Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
July 7th, 2023

Problem n.1

The dataframe `Lakes_pollution.txt` contains data related to the pollution levels in 146 different `locations` of 26 different Italian lakes (`italian_lakes`). For each location, the following quantities are registered: `depth` $\in \mathbb{R}$ (measured in meters), `mercury_conc` $\in \mathbb{R}$ (allowed concentrations for mercury (Hg): (0.1 - 1) 10^{-2} mg/L), `ph` $\in \mathbb{R}$ (ideal pH range for freshwater ecosystems: 6.5 - 8.5), `turbidity`$\in \mathbb{R}$ (range for clear water: 1 - 10 Nephelometric Turbidity Units (NTU); higher values indicate increased turbidity), `wastewater_discharge` $\in$ {Yes, No} and `BOD` (Biological Oxygen Demand) $\in \mathbb{R}$, measured in milligrams of oxygen consumed per litre of water (mg/L).

`BOD` is a water quality parameter that measures the amount of dissolved oxygen ~~(DO)~~ consumed by microorganisms during the decomposition of organic matter in water. Modelling `BOD` is important because it indicates the level of organic pollution or the amount of biodegradable organic material present in water. Lower `BOD` values indicate cleaner water with less organic pollution, while higher `BOD` values indicate more significant organic pollution and greater oxygen demand. Consider the following linear model:

$$\texttt{BOD}_{ij} = \beta_0 + \beta_1\,\texttt{depth}_{ij} + \beta_2\,\texttt{mercury\_conc}_{ij} + \beta_3\,\texttt{ph}_{ij} + \beta_4\,\texttt{turbidity}_{ij} + \beta_5\,\texttt{wastewater\_discharge}_{ij} + \epsilon_{ij}$$

for $i \in$ `italian_lakes`, $j \in$ `locations` and with $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

a) - Fit the model and estimate the model unknowns, assuming ϵ_{ij} i.i.d.

- Is it possible to assume the homoscedasticity of the residuals? Provide plots in support of your answer.

b) - What is the **percentage** of unexplained variability?

- Report the mean difference of `BOD` between the locations with wastewater discharge with respect to the ones that are not discharged.

c) Within the context of homoscedastic and correlated residuals, introduce a Compound Symmetry Correlation Structure within each $i \in$ `italian_lakes` (let ρ be the extra diagonal term in the correlation matrix).

- Report the estimated ρ and σ and a **95**% confidence interval for both. Draw your conclusions.

d) Consider now the variable `italian_lakes` as a random intercept.

- Fit a suitable model, report the parameterization and provide the model unknowns.

- Compare the model at point c) and the model at point d).

e) Report the dot plot of the estimated random intercepts. Net to the effect of fixed effect covariates, which is the lake associated with the *highest* concentration of `BOD`?

Upload your solution here