Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
January 25th, 2023

Problem n.1

The file `footprints.txt` contains the carbon footprints of 3000 individuals who have enrolled in an app tracking program in 30 different European cities, identified by the variable `IDCity`. The app records various lifestyle habits, such as purchasing behaviours, travels, etc. Consider the following linear mixed-effects model:

$$\texttt{carbon\_footprint}_i = \beta_0\,\mathbb{1}_i + \beta_1\,\texttt{purchases}_i + \beta_2\,\texttt{heating}_i + \beta_3\,\texttt{flights}_i + b_{0i}\,\mathbb{1}_i + \boldsymbol{\epsilon}_i \qquad \textbf{(M1)}$$

for $i \in$ `IDCity` with $\boldsymbol{\epsilon}_i$ i.i.d. such that $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_i\boldsymbol{\mathcal{C}}_i\boldsymbol{\Lambda}_i)$ and b_{0i} i.i.d. such that $b_{0i} \sim \mathcal{N}(0, \sigma^2 d_{11})$, where

- $\texttt{carbon\_footprint}_i$ is the 100-dimensional vector of the carbon footprints produced by individuals in city i;
- $\texttt{purchases}_i$ is the 100-dimensional vector of the normalized quantity of new purchases (clothing, devices, and furniture) per month made by individuals in city i;
- $\texttt{heating}_i$ is the 100-dimensional vector of binary variables indicating the sources of energy used for home heating by individuals in city i (`1`=renewables and `0`=fossil);
- $\texttt{flights}_i$ is the 100-dimensional vector of binary variables indicating whether individuals in city i used the plane more than three times a year (`1`=yes and `0`=no).

a) Assuming *homoscedastic* residuals, fit the model **M1** briefly detailing the implementation choices and estimate β_0, β_1, σ^2, $\sigma^2 \cdot d_{11}$ and the PVRE for **M1**.

b) On top of **M1**, fit now a model **M2**, introducing *heteroscedastic* residuals with

$$\boldsymbol{\Lambda}_i = \begin{bmatrix} \lambda_1^{(i)} & 0 & .. & 0 \\ 0 & \lambda_2^{(i)} & .. & 0 \\ .. & .. & .. & 0 \\ 0 & 0 & .. & \lambda_{100}^{(i)} \end{bmatrix}$$

and $\boldsymbol{\lambda}^{(i)} = [\lambda_1^{(i)} \quad \lambda_2^{(i)} \quad ... \quad \lambda_{100}^{(i)}]' = |\texttt{purchases}_i|^{\delta}$, for $i \in$ `IDCity`.
Briefly detail the implementation choices for **M2** and compute β_0, β_1, δ, σ^2, $\sigma^2 \cdot d_{11}$ and the PVRE for **M2**.

c) Comment on the obtained value of δ.

d) Compare **M1** and **M2** in terms of σ^2, AIC, PVRE and likelihood ratio test (if possible) and comment on the obtained results. Which model would you choose?

e) On top of the selected model, net of the impact of fixed effect covariates, which are the `IDCity` associated with the lowest and highest carbon footprint? How did you get these answers?

Upload your solution here **https://forms.office.com/e/URYMLXVTyH**