Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
Exam Demo - Part B - 2023/2024

Problem 1: Asthma prevalence in Italian regions

The dataset `asthma.txt` contains data regarding asthma prevalence across the 110 Italian provinces, identified by `province_id` and grouped by their regional code `region_id`.

For each province, the dataset includes the following variables: average `age`, air `pollution` index, average number of `sunny` days per month, average `tobacco` consumption, average household `income`, average `education` level, and `asthma` prevalence rate, standardized per 10,000 residents. Additionally, the dataset includes a categorical variable `urban` indicating whether the province is categorized as urban or not.

All numerical variables have been scaled to have a mean 0 and a standard deviation 1.

a) Implement the following linear regression model **M0**:

$$\texttt{asthma} = \beta_{0,k} + \beta_1\,\texttt{age} + \beta_2\,\texttt{pollution} + \beta_3\,\texttt{sunny} + \beta_4\,\texttt{tobacco} + \beta_5\,\texttt{income} + \beta_6\,\texttt{education} + \epsilon \qquad (1)$$

with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and k the grouping variable induced by `urban` ($k = \texttt{urban}$).

Report the estimates of the parameters of the model fitted with Ordinary Least Squares and verify the model assumptions (required for both estimation and inference).

b) Can we affirm at 90% confidence level that the age has a positive effect on asthma prevalence? Additionally, provide an 95% confidence interval for the mean difference between the asthma prevalence in an urban province and in a non-urban one

c) After having reduced the model **M0**, if appropriate, update it by introducing a compound-Symmetry Correlation Structure using the region as a grouping factor (model **M1**). Provide a 99% confidence interval for the parameters ρ and σ of the compound symmetry.

d) From the possibly reduced version of the model **M0**, update it now by introducing a random intercept related to the regional grouping factor (model **M2**). What do you observe? Provide the estimate of the standard deviation of the random intercept along with the one of the error term.

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

Problem 2

Problem statement

a) First question

b) Second question

...

Upload your results here: `https://forms.office.com/path/to/form`

Problem 3 [only for 10 CFU version]

Problem statement

a) First question

b) Second question

...

Upload your results here: `https://forms.office.com/path/to/form`