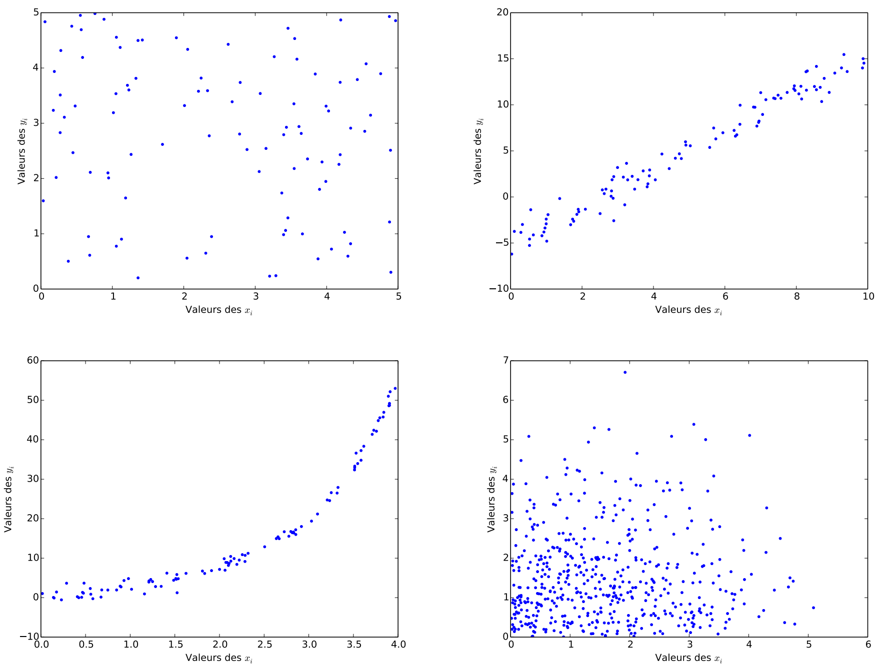
TP n° 07 – Corrélation, régression linéaire et méthode des moindres carré

I Introduction

On cherche à savoir si deux quantités physiques X et Y sont liées entre elles. Exemple : dans un circuit l'intensité I et la tension U.
On dit qu'il y a corrélation entre deux variables lorsqu'elles ont tendance à varier toujours dans le même sens (si X augmente, Y a tendance à augmenter) soit toujours dans le sens inverse (si X augmente, Y a tendance à diminuer).
Pour établir le lien possible entre X et Y, on effectue n mesures qui nous donnent n couples (x_i, y_i).
Objectif :

1. à partir de cet échantillon, on va quantifier la corrélation entre X et Y.
2. si la liaison est linéaire, c'est-à-dire si $Y = aX + b$, on va estimer l'équation de la droite.

Exemple : on a tracé le nuage de points pour quatre échantillons différents.
Pour quels exemples y'a-t-il liaison entre X et Y ? Et dans ce cas, la liaison est-elle linéaire, non-linéaire ?



II Coefficient de corrélation linéaire

II.1 La théorie

Définition. Soit X et Y deux variables. Le coefficient de corrélation linéaire est :

$$\rho(X, Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sigma(X)\sigma(Y)} \quad [1]$$

Propriété.

1. ρ est un nombre compris entre -1 et 1.
2. Il sert à quantifier la corrélation linéaire entre X et Y :
   (a) si $\rho > 0$, les variables ont tendance à varier dans le même sens
   (b) si $\rho < 0$, les variables ont tendance à varier dans le sens inverse
   (c) si $\rho = 0$, les variables ne sont pas corrélées
   (d) $\rho = \pm 1$ si et seulement si $Y = aX + b$ (si $\rho = 1, a \geqslant 0$, si $\rho = -1, a \leqslant 0$)

1. Pas d'inquiétude, nous n'avez pas besoin de savoir ce qu'est $\mathbb{E}(X)$ et $\sigma(X)$ pour faire la suite

Un estimateur de ρ est :

$$r = \frac{n\sum_{i=0}^{n-1} x_i y_i - \left(\sum_{i=0}^{n-1} x_i\right)\left(\sum_{i=0}^{n-1} y_i\right)}{\sqrt{n\sum_{i=0}^{n-1} x_i^2 - \left(\sum_{i=0}^{n-1} x_i\right)^2}\sqrt{n\sum_{i=0}^{n-1} y_i^2 - \left(\sum_{i=0}^{n-1} y_i\right)^2}}$$

II.2 La pratique

Exercice 1. Soit X une liste de nombres. Écrire une fonction `som` qui renvoie la somme des éléments de X :

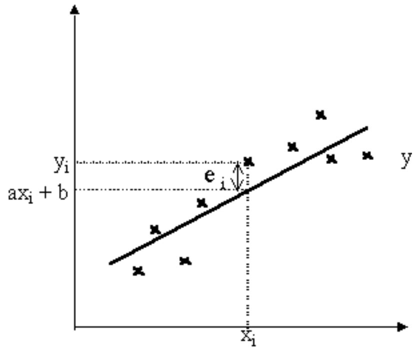
```
>>>X=[1,2.1,7]
>>>som(X)
10.1
```

Exercice 2.
1. Écrire une fonction `som_carre` qui renvoie la somme des carrés des éléments d'une liste.
2. Écrire une fonction `som_prod` qui a comme entrée deux listes de même longueur $[x_0, \cdots, x_{n-1}]$ et $[y_0, \cdots, y_{n-1}]$ et renvoie $\sum_{i=0}^{n-1} x_i y_i$.
3. Enfin écrire une fonction `coeff_corr` qui a comme entrée deux listes de même longueur $[x_0, \cdots, x_{n-1}]$ et $[y_0, \cdots, y_{n-1}]$ et renvoie l'estimation du coefficient de corrélation linéaire r. On utilisera les fonctions des questions précédentes.
4. Que renvoie `coeff_corr(X,X)` ? Pourquoi ?

III Méthode des moindres carrés

Si le coefficient de corrélation linéaire est "assez proche" de 1 ou -1, on cherche la droite qui passe "au plus près" du nuage de points. Pour ça, il faut se fixer un critère d'ajustement.

On projète chaque point (x_i, y_i) sur la droite parallèlement à (Oy) et on note ϵ_i l'écart obtenu.



Dans la méthode des moindres carrés, on choisit la droite qui minimise la somme des carrés des écarts : $\sum_{i=0}^{n-1} \epsilon_i^2$.

La droite obtenue est appelée "droite de régression de Y sur X".

Si la droite de régression s'écrit $y = ax + b$, de "bons" estimateurs de a et b sont :

$$\hat{a} = \frac{n\sum_{i=0}^{n-1} x_i y_i - \left(\sum_{i=0}^{n-1} x_i\right)\left(\sum_{i=0}^{n-1} y_i\right)}{n\sum_{i=0}^{n-1} x_i^2 - \left(\sum_{i=0}^{n-1} x_i\right)^2} \qquad \hat{b} = \frac{\left(\sum_{i=0}^{n-1} x_i^2\right)\left(\sum_{i=0}^{n-1} y_i\right) - \left(\sum_{i=0}^{n-1} x_i\right)\left(\sum_{i=0}^{n-1} x_i y_i\right)}{n\sum_{i=0}^{n-1} x_i^2 - \left(\sum_{i=0}^{n-1} x_i\right)^2}$$

Exercice 3. Écrire une fonction `dte_reg` qui a comme entrée deux listes de même longueur $[x_0, \cdots, x_{n-1}]$ et $[y_0, \cdots, y_{n-1}]$ et renvoie l'estimation de $\hat{a}$ et $\hat{b}$.

On utilisera les fonctions précédentes.
Testez la sur (X, X).

Exercice 4. Application :

On a les relevés suivants :

U (V)	1	2.1	2.9	4
I (mA)	0.99	2.06	3.05	3.98

Les quantités I et U sont-elles corrélées ? Si oui, quelle est l'équation de la droite de régression linéaire de U sur I ?
Que se passe-t-il si on échange I et U ? A votre avis, pourquoi ?

Correction TP n° 07 – Corrélation, régression linéaire et méthode des moindres carré

Solution 1.

```
def som(X):
    som=0
    for i in X:
        som=som+i
    return(som)
```

Solution 2.

1.
```
def som_carre(X):
    som_carre=0
    for i in X:
        som_carre=som_carre+i**2
    return(float(som_carre))
# on renvoie un flottant pour eviter ensuite de diviser par un entier
```

2.
```
def som_prod(X,Y):
    som_prod=0
    for i in range(len(X)):
        som_prod=som_prod+X[i]*Y[i]
    return(float(som_prod))
```

3.
```
def coeff_corr(X,Y):
    n=len(X)
    r=(n*som_prod(X,Y)-som(X)*som(Y))/sqrt(n*som_carre(X)-(som(X))**2)/
    sqrt(n*som_carre(Y)-(som(Y))**2)
    return(r)
```

4. On trouve `coeff_corr(X,X)=1` car X est évidemment fortement corrélé à X.

Solution 3.

```
def dte_reg(X,Y):
    n=float(len(X))
    a=(n*som_prod(X,Y)-som(X)*som(Y))/(n*som_carre(X)-(som(X))**2)
    b=(som_carre(X)*som(Y)-som(X)*som_prod(X,Y))/(n*som_carre(X)-(som(X))**2)
    return(a,b)
```

`dte_reg(X,X)` renvoie `(1,0)` car $X = 1 \times X + 0$.

Solution 4. On trouve un coefficient de corrélation linéraire égal à $\approx 0,99775$: les variables sont fortement corrélées.
Les coefficients de la droite de régression linéaire sont $a \approx 1,012$ et $b \approx -0.011$ Si on échange U et I, le coefficient de corrélation n'est pas modifié : le lien entre X et Y est le même que celui entre Y et X.
Par contre, la droite de régression n'est pas la même.
En effet, dans les formules, X et Y n'ont pas un rôle symétrique. Cela vient du fait qu'on projette parallèlement à l'axe Oy sur la droite de régression.
On a plusieurs choix pour la droite qui passe au plus près du nuage de points. Dans la méthode des moindres carrés présentée ici, X est dite la variable prédictive ou explicative et Y la variable à expliquer.