# MA324 : TP1

You may use TP1.py using your favorite Python IDE (e.g. spyder).

**Exercice 1** *(Simulating data)*
The first step to try the formula seen during the lecture 1 is to simulate data.

1. Using the numpy/linspace create an 1D array $x$ that contains 100 samples uniformly distributed on $[0, 1]$
2. Simulate the noisy linear model say $y = 10 + 2x + \varepsilon$, where $\varepsilon$ contains iid realizations of a random uniformly distributed random variable on $[-1, 1]$. You may use a numpy/random function.
3. Using the code given in TP1.py check graphically that your simulation is indeed correct

**Exercice 2** *(Computing the LS estimator $\beta_1$ and $\beta 2$)*

1. Compute $C_{XY}$ and $\sigma_X^2$ as in lecture 1. Hint : You may use a dot products.
2. Compute $\beta_1$ and $\beta_2$ as in lecture 1 and check graphically that the LS model $\beta_1 + \beta_2 x$ fits the simulated data. The code for the plot is given.
3. What are the predicted value of the model for $x = 1/2$, $x = 0$, $x = 1$. Is this value plausible knowing the true underlying model. What happens if you increase/decrease the number of observed samples ?
4. During lecture 1, one of you asked if it was possible to compute the above estimators $\beta_1$ and $\beta_2$ even if the underlying model of the data is not linear. Try with the non-linear model $y = 10 + 2x^2 + \varepsilon$ on say $[-1, 1]$. In this case, what is the predicted value of the model for $x = 0$. Is this value plausible known the true underlying model ? In this case, what happens if you increase/decrease the number of observed samples ?

**Exercice 3** *(Model estimation)*

1. Simulate the model as in ex1, but change the distribution of the noise $\varepsilon$ for iid realizations of $\mathcal{N}(0, 1)$
2. We would like to see if the estimator $\beta_2$ is biased or not [1] and to have some intuition of the answer we wish to do some simulations. The simplest way to do so is to repeat the estimations with different noise realizations and draw an histogram. Write a code that stores the different values for the estimated $\beta_2$. The code to display histograms is given.

**Exercice 4** *(Effect of outliers)*

1. We want to observe the effect of outliers ie points that differ significantly from the rest of observations. To do so, we generate data as in exercise 1 then randomly corrupt few samples. You may use `https://www.w3schools.com/python/ref_random_randint.asp` if you need.
2. Conclude on the effect of outliers in practice

1. Unbiased means that its expectation is equal to the true value $\beta_2^*$