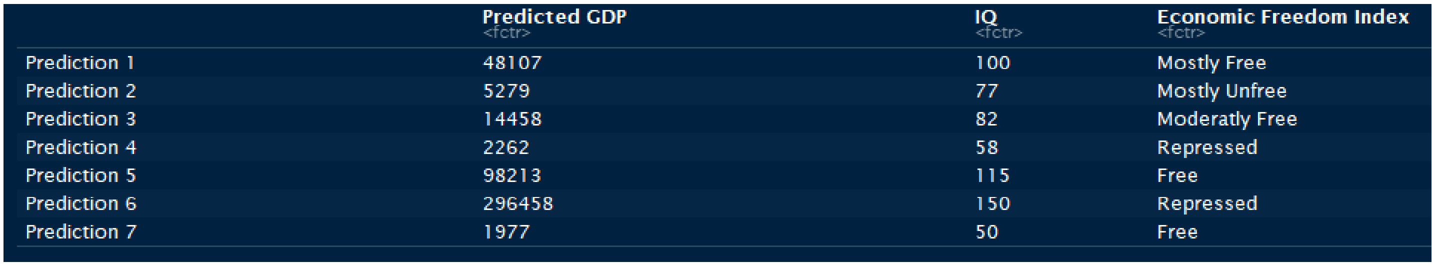


```
                                              2.5 %      97.5 %
(Intercept)                              6.13088514  8.49389662
Data3$IQ                                 0.02648317  0.04625603
Data3$EconfreeCategoryModeratly Free -1.40567233 -0.02566346
Data3$EconfreeCategoryMostly Free    -0.86481655  0.52847554
Data3$EconfreeCategoryMostly Unfree  -2.25013794 -0.83244712
Data3$EconfreeCategoryRepressed      -2.45382852 -0.94168116
```

We have an estimated slope for IQ of 0.0363 with 95% confidence the true slope is between 0.0264 and 0.0462.

PREDICTING VALUES FROM THE MODEL #3

Having a model allow to predict new values for GDP per capita by creating values of IQ and Economic Freedom Index and place them into the equation. In this section a list of 7 IQ's and Economic Freedom Indexes were randomly created to place them into the equation and see how the model predict the GDP per capita for each pair.



	Predicted GDP <fctr>	IQ <fctr>	Economic Freedom Index <fctr>
Prediction 1	48107	100	Mostly Free
Prediction 2	5279	77	Mostly Unfree
Prediction 3	14458	82	Moderatly Free
Prediction 4	2262	58	Repressed
Prediction 5	98213	115	Free
Prediction 6	296458	150	Repressed
Prediction 7	1977	50	Free

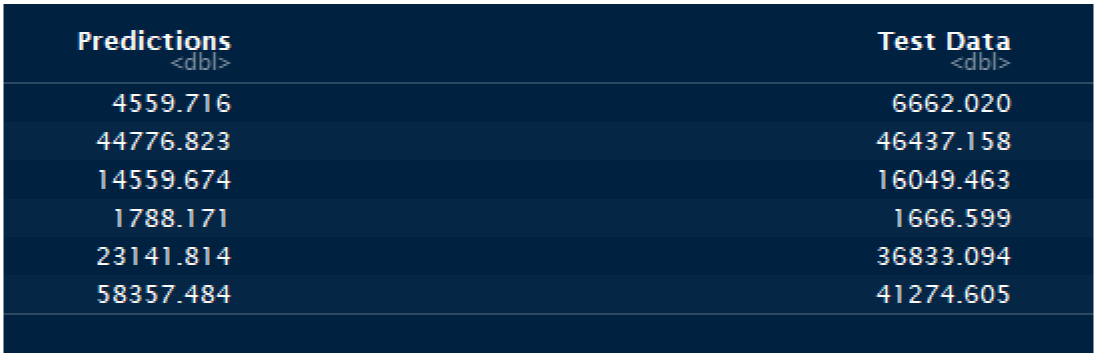
By looking at the table above, there are 7 predicted GDP per capita calculated. It is important and curious to notice that the Economic Freedom Index have huge impact on the Predicted GDP per capita, more than I was expected but IQ is statistically more significant.

The Prediction 6 and 7 were made to understand in a visual way the impact of IQ and Economic Freedom Index in the GDP per capita. The Prediction for a Country with IQ level of 150 with Economic Freedom "Repressed" is $296.459,00 which is almost double than the outliers that were eliminated. On the other hand a Country with very low IQ but with Economic Freedom has only $1.977,00 GDP per capita.

PREDICTING VALUES FROM A TRAINING AND TEST SET

The data was divided into a training set and a test set in order to apply the model and see how it predict values for GDP per capita and compares the test set and the predictions. The data was split with a ratio of 80-20.

In the table below it is possible to see and compare the test values of GDP per capita to the Prediction values for GDP per capita.



Predictions	Test Data
<dbl>	<dbl>
4559.716	6662.020
44776.823	46437.158
14559.674	16049.463
1788.171	1666.599
23141.814	36833.094
58357.484	41274.605

Some of the values are really close meaning that the Model #3 is a good fit for the data, but there still some minor differences in some cases and in other cases the model predicted a value way outside of the range, this might be because of the confidence intervals, when a value is outside the range, the prediction has more changes to have greater margin of error.

ASIDE Q

For this project I used at least 60 hypotheses, and I am sure that there are probably more that I missed. If I multiply 60 to every p-value used, most of the conclusions would still very similar with little change, some of them might change and make some difference, but the fundamental analysis and conclusion would remain the same.

CONCLUSION

After performing a semi-extensive analysis on the impact of IQ, Economic Freedom Index and Religious Freedom to the GDP per capita of all countries in the world, it was found that the Religious Freedom does not have statistical significance or influence on the GDP per capita. On the other hand, the IQ has the most influence on the GDP per capita followed by the Economic Freedom Index.

Predicting the GDP per capita for a Country needs more predictors that the ones that were used on this project and it would be recommended for further investigation to consider some variables like natural resources, alliances with other rich countries, continent and area, ports and airports, type of religious worshiped in each Country, and more.

Finding a good fit model for the data was not very difficult because there were not a lot of variables taken into consideration, that lead to less accurate predictions and a bigger margin of error. On the other hand we found out that the IQ really has a statistical significance on the GDP per capita of the Country, and it was found that there is a correlation between the two of about 0.52. Saying that, in social sciences it is hard to have a greater correlation because social science problems are multivariate which was not the scope of this project.

It was interesting seeing that the GDP per capita needed a logarithmic transformation in order to behave as a normal distribution, this indicates, by looking at the histogram of the GDP per capita, that the Pareto Distribution was present for the GDP per capita.

It was fun and enlighten to work in this project to understand better the influence of “Intelligence” on the wellbeing of a Country. As further studies it would be interesting to see what is the correlation and influence that the IQ have on personal success, income and happiness, but that’s for the future.

REFERENCES

Alon, I., & Chase, G. (2016). *In God's Name: Why Should Religious Freedom Affect Economic Prosperity?* Georgetown's Berkley Center: religious freedom institute.

Bartels, J. M., Ryan, J. J., Urban, L. S., & Glass, L. A. (2009). Correlation between estimates of state IQ and FBI crime statistics. *Personality and Individual Differences*, 579-583.

Dansen, L., Lee, T. A., & Detternan, D. (2003). The causal factor underlying the correlation between psychometric g and scholastic performance. *Intelligence 31*, 67-83.

Devlin, B. (1997). *Intelligence, Genes, and Success: Scientists Respond to The Bell Curve.* New York: Copernicus.

Flynn, J. R. (1994). *IQ gains over time.* New York: Macmillan.

Jencks, C. (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America.* New York: Basic Books.

Jensen, A. R. (1998). *The g Factor: The Science of Mental Ability.* Wesport: Praeger.

Lynn, R. M. (2007). National IQs predict differences in scholastic achievement in 67 countries. *Journal of Biosocial Science*, 861 - 874.

Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations.* Westport, CT: Praeger.

Lynn, R., & Vanhanen's, T. (2000). *Intelligence and the Wealth and Poverty of Nations.* Northen Ireland: Washington Summit Publisher.

Mackintosh, N. J. (1998). IQ and Human Intelligence. *Oxford University Press*.

Miller, T., Kim, A. B., & Roberts, J. M. (2018). *2018 Index of Economic Freedom.* New York : The Heritage Foundation.

Resing, W. and P. Drenth. (2007). *Intelligentie weten en meten.* Amsterdam: Uitgeverij Nieuwezijds.

Richard J. Herrnstein, C. M. (2010). *The Bell Curve.* USA: Simon Schuster.

Strenze, T. (2006). Intelligence and socioeconomic success: A meta-analytic. *Intelligence 35*, 402-423.

The World Bank. (2018). Retrieved from https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2018&start=2018

Ulric, N. (1998). *The Rising Curve: Long-Term Gains in IQ and Related Measures.* Washington: American Psychological Association.

Code will be posted on: https://github.com/ArMnKnows/GDPvsIQ