Proposal

Alice Huang

29/03/2023

Data

Since the Industrial Revolution, Carbon Dioxide (CO2) Emissions have been increasing in Europe. High CO2 emissions have been associated with greenhouse gas effects and warming climates. High CO2 emissions and their resulting greenhouse gas effects have significant impacts on wildlife, human respiratory health, natural disasters, crop yields, and more. Thus understanding the impact of different factors associated with CO2 emissions is of great interest to many governments, businesses and individuals.

In this paper, we propose a Bayesian model for the 2019 Carbon Dioxide (CO2) emissions of European countries. The dataset was taken from Kaggle and originally pulled from the US Energy Commission. The original dataset had CO2 emissions data for world countries from 1980-2019, and stratified by energy type. However, we consider just countries from Europe so that the countries are of more comparable geographical size and population and use similar energy sources. We also combined the emissions from all energy types together for the sake of simplifying the model. We removed Iceland because there was missing data. We removed Russia because it was an extreme outlier in terms of CO2 emissions. We are interested in the influence of the following covariates on the 2019 Carbon Dioxide (CO2) emissions (in MMtonnes) of European countries: - `Energy_consumption`: amount of energy consumed (quad Btu) - `Energy_production`: amount of energy produced (quad Btu) - `GDP`: nation's purchasing power, measure of economic wealth (Billion 2015$ PPP) - `Population`: Number of people in country (Millions) - `Energy_intensity_per_capita`: measure of energy efficiency per unit of GDP (units: 1000 Btu/2015$ GDP PPP) - `Energy intensity_by_GDP`: measure of energy efficiency per person (MMBtu/person)

We centered all the covariates, to reduce correlation between variables.

Summary Statistics

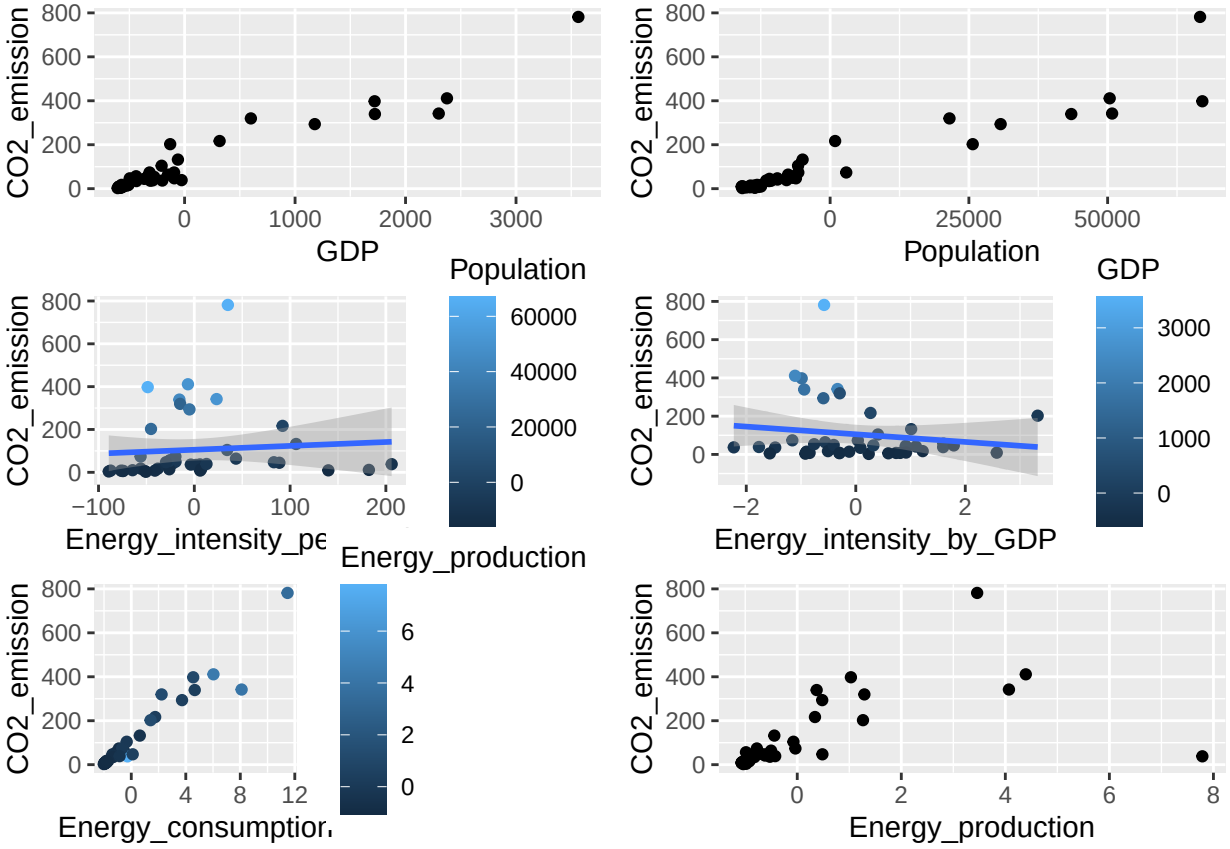
Here are the summary statistics of the variables.

Energy_consumption	Energy_production	GDP	Population	Energy_intensity_per_capita	Energy_intensity_by_GDP
Min. :-2.016506	Min. :-1.0687	Min. :-605.89	Min. :-15944	Min. :-89.33	Min. :-2.2275
1st Qu.:-1.861177	1st Qu.:-1.0013	1st Qu.:-563.89	1st Qu.:-13559	1st Qu.:-47.80	1st Qu.:-0.8850
Median :-1.213861	Median :-0.7380	Median :-341.08	Median :-10078	Median :-17.57	Median :-0.2951
Mean : 0.000000	Mean : 0.0000	Mean : 0.00	Mean : 0	Mean : 0.00	Mean : 0.0000
3rd Qu.: 0.009343	3rd Qu.: 0.2467	3rd Qu.: -93.68	3rd Qu.: -5155	3rd Qu.: 20.63	3rd Qu.: 0.7923

Energy_consumption	Energy_production	GDP	Population	Energy_intensity_per_capita	Energy_intensity_by_GDP
Max. :11.468261	Max. : 7.7868	Max. :3564.18	Max. : 67066	Max. :206.05	Max. : 3.3223

Plotting the Covariates Against CO2 Emissions

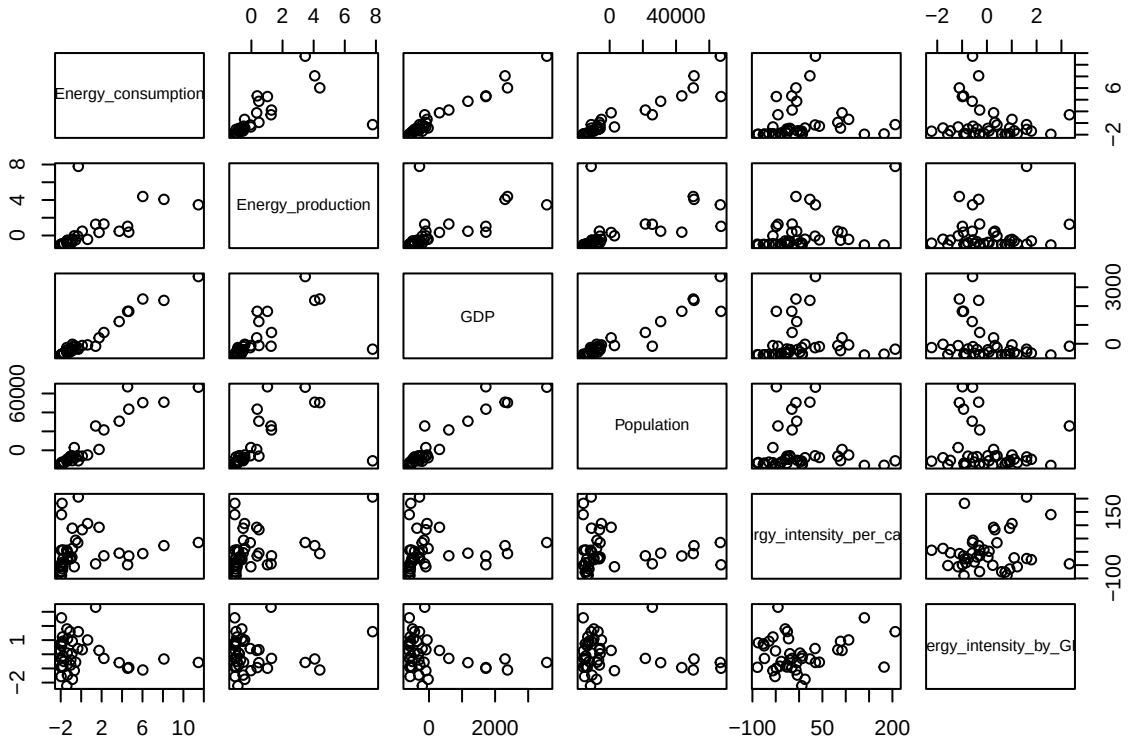
We noted that generally, as energy production increased, CO2 emissions increased. Similarly, as energy consumption increased, CO2 emissions increased. As population increased, CO2 emissions increased. On the other hand, `Energy_intensity_per_capita` and `Energy intensity_by_GDP` did not seem to significantly influence CO2 emissions.



Checking for Correlation Between Variables

The columns of the correlation matrix correspond to `Energy_consumption`, `Energy_production`, `GDP`, `Population`, `Energy_intensity_per_capita`, `Energy_intensity_by_GDP` from left to right.

```
##            [,1]       [,2]       [,3]        [,4]        [,5]        [,6]
## [1,]  1.0000000 0.66593422  0.98479391  0.94480341  0.13452921 -0.15008117
## [2,]  0.6659342 1.00000000  0.63147832  0.59161564  0.40516009  0.07041958
## [3,]  0.9847939 0.63147832  1.00000000  0.94898259  0.08846206 -0.26494276
## [4,]  0.9448034 0.59161564  0.94898259  1.00000000 -0.02417416 -0.16073089
## [5,]  0.1345292 0.40516009  0.08846206 -0.02417416  1.00000000  0.16477103
## [6,] -0.1500812 0.07041958 -0.26494276 -0.16073089  0.16477103  1.00000000
```



There is strong correlation between the following pairs of covariates:

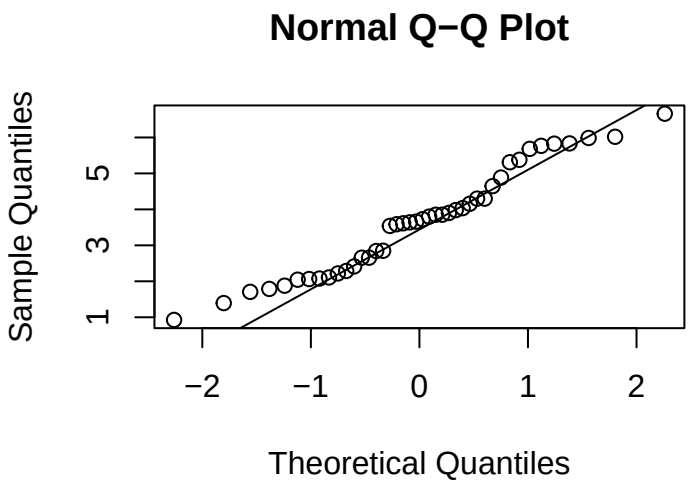
- `Energy_consumption`, `Energy_production`
- `Energy_consumption`, `GDP`
- `Energy_consumption`, `Population`
- `Energy_production`, `Population`
- `Energy_production`, `GDP`
- `GDP`, `Population`

The correlation between `Energy_consumption`, `GDP` is very close to 1, so we decide to just include `GDP` instead of both variables. Thus we will be interested in the following interactions:

- `Energy_production`, `GDP`
- `GDP`, `Population`

Considering What Type of Likelihood Model to Use

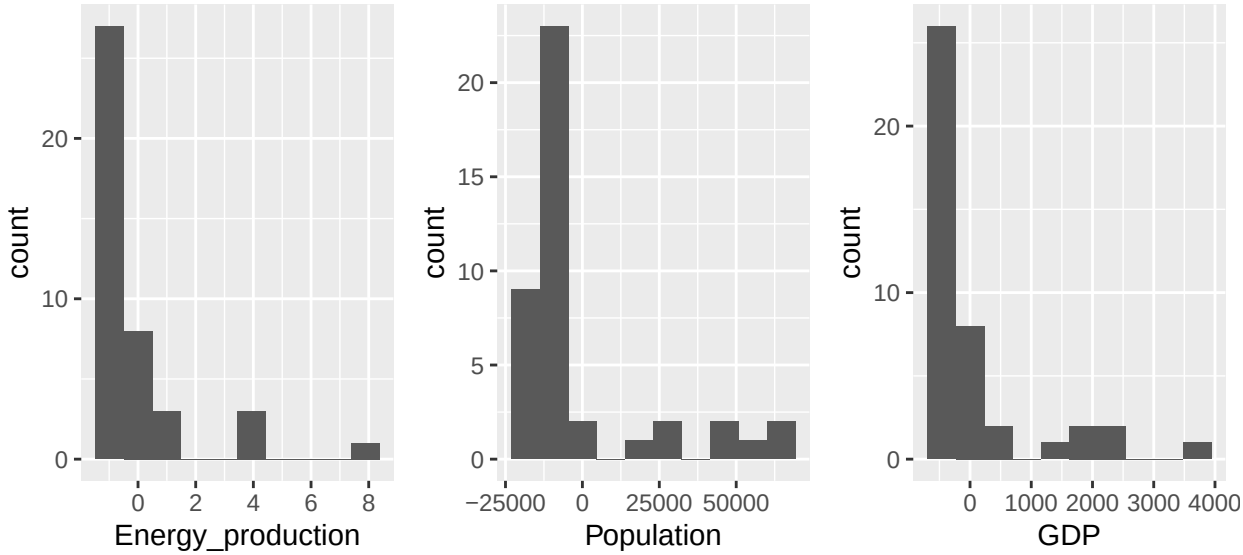
We consider a log-linear model. We use the Shapiro-Wilk test to test whether `log(CO2_emission)` is normally distributed. The p-value is $0.1566 > 0.05$ so we do not reject the null hypothesis that `log(CO2_emission)` is normally distributed. Thus a log-linear model may be a viable option.




```
##
##  Shapiro-Wilk normality test
##
## data:  log(europe2019$CO2_emission)
## W = 0.96071, p-value = 0.1566
```

Histograms of Covariates For Determining Appropriate Prior Distributions

We see that the covariates' histograms are all heavily skewed to the right.



We use the 2018 emissions data to help determine what might be reasonable priors for the 2019 data. We come up with estimates for the prior means by plotting the 2018 log CO2 emissions against each covariate and interaction term, in order to estimate the expected change in log CO2 emissions for each unit covariate change. We chose the mean of the prior for the 2019 coefficient to be the slope of the trendline on the 2018 log CO2 emissions against the 2018 coefficient. We put higher variances on the variables that seem to have more variation.

The standard deviation of log CO2 emissions has mean close to 1.5 so we consider a half normal prior with mean 1.5 and choose a standard deviation of 1.

Proposed Model

We propose the following model:

$\log(Y_i)|\mu_i, \sigma_i^2 \sim N(\mu_i, \sigma_i^2)$

$\log(\mu_i) = \beta_0 + \beta_1 P_i + \beta_2 E_i + \beta_3 G_i + \beta_4 E_i G_i + \beta_5 G_i P_i$

$\sigma^2 \sim N^+(1.5, 2)$

$\beta_0 \sim N(0, 10)$

$\beta_1 \sim N(0, 5)$

$\beta_2 \sim N(0.5, 2)$

$\beta_3 \sim N(0.0025, 5)$

$\beta_4 \sim N(0, 4)$

$\beta_5 \sim N(0, 4)$

where P_i is the population of country i, E_i is the energy production of country i and G_i is the GDP of country i.

Reference:

https://www.kaggle.com/datasets/lobosi/c02-emission-by-countrys-grouth-and-population