Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
June 15th, 2020

Problem n.1

The file `pollution.txt` reports the measurements of two air pollutants (PM2.5 and PM10) collected in 100 days by an air-quality monitoring site of a Chinese city.

a) Perform a statistical test at level 5% to verify if the mean of the two air pollutants is significantly different from (50,50). Verify the assumptions required to perform the test and provide the p-value.

b) Find an elliptical confidence region at level 95% for the mean value of the two air pollutants. Provide a plot and report: the analytical expression of the region, its centre, the direction and the length of the principal axes of the ellipse.

c) Does the vector (50,50) lies within or outside the region identified at point b)? Comment the answer highlighting the connection to the conclusion of point a).

d) Provide two T^2 simultaneous confidence intervals (global confidence 95%) for the mean of PM2.5 and the mean of PM10.

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
June 15th, 2020

Problem n.2

Prehistoric men crafted stone tools by striking raw stones to obtain the desired shape. Archaeologists rarely find tools, but they often find stone flakes, the waste of the crafting process. The file `stoneflakes.txt` contains the lengths and the widths of stone flakes collected in 75 different archaeological sites.

a) Identify possible clusters within the data using a hierarchical clustering algorithm (Euclidean distance, Ward linkage). Provide the plot of the dendrogram and qualitatively identify the optimal number of clusters.

b) Assuming that the clusters identified at point a) have the same covariance structure, formulate a MANOVA model for the geometrical features (width and length) of the stone flakes as a function of the clustering membership. Verify the assumptions of the model. Is there statistical evidence to state that the membership to a cluster has an effect on the mean features of the stone flakes?

c) Provide confidence intervals for the differences between the mean features of stone flakes belonging to the identified clusters. Use a Bonferroni correction to ensure a 90% global level. Use the computed intervals to comment about the differences among the clusters.

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
June 15th, 2020

Problem n.3

A newly designed airfoil is tested with a series of aerodynamic and acoustic tests in an anechoic wind tunnel. The file `airfoil.txt` reports the sound level (in decibel) measured under different experimental conditions characterized by different air stream frequencies (in hertz) and velocities (labelled H and L for high velocity and low velocity respectively). Consider the following linear model for the sound level (Y), which accounts for the air stream frequency x and for the air stream velocity:

$$Y = \beta_{0,g} + \beta_{1,g} \cdot x + \epsilon,$$

with $\epsilon \sim N(0, \sigma^2)$ and g the grouping structure induced by the velocity of the air stream.

a) Provide the pointwise estimates of the parameters of the model and verify the model assumptions.

b) Perform three statistical tests to verify if

   - there is a significant dependence of the mean sound level on the air stream frequency,
   - there is a significant dependence of the mean sound level on the air stream velocity,
   - the increase in the mean sound level induced by a unitary increase in the frequency is significantly different for high and low air stream velocities.

c) Based on the results of the previous point, reduce the model and update the parameters.

d) Using the model at point c), provide a confidence interval for the mean of the sound level of a new test performed with air stream frequency of 15 000 Hz and high air stream velocity.

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
June 15th, 2020

Problem n.4

The file `revenues.txt` collects the average daily revenues y [k€] during the lockdown of 70 minimarkets located in Milan. The dataset also reports the UTM coordinates s_i of the shops, the resident population in the neighborhood around the shop $p(s_i)$, and the Euclidean distance $d(s_i)$ [m] between the location of the shop and the Duomo $d(s_i) = \|s_i - s_d\|$, with $s_d = (514711.6, 5033903.0)$. Consider for the revenue $y(s_i)$, $i = 1, ...70$, the following model

$$y(s_i) = a_0 + a_1 \cdot p(s_i) + \delta(s_i),$$

with $\delta(s_i)$ a stationary residual.

a) Estimate via generalized least squares the parameters a_0, a_1 of the model. Report the model estimated for $\delta(s_i)$, and discuss the model assumptions.

b) Provide a kriging prediction $y^*(s_0)$ of the revenues at a shop located in the Brera district at location $s_0 = (514703.8, 5035569.3)$. For this purpose, use a point estimate of the resident population $p(s_0)$ obtained through a linear model in the variable *distance from the Duomo* (detail the model assumptions for $p(s_0)$ and its point estimate).

c) Report the kriging variance $\sigma^2(s_0)$ of the point prediction at point (b). Would you deem the variance $\sigma^2(s_0)$ to be fully representative of the uncertainty associated with the prediction $y^*(s_0)$?