Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
Academic Year 2024/2025
Exam 2025-01-17 - Part B

Problem 2: Factors Affecting Crop Yields in Agricultural Regions of Italy

The dataset `crop_yields.txt` contains data regarding wheat crop yields in tons per hectare (tons/ha) across 100 agricultural areas in Italy, each identified by `area_id` and categorized by their respective province `province_id`.

For each area, the dataset includes the following variables: average `temperature`, cumulated mm of `rainfall`, average number of `sunny` days per month, average `soil_quality` index, and average `fertilizer` usage. Additionally, the dataset includes a categorical variable `irrigation` indicating whether the region has irrigation systems in place (coded as 1 for regions with irrigation, 0 for regions without).

All continuous variables have been standardized to have a mean of 0 and a standard deviation of 1.

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

$$\texttt{crop\_yield} = \beta_{0,k} + \beta_1\,\texttt{temperature} + \beta_2\,\texttt{rainfall} + \beta_3\,\texttt{sunny} + \beta_4\,\texttt{soil\_quality} + \beta_5\,\texttt{fertilizer} + \epsilon \quad (1)$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and k represents the grouping factor induced by `irrigation`.

Estimate the parameters of the model using Ordinary Least Squares (OLS) and assess the assumptions underlying the model.

b) Test whether we can affirm with 99% confidence that temperature has a negative effect on crop yields. Additionally, provide a 95% confidence interval for the mean difference in crop yields between areas with and without irrigation.

c) Update **M0** by introducing a compound-symmetry correlation structure, with the province as a grouping factor (model **M1**). Report the 99% confidence interval for the parameters ρ and σ associated with the compound symmetry.

d) Update **M0** by incorporating a random intercept based on the province grouping factor (model **M2**). Compare the two models. What do you observe? Comment.

Upload your results here: https://forms.office.com/e/hsm9pRCHyS