# CO-2 data analysis

Bortolotti Simone, Franzè Lorenzo, Stancioi Ioana-Ruxandra

# Content

- Task and dataset
- Preliminary analysis
- BAS analysis and model selection
- Models (JAGS)
- Clustering
- Conclusions

# Task and dataset

Section 1

# Task and dataset

- ➔ Explain the $CO_2$ emissions with other variables
- ➔ Analysis of the relation between GDP and $CO_2$

  (checking whether the richer we are, the more $CO_2$ we emit)

| | country | y | EnergyUse | GDP | pop | co2percap | Lowcarbor | internet | urb |
|---|---|---|---|---|---|---|---|---|---|
| 6 | Algeria | 2005 | 11373.61 | 10504.86 | 33149720 | 3.2119 | 0.408 | 1945357 | 63.83 |
| 13 | Argentina | 2005 | 20011.44 | 19426.44 | 38892924 | 4.1507 | 15.966 | 6936809 | 90.031 |
| 19 | Australia | 2005 | 64710.33 | 42217.14 | 20178543 | 19.2086 | 4.01 | 12750509 | 84.582 |
| 21 | Austria | 2005 | 47515.15 | 49316.26 | 8253656 | 9.5797 | 26.583 | 4787117 | 58.813 |
| 22 | Azerbaijan | 2005 | 18607.54 | 7222.036 | 8538610 | 4.0155 | 4.872 | 685682 | 52.389 |
| 25 | Banglades | 2005 | 1906.654 | 2279.531 | 1.39E+08 | 0.271 | 1.067 | 346583 | 26.809 |
| 28 | Belgium | 2005 | 64611.81 | 46341.29 | 10546885 | 11.9126 | 17.767 | 5887272 | 97.403 |

# Preliminary analysis - Normalization

Per capita perspective

- Co2percap = unnormalized$co2percap,
- EnergyUse = normalized$EnergyUse * (1-unnormalized_data$Lowcarbon_energy/100),
- GDP = normalized$GDP,
- Pop = normalized$pop,
- Internet = unnormalized$internet/unnormalized$pop,
- Urb = unnormalized$urb/100

# Preliminary analysis

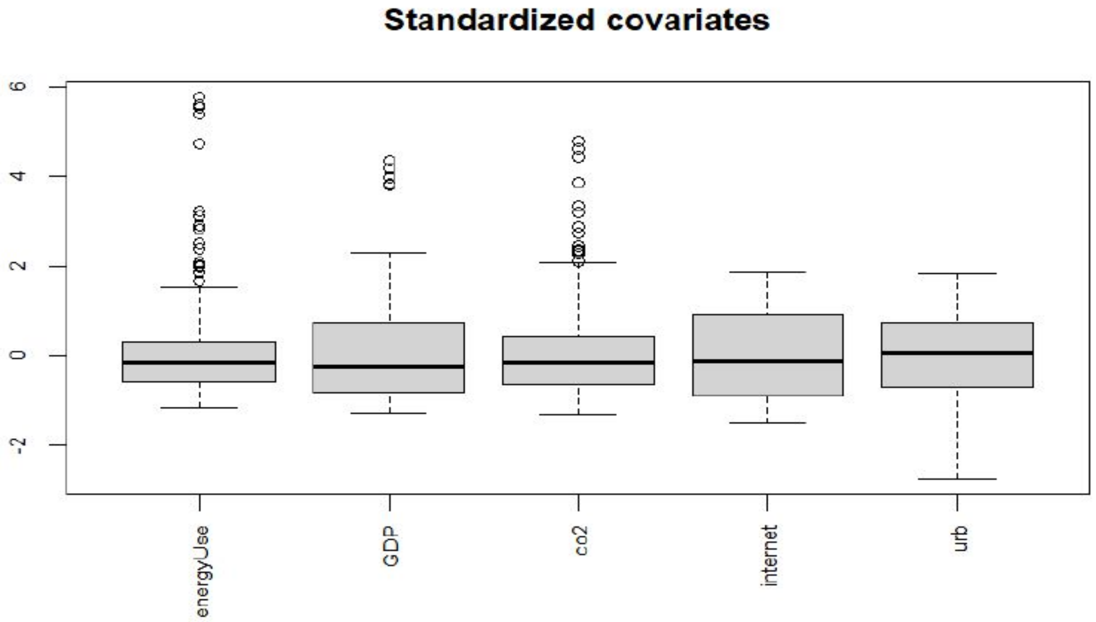
Section 2

# Preliminary analysis - Normalization

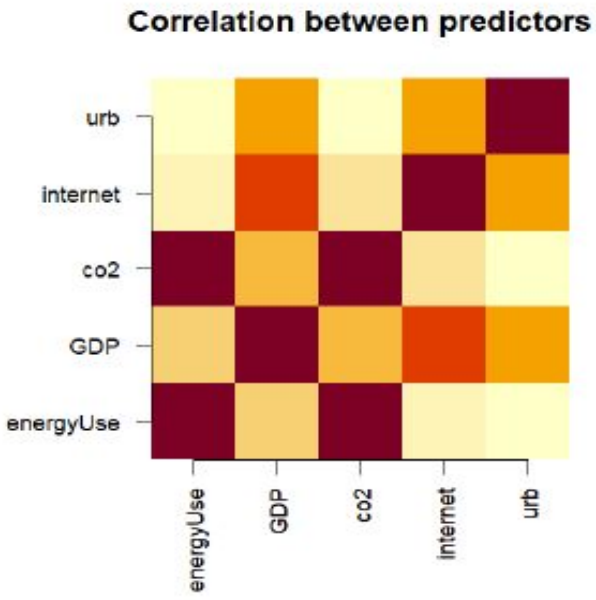
High difference in scale and distribution

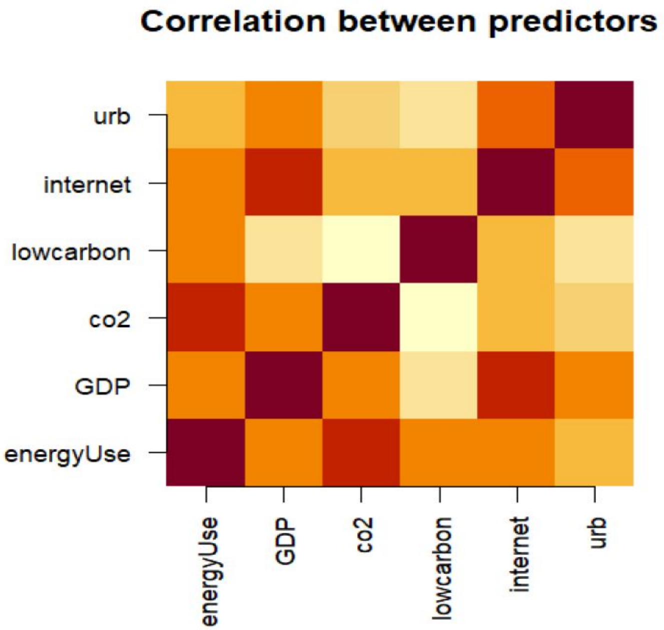
ZScore
+
MinMax

# Preliminary analysis - Covariates

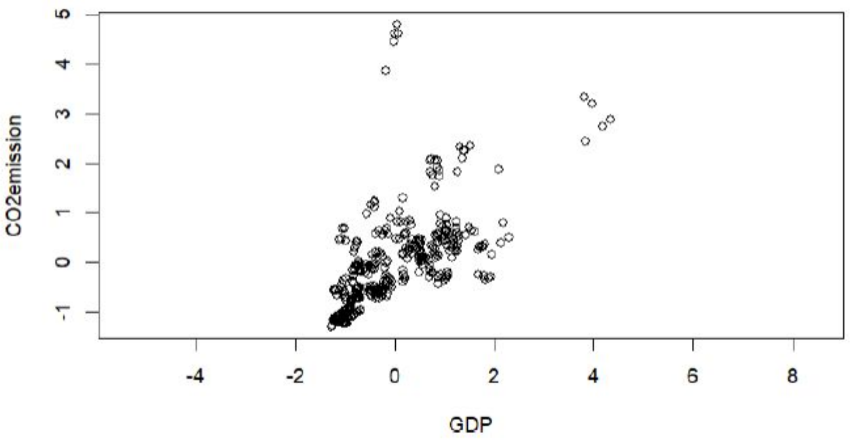


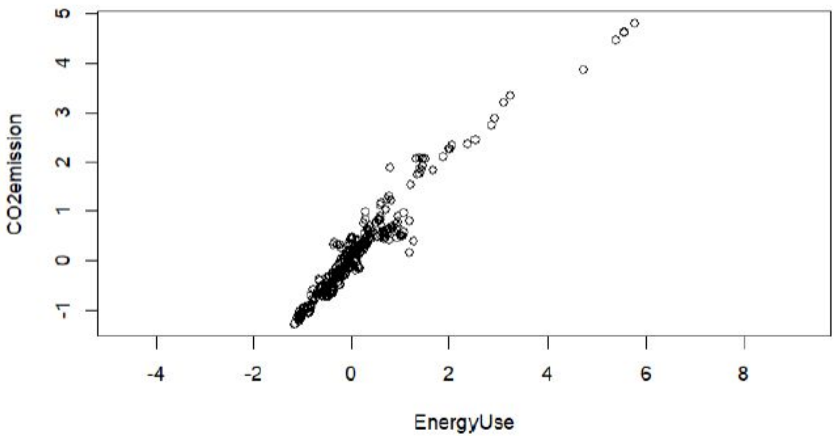
# Preliminary analysis - Correlation matrix



# Preliminary analysis - GDP & EnergyUse



# BAS analysis and feature selection

Section 3

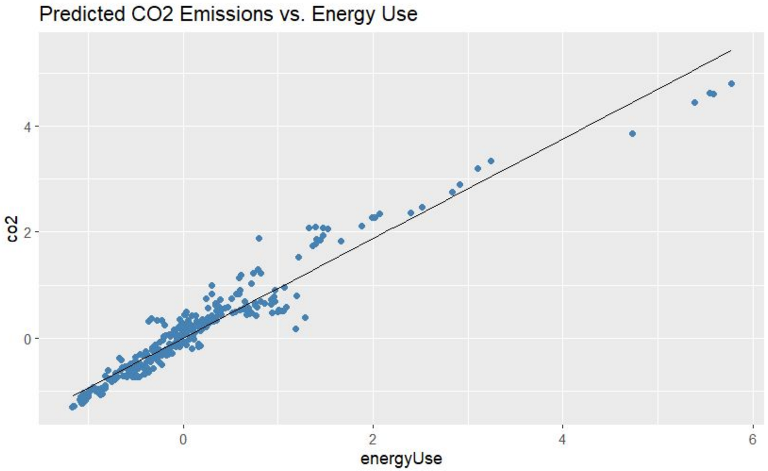
- Before considering more complex models
- Split in train and test

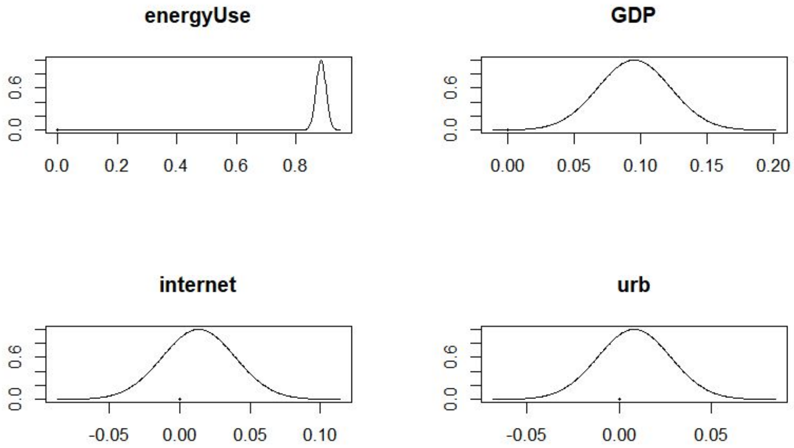
**prediction using only energyUse**

```
g-prior | alpha = n
[1] "Train Mean sum of squared error
is: 0.0666247331734019"
[1] "Test Mean sum of squared error is:
0.0591558419947747"
```

**prediction using all the covariates**

```
g-prior | alpha = n
[1] "Train Mean sum of squared error
is: 0.0576266600005661"
[1] "Test Mean sum of squared error
is: 0.0543753086310353"
```



## Jeffreys-Zellner-Siow (JZS) priors

```
prior="JZS" | alpha=1

Showing results only for HPM

[1] "Intercept" "energyUse" "GDP"

[1] "Train Mean sum of squared error is:
0.0577376077419626"

[1] "Test Mean sum of squared error is:
0.0541967365249909"
```

- Features selection
- Showing results only for HPM
- we decided not to use a BMA since we are interested in the best model and not in the average of the models

| | P(B != 0 \| Y) | model 1 | model 2 | model 3 | model 4 | model 5 |
|---|---|---|---|---|---|---|
| Intercept | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| energyUse | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GDP | 0.998 | 1.000 | 1.000 | 1.000 | 0.000 | 1.000 |
| internet | 0.029 | 0.000 | 1.000 | 0.000 | 1.000 | 1.000 |
| urb | 0.026 | 0.000 | 0.000 | 1.000 | 0.000 | 1.000 |
| BF | NA | 1.000 | 0.028 | 0.027 | 0.002 | 0.001 |
| PostProbs | NA | 0.946 | 0.026 | 0.025 | 0.002 | 0.001 |
| R2 | NA | 0.943 | 0.943 | 0.943 | 0.940 | 0.943 |
| dim | NA | 3.000 | 4.000 | 4.000 | 3.000 | 5.000 |
| logmarg | NA | 405.047 | 401.471 | 401.423 | 398.607 | 397.930 |

```
Marginal Posterior Summaries of Coefficients:

Using  HPM

Based on the top  1 models
```

| | post mean | post SD | post p(B != 0) |
|---|---|---|---|
| Intercept | 0.02865 | 0.01416 | 1.00000 |
| energyUse | 0.88544 | 0.01624 | 1.00000 |
| GDP | 0.11161 | 0.01678 | 0.99841 |
| internet | 0.00000 | 0.00000 | 0.02884 |
| urb | 0.00000 | 0.00000 | 0.02607 |

**Adding more covariates**

- non-linear relationships
- GDP^2 and energyUse^2
- Feature selection

## Jeffreys-Zellner-Siow (JZS) priors

```
prior="JZS" | alpha=1

Showing results only for HPM

[1] "Intercept" "energyUse" "GDP"

[1] "Train Mean sum of squared error is:
0.0577376077419626"
[1] "Test Mean sum of squared error is:
0.0541967365249909"
```

| | P(B != 0 \| Y) | model 1 | model 2 | model 3 | model 4 | model 5 |
|---|---|---|---|---|---|---|
| Intercept | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| energyUse | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GDP | 0.438 | 0.000 | 1.000 | 0.000 | 1.000 | 0.000 |
| internet | 0.045 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 |
| urb | 0.037 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 |
| energyUse2 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GDP2 | 0.035 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 |
| BF | NA | 1.000 | 0.781 | 0.060 | 0.045 | 0.040 |
| PostProbs | NA | 0.499 | 0.390 | 0.030 | 0.022 | 0.020 |
| R2 | NA | 0.948 | 0.949 | 0.948 | 0.949 | 0.948 |
| dim | NA | 3.000 | 4.000 | 4.000 | 5.000 | 4.000 |
| logmarg | NA | 417.241 | 416.993 | 414.426 | 414.138 | 414.021 |

```
 Marginal Posterior Summaries of Coefficients:

 Using  HPM

 Based on the top  1 models
```

| | post mean | post SD | post p(B != 0) |
|---|---|---|---|
| Intercept | 0.028655 | 0.013570 | 1.000000 |
| energyUse | 1.078736 | 0.020586 | 1.000000 |
| GDP | 0.000000 | 0.000000 | 0.437515 |
| internet | 0.000000 | 0.000000 | 0.044818 |
| urb | 0.000000 | 0.000000 | 0.036636 |
| energyUse2 | -0.046170 | 0.005378 | 0.999997 |
| GDP2 | 0.000000 | 0.000000 | 0.035117 |

```
[1] "Intercept"  "energyUse"  "energyUse2"
[1] "Train Mean sum of squared error is: 0.0530340808788469"
[1] "Test Mean sum of squared error is: 0.0431591495882435"
```

**Conclusion of regression with BAS**

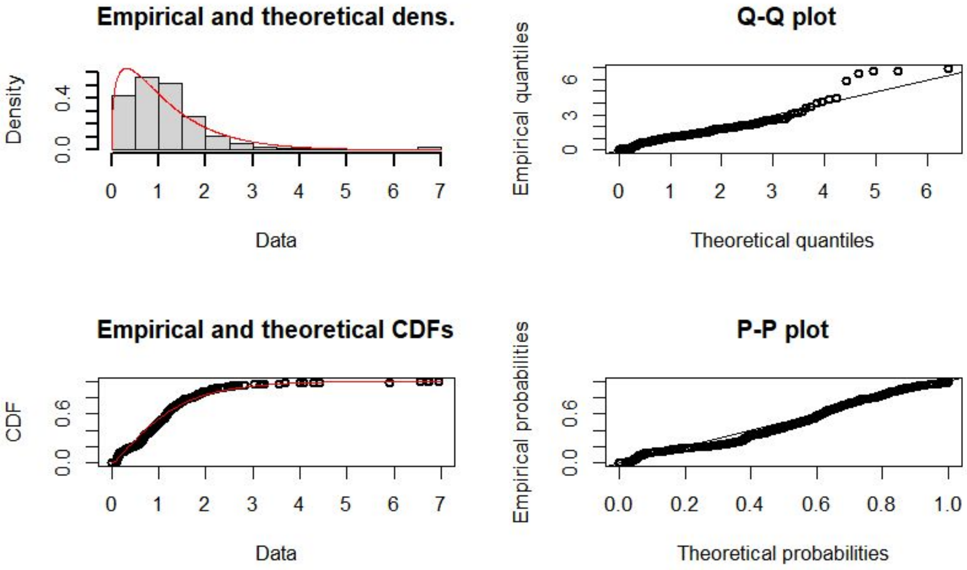
- If we avoid adding more covariates, the model with the JZS prior has the best performance and among the selected the one with only energyUse and GDP as covariates is the best. It has a lower BIC compared to the other subset of models. The model doesn't overfit on the test set and has a good performance. Lastly it shows a strong relashionship between co2 and GDP since the posterior probability is close to 1.
- If instead we add more covariates we can see that the model with the squared of energyUse and energy has a lower BIC compared to the other subset of models and get better overall performances. In this case the importance of GDP with co2 decays.

# Models (JAGS)

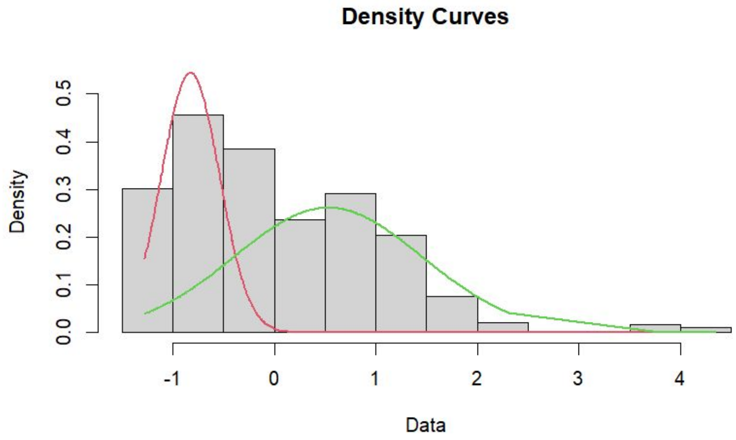
Section 4.0 - 4.1

- This part was only a trial
- Study in deep the distribution of the covariates
- Add  this information into the model to see if the prediction improves
- Get the distribution shape (keeping in mind variables have been standardized and it doesn't reflect the true distribution)

**EnergyUse**




**GDP**



## JAGS model considering covariate distribution energyUse

$$
\begin{aligned}
y_i &\sim \mathcal{N}(\beta_o + X_i\beta, \sigma_y^2)\\
X_1 &\sim \mathcal{G}(alphaEnergy, betaEnergy)\\
\beta_0 &\sim \mathcal{N}(0, 100)\\
\beta_i &\sim \mathcal{N}(0, 100)\\
\sigma_y^{-2} &\sim \mathcal{G}(0.01, 0.01)\\
alphaEnergy &\sim \mathcal{G}(1, 1)\\
betaEnergy &\sim \mathcal{G}(1, 1)
\end{aligned}
$$

| | Mean | SD | Naive SE | Time-series SE |
|---|---|---|---|---|
| alphaEnergy | 1.35769 | 0.091352 | 1.055e-03 | 2.081e-03 |
| beta[1] | 0.90722 | 0.015266 | 1.763e-04 | 2.942e-04 |
| beta[2] | 0.08923 | 0.024638 | 2.845e-04 | 5.072e-04 |
| beta[3] | 0.01262 | 0.023156 | 2.674e-04 | 4.540e-04 |
| beta[4] | 0.01061 | 0.017120 | 1.977e-04 | 2.286e-04 |
| beta0 | -1.05792 | 0.021904 | 2.529e-04 | 4.091e-04 |
| betaEnergy | 1.16403 | 0.093768 | 1.083e-03 | 2.114e-03 |
| var.y | 0.05773 | 0.004384 | 5.062e-05 | 5.093e-05 |

2. Quantiles for each variable:

| | 2.5% | 25% | 50% | 75% | 97.5% |
|---|---|---|---|---|---|
| alphaEnergy | 1.18667 | 1.293468 | 1.35418 | 1.41789 | 1.54281 |
| beta[1] | 0.87681 | 0.897028 | 0.90730 | 0.91765 | 0.93662 |
| beta[2] | 0.04128 | 0.072276 | 0.08916 | 0.10571 | 0.13721 |
| beta[3] | -0.03205 | -0.002955 | 0.01253 | 0.02858 | 0.05688 |
| beta[4] | -0.02246 | -0.001153 | 0.01080 | 0.02199 | 0.04451 |
| beta0 | -1.10080 | -1.072313 | -1.05802 | -1.04332 | -1.01375 |
| betaEnergy | 0.98633 | 1.100396 | 1.15981 | 1.22482 | 1.35676 |
| var.y | 0.04979 | 0.054642 | 0.05755 | 0.06053 | 0.06698 |

## JAGS model considering covariate distribution energyUse and GDP

Likelihood:

$y_i \sim N(\mu_i, \sigma_y^2)$,

$\mu_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$,

$x_{i1} \sim \mathrm{Gamma}(\alpha_{\mathrm{Energy}}, \beta_{\mathrm{Energy}})$, for energyUse covariate

$z_i \sim \mathrm{Bernoulli}(p_{\mathrm{GDP}})$,

$x_{i2} \sim N(\mu_1 z_i + \mu_2(1 - z_i), \tau_3 z_i + \tau_4(1 - z_i))$, for GDP cova

Priors:

$\beta_j \sim N(0, 100)$, for $j = 1, \ldots, p$

$\beta_0 \sim N(0, 100)$,

$\sigma_y^{-2} \sim \mathrm{Gamma}(0.01, 0.01)$, where $\sigma_y^2 = \frac{1}{\sigma_y^{-2}}$,

$\alpha_{\mathrm{Energy}} = 2$,

$\beta_{\mathrm{Energy}} = 2$,

$\mu_1 \sim N(-0.8, 100)$,

$\mu_2 \sim N(0.5, 100)$,

$\tau_3 \sim \mathrm{Gamma}(0.01, 0.01)$,

$\tau_4 \sim \mathrm{Gamma}(0.01, 0.01)$,

$p_{\mathrm{GDP}} \sim \mathrm{Beta}(1, 1)$.

| | Mean | SD | Naive SE | Time-series SE |
|---|---|---|---|---|
| beta[1] | 0.90737 | 0.015500 | 0.0001790 | 0.0003201 |
| beta[2] | 0.08868 | 0.024702 | 0.0002852 | 0.0005059 |
| beta[3] | 0.01345 | 0.023294 | 0.0002690 | 0.0004602 |
| beta[4] | 0.01030 | 0.017080 | 0.0001972 | 0.0002306 |
| beta0 | -1.05835 | 0.022125 | 0.0002555 | 0.0004334 |
| mu1 | -0.83426 | 0.051364 | 0.0005931 | 0.0019826 |
| mu2 | 0.51722 | 0.101250 | 0.0011691 | 0.0031127 |
| pGDP | 0.38227 | 0.052767 | 0.0006093 | 0.0019452 |
| var.y | 0.05769 | 0.004278 | 0.0000494 | 0.0000493 |

2. Quantiles for each variable:

| | 2.5% | 25% | 50% | 75% | 97.5% |
|---|---|---|---|---|---|
| beta[1] | 0.87672 | 0.896838 | 0.90732 | 0.91766 | 0.93752 |
| beta[2] | 0.04132 | 0.071764 | 0.08839 | 0.10520 | 0.13775 |
| beta[3] | -0.03286 | -0.002063 | 0.01378 | 0.02894 | 0.05911 |
| beta[4] | -0.02324 | -0.001317 | 0.01024 | 0.02186 | 0.04405 |
| beta0 | -1.10182 | -1.073324 | -1.05842 | -1.04350 | -1.01585 |
| mu1 | -0.94126 | -0.868485 | -0.83167 | -0.79826 | -0.73958 |
| mu2 | 0.32106 | 0.448521 | 0.51691 | 0.58484 | 0.71749 |
| pGDP | 0.27716 | 0.345979 | 0.38372 | 0.42015 | 0.48108 |
| var.y | 0.04990 | 0.054692 | 0.05748 | 0.06043 | 0.06678 |

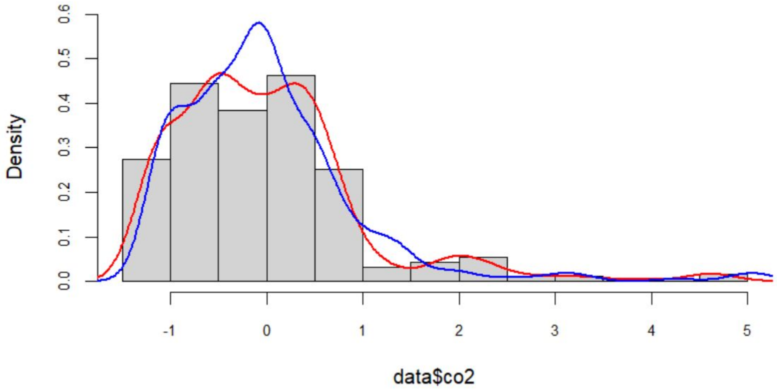
## JAGS model no threshold on GDP

$$y_i \sim \mathcal{N}(\beta_o + X_i\beta, \sigma_y^2)$$
$$\beta_0 \sim \mathcal{N}(100)$$
$$\beta_i \sim \mathcal{N}(0, 100)$$
$$\sigma_y^{-2} \sim \mathcal{G}(0.01, 0.01)$$

| | Mean | SD | Naive SE | Time-series SE |
|---|---|---|---|---|
| beta[1] | 0.90697 | 0.015372 | 1.775e-04 | 3.116e-04 |
| beta[2] | 0.08965 | 0.024734 | 2.856e-04 | 5.249e-04 |
| beta[3] | 0.01254 | 0.023072 | 2.664e-04 | 4.495e-04 |
| beta[4] | 0.01069 | 0.016864 | 1.947e-04 | 2.284e-04 |
| beta0 | -1.05752 | 0.021857 | 2.524e-04 | 4.280e-04 |
| var.y | 0.05763 | 0.004394 | 5.074e-05 | 5.074e-05 |

2. Quantiles for each variable:

| | 2.5% | 25% | 50% | 75% | 97.5% |
|---|---|---|---|---|---|
| beta[1] | 0.87713 | 0.8966404 | 0.90687 | 0.91714 | 0.93736 |
| beta[2] | 0.04167 | 0.0729711 | 0.08969 | 0.10657 | 0.13861 |
| beta[3] | -0.03235 | -0.0029236 | 0.01244 | 0.02774 | 0.05806 |
| beta[4] | -0.02303 | -0.0006827 | 0.01049 | 0.02227 | 0.04367 |
| beta0 | -1.10026 | -1.0717167 | -1.05756 | -1.04309 | -1.01424 |
| var.y | 0.04962 | 0.0545871 | 0.05744 | 0.06042 | 0.06692 |




- The prediction of previous models doesn't improve but it allowed us to extract the parameters of GDP and energyUse

# Bayesian LASSO prior

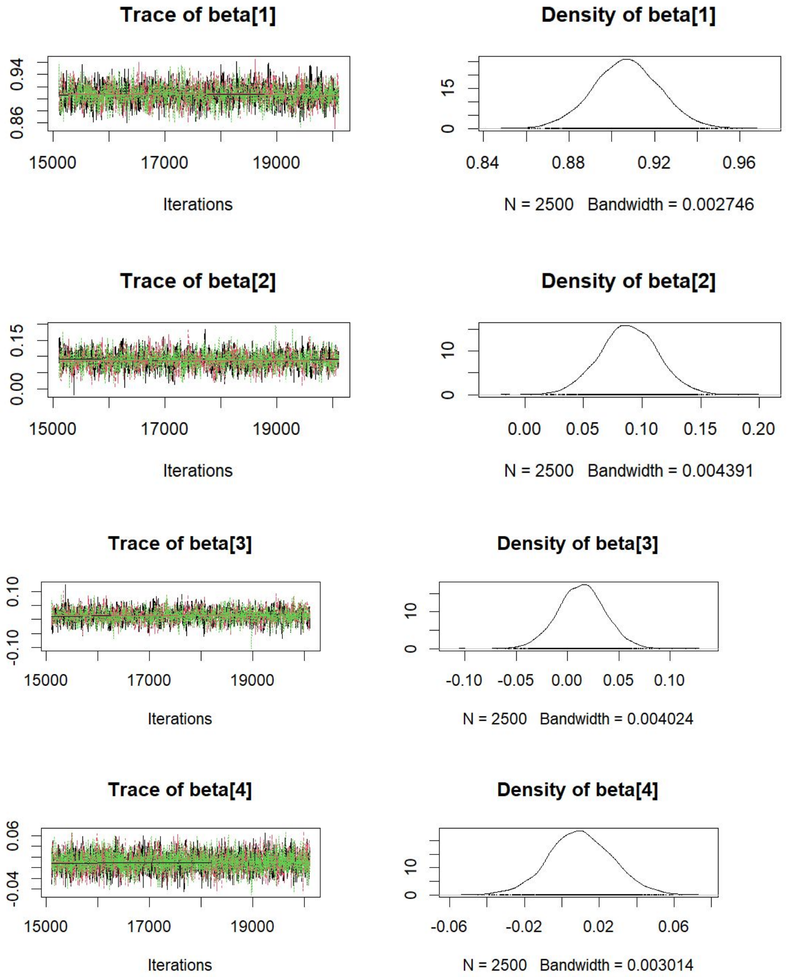
We use now lasso regression to select again the most important covariates in the model, and see if the model without some covariates can improve without degrading too much.

$$
\begin{aligned}
y_i &\sim N(\beta_0 + X_i\beta, \sigma_y^2)\\
\beta_0 &\sim N(0, 100)\\
\beta_i &\sim \mathrm{DE}(0, \sigma_b^2\sigma^2)\\
\sigma_y^2 &\sim \mathrm{IG}(0.01, 0.01)\\
\sigma_b^2 &\sim \mathrm{IG}(0.01, 0.01)
\end{aligned}
$$

```
           Mean      SD  Naive SE Time-series SE
beta[1]  0.90688 0.01543 0.0001782      0.0003694
beta[2]  0.08862 0.02468 0.0002849      0.0005653
beta[3]  0.01353 0.02295 0.0002650      0.0005160
beta[4]  0.01014 0.01694 0.0001956      0.0002715
beta0   -1.05745 0.02201 0.0002541      0.0004957

2. Quantiles for each variable:

            2.5%       25%       50%      75%    97.5%
beta[1]  0.87623  0.896427  0.906878  0.91717  0.93735
beta[2]  0.03993  0.072306  0.088606  0.10543  0.13740
beta[3] -0.03169 -0.001674  0.013657  0.02863  0.05913
beta[4] -0.02284 -0.001261  0.009799  0.02150  0.04427
beta0   -1.10094 -1.072499 -1.057443 -1.04231 -1.01483
```



# Threshold

Section 4.2.1

- For this analysis different models have been deployed

## Find the threshold in GDP considering other variables

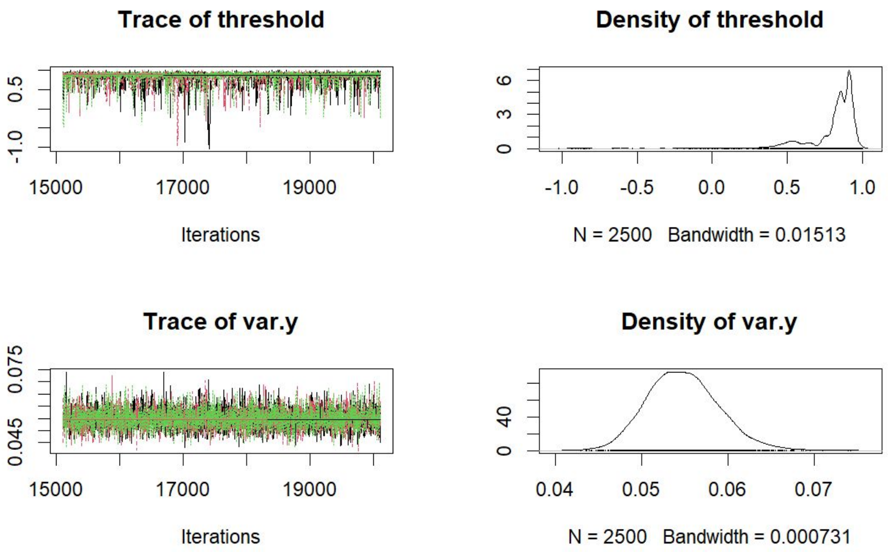
The model has a form:

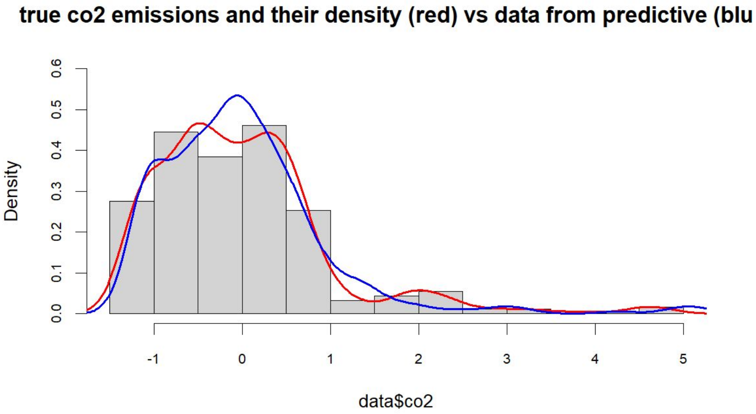
$$
\begin{aligned}
& y_i \sim \mathcal{N}(\mu_i, \sigma_y^2) \\
& \mu_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j + betapoor \cdot \mathrm{GDP}_i \cdot I(\mathrm{GDP}_i < \mathrm{threshold}) + betarich \cdot \mathrm{GDP}_i \cdot I(\mathrm{GDP}_i >= \mathrm{threshold}) \\
& \beta_0 \sim \mathcal{N}(100) \\
& \beta_i \sim \mathcal{N}(0, 100) \\
& betapoor \sim \mathcal{N}(0, 100) \\
& betarich \sim \mathcal{N}(0, 100) \\
& \sigma_y^{-2} \sim \mathcal{G}(0.01, 0.01) \\
& \mathrm{threshold} \sim \mathcal{U}(-2, 1)
\end{aligned}
$$

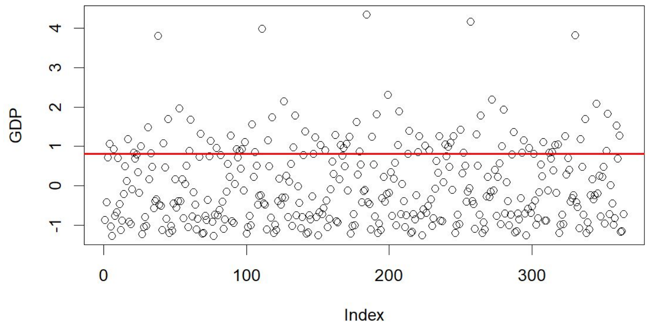
| | Mean | SD | Naive SE | Time-series SE |
|---|---|---|---|---|
| beta[1] | 0.89524 | 0.015315 | 1.768e-04 | 3.484e-04 |
| beta[2] | -0.03050 | 0.025109 | 2.899e-04 | 5.835e-04 |
| beta[3] | -0.00564 | 0.017090 | 1.973e-04 | 2.486e-04 |
| beta0 | -0.99153 | 0.028691 | 3.313e-04 | 7.761e-04 |
| beta poor | 0.23855 | 0.044293 | 5.114e-04 | 1.211e-03 |
| beta rich | 0.07283 | 0.024845 | 2.869e-04 | 4.626e-04 |
| threshold | 0.80185 | 0.194603 | 2.247e-03 | 5.389e-03 |
| var.y | 0.05478 | 0.004112 | 4.748e-05 | 5.029e-05 |

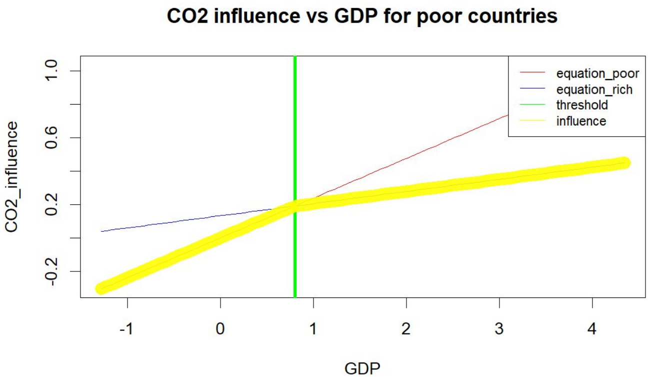
2. Quantiles for each variable:

| | 2.5% | 25% | 50% | 75% | 97.5% |
|---|---|---|---|---|---|
| beta[1] | 0.86507 | 0.88506 | 0.895469 | 0.905597 | 0.92454 |
| beta[2] | -0.07941 | -0.04786 | -0.030390 | -0.013649 | 0.01849 |
| beta[3] | -0.03925 | -0.01729 | -0.005745 | 0.006217 | 0.02823 |
| beta0 | -1.04490 | -1.01105 | -0.993111 | -0.973394 | -0.93136 |
| beta poor | 0.15322 | 0.20885 | 0.237844 | 0.267334 | 0.32775 |
| beta rich | 0.02456 | 0.05594 | 0.073168 | 0.089727 | 0.12164 |
| threshold | 0.32895 | 0.79624 | 0.859569 | 0.910171 | 0.95570 |
| var.y | 0.04734 | 0.05190 | 0.054606 | 0.057402 | 0.06348 |






- The coefficient beta_poor for poor countries is higher than the coefficient beta_rich for rich countries. This means that the relationship between $CO_2$ emissions and GDP is stronger for poor countries than for rich countries.
- However because of all the parameters that are free we think this model is too general and the threshold refers for countries with very high GDP (40000 )
- In this we use a single intercept: the assumption is the baseline $CO_2$ emissions level (intercept) is the same across all GDP levels, but the rate of change (slope) with respect to GDP changes at the threshold. Instead all the intercepts are summarized in the first parameter

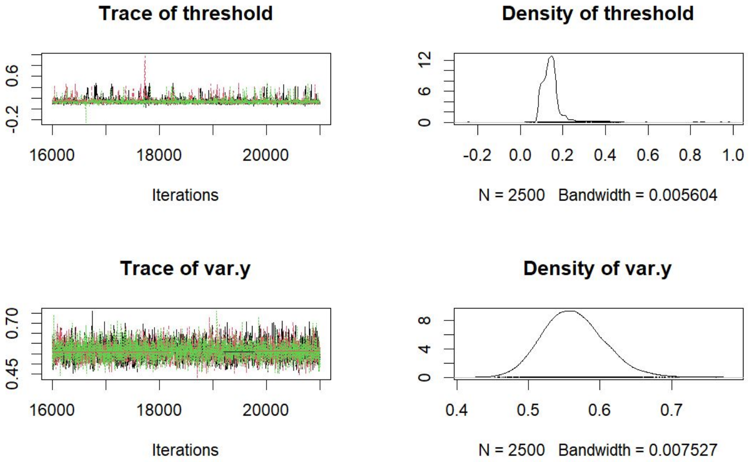
## Find the threshold in GDP considering only GDP

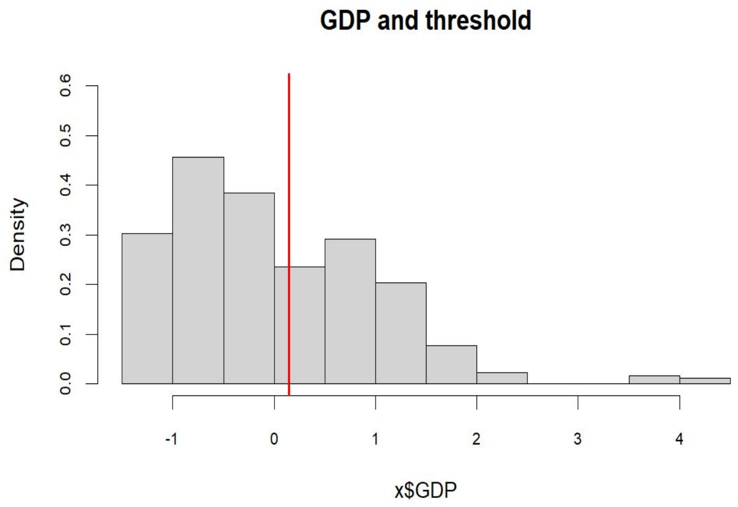
$$
\begin{aligned}
&y_i \sim \mathcal{N}(\mu_i, \sigma_y^2)\\
&\mu_i = (betapoor \cdot \mathrm{GDP}_i + betapoor_0) \cdot I(\mathrm{GDP}_i < \mathrm{threshold}) + (betarich \cdot \mathrm{GDP}_i + betarich_0) \cdot I(\mathrm{GDP}_i >= \mathrm{threshold})\\
&betapoor \sim \mathcal{N}(0, 100)\\
&betarich \sim \mathcal{N}(0, 100)\\
&betapoor_0 \sim \mathcal{N}(0, 100)\\
&betarich_0 \sim \mathcal{N}(0, 100)\\
&\sigma_y^{-2} \sim \mathcal{G}(0.01, 0.01)\\
&\mathrm{threshold} \sim \mathcal{U}(-1.5, 1)
\end{aligned}
$$

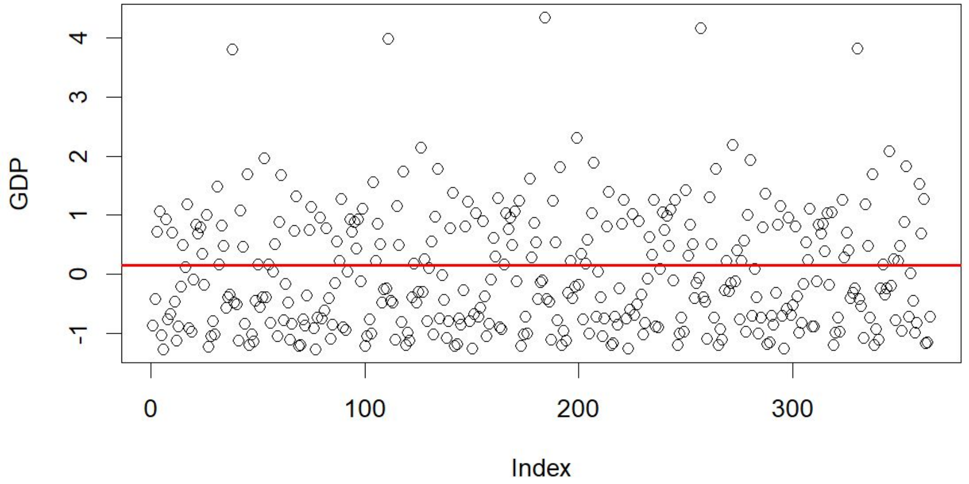
| | Mean | SD | Naive SE | Time-series SE |
|---|---|---|---|---|
| beta_poor | 1.41552 | 0.14797 | 0.0017086 | 0.003881 |
| beta_poor_0 | 0.58793 | 0.11624 | 0.0013422 | 0.002984 |
| beta_rich | 0.53322 | 0.08458 | 0.0009766 | 0.001601 |
| beta_rich_0 | 0.01984 | 0.10842 | 0.0012519 | 0.002093 |
| threshold | 0.14651 | 0.05753 | 0.0006643 | 0.001518 |
| var.y | 0.56299 | 0.04235 | 0.0004890 | 0.000489 |

2. Quantiles for each variable:

| | 2.5% | 25% | 50% | 75% | 97.5% |
|---|---|---|---|---|---|
| beta_poor | 1.11997 | 1.31635 | 1.41820 | 1.5160 | 1.6991 |
| beta_poor_0 | 0.35660 | 0.51157 | 0.59076 | 0.6674 | 0.8118 |
| beta_rich | 0.36989 | 0.47665 | 0.53184 | 0.5893 | 0.7027 |
| beta_rich_0 | -0.20049 | -0.04954 | 0.02132 | 0.0933 | 0.2267 |
| threshold | 0.08693 | 0.11585 | 0.13877 | 0.1581 | 0.3309 |
| var.y | 0.48722 | 0.53313 | 0.56090 | 0.5898 | 0.6537 |





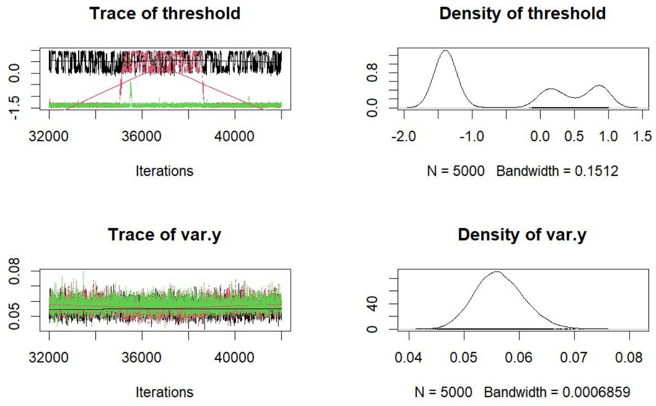

- This model is more robust threshold at 30000 (standardized variables)

# Threshold with different intercepts - Unstable

$$y_i \sim \mathcal{N}(\mu_i, \sigma_y^2)$$

$$\mu_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j + (betapoor \cdot \mathrm{GDP}_i + betapoor_0) \cdot I(\mathrm{GDP}_i < \mathrm{threshold}) + (betarich \cdot \mathrm{GDP}_i + betarich_0) \cdot I(\mathrm{GDP}_i >= \mathrm{threshold})$$

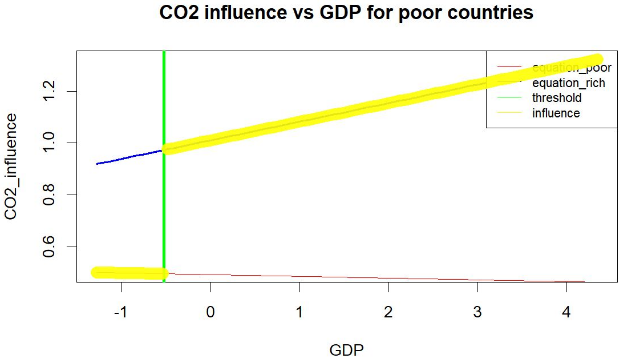
$$\beta_0 \sim \mathcal{N}(100)$$
$$\beta_i \sim \mathcal{N}(0, 100)$$
$$betapoor \sim \mathcal{N}(0, 100)$$
$$betarich \sim \mathcal{N}(0, 100)$$
$$betapoor_0 \sim \mathcal{N}(0, 100)$$
$$betarich_0 \sim \mathcal{N}(0, 100)$$
$$\sigma_y^{-2} \sim \mathcal{G}(0.01, 0.01)$$
$$\mathrm{threshold} \sim \mathcal{U}(-1.5, 1)$$

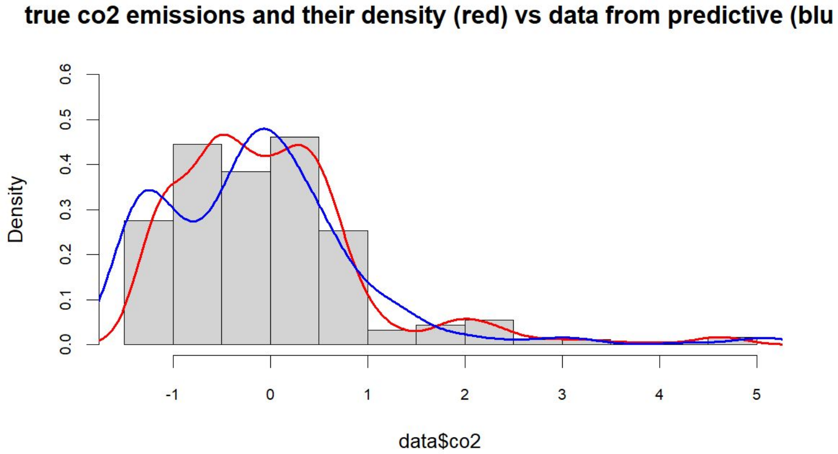


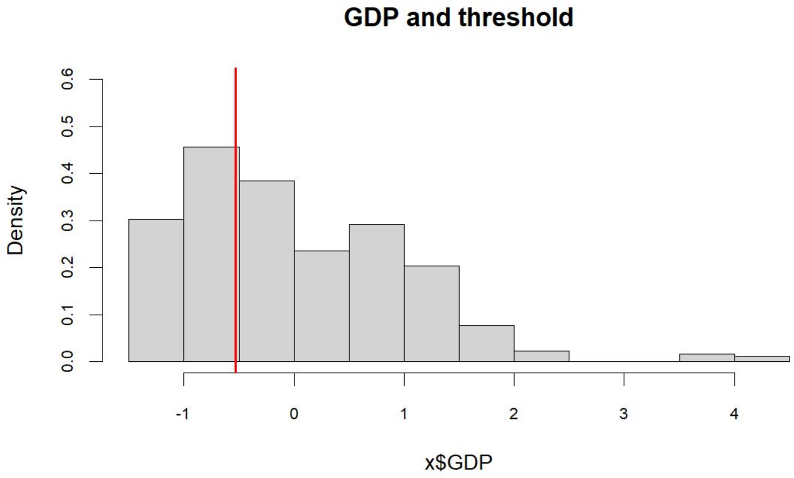

| | Mean | SD | Naive SE | Time-series SE |
|---|---|---|---|---|
| beta[1] | 0.902602 | 0.016226 | 1.325e-04 | 3.037e-04 |
| beta[2] | -0.008552 | 0.034486 | 2.816e-04 | 2.376e-03 |
| beta[3] | 0.003648 | 0.018780 | 1.533e-04 | 5.034e-04 |
| beta0 | -2.017844 | 1.212530 | 9.900e-03 | 1.380e-01 |
| beta_poor | -0.006683 | 7.514279 | 6.135e-02 | 6.063e-02 |
| beta_poor_0 | 0.492876 | 7.327803 | 5.983e-02 | 1.181e-01 |
| beta_rich | 0.071660 | 0.038503 | 3.144e-04 | 1.310e-03 |
| beta_rich_0 | 1.010556 | 1.218549 | 9.949e-03 | 1.245e-01 |
| threshold | -0.528305 | 0.975763 | 7.967e-03 | 1.471e-01 |
| var.y | 0.056470 | 0.004435 | 3.621e-05 | 7.808e-05 |

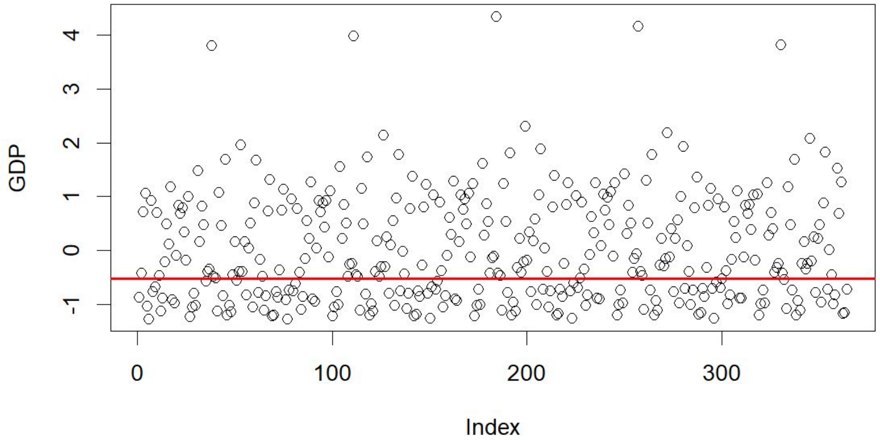
2. Quantiles for each variable:

| | 2.5% | 25% | 50% | 75% | 97.5% |
|---|---|---|---|---|---|
| beta[1] | 0.87124 | 0.891503 | 0.902701 | 0.91368 | 0.93448 |
| beta[2] | -0.07634 | -0.033781 | -0.005914 | 0.01701 | 0.05251 |
| beta[3] | -0.03301 | -0.009233 | 0.003680 | 0.01625 | 0.04005 |
| beta0 | -4.51276 | -3.152240 | -1.477989 | -1.10852 | -0.35529 |
| beta_poor | -17.40161 | -1.182368 | 0.222070 | 0.85561 | 17.11251 |
| beta_poor_0 | -16.75105 | -0.877324 | 0.581135 | 2.68696 | 16.51105 |
| beta_rich | -0.01222 | 0.047735 | 0.076237 | 0.09879 | 0.13790 |
| beta_rich_0 | -0.70674 | 0.099705 | 0.499394 | 2.13322 | 3.45284 |
| threshold | -1.48994 | -1.400170 | -1.296270 | 0.40722 | 0.92799 |
| var.y | 0.04834 | 0.053384 | 0.056238 | 0.05932 | 0.06562 |


CO2 influence vs GDP for poor countries
equation_poor
equation_rich
threshold
influence
CO2_influence
0.6
0.8
1.0
1.2
-1
0
1
2
3
4
GDP


true co2 emissions and their density (red) vs data from predictive (blu
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
-1
0
1
2
3
4
5
data$co2


GDP and threshold
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
-1
0
1
2
3
4
x$GDP


GDP
-1
0
1
2
3
4
0
100
200
300
Index

# Section 4.2.1: Threshold

To answer the thesis of a threshold in the GDP above which the GDP does not influence the CO2 production we start by putting a **uniform prior** on the threshold.

Define the linear predictor, only for $\mu$ :
$\mu[i] = \langle X[i,], \beta \rangle$

Normal model:
$\text{co2percap}[i] \sim \mathcal{N}(\mu[i], \beta_\gamma)$

Indicator variable:
$\text{indicator}[i] = (\text{unGDP}[i] \geq \text{threshold})$

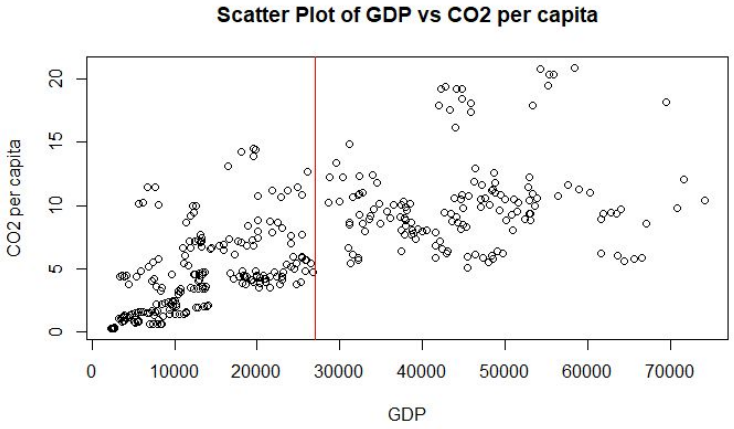
Covariate matrix $X[i,]$ :
$X[i,1] = (1 - \text{indicator}[i])$
$X[i,2] = \text{indicator}[i]$
$X[i,3] = \text{GDP}[i] \cdot (1 - \text{indicator}[i])$
$X[i,4] = \text{GDP}[i] \cdot \text{indicator}[i]$

Prior for the variance $\beta_\gamma$ :
$\beta_\gamma \sim \text{Gamma}(0.001, 0.001)$

Uninformative priors for $\beta$ parameters:
$\beta[1] \sim \mathcal{N}(0, 0.01)$
$\beta[2] \sim \mathcal{N}(0, 0.01)$
$\beta[3] \sim \mathcal{N}(0, 0.01)$

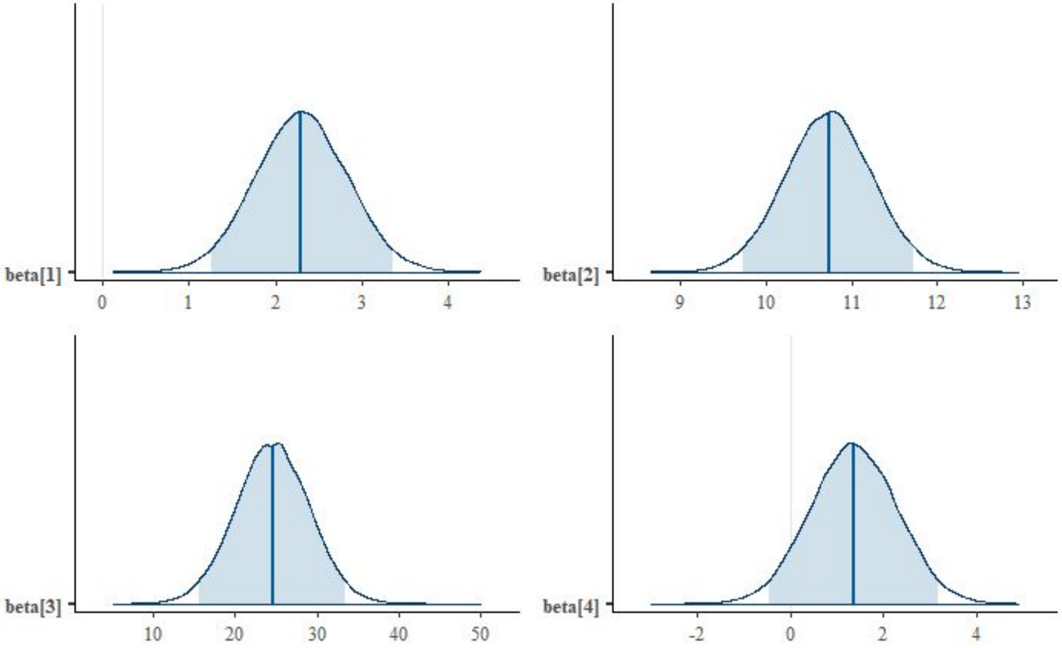
Informative prior:
$\beta[4] \sim \mathcal{N}(0, 1)$

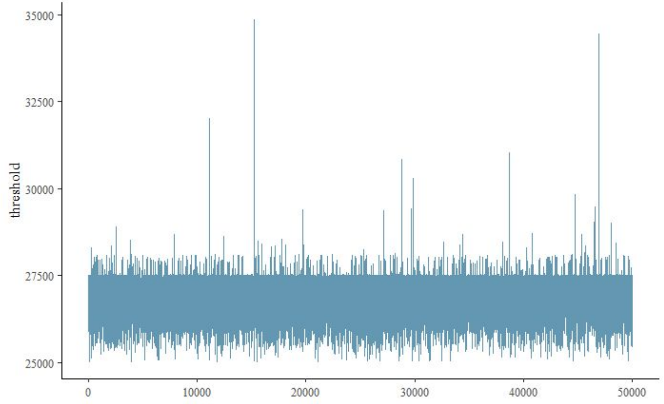
Prior for threshold:
$\text{threshold} \sim \text{Uniform}(25000, 35000)$

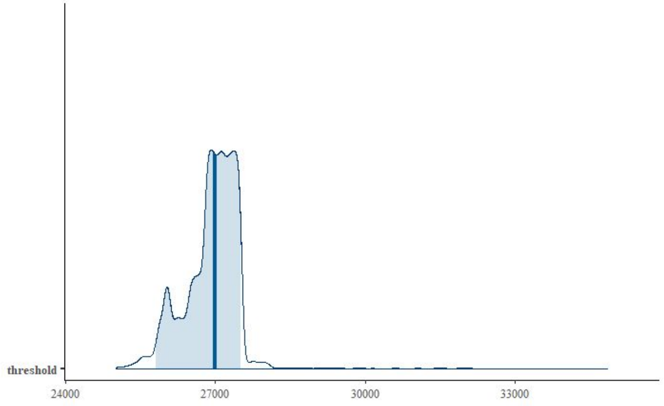



The threshold has been found around **27000**. This will have to be confirmed by a **broader** model.

# Section 4.2.1: Threshold




Threshold 97.5% credible interval: 25800-27510

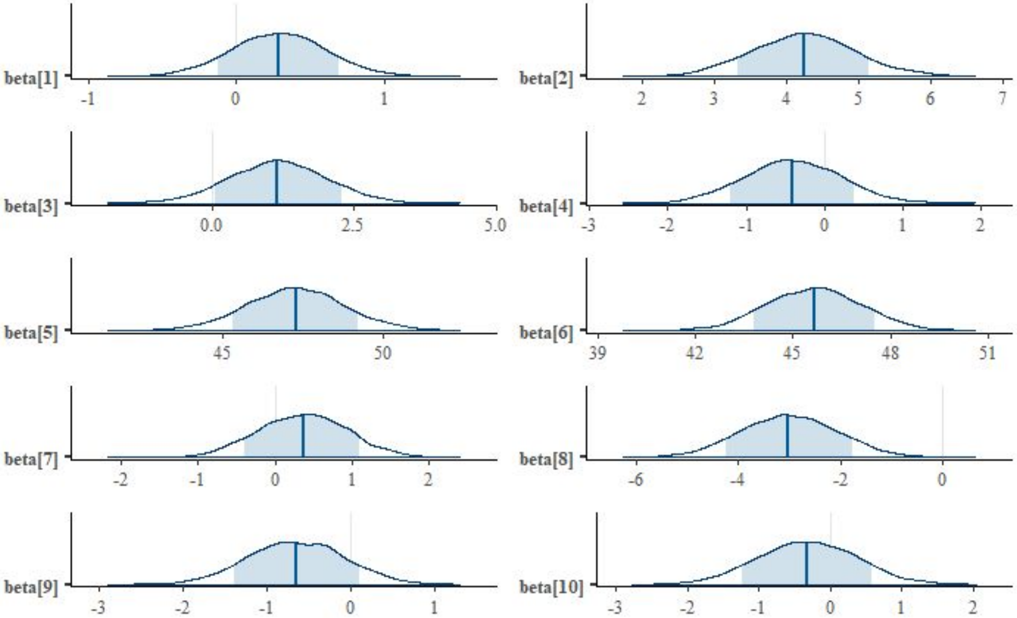
# Normal model with threshold

Section 4.2.2

# Section 4.2.2: Normal model with threshold

We now add all more relevant covariates in a Normal model, keeping a flat prior, to find out if our thesis hold.

$$
\begin{aligned}
\mu_i &\leftarrow \text{inprod}(X_i, \beta) \\
\text{co2percap}_i &\sim \mathcal{N}(\mu_i, \beta_\gamma) \\
\text{indicator}_i &\leftarrow (\text{unGDP}_i \geq \text{threshold}) \\
X_{i,1} &\leftarrow (1 - \text{indicator}_i) \\
X_{i,2} &\leftarrow \text{indicator}_i \\
X_{i,3} &\leftarrow \text{GDP}_i \cdot (1 - \text{indicator}_i) \\
X_{i,4} &\leftarrow \text{GDP}_i \cdot \text{indicator}_i \\
X_{i,5} &\leftarrow \text{EnergyUse}_i \cdot (1 - \text{indicator}_i) \\
X_{i,6} &\leftarrow \text{EnergyUse}_i \cdot \text{indicator}_i \\
X_{i,7} &\leftarrow \text{urb}_i \cdot (1 - \text{indicator}_i) \\
X_{i,8} &\leftarrow \text{urb}_i \cdot \text{indicator}_i \\
X_{i,9} &\leftarrow \text{internet}_i \cdot (1 - \text{indicator}_i) \\
X_{i,10} &\leftarrow \text{internet}_i \cdot \text{indicator}_i
\end{aligned}
$$



# Section 4.2.2: Normal model with threshold

It should be noticed we **separated** the two dataset completely, without assuming that the other covariates would have remained the same below and above the threshold. We did this since we noticed how in our more general model **this was already the case**, without the need of assuming it. This is a stronger way of demonstrating our thesis, since we don’t need to leave out a level of complexity to let emerge the difference in the GDP parameter below and above.

Indeed, if we used the assumption of a single parameter for the remaining covariate in the two groups, we could have **missed** how **another covariate** was explaining better the variance in the observation with respect to the GDP parameter, making the difference in the GDP parameter near zero.

# Gamma model with threshold

Section 4.2.4

# Section 4.2.4: Gamma model with threshold

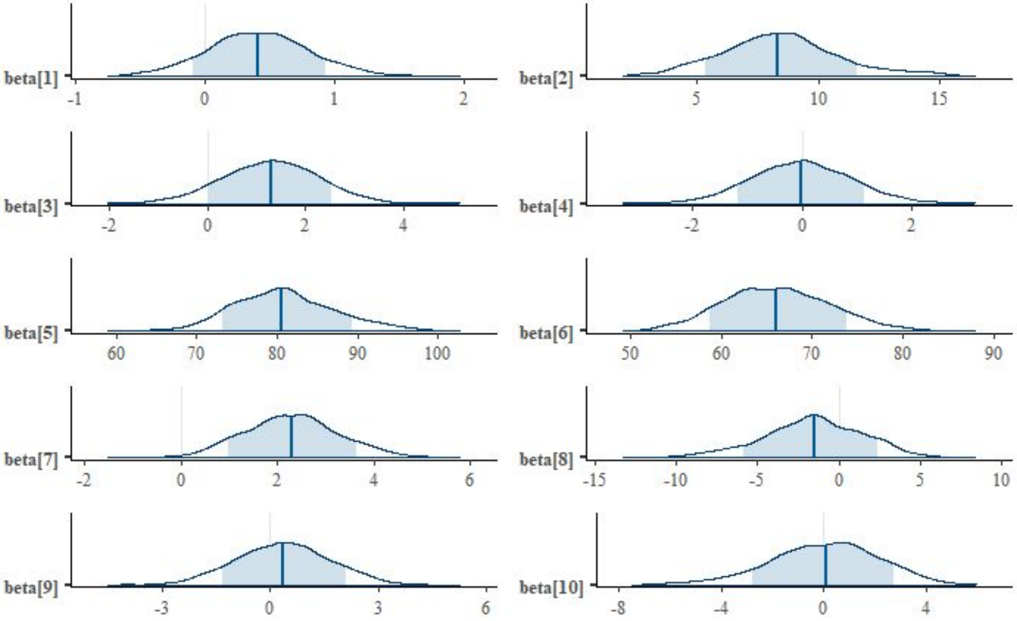
In order to establish the **robustness** of the Normal model, we tried also an analogous Gamma model.

$$\mu_i \leftarrow \langle X[i,], \beta \rangle$$

$$\text{co2percap}[i] \sim \text{Gamma}(\mu_i + 0.0001, \beta_\gamma)$$

$$\text{indicator}[i] \leftarrow (\text{unGDP}[i] \geq \text{threshold})$$

$$\begin{aligned}
X[i,1] &\leftarrow (1 - \text{indicator}[i]) \\
X[i,2] &\leftarrow \text{indicator}[i] \\
X[i,3] &\leftarrow \text{GDP}[i] \cdot (1 - \text{indicator}[i]) \\
X[i,4] &\leftarrow \text{GDP}[i] \cdot \text{indicator}[i] \\
X[i,5] &\leftarrow \text{EnergyUse}[i] \cdot (1 - \text{indicator}[i]) \\
X[i,6] &\leftarrow \text{EnergyUse}[i] \cdot \text{indicator}[i] \\
X[i,7] &\leftarrow \text{urb}[i] \cdot (1 - \text{indicator}[i]) \\
X[i,8] &\leftarrow \text{urb}[i] \cdot \text{indicator}[i] \\
X[i,9] &\leftarrow \text{internet}[i] \cdot (1 - \text{indicator}[i]) \\
X[i,10] &\leftarrow \text{internet}[i] \cdot \text{indicator}[i]
\end{aligned}$$

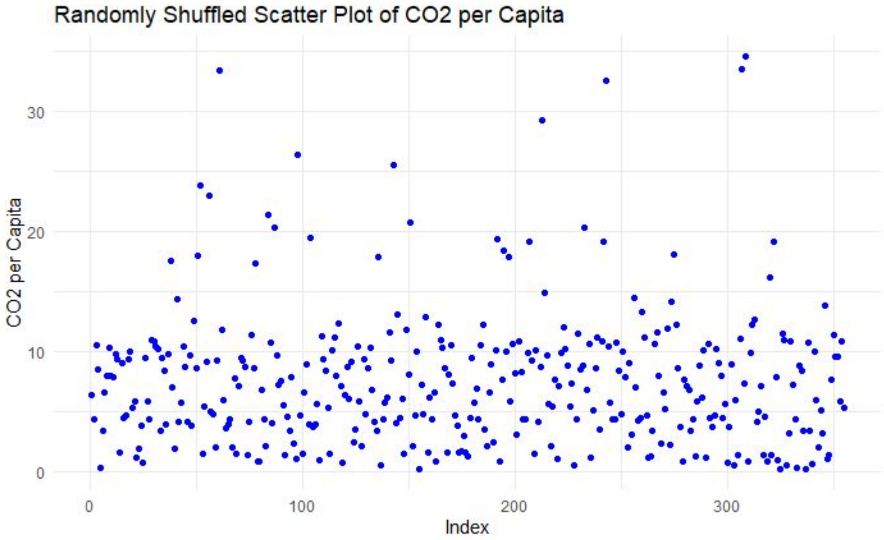


# Time series model

Section 4.3

# Section 4.3: Time series model

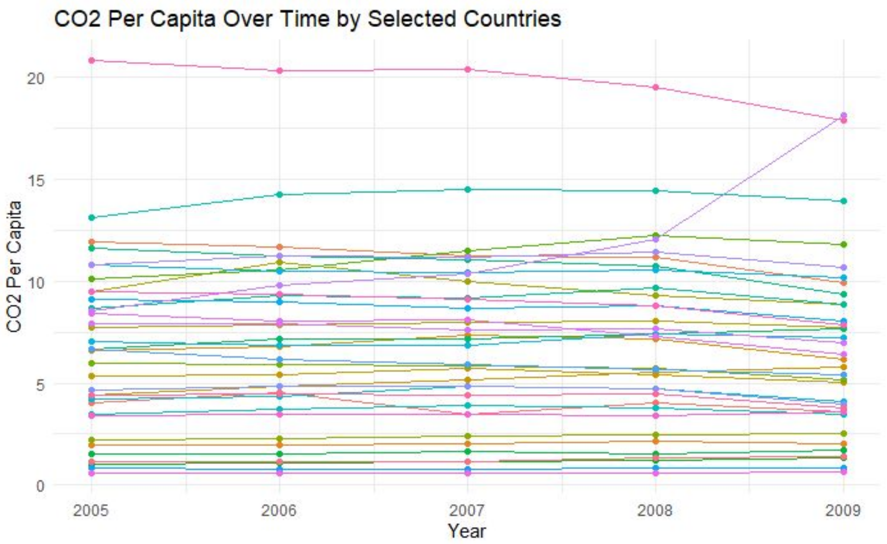
Let's consider a Bayesian AR(1) model, where the current observation in the series is based on the **immediately preceding observation**, adjusted by a stochastic term. We insert also the **main covariates**.




The resulting time series have **much more stable** trajectory, demonstrating the potentially great effectiveness of this data organization in the case of this dataset.

# Section 4.3: Time series model

Let's consider a Bayesian AR(1) model, where the current observation in the series is based on the **immediately preceding observation**, adjusted by a stochastic term. We insert also the **main covariates**.



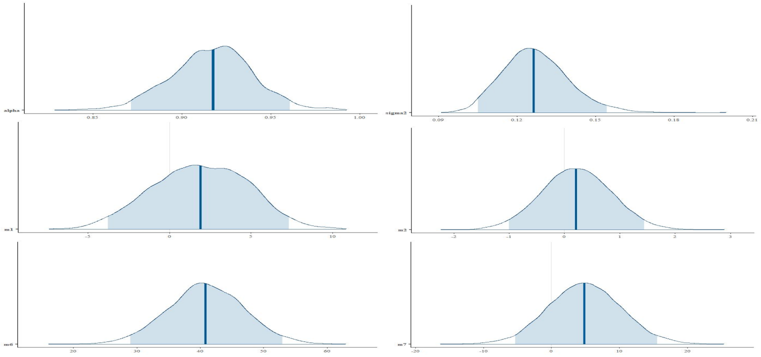

The resulting time series have **much more stable** trajectory, demonstrating the potentially great effectiveness of this data organization in the case of this dataset.

# Section 4.3: Time series model

The structure is composed by a **Normal** model, where the mean is a linear combination of the covariates and flat priors for the parameters.

$$Y[i] \sim N(\mu[i], \tau)$$
$$\mu[i] \leftarrow \alpha \cdot Y[i-1] + m_1 \cdot x_1[i-1] + m_2 \cdot x_2[i-1] + m_6 \cdot (x_1[i] - x_1[i-1]) + m_7 \cdot (x_2[i] - x_2[i-1])$$

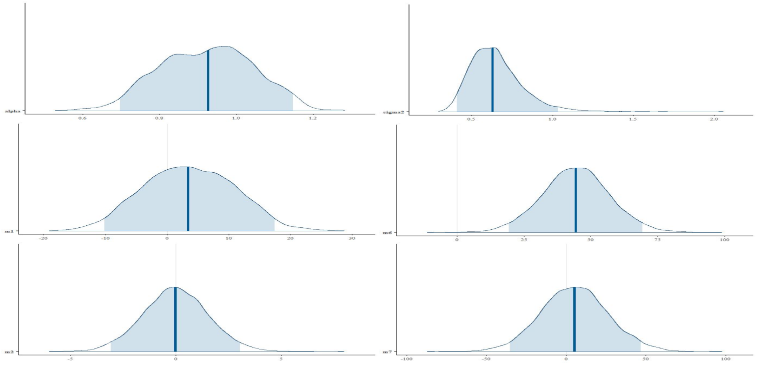
$$\alpha \sim \text{Uniform}(-1.5, 1.5)$$
$$\tau \sim \text{Gamma}(0.1, 10)$$
$$m_1 \sim N(0.0, 1.0 \times 10^{-4})$$
$$m_2 \sim N(0.0, 1.0 \times 10^{-4})$$
$$m_6 \sim N(0.0, 1.0 \times 10^{-4})$$
$$m_7 \sim N(0.0, 1.0 \times 10^{-4})$$



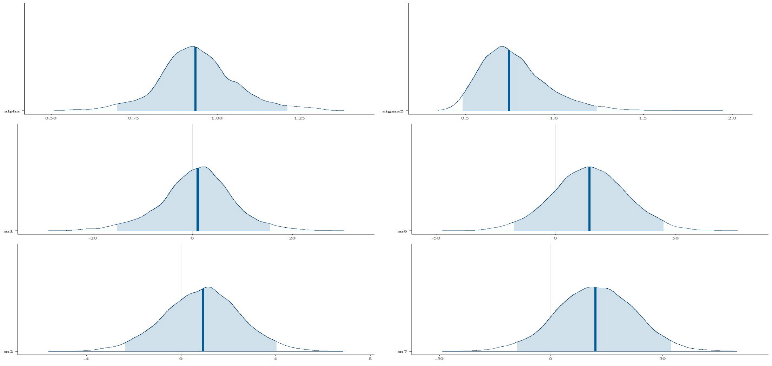
**Previous CO2 per capita** *Y[i-1]*: previous year observation.
**EnergyUse** *x1[i-1]*: previous year total energy consumption.
**EnergyUse-difference** *x1[i]-x1[i-1]*: change in energy consumption from previous year.
**GDP** *x2[i-1]*: previous year GDP.
**GDP-difference** *x2[i]-x2[i-1]*: change in GDP from previous year.

# Section 4.3: Time series model

Two subset of data were examinated according to the time **period**: pre crisis and during the economic crisis.



2005-2007



2007-2009

We assumed that **more variability** could be present when the economic crisis happened, this was in a way confirmed by the data.

# Section 4.3: Time series model

The following are the **In-sample** and **Out-of-sample prediction** of the first model (period 2005-2008 used to predict year 2008-2009), we noticed for all combinations of data used in the model and in the prediction a small MSE, which brought to a **satisfactory** prediction.

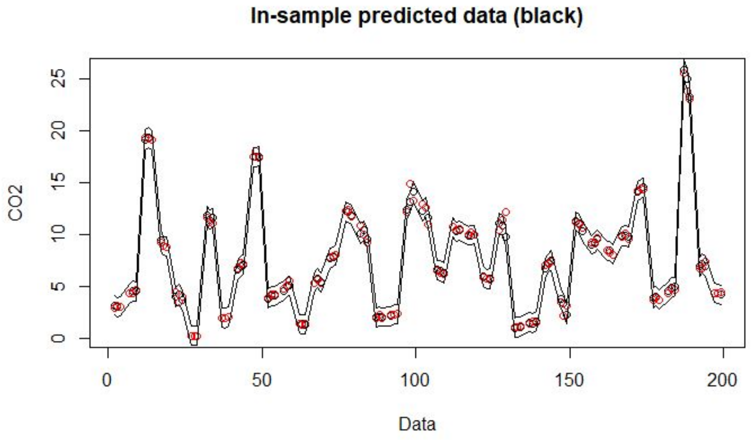
| | 05-06 06-09 | 05-07 07-09 | 05-08 08-09 |
|---|---|---|---|
| MSE: | 0.439 | 0.551 | 0.924 |
| R2: | 0.986 | 0.982 | 0.969 |

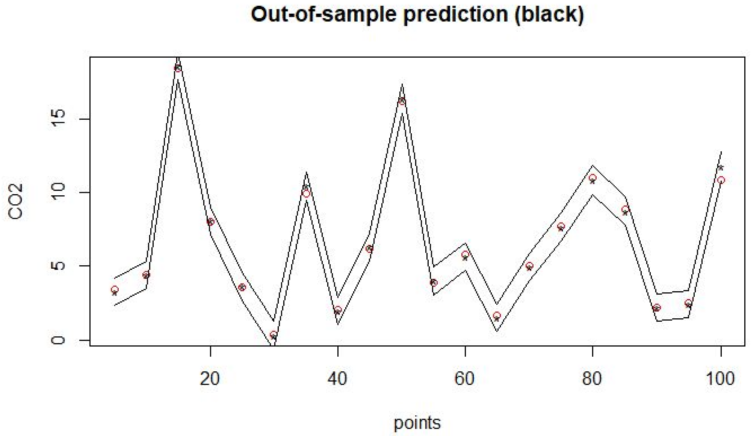
---

| | 05-06 06-07 |
|---|---|
| MSE: | 0.173 |
| R2: | 0.994 |

| | 05-06 07-08 | 05-07 07-08 |
|---|---|---|
| MSE: | 0.154 | 0.150 |
| R2: | 0.995 | 0.995 |

| | 05-06 08-09 | 05-07 08-09 | 05-08 08-09 |
|---|---|---|---|
| MSE: | 0.990 | 0.952 | 0.924 |
| R2: | 0.967 | 0.968 | 0.969 |



# Section 4.3: Time series model

Two subset of data were examinated according to the **GDP**: Low-GDP countries and High-GDP countries.

```
All-data 05-08 to predict 08-09:
MSE:  0.924
R2:   0.969

High-GDP 05-08 to predict 08-09:
MSE:  0.041
R2:   0.995

Low-GDP 05-08 to predict 08-09:
MSE:  0.070
R2:   0.995
```

High-GDP 05-08 to predict 08-09:

| | mean | sd |
|---|---|---|
| alpha | 0.93706022 | 0.03886073 |
| sigma2 | 0.23362421 | 0.03112142 |
| m1 | 2.35499572 | 1.37874488 |
| m2 | 0.03115735 | 0.19312618 |
| m6 | 32.27048530 | 4.24668766 |
| m7 | 0.80912565 | 2.12125569 |

Low-GDP 05-08 to predict 08-09:

| | mean | sd |
|---|---|---|
| alpha | 0.97281260 | 0.01882084 |
| sigma2 | 0.46537691 | 0.07556155 |
| m1 | 1.25447141 | 1.45214557 |
| m2 | 0.07163004 | 0.51977068 |
| m6 | 30.77443218 | 6.49935315 |
| m7 | 4.10598720 | 4.49939565 |

For both cases the results are indeed **more accurate** than with all data together, this could mean that the two groups present different characteristics, but not more complexity. In particular the Highest-GDP group has the highest accuracy.

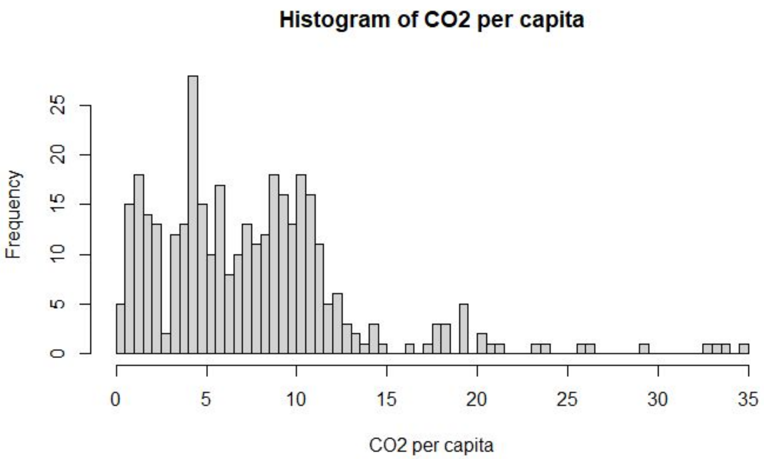
# Clustering

Section 5

# Section 5: Clustering

One-dimension mixture model, CO2 per capita clustered in 3 Normal curves.

$$
\begin{aligned}
&\#\,\text{Likelihood}\\
&z_i \sim \text{Categorical}(w)\\
&y_i \sim N(\mu_{z_i}, \sigma^2_{z_i})\\
\\
&\#\,\text{Prior}\\
&\mu_i \sim N(0, 0.01)\\
&\sigma^2_i \sim \text{Gamma}(0.01, 0.01)\\
\\
&w \sim \text{Dirichlet}(a)
\end{aligned}
$$

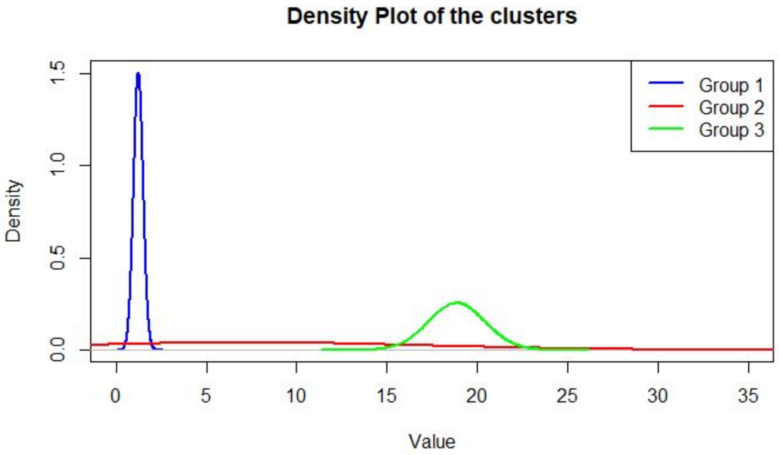



The resulting clusters have mean 1.25, 7.24 and 18.87. The largest is the the **middle** one (82%) containing most variability, followed by the **first** one of the less polluting countries (13%), finally the **third** group contains some “super-polluters” (5%).

# Section 5: Clustering

One-dimension mixture model, $CO_2$ per capita clustered in 3 Normal curves.

$$
\begin{aligned}
&\# \text{ Likelihood}\\
&z_i \sim \text{Categorical}(w)\\
&y_i \sim N(\mu_{z_i}, \sigma^2_{z_i})\\
\\
&\# \text{ Prior}\\
&\mu_i \sim N(0, 0.01)\\
&\sigma^2_i \sim \text{Gamma}(0.01, 0.01)\\
\\
&w \sim \text{Dirichlet}(a)
\end{aligned}
$$

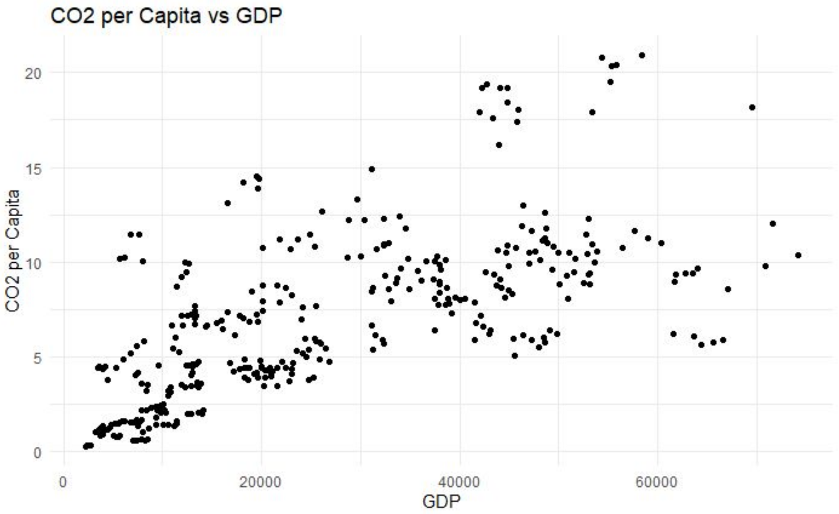



The resulting clusters have mean 1.25, 7.24 and 18.87. The largest is the the **middle** one (82%) containing most variability, followed by the **first** one of the less polluting countries (13%), finally the **third** group contains some “super-polluters” (5%).

# Section 5: Clustering

Two-dimension mixture model, CO2 per capita clustered in 2 Normal bivariate curves.

$$
\begin{aligned}
&z_i \sim \text{Bernoulli}(w) \\
&y_{i,1:2} \sim \text{Multivariate Normal}\left(\mu_{z_i+1,1:2}, \sigma_{z_i+1,1:2,1:2}\right) \\
\\
&\mu_{1,1} \sim N(0, 0.1) \\
&\mu_{1,2} \sim N(0, 0.1) \\
&\mu_{2,1} \sim N(0, 0.1) \\
&\mu_{2,2} \sim N(0, 0.1) \\
&w \sim \text{Beta}(1, 1) \\
\\
&\sigma^2_{1,1,1} \sim \text{Gamma}(1, 1) \\
&\sigma_{1,1,2} \sim N(0, 0.1) \\
&\sigma_{1,2,1} = \sigma_{1,1,2} \\
&\sigma^2_{1,2,2} \sim \text{Gamma}(1, 1) \\
\\
&\sigma^2_{2,1,1} \sim \text{Gamma}(1, 1) \\
&\sigma_{2,1,2} \sim N(0, 0.1) \\
&\sigma_{2,2,1} = \sigma_{2,1,2} \\
&\sigma^2_{2,2,2} \sim \text{Gamma}(1, 1)
\end{aligned}
$$




The two groups found have average GDP around 23000 and 40000, with the **first** group accounting for more than three times the point in the second group. The **second** group is also responsible for an average CO2 per capita production of 12.5, way higher than the 5.5 of the first group.

# Section 5: Clustering

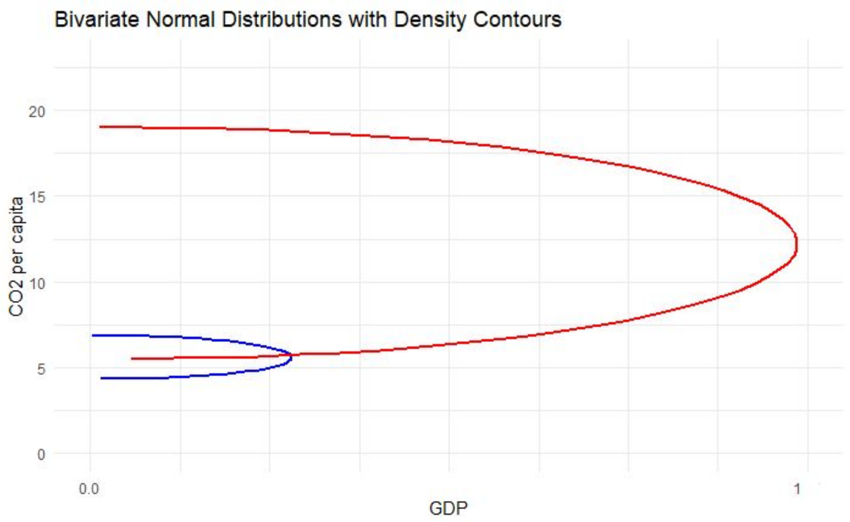
Two-dimension mixture model, CO2 per capita clustered in 2 Normal bivariate curves.

$$z_i \sim \text{Bernoulli}(w)$$
$$y_{i,1:2} \sim \text{Multivariate Normal}\left(\mu_{z_i+1,1:2}, \sigma_{z_i+1,1:2,1:2}\right)$$

$$\mu_{1,1} \sim N(0, 0.1)$$
$$\mu_{1,2} \sim N(0, 0.1)$$
$$\mu_{2,1} \sim N(0, 0.1)$$
$$\mu_{2,2} \sim N(0, 0.1)$$
$$w \sim \text{Beta}(1, 1)$$

$$\sigma^2_{1,1,1} \sim \text{Gamma}(1, 1)$$
$$\sigma_{1,1,2} \sim N(0, 0.1)$$
$$\sigma_{1,2,1} = \sigma_{1,1,2}$$
$$\sigma^2_{1,2,2} \sim \text{Gamma}(1, 1)$$

$$\sigma^2_{2,1,1} \sim \text{Gamma}(1, 1)$$
$$\sigma_{2,1,2} \sim N(0, 0.1)$$
$$\sigma_{2,2,1} = \sigma_{2,1,2}$$
$$\sigma^2_{2,2,2} \sim \text{Gamma}(1, 1)$$




The two groups found have average GDP around 23000 and 40000, with the **first** group accounting for more than three times the point in the second group. The **second** group is also responsible for an average CO2 per capita production of 12.5, way higher than the 5.5 of the first group.

# Conclusions

- Relationship between $CO_2$ emissions and GDP is non linear.
- Strength of autoregression in country-wide $CO_2$ prediction.
- Usefulness of EnergyUse with respect to the other covariates