Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
June 28, 2024

Problem n.1

The file `FarmData.txt` contains data regarding 2000 crops grown across 25 different farms. The crops are tested for yield using a standardized measurement across farms. The response variable `Yield` $\in \mathbb{R}$ is the normalized yield score. Three explanatory variables at the soil level are provided: `SoilQuality` $\in \mathbb{R}$: a scale centered at 0 for soil quality; `Irrigated` $\in \{0,1\}$: 1 if the soil is irrigated, 0 otherwise; `Fertilized` $\in \{0,1\}$: 1 if the soild is fertilized, 0 otherwise. Moreover, `IDFarm` $\in 1,...,25$ provides the anonymous farm identification number.

Consider the following linear mixed-effects model:

$$\texttt{Yield}_i = \beta_0\, \mathbb{1}_i + \beta_1\, \texttt{Irrigated}_i + \beta_2\, \texttt{SoilQuality}_i + \beta_3\, \texttt{Fertilized}_i + b_{0i}\, \mathbb{1}_i + b_{1i}\, \texttt{SoilQuality}_i + \boldsymbol{\epsilon}_i \quad \text{for} \quad i \in \texttt{IDFarm}$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i \boldsymbol{\mathcal{C}}_i \boldsymbol{\Lambda}_i)$ and

$$\boldsymbol{b}_i = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim \mathcal{N}\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix} \Big)$$

a) Assume homoscedastic residuals and $d_{12} = d_{21} = 0$. Fit the model (**M1**), briefly detail its implementation reporting also the relevant R code, and compute β_1, β_3, σ^2, $\sigma^2 \cdot d_{11}$ and $\sigma^2 \cdot d_{22}$.

b) Compute the PVRE for **M1**.

c) Fit now a model (**M2**) with the same fixed part of **M1**, but in which there is no random slope. Moreover, assume uncorrelated heteroscedastic residuals with

$$\boldsymbol{\Lambda}_i = \begin{bmatrix} \lambda_1^{(i)} & 0 & .. & 0 \\ 0 & \lambda_2^{(i)} & .. & 0 \\ .. & .. & .. & 0 \\ 0 & 0 & .. & \lambda_{80}^{(i)} \end{bmatrix}$$

and $\boldsymbol{\lambda}^{(i)} = [\lambda_1^{(i)} \quad \lambda_2^{(i)} \quad ... \quad \lambda_{80}^{(i)}]' = \sqrt{\texttt{SoilQuality}_i}$ ($\boldsymbol{\lambda}_i$'s known case), for $i \in$ `IDFarm`.
Briefly detail the implementation of **M2** reporting also the relevant R code. Compute β_1, β_3, $\sigma^2 \cdot d_{11}$.

d) Report the dot plot of the estimated random intercepts in **M2**. Net of the impact of fixed effect covariates, which is the `IDFarm` associated to the highest and lowest yield?

Upload your solution **https://forms.office.com/e/fS6PCQAtEj**