Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
February 8th, 2023

Problem n.3

The file `Cancer_mortality.txt` contains data regarding the 107 italian provinces (`ID_prov`) grouped by their regional code (`ID_reg`). For each province, the observed values of the following variables are provided: average age (`avg_age` $\in \mathbb{R}$), the productivity index (`productivity_index` $\in \mathbb{R}$), the average sunny days (`avg_sunny_days` $\in \mathbb{R}$), the average alcohol consumption (`avg_alcohol_consumption` $\in \mathbb{R}$), the average household income (`avg_household_income` $\in \mathbb{R}$), the average education level (`avg_education_level` $\in \mathbb{R}$), and the cancer mortality (20-64 years), standardized per 10000 residents (`cancer_mortality` $\in \mathbb{R}$). All variables have been scaled to have mean 0 and standard deviation 1. Moreover, the values of the 2-level variable `polluted` $\in$ {Yes, No} are also provided, indicating whether a province is polluted or not.

a) Formulate a classical linear regression model (**M0**) for `cancer_mortality` as a function of all the other variables, with no interaction. Report the model and its parameterization, together with the estimates of the parameters of the model and the standard deviation σ of the error term $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Moreover, report the plot of the standardized residuals and comment on it.

b) Does the average education level have a positive or negative effect on cancer mortality? Is that significant? Moreover, report the mean difference of the cancer mortality of the polluted provinces with respect to the not-polluted ones.

c) Update the model **M0**, introducing a compound-Symmetry Correlation Structure and using the region as grouping factor (model **M1**). Report the model and its parameterization. Compute and report the 99% confidence intervals for rho and sigma; comment on the obtained values and draw your conclusions.

d) Fit a mixed-effects model with random intercept (model **M2**) starting from model **M0** and introducing a random intercept related to the regional grouping factor. Report the model and its parameterization, together with the estimates of the parameters of the model and the standard deviations of the random intercept and the error term. Compare **M1** and **M2** and draw your conclusions.

Upload your solution here