Course:	12741 Data Management		
Instructor:	Matteo Pozzi;	*TA*:	Fabricio Flores
HW#2	Posted on Friday Nov. 1	Due on **Thursday Nov. 8**	

Please upload an electronic version of your work on Canvas, and submit a hard copy to the instructor (in class or in his mail box).

1. [15 points]

Find in the internet the time history of the annual rainfall on the city of Pittsburgh from years 1960 to 2012.

a) Plot the time history, estimate mean and standard deviation. [5 points]

b) Apply linear regression analysis to that time history, and comment the results you find. [10 points]

2. [60 points]

File "WaterTempStrain.txt" reports the recordings or 3 sensors, during a one-year campaign, with a sampling period Δt of 144 minutes (corresponding to 10 measures per day). The file has 3651 rows.

Column 1 reports the water level in the soil under a building (in meters), Column 2 the temperature (in Celsius), and Column 3 the strain on a column of the building (in micro-strains).

You can process the measures by using the following model:

$$y_i = q_i + \gamma T_i + \delta W_i + n_i \quad \text{with} \quad q_i = \alpha\, t_i + \beta$$

where y indicates the strain measure, T the temperature measure, W the water level measure, q the actual strain component not depending on W or T, t the time, n the noise, and sub-script i refers each variables to the i^{th} recording, at time t_i. $\alpha, \beta, \gamma, \delta$ are the model parameters.

a) By using linear regression, find the best fitting parameter vector $\hat{\mathbf{w}} = [\hat{\alpha} \quad \hat{\beta} \quad \hat{\gamma} \quad \hat{\delta}]^{\mathrm{T}}$, specifying the corresponding physical unit. [10 points]

b) Using the model identified in question (a), infer q_{1500}, the strain at time t_{1500}. [10 points]

c) Using the model identified in question (a), predict $q_{@3\mathrm{y}}$ the actual strain component not depending on water level or temperature, at time 3 years, i.e. 2 years after the end of the monitoring campaign recorded in the database. [10 points]

d) The strain measures may contain some outliers. Plot the residuals obtained from the model fitted in part (a), and report the mean value and standard deviation of the residuals. [5 points]

e) By using the "Chauvenet's criterion", marked as outliers and remove from dataset all points for which the residual is more than 3 standard deviations above or below the mean. Repeat the regression analysis and the outlier removal until converge. During these iterations, report the value of the removed outliers and the position in the original dataset (i.e. the row number). Report the value of $\hat{\mathbf{w}}$ after having removed all outliers. [10 points]

f) Using the model identified in question (e), infer q_{1500} and predict again $q_{@3\mathrm{y}}$, as defined above, in questions (b-c). [10 points]

g) Suppose now you do not have access to the recording of the temperature sensor. Repeat the analysis, finding $\hat{\mathbf{w}}$, q_{1500} and $q_{@3\mathrm{y}}$. [5 points]

3. [25 points]

Consider the third column of the file “WaterTempStrain.txt”, which you have used in Ex.2, reporting the strain measures.

a) Apply a moving average smoother to that time history, using three value of p: $p = 1; 4; 20$ (corresponding to $n = 3; 9; 41$ respectively). If you can, use the time history after having removed the outliers, which you have got at question (e)in Ex.2. Plot the time histories of the smoothed data for each of the three values of p. Comment on the differences among the three time histories. [15 points]

b) Report the value at row 1500, for each time history. Do you think that the smoothed datum is a good estimation of q_{1500}, as defined in Ex.2? Support your claim. [10 points]