Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
January 20th, 2021

Problem n.1

The file `wine.txt` reports the alcohol content of 150 Italian wines. The dataset also reports the color of the wine (red or white) and the region of production (Piemonte, Toscana or Veneto).

a) Build a complete ANOVA model for the alcohol content as a function of the factors *color* (red or white) and *region* (Piemonte, Toscana or Veneto). Report and verify the assumptions of the model.

b) Perform tests for the significance of the factors and of their interaction and comment the results. If needed, propose a reduced model and report the estimates of the parameters of the model.

c) Build Bonferroni confidence intervals (global level 99%) for the means and variances of the groups identified at point (b). Comment the results.

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUM1JHUOpXUlkyVFhRNkpFRjNOWjFaQTVONi4u`

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
January 20th, 2021

Problem n.2

The file `activity.txt` collects the results of an experiment carried out with a group of subjects performing some daily activities (walking, sitting, laying) while carrying a smartphone with embedded inertial sensors. The dataset reports the mean linear acceleration (estimated from the signal recorded by the accelerometer) and the mean angular velocity (estimated from the signal recorded by the gyroscope). Knowing that, on the average day, the typical smartphone user spends 3 hours walking, 12 hours sitting and 9 hours laying, answer the following questions.

a) Build a classifier for the variable *activity* based on the available quantitative features. Report the mean within the groups identified by the variable *activity* and a plot of the classification regions. Introduce and verify the appropriate assumptions.

b) Compute the APER of the classifier.

c) If the sensors record a mean body linear acceleration of 0.45 and a mean angular velocity of 0.52, what activity would the subject be performing according to the classifier built at point (a)?

d) Build a k-nearest neighbor classifier for the *activity* choosing the parameter k equal to 5. Report a plot of the classification regions and compute the error rate on the training set. Compare the results with points (a) and (b) and comment on the performances of the classifiers.

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUREJQN1pZTDRPWlFFVEdBSEJXSlkyVEk1My4u`

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
APPLIED STATISTICS
January 20th, 2021

Problem n.3

The file `bikes.txt` reports the data on public bikes rented in 50 days, the corresponding weather data (temperature [°C] and wind [m/s]) and holiday information. The dataset reports, for each day, the number of bikes rented (b), the mean temperature (t), the mean wind speed (w) and the holiday information (Holiday/No holiday). Consider the following model

$$b_{i,g} = \beta_{0,g} + \beta_{1,g} t_i + \beta_{2,g} w_i + \epsilon;$$

where $g \in \{1, 2\}$ indicates the group according to the holiday information (g = 1 for holiday, g = 2 for no holiday) and $\epsilon \sim N(0; \sigma^2)$.

a) Estimate the parameters of the model and report the estimated values.

b) Verify the needed model assumptions and perform two statistical tests at level 5% to verify if

- there is statistical evidence of a dependence of the mean number of bikes rented on weather information;
- there is statistical evidence of a dependence of the mean number of bikes rented on holiday information.

Report the hypothesis and the p-values of the test performed.

c) Comment on possible model weaknesses and, if needed, reduce the model, and update the parameter estimates.

d) Based on the model obtained at point (c), provide a pointwise estimate and a prediction interval (probability 95%) for the number of public bikes rented on a holiday with mean temperature 2°C and mean wind speed 3 m/s.

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPg`
`lUM0dNMTNPQzlFVjhUNDUwQzZMU1o2R1FXWC4u`

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
January 20th, 2021

Problem n.4

A satellite carrying an infrared spectrometer is collecting data of the surface of an asteroid in order to determine its chemical composition. The file `spectra.txt` reports the reflectance spectra of 10 independent measurements collected at different positions of the satellite along its orbit. Each spectrum has been sampled on the same grid of wavelengts (`wl1`,...,`wl80`). It is known that the spectrometer measurements are affected by small errors. Answer the following questions, considering a functional data analysis approach.

a) Consider the first spectrum and perform a smoothing of this datum using a B-spline basis of degree 3. Choose the number of basis functions using a generalized cross-validation (GCV) criterion and provide a plot of the value of the GCV statistic versus the number of basis elements considered. Report the number of basis functions chosen and the first 3 coefficients of the basis expansion.

b) Perform a smoothing of the other spectra using the basis chosen at point (a). Provide a plot of the smoothed data.

c) Use the k-mean alignment algorithm to simultaneously cluster (k=3) and align the data allowing affine transformation for the abscissas. Use the correlation between the curves as similarity measure. Provide a plot of the aligned data colored according to the cluster assignment and a plot of the warping functions colored according to the cluster assignment. Comment on the results.

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPg`
`lUMFFTNE5HUDJaWlFKVUpORkNYWkFMUklZSi4u`