Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
June 16th, 2023

Problem 3: Dissolved oxygen in Italian lakes

The dataframe `Lakes_pollution.txt` contains data related to the pollution levels in 125 different `locations` of 20 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): 0.001 - 0.01 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}, `DO` (Dissolved Oxygen) $\in \mathbb{R}$.

Being able to model `DO` is important because low levels can be stressful for aquatic organisms and can lead to fish kills. Consider the following linear model:

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

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

a) Fit the model and provide the estimates of the model unknowns, after having eventually reduced it. What is the percentage of unexplained variability? In your opinion, the homoscedasticity of the residuals can be assumed?

b) What is the average increase of `DO` due to increment of 3 NTU in `turbidity`? Is that significant? Report the mean difference of `DO` 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 of them. Draw your conclusions.

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

   Compute and report the PVRE index and comment on the obtained result.

   [**Bonus**] Make a comparison between the model at point c) and the model at point d).

e) Report the dot plot of the estimated random intercepts. Ignoring the effect of fixed effect covariates, which is the lake associated with the lowest concentration of `DO`?

Upload your results here:
https://forms.office.com/e/YFttfWTTbz