TP 7

Statistique descriptive bivariée

Compétences attendues.

- ✓ Représenter un nuage de points associé à une série statistique double.
- ✓ Représenter la droite des moindres carrés.
- ✓ Calculer le coefficient de corrélation linéaire et interpréter sa valeur.

Objectifs. Les données statistiques ne vont pas toujours toutes seules, et pour un même individu, il est possible de s'intéresser à plusieurs caractères. Dans ce TP, nous nous limiterons à l'étude simultanée de deux caractères. Nous nous poserons alors la question suivante : peut-on exprimer l'un de ces caractères en fonction de l'autre ? Plus précisément, l'un est-il une fonction affine de l'autre ? De cette recherche de correspondances peuvent découler des analyses fines, explicatives voire prédictives, ou au contraire mettre en évidence des absences de corrélation entre ces caractères.

Mathieu Mansuy - Professeur de Mathématiques en ECS2 au Lycée Louis Pergaud (Besançon)
mansuy.mathieu@hotmail.fr

1 Série statistique double ou bivariée

1.1 Définition

Soit $\Omega = \{\omega_1, \dots, \omega_n\}$ une population d'effectif n, sur laquelle nous étudions deux caractères quantitatifs $X, Y : \Omega \to \mathbb{R}$ avec X supposé non constant. Pour tout $i \in [\![1, n]\!]$, on note :

- $x_i = X(\omega_i)$ la modalité de X prise par l'individu ω_i,
- $y_i = Y(\omega_i)$ la modalité de Y prise par l'individu ω_i.

Définition.

On appelle *série statistique double (ou bivariée)* de la population Ω pour le couple de caractères (X, Y) la donnée du n-uplet $((x_i, y_i))_{1 \leq i \leq n}$ des modalités de (X, Y) sur Ω.

Exemple. 10 enfants de 6 ans d'une même classe sont mesurés et pesés. On note X la variable désignant la taille de l'enfant (en centimètres) et Y celle désignant le poids de l'enfant (en kilogrammes). On obtient la série statistique double suivante :

Enfant	1	2	3	4	5	6	7	8	9	10
X	121	123	108	118	111	109	114	103	110	115
Y	25	22	19	24	19	18	20	15	20	21

Représentation informatique. On représentera une série statistique double sur `Scilab` par deux vecteurs `x` et `y` de taille n, où `(X(i),Y(i))` est la modalité $(X, Y)(\omega_i)$.

1.2 Représentation graphique

On représente une série statistique double à l'aide d'un *nuage de points*. C'est l'ensemble des points M_i du plan de coordonnées (x_i, y_i) pour tout $1 \leq i \leq n$.

Pour tracer un nuage de points sur `Scilab`, on utilise l'instruction `plot2d`.

Définition.

Soient `x` et `y` deux vecteurs de même taille.

L'instruction `plot2d(x,y,-1)` trace le nuage de points dont les abscisses sont données par `x` et les ordonnées par `y`.

Remarque. L'option -1 a pour effet de ne pas relier les points. On peut également utiliser d'autres valeurs négatives $(-2, -3, \dots)$ pour changer la forme des points.

Exercice 1 (★)

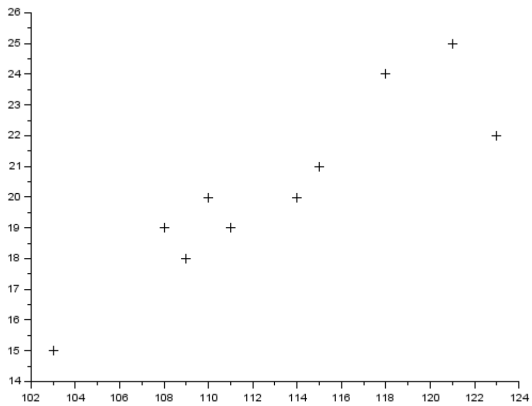
Représenter le nuage de points associé à la série statistique double proposée en exemple.

Les points sont-ils sur une même droite ? Si non, proposer une droite qui passe « très près » de tous ces points.

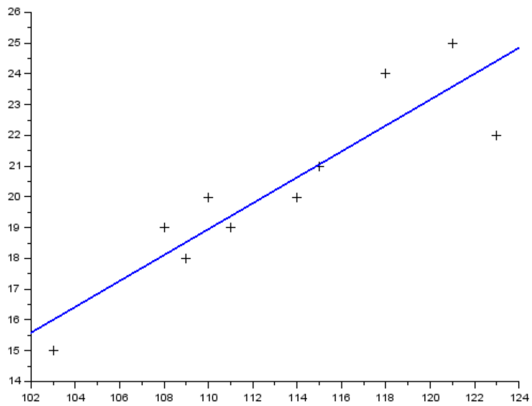
On utilise le code suivant pour représenter le nuage de points :

```
X = [121 123 108 118 111 109 114 103 110 115]
Y = [25 22 19 24 19 18 20 15 20 21]
plot2d(X,Y,-1)
```

On obtient :



Les points ne sont pas sur la même droite. On peut proposer approximativement la droite suivante qui semble passer « la plus près » de tous ces points :



2 Régression linéaire

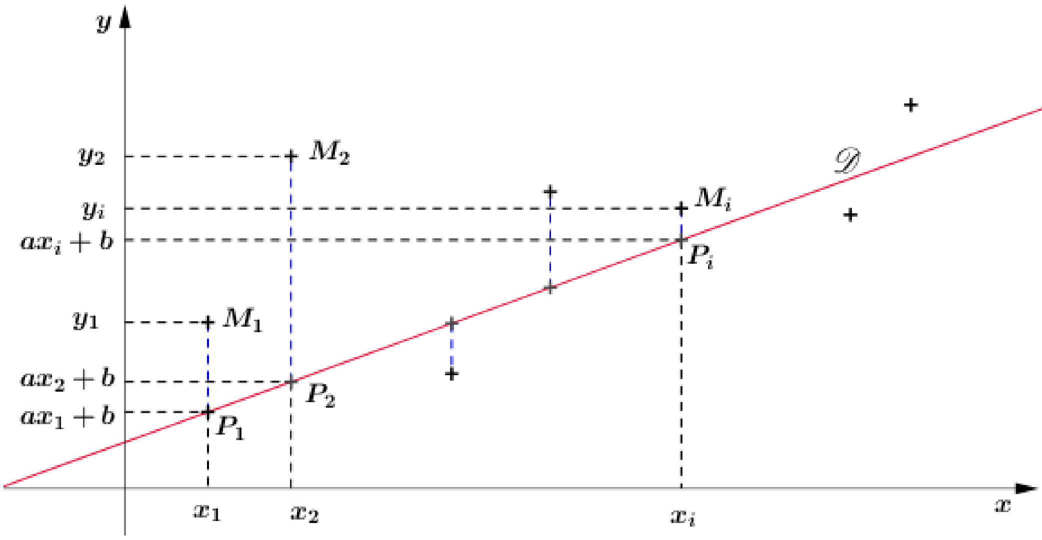
2.1 Méthode des moindres carrés

Dans cette section, on essaie d'expliquer Y (dit caractère *expliqué*) à partir de X (dit caractère *explicatif*). On étudie pour cela l'existence d'une relation linéaire pour la série statistique double $((x_i, y_i))$, c'est-à-dire qu'on souhaite « placer » tous les points M_i de coordonnées (x_i, y_i) sur une même droite $\mathscr{D}$. On cherche donc $(a, b) \in \mathbb{R}^2$ tel que :

$$\forall i = 1, \ldots, n, \quad y_i = ax_i + b.$$

Seulement, il y a très peu de chance qu'une telle droite existe, nos points n'étant très probablement pas alignés.

On va chercher la « meilleure » droite $\mathscr{D}$ approchant l'ensemble des points M_i au sens suivant : pour tout $1 \leq i \leq n$, on mesure la distance M_iP_i entre M_i et le point $P_i \in \mathscr{D}$ d'abscisse x_i.



On cherche $(a,b) \in \mathbb{R}^2$ rendant minimale la quantité[1] :

$$\sum_{i=1}^{n} M_iP_i^2 = \sum_{i=1}^{n}(y_i - ax_i - b)^2. \tag{$*$}$$

Une telle droite, si elle existe, est appelée *droite des moindres carrés.*

2.2 Existence et unicité de la droite des moindres carrés

On réécrit matriciellement ce problème. Notons $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ et $A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$.

Les points M_i sont tous sur la même droite si et seulement s'il existe $U = \begin{pmatrix} a \\ b \end{pmatrix} \in \mathscr{M}_{2,1}(\mathbb{R})$ tel que :

$$Y = AU.$$

Si cette équation admet une solution, c'est à dire si Y appartient à $\text{Im}(A)$, alors c'est bon : la droite $y = ax + b$ est solution du problème.

Sinon, et c'est ce cas qui est intéressant puisque les points ne sont pas alignés en général, on cherche à minimiser la quantité $(*)$ qui se réécrit :

$$\sum_{i=1}^{n}(y_i - ax_i - b)^2 = \|Y - AU\|^2.$$

On se retrouve dans la situation du cours de *recherche de pseudo-solutions d'un système linéaire* : on sait (si A est de rang 2, ce qu'on peut supposer car les x_i ne sont pas tous égaux) que ce problème admet une unique solution $U = \begin{pmatrix} a \\ b \end{pmatrix}$. On obtient donc le théorème suivant.

Théorème 1 (Problème des moindres carrés : Régression linéaire)

Considérons une série statistique double $((x_i, y_i))_{1\leq i\leq n}$.

Il existe une et une seule droite minimisant la quantité $(*)$. On l'appelle *la droite des moindres carrés associée à la série statistique double* $((x_i, y_i))_{1\leq i\leq n}$.

[1]Pourquoi vouloir minimiser cette quantité en particulier ? Pourquoi pas une autre, comme par exemple la somme des longueurs, ou encore la plus grande des longueurs ? Une des raisons pour lesquelles on s'intéresse à la somme des carrés est qu'on dispose alors d'un résultat garantissant l'existence et l'unicité d'une telle droite, comme on va le voir dans la section suivante.

Remarque. La droite des moindres carrés est la droite qui passe « la plus près » de tous les points du nuage de points au sens des moindres carrés (c'est-à-dire au sens où elle minimise la quantité $(*)$).

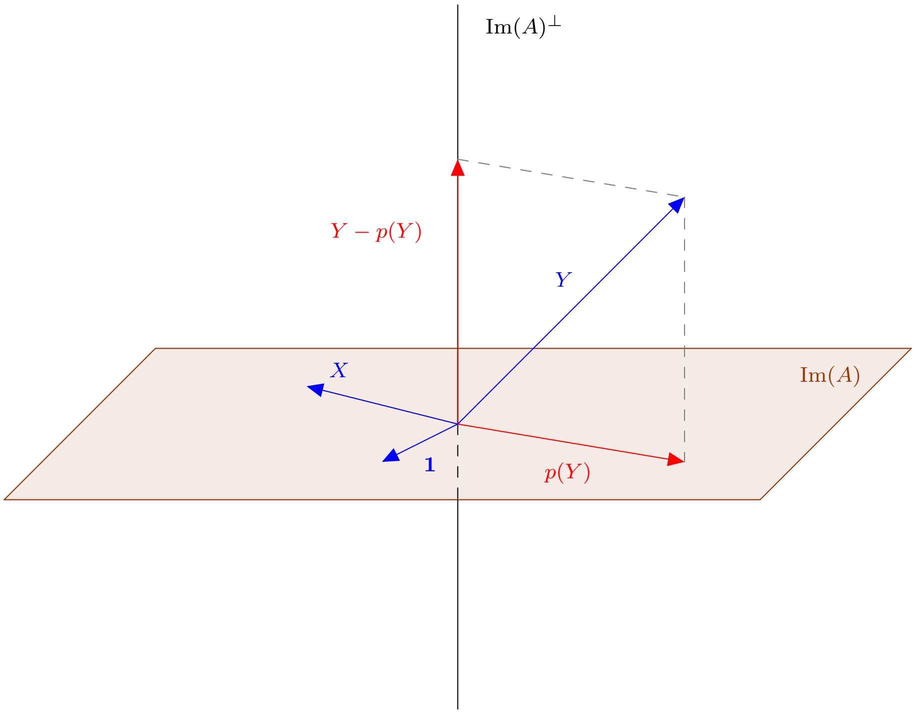
2.3 Détermination de la droite des moindres carrés

Précisons le théorème précédent. On sait que ce minimum est atteint en un unique point $p(Y)$:

$$\min_{V\in\mathscr{M}_{2,1}(\mathbb{R})} ||Y-AV|| = ||Y-p(Y)||$$

où p est la projection orthogonale sur le sous-espace F de $\mathscr{M}_{n,1}(\mathbb{R})$ suivant :

$$F=\{AV,\ V\in\mathscr{M}_{2,1}(\mathbb{R})\}=\mathrm{Im}(A)=\mathrm{Vect}\left(A\begin{pmatrix}1\\0\end{pmatrix},A\begin{pmatrix}0\\1\end{pmatrix}\right)=\mathrm{Vect}\left(\begin{pmatrix}x_1\\\vdots\\x_n\end{pmatrix},\begin{pmatrix}1\\\vdots\\1\end{pmatrix}\right).$$



On cherche donc $U\in\mathscr{M}_{2,1}(\mathbb{R})$ tel que $p(Y)=AU$. On a :

$$\begin{aligned}
Y-p(Y)\in Im(A)^{\perp}=Vect\left(\begin{pmatrix}x_1\\\vdots\\x_n\end{pmatrix},\begin{pmatrix}1\\\vdots\\1\end{pmatrix}\right)^{\perp} &\Leftrightarrow \begin{cases}<X,Y-p(Y)>=0\\<\mathbf{1},Y-p(Y)>=0\end{cases}\\
&\Leftrightarrow \begin{cases}{}^tX\times(Y-p(Y))=0\\{}^t\mathbf{1}\times(Y-p(Y))=0\end{cases}\\
&\Leftrightarrow \begin{cases}\begin{pmatrix}x_1 & \dots & x_n\end{pmatrix}\times(Y-p(Y))=0\\\begin{pmatrix}1 & \dots & 1\end{pmatrix}\times(Y-p(Y))=0\end{cases}\\
&\Leftrightarrow \begin{pmatrix}x_1 & \dots & x_n\\1 & \dots & 1\end{pmatrix}\times(Y-p(Y))=\begin{pmatrix}0\\0\end{pmatrix}\\
&\Leftrightarrow {}^tA(Y-p(Y))=0_{2,1}\\
&\Leftrightarrow {}^tAp(Y)={}^tAY\\
&\Leftrightarrow {}^tA\times AU={}^tAY
\end{aligned}$$

D'où finalement si ${}^tA \times A \in \mathscr{M}_2(\mathbb{R})$ est inversible, la solution au sens des moindres carrés est :

$$U = ({}^tA \times A)^{-1} \times^t AY.$$

Cette formule n'est pas au programme d'ECS.

Remarque. La matrice $({}^tA \times A)$ est bien inversible car :

$${}^tAA = \begin{pmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} = \begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix}$$

et donc $\det({}^tA \times A) = n \sum x_i^2 - (\sum x_i)^2 > 0$. En effet par l'inégalité de Cauchy-Schwarz, on a :

$$(\sum x_i \times 1)^2 \leq \left(\sum 1\right) \times \left(\sum x_i^2\right) = n \sum x_i^2.$$

De plus, on a égalité si et seulement si X et $\mathbf{1}$ sont colinéaires, c'est à dire si les x_i sont égaux, ce qui n'est pas le cas par hypothèse.

Exercice 2 (★★)

Reprenons la série statistique de l'exemple de départ.

1. Appliquer la méthode qui précède afin d'obtenir les valeurs de a et b.
2. Représenter la droite des moindres carrés sur le même graphique que le nuage de points. Le résultat est-il conforme à vos attentes ?
3. Quel est le signe du coefficient directeur de cette droite ? Comment l'interprétez vous ?

1. On doit tout d'abord construire la matrice $A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$. On peut le faire à l'aide de la commande (`X'` désigne la transposée du vecteur ligne `X`) :

```
A = [X', ones(length(X),1)]
```

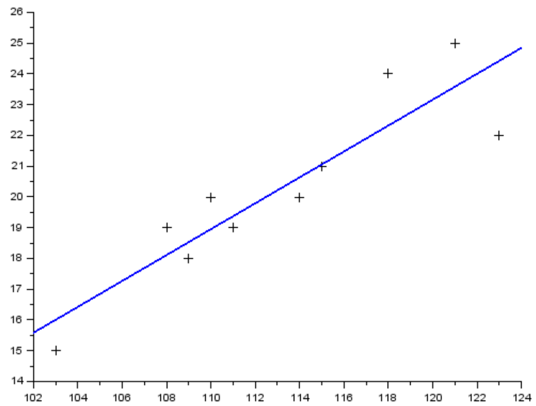
   On calcule ensuite la matrice $U = ({}^tA \times A)^{-1} \times^t AY$ permettant d'obtenir a et b pour construire la droite des moindres carrés :

```
B = A'*A
U = inv(B)*(A')*Y'
a = U(1); b = U(2);
```

   Reste alors à la tracer avec le nuage de points :

```
function y = f(x)
    y = a*x+b
endfunction

x = linspace(102,124,10)
fplot2d(x,f,2)
```

On obtient :



ce qui correspond (environ) à la droite à laquelle on pensait.

2. Le coefficient directeur est de signe positif (la droite est croissante), ce qu'on peut interpréter par le fait que les caractères quantitatifs X et Y ont tendance à varier dans le même sens. Cela parait conforme à l'intuition, le poids et la taille aillant tendance à varier effectivement dans le même sens.

On peut également obtenir l'équation de la droite des moindres carrés à l'aide de la commande `Scilab` suivante.

Définition.

Soient `x`, `y` des vecteurs de même taille, `x` ayant au moins deux coefficients distincts.

La commande `[a,b] = reglin(x,y)` renvoie deux réels `a,b` tels que $y = ax + b$ est l'équation de la droite des moindres carrés pour la série statistique double $((x_i, y_i))$.

Le saviez vous ?

Cérès est la plus petite planète naine du système solaire. Avec un diamètre d'environ 950 kilomètres, il s'agit de l'objet le plus grand et le plus massif de la ceinture d'astéroïdes située entre les orbites des planètes Mars et Jupiter. Elle fut découverte le 1^{er} janvier 1801 par Giuseppe Piazzi, astronome italien, qui pu suivre sa trajectoire jusqu'au 14 février 1801, date à laquelle l'astéroïde s'approcha trop près du Soleil pour continuer à être observé. Difficile alors pour les astronomes de l'époque de prédire la position exacte de Cérès sur la base des seules observations de Piazzi afin de confirmer sa découverte.

Afin de retrouver l'astéroïde, Carl Friedrich Gauss (mathématicien allemand, 1777 - 1855) développa la méthode des moindres carrés permettant de comparer les données expérimentales de Piazzi au modèle mathématique censé décrire ces données. Il obtient ainsi un résultat suffisamment précis pour permettre à Zach, un astronome allemand, de localiser à nouveau Cérès à la fin de l'année 1801.

La méthode des moindres carrés de Gauss, dont la régression linéaire est un exemple d'application, ne fut publiée qu'en 1809. Elle fut indépendamment élaborée par le mathématicien français Adrien - Marie Legendre (1752 - 1833) en 1805.

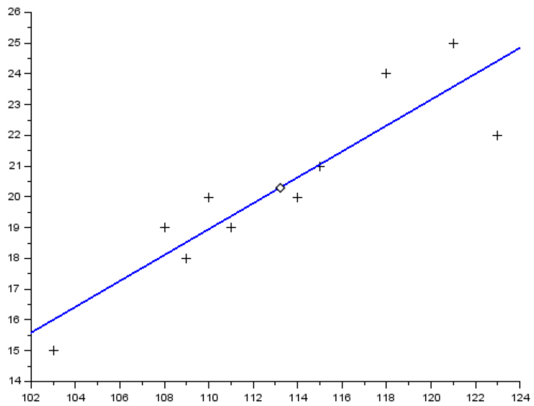
2.4 Covariance, corrélation linéaire

Soient $x = (x_1, \dots, x_n)$ et $y = (y_1, \dots, y_n)$ deux séries statistiques. On rappelle que la *moyenne* de la série statistique x est donnée par :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

On appelle *point moyen* de la série statistique double $((x_i, y_i))$ le point de coordonnées $(\overline{x}, \overline{y})$.

Exercice 3 (★)
Représenter le point moyen dans notre exemple (on pourra utiliser la commande `plot2d(·, ·, -5)` pour le différentier des autres points). Que constate-t-on ?

> On représente le point moyen à l'aide des instructions suivantes :
>
> ```
> xm = mean(X)
> ym = mean(Y)
> plot2d(xm,ym,-5)
> ```
>
> On obtient le graphe suivant :
>
> 
>
>
> On remarque que le point moyen se trouve sur la droite des moindres carrés. On verra que c'est en fait toujours le cas. Ce qui peut permettre de tracer la droite des moindres carrés plus facilement.

Définition.
On définit :

- l'*écart-type* de la série statistique x par :

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

- la *covariance* (empirique) de la série statistique double $((x_i, y_i))$ par :

$$Cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

Exercice 4 (★ - Formules de Huygens)

Montrer que :

$$\sigma_x^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2 \quad \text{et} \quad Cov(x,y) = \overline{xy} - \overline{x}\cdot\overline{y}.$$

> Il suffit de développer. À vous de jouer !

Propriété 2 (Équation de la droite des moindres carrés - Hors Programme)

Soit $((x_i, y_i))$ une série statistique double. La droite des moindres carrés a pour équation :

$$y - \overline{y} = \frac{Cov(x,y)}{\sigma_x^2}(x - \overline{x}).$$

En particulier, elle passe par le point moyen $(\overline{x}, \overline{y})$.

Exercice 5 (★★★)

Montrer ce résultat en effectuant le calcul $U = ({}^tA \times A)^{-1} \times {}^tAY$.

> On a déjà fait une partie du calcul :
>
> $${}^tAA = \begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix} \quad \text{et} \quad \det({}^tA \times A) = n\sum x_i^2 - (\sum x_i)^2 = n^2\sigma_x^2$$
>
> par la formule de Huygens. On obtient :
>
> $$({}^tAA)^{-1} = \frac{1}{n^2\sigma_x^2}\begin{pmatrix} n & -\sum x_i \\ -\sum x_i & \sum x_i^2 \end{pmatrix} = \frac{1}{n\sigma_x^2}\begin{pmatrix} 1 & -\overline{x} \\ -\overline{x} & \overline{xx} \end{pmatrix}.$$
>
> D'autre part, on a :
>
> $${}^tAY = \begin{pmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix} = n\begin{pmatrix} \overline{xy} \\ \overline{y} \end{pmatrix}.$$
>
> D'où finalement :
>
> $$U = ({}^tAA)^{-1}\,{}^tAY = \frac{1}{\sigma_x^2}\begin{pmatrix} Cov(x,y) \\ \overline{xx}\cdot\overline{y} - \overline{x}\cdot\overline{xy} \end{pmatrix}.$$
>
> On obtient bien $a = \dfrac{Cov(x,y)}{\sigma_x^2}$ et :
>
> $$b = \frac{\overline{xx}\cdot\overline{y} - \overline{x}\cdot\overline{xy}}{\sigma_x^2} = \frac{\overline{xx}\cdot\overline{y} - \overline{x}^2\cdot\overline{y} + \overline{x}^2\cdot\overline{y} - \overline{x}\cdot\overline{xy}}{\sigma_x^2} = \overline{y} - \overline{x}\frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sigma_x^2} = \overline{y} - \overline{x}\frac{Cov(x,y)}{\sigma_x^2}$$

Définition.

On appelle *coefficient de corrélation linéaire* de x et y le réel définit par :

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x\sigma_y}.$$

Propriété 3

Le coefficient de corrélation linéaire vérifie les propriétés suivantes :

- $|\rho_{x,y}| \leq 1$;
- $\rho_{x,y} = \pm 1$ si et seulement s'il existe a et b tels que $Y = aX + b$. Dans ce cas, le signe de a est le même que celui de $\rho_{x,y}$.

Exercice 6 (★★)

À l'aide de l'inégalité de Cauchy-Schwarz dans $\mathbb{R}^n$, démontrer la propriété précédente.

> Ca a déjà été fait en cours, à vous de jouer !

À retenir.

- Un coefficient de corrélation linéaire proche de ± 1 indique que la droite des moindres carrés approche bien le nuage de points.

  Si de plus il est positif, c'est que x_i « a tendance » à augmenter lorsque y_i augmente, alors que s'il est négatif, x_i diminue lorsque y_i augmente.
- Une corrélation linéaire proche de 0 indique une absence de relation de dépendance **linéaire** entre x et y.

En général, on estime que la corrélation linéaire entre les séries X et Y est forte quand $|\rho_{X,Y}| \geq 0,9$. Lorsque c'est le cas, la droite des moindres carrés va permettre de faire des prédictions.

On peut utiliser les commandes `Scilab` suivantes pour calculer le coefficient de corrélation linéaire.

Définition.

Soient `x`, `y` des vecteurs de même taille n.

- La commande `stdev(x)` renvoie l'écart-type de la série statistique `x`.
- La commande `corr(x,y,1)` renvoie la covariance de la série statistique double $((x_i, y_i))$.
- La commande `correl(x,y,eye(n,n))` renvoie le coefficient de corrélation linéaire de la série statistique double $((x_i, y_i))$.

Remarque. La matrice `eye(n,n)` indique l'effectif de chaque modalité de la série double, en l'occurrence 1 pour la modalité (x_i, y_i) et 0 pour (x_i, y_j) lorsque $i \neq j$.

Exercice 7 (★)

1. Calculer le coefficient de corrélation linéaire pour notre série statistique double. Cela correspond-il à ce que vous vous attendiez ?

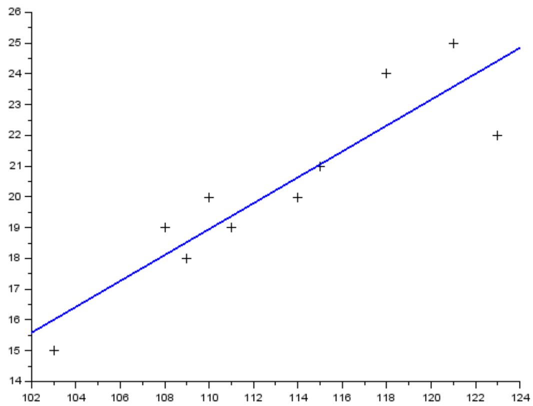
2. Estimer graphiquement le poids d'un enfant de 6 ans qui mesure 1m20.

1. On calcule le coefficient de corrélation linéaire de notre série statistique double grâce à la commande :

```
n = length(X)
disp(correl(X,Y,eye(n,n)))
```

   On obtient `0.9001375`. La corrélation linéaire entre les séries X et Y est donc forte. Ca correspond à ce qu'on s'attendait puisqu'une dépendance linéaire forte était prévisible étant donné l'allure du nuage de points.

2. Étant donné cette corrélation linéaire forte, notre modèle linéaire est justifié. Ce qui nous permet de faire des analyses prédictives à partir du graphe :



   On peut ainsi estimer qu'un enfant de 6 ans mesurant 1m20 pèse 23kg.

Exercice 8 (★ - Sensibilités aux valeurs extrêmes)
Le petit Yanis arrive en retard en classe ce matin là. Il prétend mesurer 105 cm et peser 27 kg.

1. Calculer le coefficient de corrélation linéaire pour la série statistique double associée aux 11 enfants. Comparer cette valeur avec celle obtenue dans l'exercice précédent.
2. Représenter le nouveau nuage de points, en distinguant les points correspondants aux 10 enfants avec celui représentant Yanis.
3. Tracer la droite des moindres carrés correspondant à la nouvelle série statistique. Que constatez vous ?

1. On calcule le nouveau coefficient de corrélation linéaire :

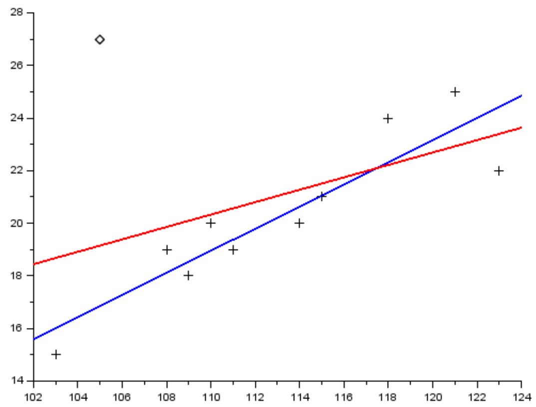
```
X1 = [X,105]
Y1 = [Y,27]

n = length(X1)
disp(correl(X1,Y1,eye(n,n)))
```

On obtient cette fois un coefficient de corrélation linéaire de `0.4412470`. Avec ce dernier point, le coefficient de corrélation linéaire a été divisé par 2 ! On peut douter de ce qu'affirme Yanis. On notera aussi qu'un seul point très éloigné de la droite des moindres carré a une grosse influence sur le coefficient de corrélation linéaire.

2. et 3. On utilise pour cela les instructions suivantes :

```
//Nuage de points
plot2d(X,Y,-1)
plot2d([105],[27],-5)

//Regression linéaire pour 10 élèves
[a,b] = reglin(X,Y)

function y = f(x)
    y = a*x+b
endfunction

x = linspace(102,124,10)
fplot2d(x,f,2)

//Regression linéaire pour 11 élèves
[a1,b1] = reglin(X1,Y1)

function y = f(x)
    y = a1*x+b1
endfunction

x = linspace(102,124,10)
fplot2d(x,f,5)
```

On obtient le graphe suivant, avec en bleu la régression linéaire pour les 10 élèves, et en rouge celle pour les 11 élèves :



On remarque que cette valeur erronée a beaucoup fait dévier la droite des moindres carrés, et rend la régression linéaire inadaptée et inutilisable pour faire des prévisions éventuelles.

À retenir.
La méthode des moindres carrés est très sensible aux valeurs extrêmes : une seule valeur très éloignée de la droite des moindres carrés a une « grosse » influence sur le coefficient de corrélation linéaire et également sur la position de la droite des moindres carrés. Pour y remédier, il peut être avantageux d'exclure au préalable les valeurs aberrantes des séries statistiques avant d'en faire l'étude.

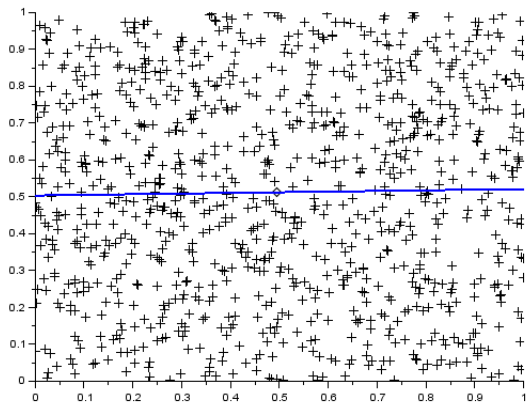
Exercice 9 (★★ - Indépendance et corrélation linéaire)

1. (a) Créer deux vecteurs `x` et `y` contenant chacun 1000 nombres tirés au hasard suivant une loi uniforme sur $[0, 1]$.
   (b) Représenter le nuage de points ainsi que le point moyen et la droite des moindres carrés (à l'aide de la commande `reglin`). Qu'en pensez vous ?
   (c) Calculer le coefficient de corrélation linéaire. Comment expliquer le résultat obtenu ?
2. On pose à présent `x = grand(1,100,'unf',-1,1)` et `y = x.^2`.
   (a) Représenter le nuage de points associés à ces séries statistiques ainsi que la droite des moindres carrés (à l'aide de la commande `reglin`). Qu'en pensez vous ?
   (b) Calculer le coefficient de corrélation linéaire. Les variables `x` et `y` sont-elles indépendantes ?

1. On utilise le code suivant :

```
//Nuage de points
x = rand(1,1000) ; y = rand(1,1000) ;
plot2d(x,y,-1)

//Point moyen
xm = mean(x) ; ym = mean(y) ;
plot2d(xm,ym,-5)

//Régression linéaire
[a,b] = reglin(x,y)

function y = f(t)
    y = a*t+b
endfunction

t = linspace(0,1,10)
fplot2d(t,f,2)

//Coefficient de corrélation linéaire
disp(correl(x,y,eye(1000,1000)))
```

On obtient le graphe suivant :



La droite de régression linéaire est quasi horizontale. Elle ne semble pas du tout adaptée pour approximer notre nuage de points, il ne semble pas y avoir de corrélation linéaire entre les deux séries statistiques.

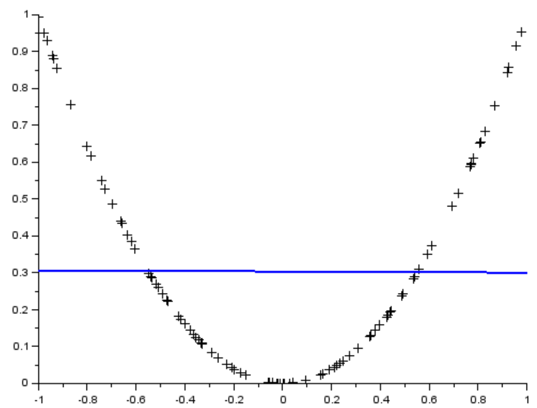
Le coefficient de corrélation linéaire obtenu est `0.0181605`. Il confirme l'indépendance linéaire qu'on avait constaté auparavant. Il explique également pourquoi la droite de régression linéaire est horizontale. Rappelons en effet que le coefficient directeur de cette droite fait intervenir la covariance, qui dans notre cas est voisin de 0.

Tous ces résultats ne sont pas étonnants : les séries statistiques `x` et `y` ont été choisies de manière indépendante l'une de l'autre, la fonction `rand` générant des réalisations indépendantes de la loi $\mathscr{U}([0,1])$. Elles sont en particulier linéairement indépendantes, ce qui implique que leur coefficient de corrélation linéaire est proche de 0.

2. On adapte le code précédent :

```
//Nuage de points
x = grand(1,100,'unf',-1,1) ; y = x.^2 ;
plot2d(x,y,-1)

//Régression linéaire
[a,b] = reglin(x,y)

function y = f(t)
    y = a*t+b
endfunction

t = linspace(-1,1,10)
fplot2d(t,f,2)

//Coefficient de corrélation linéaire
disp(correl(x,y,eye(100,100)))
```

On obtient le graphe suivant :



La droite de régression linéaire est là aussi quasi horizontale -ce qui indique un coefficient de corrélation linéaire proche de 0) et ne semble également pas du tout adaptée pour approximer notre nuage de points. Le coefficient de corrélation linéaire obtenu le confirme : il est de `-0.0056861`.

On peut conclure que les séries statistiques `x` et `y` sont linéairement indépendantes. Mais attention à ne pas conclure à l'indépendance de ces séries : elles ne le sont clairement pas puisque `y = x.^2`.

À retenir.
Si deux séries statistiques proviennent de caractères indépendants, alors le coefficient de corrélation linéaire est proche de zéro. À l'inverse, un coefficient de corrélation linéaire proche de zéro n'assure en rien l'indépendance des deux caractères étudiés : il indique seulement une indépendance **linéaire** entre ceux-ci.

Exercice 10 (★)
Considérons les séries statistiques `x` et `y` suivantes :

```
--> x = 1:40 ;
--> y = log(x)+grand(1,40,'unf',-1,1) ;
```

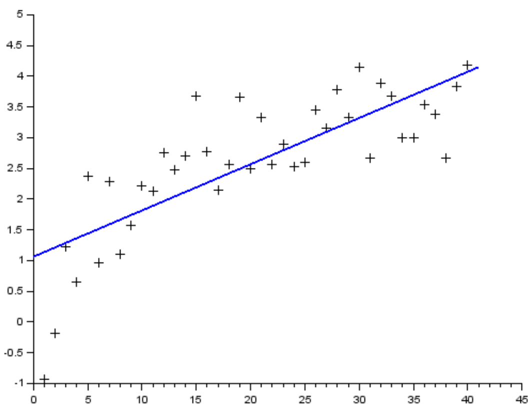
1. Représenter le nuage de points associé.
2. (a) Calculer le coefficient de corrélation linéaire. Semble-t-il bon ?
   (b) Calculer les coefficients a et b de l'équation de la droite de régression de `y` par rapport à `x`.
   (c) Superposer la droite de régression linéaire au nuage de points.
3. Vérifier que le nuage de points se superpose bien avec courbe représentative de la fonction $f : x \mapsto \ln(x)$.
4. Étudier la corrélation linéaire de `y` par rapport à `log(x)`.

1. et 2. On utilise le code suivant :

```
//Nuage de points
x = 1:40
y = log(x)+grand(1,40,'unf',-1,1)
plot2d(x,y,-1)

//Coeff de correl lin
disp(correl(x,y,eye(40,40)))

//Droite de rég lin
[a,b] = reglin(x,y)
function y = f(t)
    y = a*t+b
endfunction

t = linspace(0,41,10)
fplot2d(t,f,2)
```

On obtient `0.7797028` comme coefficient de corrélation linéaire, bien en deçà de la barre des 0.9. La corrélation linéaire entre les deux séries statistiques n'est donc pas bonne.

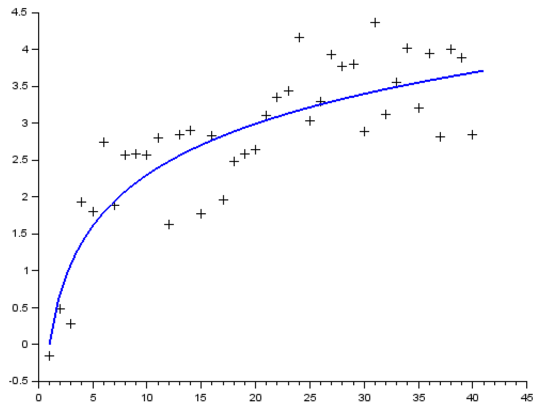
On obtient la représentation suivante du nuage de points et de la droite de régression linéaire associée :



3. Avec le code :

```
//Nuage de points
x = 1:40
y = log(x)+grand(1,40,'unf',-1,1)
plot2d(x,y,-1)

//Courbe du log
function y = f(t)
    y = log(t)
endfunction

t = linspace(1,41,50)
fplot2d(t,f,2)
```

on obtient :



Le nuage de point se superpose effectivement bien avec la courbe représentative du logarithme.

4. En exécutant la commande `disp(correl(log(x),y,eye(40,40)))`, on obtient `0.8864912` comme coefficient de corrélation linéaire. C'est proche des 0.9, ce qui confirme une forte corrélation linéaire entre `log(x)` et `y`.

À retenir.
Les relations entre les caractères X et Y ne sont pas nécessairement linéaires, elles peuvent être logarithmiques, exponentielles, ... L'étude faite dans ce TD peut cependant nous aider pour des régressions non linéaires, en étudiant la corrélation linéaire entre Y et $\ln(X)$, $\exp(X)$, ...

Exercice 11 (★ - Corrélation et causalité)

Une bonne corrélation entre deux séries de données ne signifie pas pour autant qu'il existe un lien de cause à effet entre les deux. À titre d'exemple, considérons la série statistique suivante :

Année	1996	1997	1998	1999	2000
Morts	15.85	15.7	15.39	15.32	14.85
Importations de citrons	230	280	360	410	525

Ce tableau donne le nombre de morts (pour un million d'habitants) sur les autoroutes américaines, ainsi que le nombre de tonnes de citrons mexicains importés aux États-Unis de 1996 à 2000.

Calculer le coefficient de corrélation linéaire pour cette série double. En déduisez vous une information pertinente ?

> On obtient avec le code :
> Avec le code :
>
> ```
> m = [15.85 15.7 15.39 15.32 14.85]
> i = [230 280 360 410 525]
> disp(correl(m,i,eye(5,5)))
> ```
>
> un coefficient de corrélation linéaire de `-0.9951611`. Ces deux séries statistiques sont donc fortement corrélées linéairement. Pourtant il semble bien difficile d'envisager un lien rationnel entre le nombre de morts sur les routes et les importations de citrons mexicains. Le travail du mathématicien s'arrête donc au calcul de la corrélation, l'exploitation de ces chiffres nécessite d'autres compétences !

Mise en garde.

Attention donc à l'erreur courante, notamment dans les médias, qui est de croire qu'un coefficient de corrélation linéaire élevé (en valeur absolue) induit une relation de causalité entre les deux phénomènes mesurés. Voir à ce sujet cette page des Décodeurs du `monde.fr` présentant un outil de corrélation géographique sur la base de données sans rapport, de manière à générer « vos propres cartes pour ne rien démontrer du tout ». Vous y apprendrez par exemple que la consommation de fromage est fortement corrélée au nombre de licences de football. Mais ce n'est pas pour autant que les footballeurs mangent plus de fromage.