3 Multiple Linear Regression

Exercise 3.1
We want to measure the sea level with respect to three reference locations A, B and C. Several measurements are observed and rearranged in a linear system $Ax = y$ where $x = (x_A, x_B, x_C)^T$ contains the levels of A, B and C :

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_A \\ x_B \\ x_C \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 2 \\ 1 \end{pmatrix}.$$

The solution x^* must minimize the shortcoming between Ax and y in the least squares sense, i.e., it must minimize $\| Ax - y \|_2^2$. Show that :

$$x^* = \frac{1}{4} \begin{pmatrix} 5 \\ 7 \\ 12 \end{pmatrix}.$$

Exercise 3.2
We want to approximate a cloud of points $(x_i, y_i)_{1 \leq i \leq n}$ (the x_i's are all distinct from each other) with a polynomial function of order equal or lesser than p. It is assumed that $p + 1 \leq n$. Let

$$P(x) = c_0 + c_1 x + c_2 x^2 + \ldots + c_p x^p$$

be a polynomial function. We are looking for the coefficients c_0, c_1, ..., c_p that minimize

$$E(c_0, c_1, \ldots, c_p) = \sum_{i=1}^{n} (P(x_i) - y_i)^2.$$

1. Write $E(c_0, c_1, \ldots, c_p)$ in the matrix form $\| Ac - y \|_2^2$ where you will define A, c and y.
2. Derive the normal equations.
3. Show that the matrix A is full rank.
4. Show that the matrix $A^T A$ is invertible since A is full rank.
5. Give the optimal solution of the minimization problem.

Exercise 3.3
The selling price of houses might be represented as a function of other variables. You have a file "prix-immobilier-notaire-agglomeration_cle245eb4.xls" with 215 rows of data. The content of this file is described in the companion file "description.txt" (three variables could be used to explain the mean selling price).

1. Read the data set with Pandas.

```
import pandas as pd
file='prix-immobilier-notaire-agglomeration_cle245eb4.xls'
df=pd.read_excel(file)
```

2. Make the simple linear regression of the price "Prix moyen par logement actualisé à 2006" with respect to the "Revenu imposable brut par ménage 2006". Plot the samples and the regression line. Is it a good approximation ?
   The variables can be obtained as follows :

```
x1=df['Revenu imposable brut par ménage 2006'].values
y=df['Prix moyen par logement actualisé à 2006'].values
```

3. Make the multiple linear regression of the price "Prix moyen par logement actualisé à 2006" with respect to the three regressor variables. Does it improve the approximation ?
4. Compute the three Variance Inflation Factors. Is it relevant to keep the three variables ?
5. Compute the confidence interval for each partial slope estimate at level $\alpha = 0.05 = 5\%$.
6. According to the F-test, is it relevant to use a regression model ? The threshold of the F-test can be computed by using "scipy.stats.f". It is the same principle as for the t-test in case of the simple linear regression.