Statistical Programming for Social Data Science
Coursework 1
Analysis of the relationship between ESG markers, GDP and CO_2 emissions

Candidate Number: BBLF1
Student Number: 22232992

November 30, 2022

1 Introduction

The following report analyzes the relation between the ESG indicators and the GDP with the CO_2 emissions of 191 countries around the world. Particularly, this paper is aimed at testing the hypothesis where lower GDP values lead to lower CO_2 emissions per capita, but higher CO_2 emissions per Watt. Moreover, we want to analyze the relationship between other ESG indicators (e.g. gender equity, fair distribution of well-being, government expenditure on education, control of corruption, political stability) lead to higher CO_2 emissions per capita. - gender equity - access to the Internet - government expenditure on education - control of corruption - political stability

2 Datasets

2.1 ESG dataset

The World Bank's ESG Data Draft dataset (ESG_dataset) provides information on 17 key sustainability themes spanning environmental, social, and governance categories for 239 countries. For each country, the dataset describes the trend over time of 67 ESG indicators and each ESG marker has been evaluated for the years from 1960 to 2021. Even though the dataset contains the whole time-series of the ESG indicators over time, I decided to select only one specific year because I have not taken a proper course for time-series analysis yet.

2.1.1 Selecting the Year and the ESG indicators

The dataset contains many ESG indicators and many years, but we are only interested in one year and 5-10 ESG indicators as covariates. Moreover, most of the ESG indicators and years analyzed contain many missing values, so we analyzed the distribution of NAs in order to find the year with the lowest number of missing values in the relevant ESG indicators.

The biggest challenge is that the number of missing values in the selected year depends on the subset of ESG indicator selected, and the number of missing values of the ESG indicators depends on the selected year. To solve this, we selected them in three steps.

Step 1: Select a large subset of relevant ESG indicators As the first step, we selected a subset (*list1*) of 29 ESG indicators (2-3 similar indicators for each covariate that we would like to consider), that will enable us to compute the missing values for each year while considering only the relevant markers.

Step 2: Select the Year Relatively to the year selection, the idea is to select the most recent year with the lowest number of missing values, when considering only the relevant ESG indicators, previously selected in step 1. The Figure 1 shows the barplot that displays how many missing values in the relevant ESG markers are associated with each year, along with the trend of two of the relevant markers. Based on the plot, we identified 2018 as the most convenient year.

Step 3: Select a small subset of ESG indicators After selecting the year, we analyzed the large subset of ESG markers considering:

1. the missing values of each indicator in year 2018 only (see Figure 2)
2. the missing values of each indicator when considering the mean of the years 2017-2019 (while temporarily removing NAs in the computation). This is supposed to reduce the count of missing values, but, according to Figure 3, this approach does not show a significant improvement over 1.

Therefore, we selected the best ESG indicators, based on these results and on the following criteria:

- count of missing values in year 2018

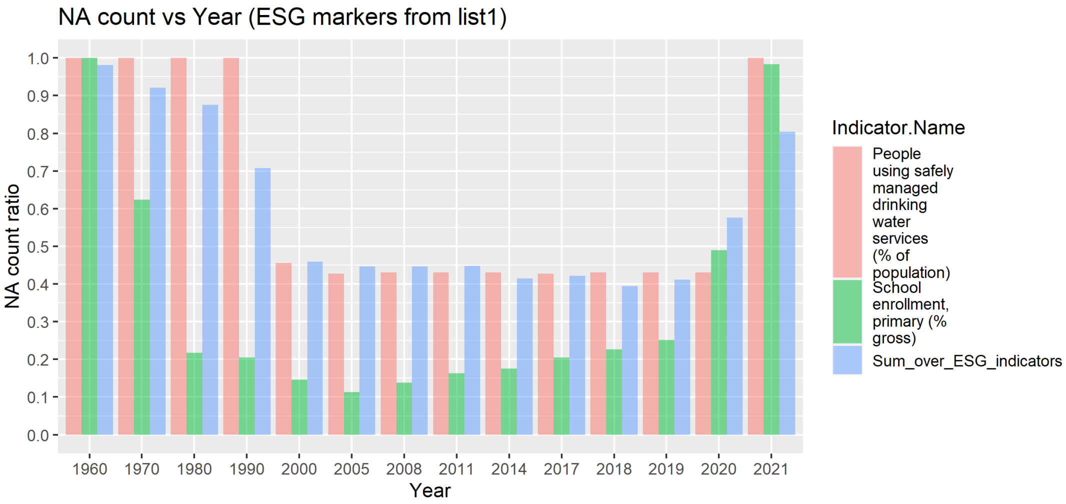


Figure 1: *Plot showing the count of missing values in a subset of ESG markers over the years 1960-2021. Particularly, the* ***Sum_over_ESG_indicators*** *serie considers the total count of missing values in the most relevant subset of ESG indicators (list_1). Notice that the ticks of the x-axis are not equally distant years, to improve readability and to focus the attention on the more recent years, which contain the lowest count of missing values.*

Category	ESG indicator	
Economy	Adjusted savings: natural resources depletion (% of GNI) Individuals using the Internet (% of population)	Natural resource depletion is the sum of net forest depletion, energy depletion, and mineral depletion
Governance	Government Effectiveness: Estimate Government expenditure on education, total (% of government expenditure) Control of Corruption: Estimate	
Human Rights	Ratio of female to male labor force participation rate (%) (modeled ILO estimate)	
Social and Health	Mortality rate, under-5 (per 1,000 live births) School enrollment, primary (% gross)	

Table 1: *Table describing the ESG markers selected and their category*

- relevance in relation to CO_2 *emission* outcome variable
- non-multiple collinearity with the other ESG indicators selected.

Finally the selected small subset of ESG indicators (*list2*) is found in Table 1

2.1.2 Cleaning the ESG dataset

To clean the ESG dataset we performed the following steps:

1. Drop the "Indicator.Code" column (we do not need a standardized code for it)
2. Fix the names of the Year columns by removing the initial "X" character
3. Analyze the ESG indicators contained in the dataset and select the most convenient ones as described in Section 2.1.1.
4. Drop the remaining missing values

2.2 CO2 emission dataset

The second dataset has been created by *Our World in Data*, a project of the Global Change Data Lab, which is a non-profit organization based in the United Kingdom. The data is built upon a number of datasets and processing steps, like the *Statistical review of world energy* from the BP company for the CO_2 emissions by energy type and the *Climate Adaptation Integration Tool (CAIT)* from World Resources Institute for the GHG emissions.

2.2.1 Cleaning the CO2 emission dataset

To clean the CO_2 dataset we followed the steps:

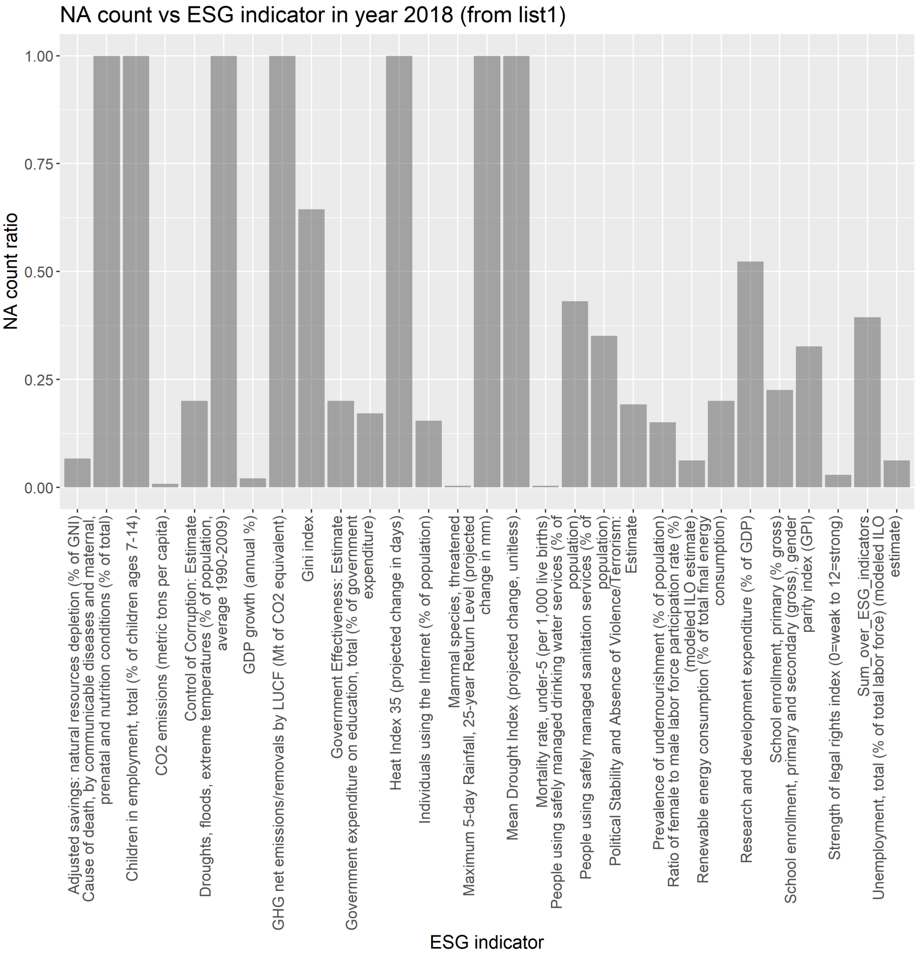


Figure 2: *Plot showing the count of missing values in the large subset of ESG markers (list1) when considering the year 2018 only.*

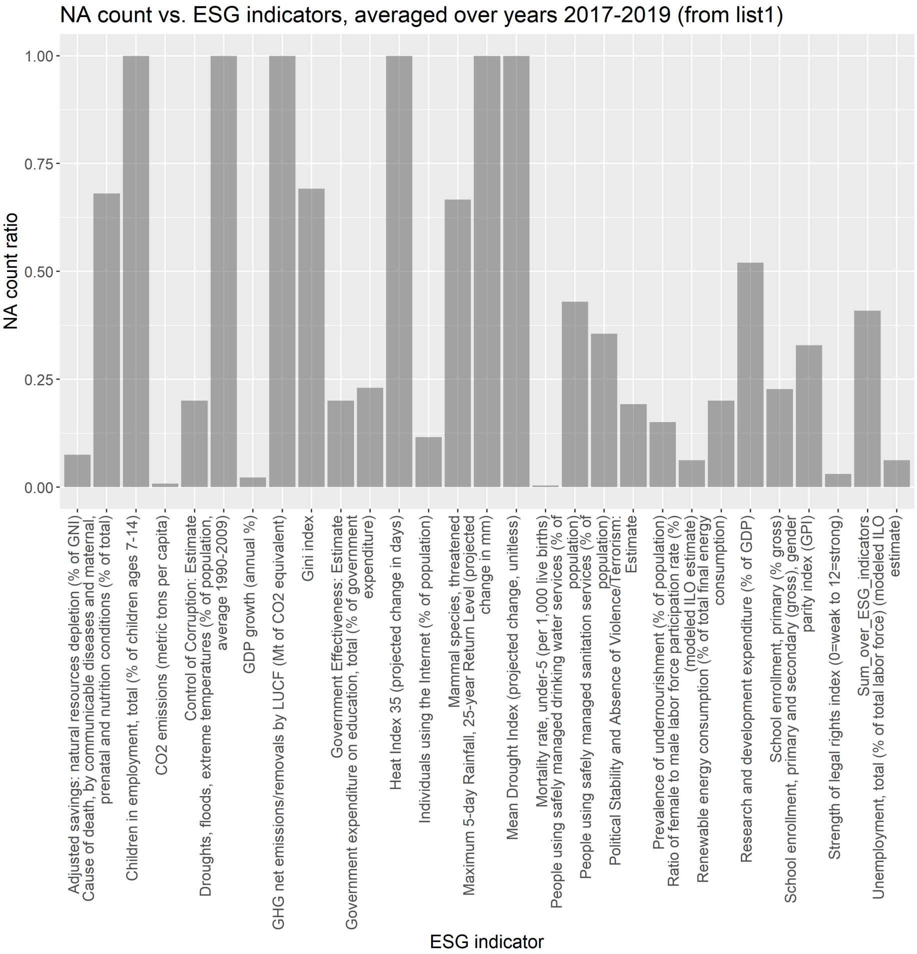


Figure 3: *Plot showing the count of missing values in a large subset of ESG markers (list1) when considering the average over the years 2017-2019. This is meant to reduce the missing value count by considering the nearest neighbours of the year 2018, when there is a missing value in 2018.*

1. Filter on the Year column based on the same year 2018 selected for the *ESG_dataset*, and then drop the "year" column
2. Filter the Country codes, based on the ones available in the ESG dataset
3. Compute the ratio of NAs for each covariate. According to the results, there are no NAs in the CO2 dataset, except for 5% in *gdp* column and 19% in *trade_co2* of missing values. For this reason, we decided to use only the columns *gdp*, *co2_per_capita*, and *total_ghg*.
4. Based on previous results, we drop *trade_co2* column and we drop NAs from *gdp* column:

3 Merging Datasets based on country codes

In order to merge the 2 datasets, we use the standardized country codes. Out of the original 191 common countries, we only have 95 countries, shared between the datasets, after filtering out the rows with NAs.

3.1 Cleaning merged dataset

To clean the resuting *merged dataset* we followed the steps:

1. Renaming the columns containing the country names in order to distinguish the original dataset and verify possible name incoherences
2. Compute the greenhouse gases emission values per capita by applying the following equation for each row:

$$value\ pro\ capita \backsim \frac{absolute\ value}{population} \tag{1}$$

3.2 Plotting the distribution of the OLS covariates

4 Regression models

4.1 Model 1: Ordinary Linear Regression (OLS) with all covariates

We fit an OLS model according to the formula:

$$co2\_per\_capita \backsim gdp + [esg\_selected\_indicators] \tag{2}$$

where the *esg_selected_indicators* are the ones listed in Table 1. The results are shown in the Figure 5 and we can see that the adjusted R-squared is , indicating that the covariates are explaining a good portion of the outcome variable. Moreover, the variables *Individuals using the Internet (% of population)* and *Adjusted savings: natural resources depletion (% of GNI)* have the lowest p-value, indicating a good correlation with the outcome variable *CO_2 emission.*

Residual plot The residual plot is shown in Figure 6 and we see that the errors are not homoskedastic and they have a quadratic shape, so in order to have a fully explanatory linear model and to use it for causal inference, we should fix these issues first.

4.1.1 Model 2: Most relevant covariates only, with quadratic terms

Based on the shape of the residual plot (Figure 6), we decided to select only the covariates with the lowest p-value, in order to reduce the complexity of the model, while adding few quadratic terms according to the formula:

$$\begin{aligned}
&co2\_per\_capita \backsim \\
&\qquad gdp\ *\ Individuals\ using\ the\ Internet\ (\%\ of\ population) + \\
&\qquad gdp\ *\ Adjusted\ savings:\ natural\ resources\ depletion\ (\%\ of\ GNI) + \\
&\qquad (Individuals\ using\ the\ Internet\ (\%\ of\ population))^2 \\
&\qquad (Adjusted\ savings:\ natural\ resources\ depletion\ (\%\ of\ GNI))^2,
\end{aligned}$$

Even though, the *adjusted R-squared* increased (see Figure 7), the residual plot did not change significantly and it highlights the same previous issues (see Figure 8).

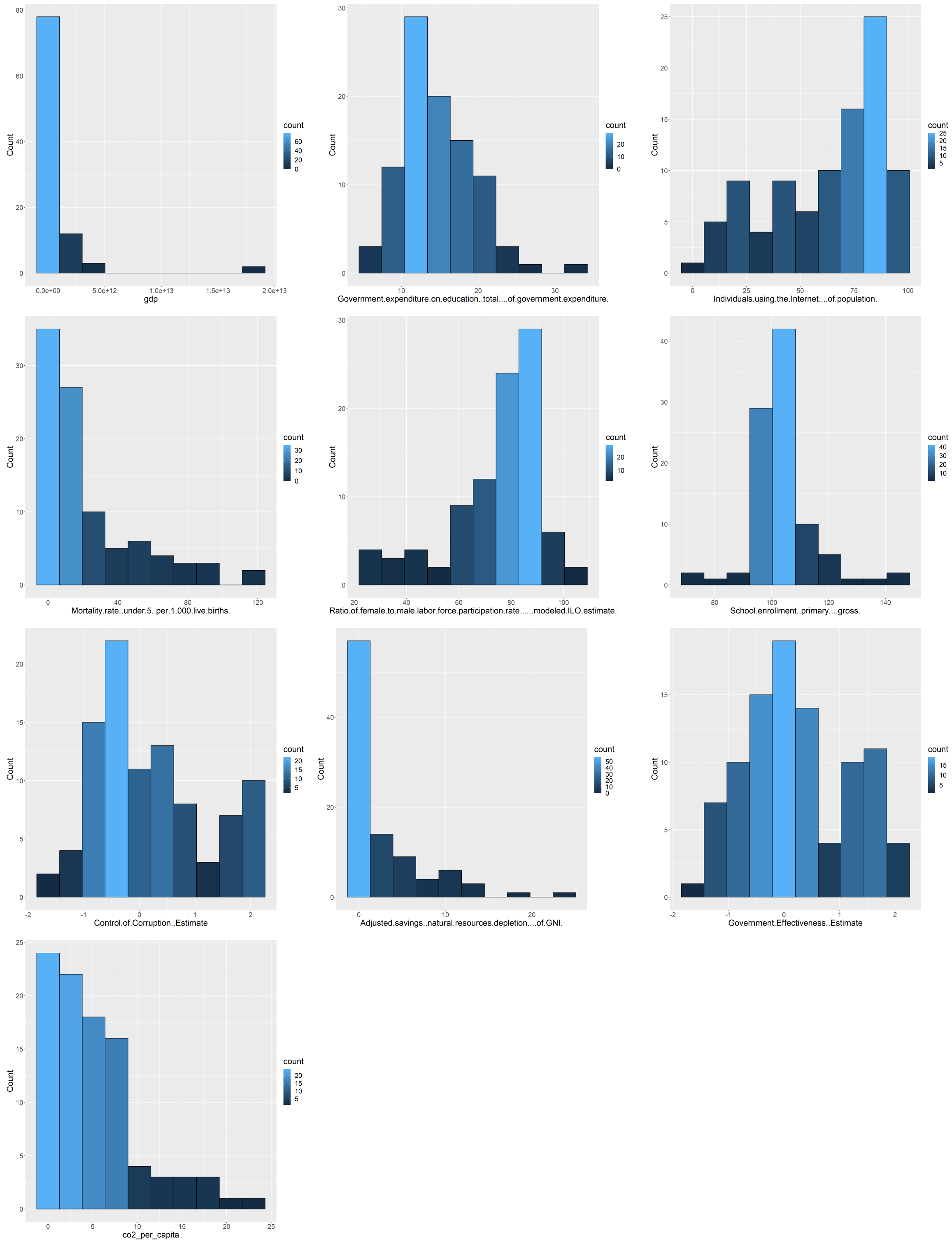
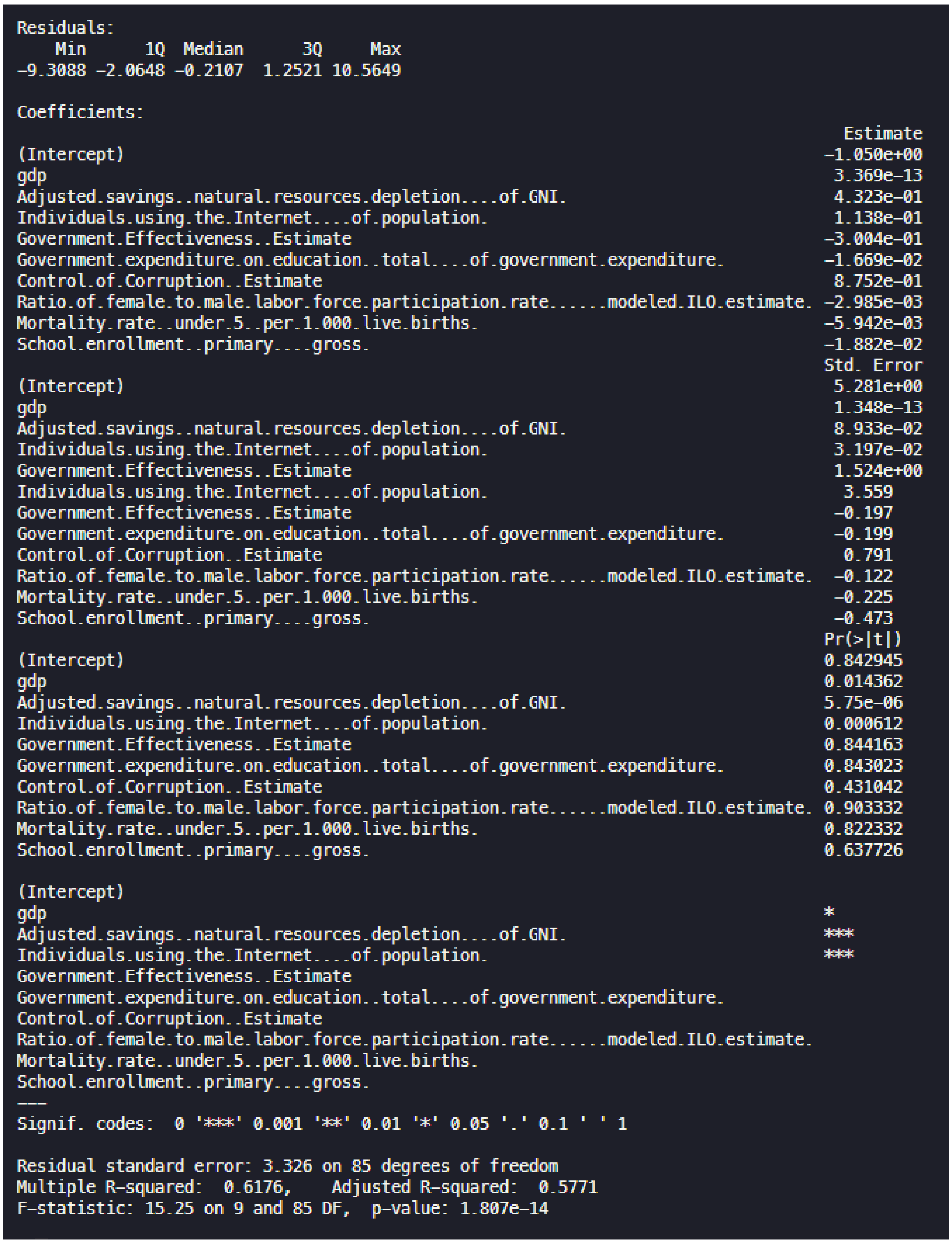


Figure 4: Distributions of the covariates and of the outcome variable *co2_per_capita*



```
Residuals:
    Min      1Q  Median      3Q     Max 
-9.3088 -2.0648 -0.2107  1.2521 10.5649 

Coefficients:
                                                                                    Estimate
(Intercept)                                                                       -1.050e+00
gdp                                                                                3.369e-13
Adjusted.savings..natural.resources.depletion....of.GNI.                           4.323e-01
Individuals.using.the.Internet....of.population.                                   1.138e-01
Government.Effectiveness..Estimate                                                -3.004e-01
Government.expenditure.on.education..total....of.government.expenditure.          -1.669e-02
Control.of.Corruption..Estimate                                                    8.752e-01
Ratio.of.female.to.male.labor.force.participation.rate......modeled.ILO.estimate. -2.985e-03
Mortality.rate..under.5..per.1.000.live.births.                                   -5.942e-03
School.enrollment..primary....gross.                                              -1.882e-02
                                                                                  Std. Error
(Intercept)                                                                        5.281e+00
gdp                                                                                1.348e-13
Adjusted.savings..natural.resources.depletion....of.GNI.                           8.933e-02
Individuals.using.the.Internet....of.population.                                   3.197e-02
Government.Effectiveness..Estimate                                                 1.524e+00
Individuals.using.the.Internet....of.population.                                    3.559
Government.Effectiveness..Estimate                                                 -0.197
Government.expenditure.on.education..total....of.government.expenditure.           -0.199
Control.of.Corruption..Estimate                                                     0.791
Ratio.of.female.to.male.labor.force.participation.rate......modeled.ILO.estimate.  -0.122
Mortality.rate..under.5..per.1.000.live.births.                                    -0.225
School.enrollment..primary....gross.                                               -0.473
                                                                                  Pr(>|t|)
(Intercept)                                                                       0.842945
gdp                                                                               0.014362
Adjusted.savings..natural.resources.depletion....of.GNI.                          5.75e-06
Individuals.using.the.Internet....of.population.                                  0.000612
Government.Effectiveness..Estimate                                                0.844163
Government.expenditure.on.education..total....of.government.expenditure.          0.843023
Control.of.Corruption..Estimate                                                   0.431042
Ratio.of.female.to.male.labor.force.participation.rate......modeled.ILO.estimate. 0.903332
Mortality.rate..under.5..per.1.000.live.births.                                   0.822332
School.enrollment..primary....gross.                                              0.637726

(Intercept)                                                                          
gdp                                                                               *  
Adjusted.savings..natural.resources.depletion....of.GNI.                          ***
Individuals.using.the.Internet....of.population.                                  ***
Government.Effectiveness..Estimate                                                   
Government.expenditure.on.education..total....of.government.expenditure.             
Control.of.Corruption..Estimate                                                      
Ratio.of.female.to.male.labor.force.participation.rate......modeled.ILO.estimate.    
Mortality.rate..under.5..per.1.000.live.births.                                      
School.enrollment..primary....gross.                                                 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.326 on 85 degrees of freedom
Multiple R-squared:  0.6176,	Adjusted R-squared:  0.5771 
F-statistic: 15.25 on 9 and 85 DF,  p-value: 1.807e-14
```

Figure 5: *Summary of the Ordinary Linear Regression model (Model 1).*

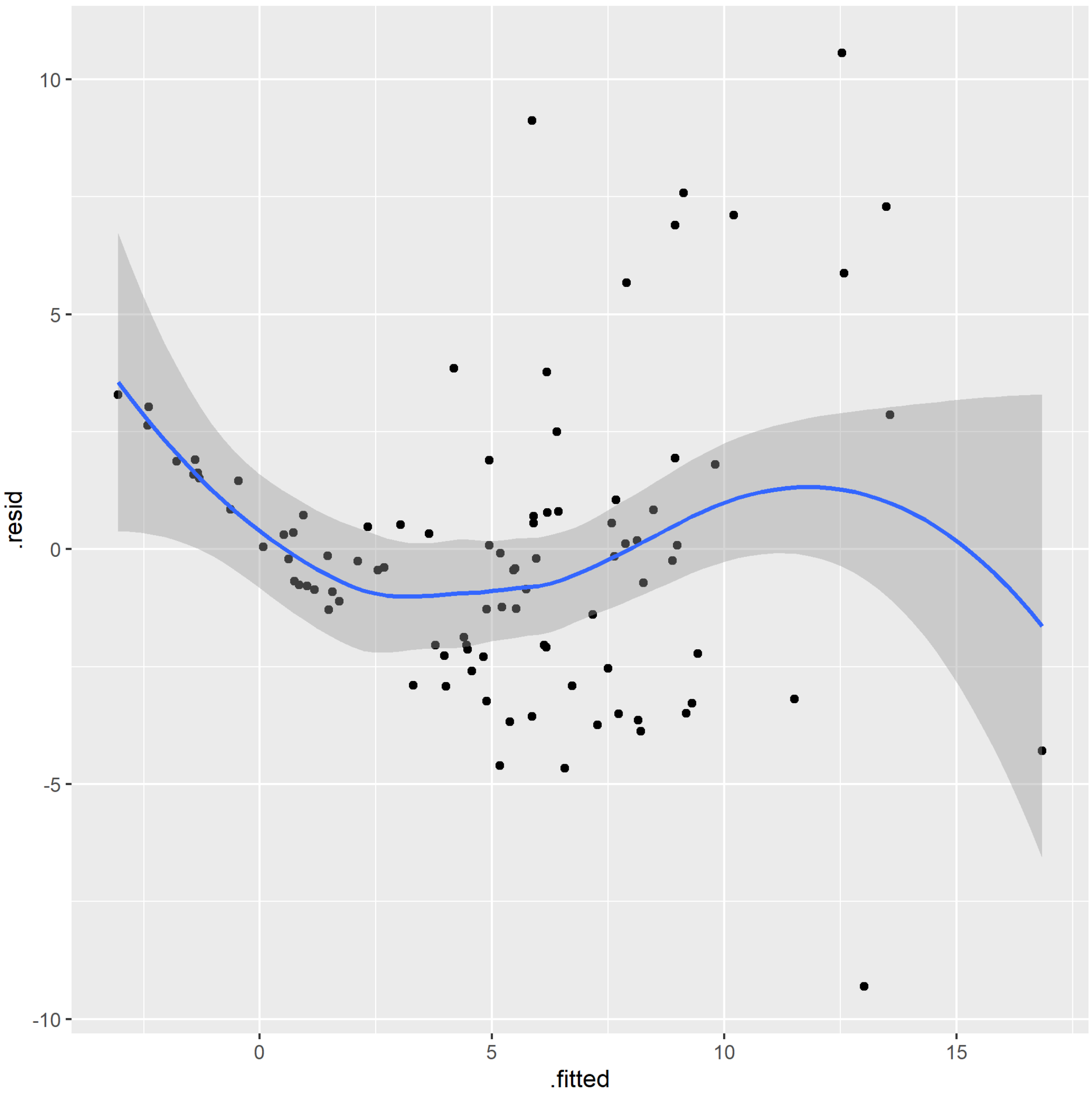
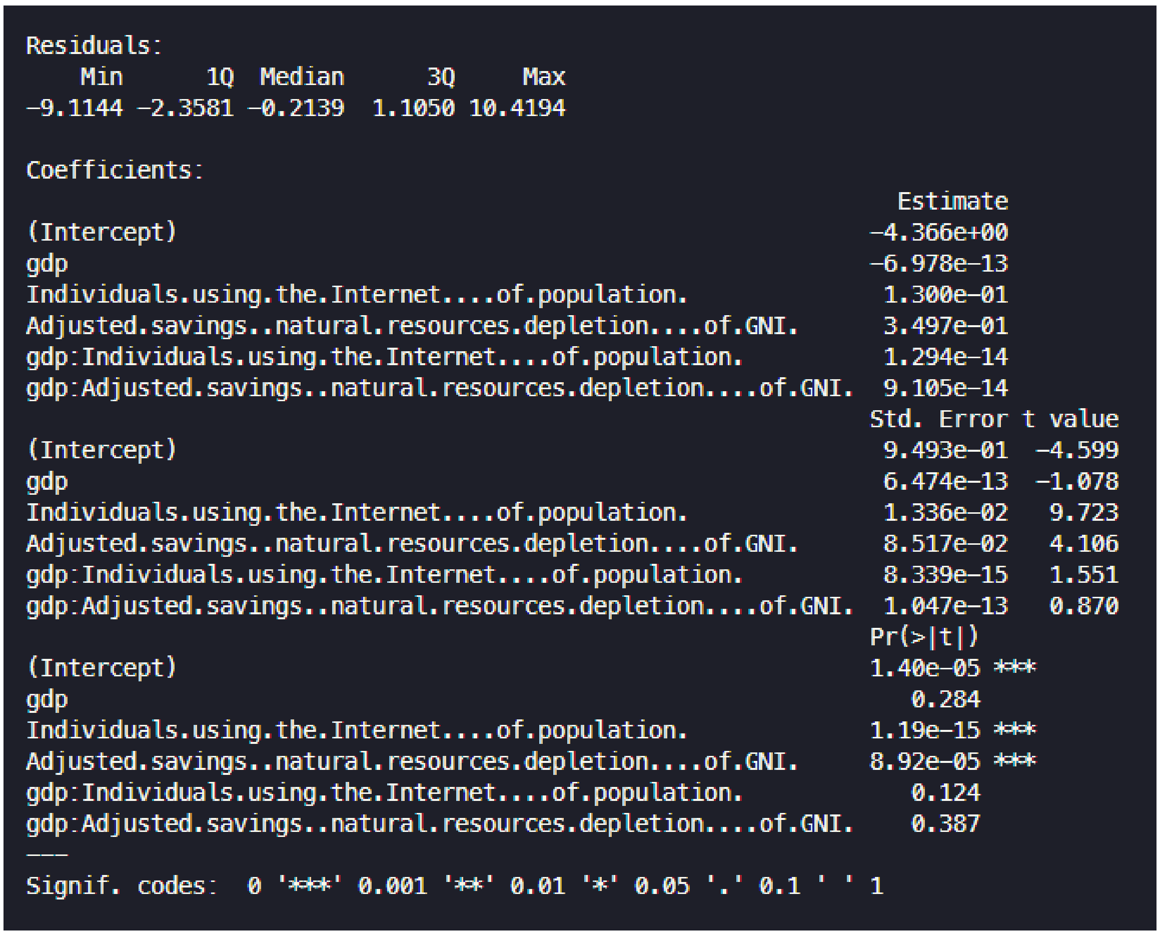


Figure 6: *The plot shows the residual values vs the fitted values for Model 1.*



```
Residuals:
    Min      1Q  Median      3Q     Max 
-9.1144 -2.3581 -0.2139  1.1050 10.4194 

Coefficients:
                                                               Estimate
(Intercept)                                                  -4.366e+00
gdp                                                          -6.978e-13
Individuals.using.the.Internet....of.population.              1.300e-01
Adjusted.savings..natural.resources.depletion....of.GNI.      3.497e-01
gdp:Individuals.using.the.Internet....of.population.          1.294e-14
gdp:Adjusted.savings..natural.resources.depletion....of.GNI.  9.105e-14
                                                             Std. Error t value
(Intercept)                                                   9.493e-01  -4.599
gdp                                                           6.474e-13  -1.078
Individuals.using.the.Internet....of.population.              1.336e-02   9.723
Adjusted.savings..natural.resources.depletion....of.GNI.      8.517e-02   4.106
gdp:Individuals.using.the.Internet....of.population.          8.339e-15   1.551
gdp:Adjusted.savings..natural.resources.depletion....of.GNI.  1.047e-13   0.870
                                                             Pr(>|t|)    
(Intercept)                                                  1.40e-05 ***
gdp                                                             0.284    
Individuals.using.the.Internet....of.population.             1.19e-15 ***
Adjusted.savings..natural.resources.depletion....of.GNI.     8.92e-05 ***
gdp:Individuals.using.the.Internet....of.population.            0.124    
gdp:Adjusted.savings..natural.resources.depletion....of.GNI.    0.387    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 7: *Summary of the Model 2.*

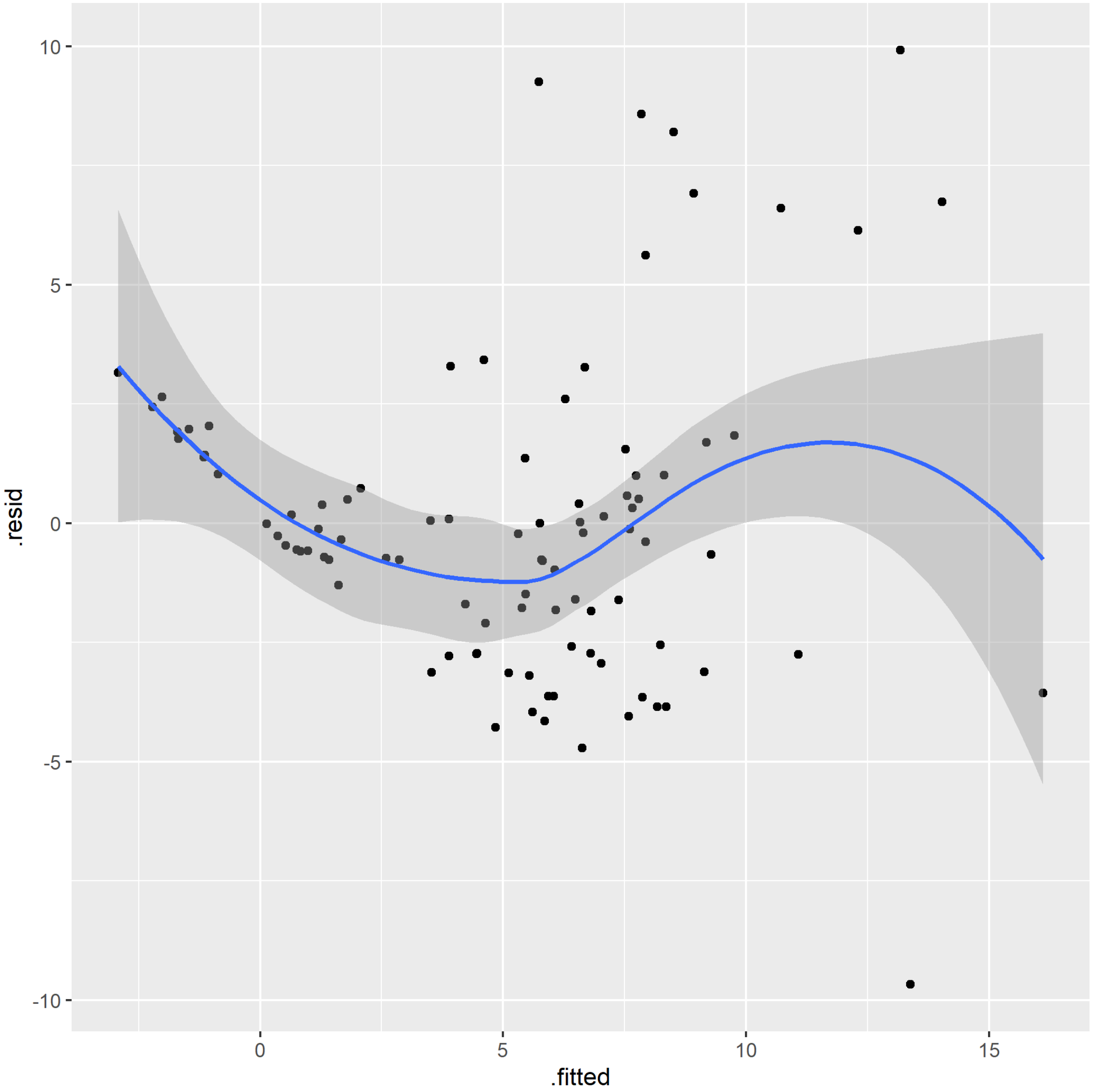


Figure 8: *The plot shows the residual values vs the fitted values for Model 2.*