Hierarchical model of and CO2 emissions and energy usage

Pim Nelissen

Introduction

Climate studies are complex, and there can be non-trivial relations between various predictors. Nonetheless, a rigorous statistical approach could lead us to some insights. The research question that we will attempt to answer here is:

> **Does a lower share of fossil fuel production in the energy mix result in a decoupling of energy usage and CO_2 emissions?**

In other words: does a high production of fossil fuels increase the effect that energy usage has on CO_2 emissions?

In the spirit of FAIR, this report, as well as it's source file, data attribution, and license, are included in a git repository which can be accessed here:

https://gitlab.com/pimnelissen/bern02-challenge-project/

Method

In this study, we will look at data consisting of CO2 emissions (tonnes, per capita), fossil fuel production (kWh, per capita) and energy usage (kWh, per capita), which will be retrieved using standrd protocols from public sources. For modelling, we will first construct a hierarchical model and then use a Bayesian hierarchical approach using the `brms` package. The entire workflow is included here and covers:

- Fetching and wrangling the data
- Defining and fitting a model
- Plotting results

The report in particular focuses on reproducability and adhering to FAIR principles (this will be discussed in more detail at the end).

Libraries

Here you may set your own working directory. Set `use_saved` to `FALSE` if you wish to re-run the model.

```
setwd("/home/pim/BERN02/challenge_project/")
use_saved = TRUE # whether to use a saved version of the model.
SEED = 2468 # seed for model fitting
```

```
library(tidyverse)
library(glue)
library(jsonlite)
library(brms)
```

Loading the data

```
# Fetch the data
df_co2 <-
 ↪  read.csv("https://ourworldindata.org/grapher/co-emissions-per-capita.csv?v=1&csvType=full

# Fetch the metadata
metadata_co2 <-
 ↪  fromJSON("https://ourworldindata.org/grapher/co-emissions-per-capita.metadata.json?v=1&cs

# Fetch the data
df_energy <-
 ↪  read.csv("https://ourworldindata.org/grapher/per-capita-energy-use.csv?v=1&csvType=full&u

# Fetch the metadata
metadata_energy <-
 ↪  fromJSON("https://ourworldindata.org/grapher/per-capita-energy-use.metadata.json?v=1&csvT

# Fetch the data
df_prod <-
 ↪  read.csv("https://ourworldindata.org/grapher/per-capita-electricity-fossil-nuclear-renewa

# Fetch the metadata
metadata_prod <-
 ↪  fromJSON("https://ourworldindata.org/grapher/per-capita-electricity-fossil-nuclear-renewa
```

Data attribution

The data has been collected from the following sources:

```
glue("CO_2 emissions dataset:\n {metadata_co2$chart$citation}\n\n",
     "Energy usage dataset:\n {metadata_energy$chart$citation}\n\n",
     "Energy production dataset:\n {metadata_prod$chart$citation}")
```

```
CO_2 emissions dataset:
 Global Carbon Budget (2024); Population based on various sources (2024)

Energy usage dataset:
 U.S. Energy Information Administration (2025); Energy Institute - Statistical Review of Wor

Energy production dataset:
 Ember (2025); Energy Institute - Statistical Review of World Energy (2025); Population base
```

Selecting data

```
sel_energy <- df_energy %>%
  select(Year, Entity, primary_energy_consumption_per_capita__kwh) %>%
  rename(energy_usage = primary_energy_consumption_per_capita__kwh)

sel_co2 <- df_co2 %>%
  select(Year, Entity, emissions_total_per_capita) %>%
  rename(co2_emissions = emissions_total_per_capita)

# calculate the share of fossil fuel production (fossil/total)
sel_prod <- df_prod %>%
  select(
    Year, Entity,

 ↪  per_capita_fossil_generation__kwh_chart_per_capita_electricity_fossil_nuclear_renewables

 ↪  per_capita_nuclear_generation__kwh_chart_per_capita_electricity_fossil_nuclear_renewables

 ↪  per_capita_renewable_generation__kwh_chart_per_capita_electricity_fossil_nuclear_renewabl
 ↪  %>%
  mutate(total =
   ↪  per_capita_fossil_generation__kwh_chart_per_capita_electricity_fossil_nuclear_renewable
   ↪  +
   ↪  per_capita_nuclear_generation__kwh_chart_per_capita_electricity_fossil_nuclear_renewabl
   ↪  +
   ↪  per_capita_renewable_generation__kwh_chart_per_capita_electricity_fossil_nuclear_renewa
   ↪  %>%
```

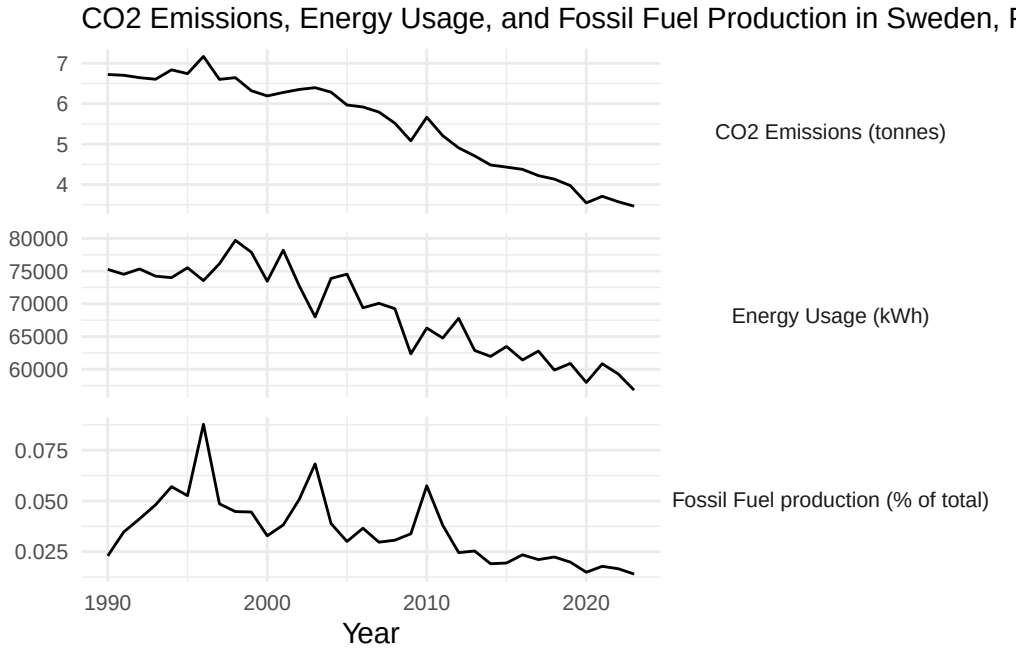
```r
  mutate(fossil =
   ↪  per_capita_fossil_generation__kwh_chart_per_capita_electricity_fossil_nuclear_renewable
   ↪  / total)

# combine all the data
combined_data <- sel_energy %>%
  left_join(sel_co2, by = c("Year", "Entity")) %>%
  left_join(sel_prod, by = c("Year", "Entity")) %>%
  drop_na()
```

A look at the data

```r
# pivot

plot_country_data <- function(country) {
  country_data <- combined_data %>%
  filter(Entity == country) %>%
  pivot_longer(
    cols = c(energy_usage, co2_emissions, fossil),
    names_to = "Variable",
    values_to = "Value"
  ) %>%
  mutate(Variable = case_when(
    Variable == "energy_usage" ~ "Energy Usage (kWh)",
    Variable == "co2_emissions" ~ "CO2 Emissions (tonnes)",
    Variable == "fossil" ~ "Fossil Fuel production (% of total)",
    TRUE ~ Variable
  ))

ggplot(country_data, aes(x = Year, y = Value)) +
  geom_line() +
  facet_grid(Variable ~ ., scales = "free_y") +
  labs(title = glue("CO2 Emissions, Energy Usage, and Fossil Fuel Production
   ↪  in {country}, Per Capita"),
       x = "Year",
       y = "") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 11),
    panel.spacing = unit(0.5, "lines"),  # Increase spacing between panels
```

```
    strip.text.y = element_text(angle = 0, size = 8),  # Adjust text size and
     ↪  angle
    axis.text.y = element_text(size = 8),  # Adjust y-axis text size
    axis.text.x = element_text(size = 8)   # Adjust x-axis text size
  )
}
```

```
plot_country_data("Sweden")
```



Standardise the data

In order to have more interpretable results and make it easier to chose priors, we standardise the data (scale between -1 and 1).

```
standardized_data <- combined_data %>%
  mutate(across(c(energy_usage, co2_emissions, fossil), ~ scale(.x)))
```

Model

We want to know whether fossil fuel share has an effect on the relation between energy usage and CO_2 emissions. Let the model for this be:

$$C_{ij} = \beta_0 + \beta_1 E_{ij} + \beta_2 F_{ij} + \beta_3 (E_{ij} \cdot F_{ij}) \tag{1}$$

$$+ b_{0j} + b_{3j}(E_{ij} \cdot F_{ij}) + \varepsilon_{ij} \tag{2}$$

$$b_{0j} \sim \mathcal{N}(0, \tau_0^2) \tag{3}$$

$$b_{3j} \sim \mathcal{N}(0, \tau_3^2) \tag{4}$$

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{5}$$

where C_{ij} is the i-th observation of the CO2 emissions per capita in country/region j. E_{ij} and F_{ij} are the i-th observation of energy usage per capita and fossil fuel share in the energy mix in region j. $E_{ij} \cdot F_{ij}$ is an interaction term between energy usage and fossil fuel share. This term will be able to quantify whether the effect of energy usage on CO2 emissions depends on the fossil fuel share, which is exactly what we want to find out to test our hypothesis. b_{0j} is the random intercept for region j and b_{3j} is the random slope for the interaction term in region j . These random intercepts and slopes make the model hierarchical, and allowing for this variability between regions will allow us to study whether the effect varies between countries/regions.

Hypothesis test

Let the two hypotheses be

$$H_0 : \beta_3 = 0 H_1 : \beta_3 > 0$$

That is, we want to know whether a higher fossil fuel share in the energy mix increases the correlation between energy usage and CO_2 emissions.

Let us say that $P(\text{rejecting } H_0 | H_0 \text{ is true}) = 0.05$ is a satisfactory condition, then we seek to show that $0 \notin \text{CI}_{\beta_3}$ where CI_{β_3} is the $(1 - \alpha) = 95\%$ confidence interval. With that in mind, let us define the model in `brms` and fit it:

```
if (use_saved) {
  bf_model <- readRDS("fitted_model.rds")
} else {
  priors <- c(
    prior(normal(0, 1), class = "b")
```

```
  )

  bf_model <- brm(
    co2_emissions ~ energy_usage * fossil + (1 + energy_usage:fossil |
↪  Entity), # variablity between regions
    data = standardized_data,
    family = gaussian(),
    prior = priors,
    cores = 4,
    iter = 4000,
    seed = SEED
  )

  saveRDS(bf_model, "fitted_model.rds")
}
```

Results

Let's print out the summary of the model.

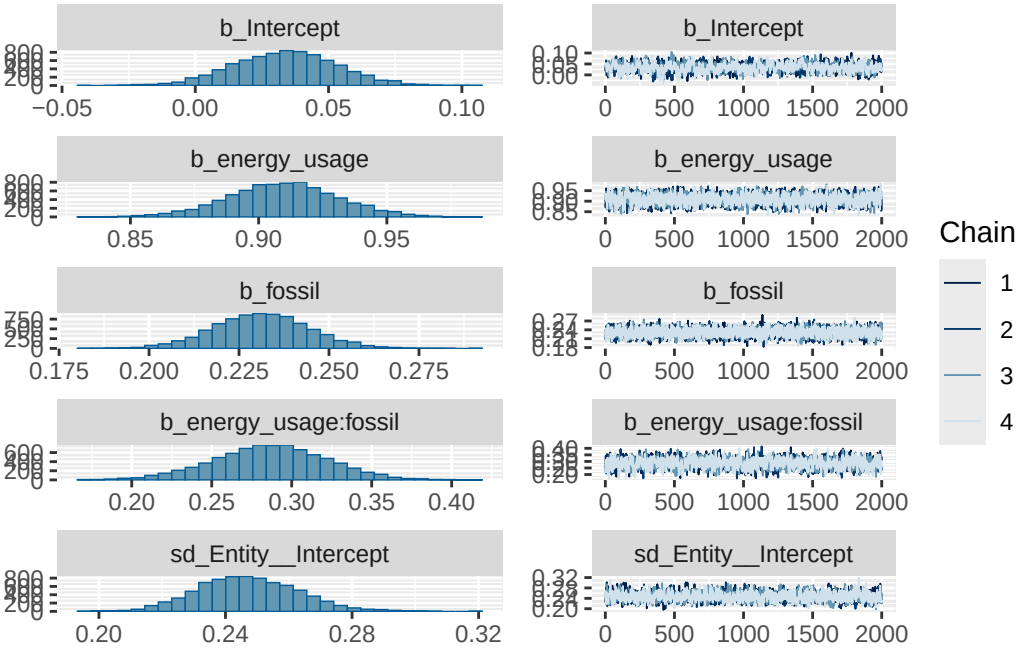
```
summary(bf_model)
```

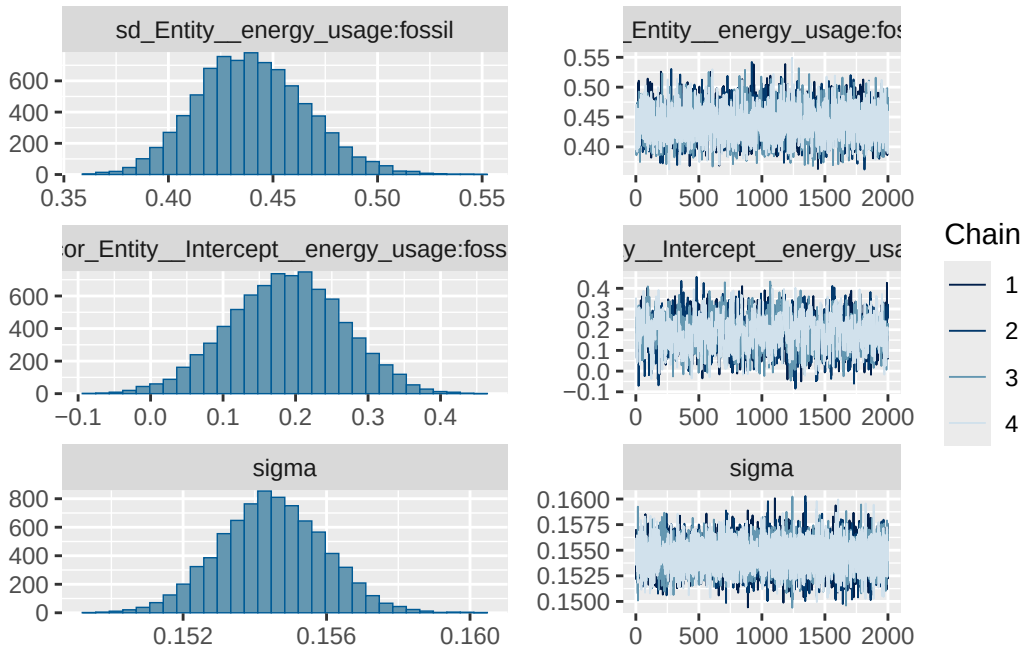
```
 Family: gaussian
  Links: mu = identity
Formula: co2_emissions ~ energy_usage * fossil + (1 + energy_usage:fossil | Entity)
   Data: standardized_data (Number of observations: 5854)
  Draws: 4 chains, each with iter = 4000; warmup = 2000; thin = 1;
         total post-warmup draws = 8000

Multilevel Hyperparameters:
~Entity (Number of levels: 214)
                                  Estimate Est.Error l-95% CI u-95% CI Rhat
sd(Intercept)                         0.25      0.02     0.22     0.28 1.00
sd(energy_usage:fossil)               0.44      0.03     0.39     0.49 1.00
cor(Intercept,energy_usage:fossil)    0.19      0.08     0.03     0.34 1.00
                                  Bulk_ESS Tail_ESS
sd(Intercept)                         1974     3678
sd(energy_usage:fossil)               3099     5144
cor(Intercept,energy_usage:fossil)    1180     2356
```

```
Regression Coefficients:
                   Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept              0.03      0.02    -0.00     0.07 1.00      811     1886
energy_usage           0.91      0.02     0.87     0.95 1.00     2862     4232
fossil                 0.23      0.01     0.21     0.26 1.00     4286     5521
energy_usage:fossil    0.29      0.04     0.22     0.35 1.00     1719     3129

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.15      0.00     0.15     0.16 1.00    14650     5300

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Let's plot all the posteriors.

```
plot(bf_model, ask = FALSE)
```





Looking at the regression coefficients, our results seem very supportive of our hypothesis. Indeed, when we check the hypothesis…

```
posterior_samples <- as.data.frame(bf_model)
prob_beta3_eq_0 <- mean(posterior_samples$`b_energy_usage:fossil` == 0)
print(glue("P(beta_3 = 0): {prob_beta3_eq_0}"))
```

```
P(beta_3 = 0): 0
```

```
prob_beta3_gt_0 <- mean(posterior_samples$`b_energy_usage:fossil` > 0)
print(glue("P(beta_3 > 0): {prob_beta3_gt_0}"))
```

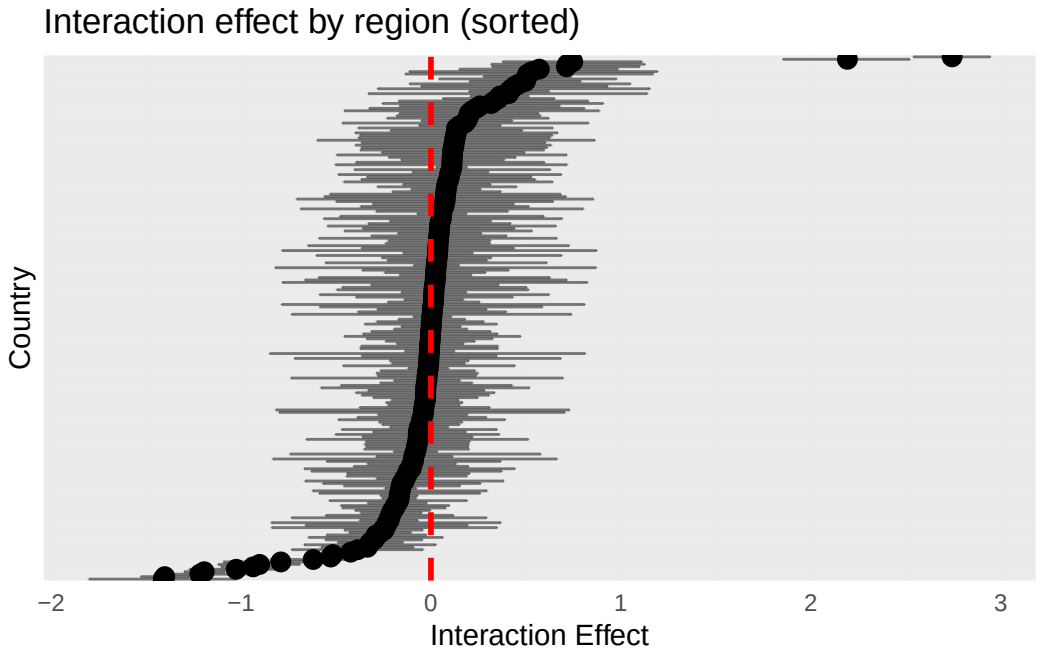
```
P(beta_3 > 0): 1
```

We find that most certainly, a reduced share in fossil fuels in the energy mix decouples CO2 emissions from energy usage. But is this really a fair conclusion? Let's explore the posterior estimates for variable slopes for each region, in order to look at the interaction effect by region.

```r
interaction_effects <- ranef(bf_model)$Entity[, , "energy_usage:fossil"]
interaction_df <- as.data.frame(interaction_effects)

# Add continent names as a column
interaction_df$Entity <- rownames(interaction_df)

# Sort the data frame by the estimate of the interaction effect
interaction_df <- interaction_df %>%
  arrange(Estimate)

# Create a dot-and-whisker plot of the selected interaction effects with
 ↪  confidence intervals
p <- ggplot(interaction_df, aes(x = reorder(Entity, Estimate), y = Estimate))
 ↪  +
  geom_point(size = 3) +  # Points for the estimates
  geom_errorbar(aes(ymin = Q2.5, ymax = Q97.5), width = 0.2, alpha = 0.5) +
   ↪  # Whiskers for the confidence intervals
  geom_hline(yintercept = 0, linetype = "dashed", color = "red", size = 1) +
  labs(x = "Country", y = "Interaction Effect", title = "Interaction effect
   ↪  by region (sorted)") +
  theme_minimal() +
  coord_flip()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```r
p + theme(axis.text.y=element_blank())
```



This highlights a problem with taking the overall effect of the interaction for all countries instead of individual ones. This dot-and-whisker plot shows great variability among the countries. This justifies the use of random slopes, and also brings into question whether the original conclusion was correct. Moreover, we have very large uncertainties for most regions. We can see that for the vast majority of regions, the interaction is not clear and the null-hypothesis cannot be rejected as the 95% credible interval includes 0 (see the red dashed line for the threshold). That is, it is not clear from this model/data that a reduction in fossil fuel in the energy mix reduces the effect that energy usage has on CO_2 levels.

Although there are large uncertainties within each region, the random effects (intercept, slope) are normally distributed and the model seems to have converged ($\hat{R} = 1$ for all parameter estimates).

Limitations of the model

A hierarchical model which allows variable slopes for each country/region is a good step in the right direction. However there are still some limitations with this model overall.

1. Few predictors, so potentially biased estimates.
2. Assumption of linearity (that CO_2 is linearly proportional to the predictors)
3. Instead of the entire dataset (more than 6000 observations), using a variable slope for each region means only a few data points per region. This results in wide confidence intervals, most of which include 0, so this makes it impossible to test the hypothesis for the majority of the regions.

Discussion of FAIR principles

- **F (Findable):** Data are described with rich metadata (from OurWorldInData). Data are registered and indexed in a searchable resource (OurWorldInData).
- **A (Accessible):** (Meta)data are retrievable by https, a standard protocol.
- **I (Interoperable):** Data is shared in CSV format.
- **R (Reusable):** Origin of the data is well-described and the procedure for obtaining and wrangling that data is described.