Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
February 10th, 2021

Problem n.1

The file `shopping.txt` contains information on the usage of an online store selling clothing in the last 24 months. For each month, it reports the number of accesses to the online store, the number of purchases of men's clothing and the number of purchases of women's clothing. Assuming each month to be independent of the others, answer the following questions.

a) Build a confidence region (level 95%) for the mean of the vector whose components are the number of accesses to the online store, the number of purchases of men's clothing and the number of purchases of women's clothing. Characterize the region by reporting its mathematical expression, its center, the direction of the axes and the length of the semi-axes. Introduce and test the hypothesis of Gaussianity; identify and discuss possible issues (if any).

b) Build four T^2-simultaneous confidence intervals (level 95%) for: the mean number of accesses to the online store, the mean number of purchases of men's clothing, the mean number of purchases of women's clothing and the mean number of total purchases.

c) Perform a test of level 95% to verify the hypothesis according to which, in mean, more than 20% of the accesses to the online store result in a purchase. Report the p-value of the test.

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUQU5ZSVBRNzUyREVSRFM2RlYwMEdVQldWTy4u`

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
February 10th, 2021

Problem n.2

In a study to develop an automated vision system able to distinguish between three species of rice, 60 images of rice grains were processed to obtain two morphological features for each grain of rice: the major axis length and the eccentricity. The data are reported in the file `rice.txt`.

a) Use a hierarchical clustering method (Euclidean distance and complete linkage) to identify the three species of rice. Provide the plot of the dendrogram. Report the number of data within each cluster, the mean within the clusters and a plot of the results.

b) Evaluate the performances of the clustering method at point (a) and identify the possible issues (if any). Propose, if needed, an alternative clustering method to cluster the data (report the number of data within each cluster, the mean within the clusters and a plot of the results).

c) Provide Bonferroni intervals (global level 95%) for the means and the variances of the major axis length of the rice grains within each of the three clusters identified. Introduce and verify the appropriate assumptions.

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUOVhNTkU4OVRJUDNHQ0Q4SDE5RzRFMlE3Ny4u`

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
February 10th, 2021

Problem n.3

The file `landslides.txt` collects data on slow moving landslides. The dataset reports, for 80 monitored landslides, the average downslope displacement rate [mm/year], the annual precipitations [mm], the hardness of subsurface rocks [in Mohs' scale], the quantity of coarse debris [g/cm^3] and the quantity of fine debris [g/cm^3].

a) Formulate a linear regression model for the average downslope displacement rate, as a function of all the other variables. Report the model and its parametrization, together with the estimates of all its parameters. Verify the model assumptions.

b) Based on appropriate test(s), reduce the model and update the model parameters.

c) Using model (b), test the hypothesis according to which the effect on the displacement rate of the increase of 1 g/cm^3 of coarse debris is two times the effect of the increase of 1 g/cm^3 of fine debris. Report the hypotheses and the p-value of the test performed. Possibly propose a new constrained model and estimate its parameters.

d) A new landslide has been identified on a site with annual precipitations 700 mm, hardness of subsurface rocks 5, quantity of coarse debris 10 g/cm^3 and quantity of fine debris 8 g/cm^3. Using the last model, compute a pointwise estimate and a confidence interval of level 99% for the mean displacement rate of the new landslide.

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUMlZFUU9QRjhTRllQOUhZWE1RSVZVVTVJVi4u`

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
February 10th, 2021

Problem n.4

The manager of the hotel *Boule de Neige* in Courmayeur is designing the pricing strategy for the carnival festivities 2021. The file `hotels.txt` collects the prices per night y [€/night] in hotels in Courmayeur and neighboring villages, as observed for 55 hotels during 2019. The dataset also reports the UTM coordinates s_i of the hotels, whether the price refers to a day during the winter season or not ($winter = yes$ or $winter = no$), and the distance of the considered hotel from the funicular connecting to the ski slopes, $d(s_i) = \|s_i - s_f\|$, with $s_f = (342362.58, 5072518.24)$. Consider for the price the following model

$$y(s_i) = a_{0,g} + a_{1,g} \cdot d(s_i) + \delta(s_i),$$

with $\delta(s_i)$ a stationary residual with spherical variogram with nugget, and $g = 1, 2$ the grouping induced by the variable $winter$ ($g = 1$ for $winter = yes$, $g = 2$ otherwise).

a) Assuming $a_{1,g} = 0$ for $g = 1, 2$ and $a_{0,1} = a_{0,2} = a_0$, estimate the parameter a_0 of the model via generalized least squares. Report the point estimate of a_0 and the model estimated for $\delta(s_i)$; discuss the model assumptions.

b) Assuming $a_{1,g} \neq 0$, estimate the parameters $a_{0,g}, a_{1,g}$ of the model via generalized least squares. Report the point estimate of $a_{0,g}, a_{1,g}$ and the model estimated for $\delta(s_i)$; discuss the model assumptions.

c) Choose the best model between those estimated at points (a) and (b). Comment on your choice.

d) Suggest to the hotel manager a pricing strategy for a stay of 4 nights at *Boule de Neige* in the period Feb. 17th to Feb. 21st ($winter = yes$, $s_0 = (342399.74, 5072272.75)$). Motivate your response and detail your assumptions.

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglURFVYODZBQjQ4SO4wROs1MDdHS1RMNEpSWi4u`