ANNEXE

C

RÉGRESSION LINÉAIRE

A - Vocabulaire fondamental en statistiques

Vocabulaire général

1 ▹ Une ***population*** (statistique) est un ensemble fini d'objets sur lesquels porte une étude statistique ; ses éléments sont appelés ***individus***.

Le nombre d'individus d'une population est appelé ***effectif total***.

Une partie de cette population est appelée un ***échantillon*** de la population.

2 ▹ Un ***caractère*** est une grandeur que l'on observe sur les individus d'une population. Un tel caractère peut être ***quantitatif*** s'il est numérique (âge, taille, poids,...) ou ***qualitatif*** s'il donne une propriété (non numérique) de chaque individu (ville de naissance, couleur des cheveux,...).

Les valeurs prises par un caractère x s'appellent les ***modalités*** de ce caractère.

Pour décrire un caractère, on peut donner la liste $x = (x_1, x_2, \dots, x_p)$ des modalités x_k de x. Par exemple, le groupe sanguin est un caractère qualitatif x dont les différentes modalités sont A, B, AB et O :

$$x = (\mathrm{A}, \mathrm{B}, \mathrm{AB}, \mathrm{O}).$$

Lorsque ces modalités sont trop nombreuses, on les regroupe en classes c'est-à-dire en intervalles disjoints et complémentaires. On peut toujours écrire $x = (x_1, x_2, \dots, x_p)$ mais, cette fois-ci, les x_k sont les intervalles correspondant aux classes de modalités. Par exemple, si l'on considère la population formée par les communes de France, alors le nombre x d'habitants de chaque commune est un caractère dont il est pertinent de regrouper les modalités par classes, par exemple de la façon suivante :

$$\big(]0, 1000],]1000, 5000],]5000, 50000],]50000, 200000],]200000, +\infty[\big).$$

Notons que lorsque le caractère est quantitatif, on ordonne les x_k.

Effectifs et fréquences

On appelle ***effectif*** de la modalité x_k, le nombre n_k d'individus pour lesquels le caractère est égal à x_k (la somme des effectifs est égale à l'effectif total).

L'***effectif cumulé*** de la modalité x_k est le nombre $n_1 + \cdots + n_k$ d'individus pour lesquels le caractère est inférieur ou égal à x_k.

La ***fréquence*** de la modalité x_k est le nombre $f_k = \frac{n_k}{\mathrm{N}}$ c'est-à-dire l'effectif de x_k divisé par l'effectif total.

La ***fréquence cumulée*** de la modalité x_k est le nombre $f_1 + \cdots + f_k$.

Une ***série statistique*** peut se visualiser à l'aide d'un ***diagramme en bâtons*** ou d'un ***histogramme*** :

- le diagramme en bâtons est une suite de rectangles (de largeur fine) dont les hauteurs correspondent aux différentes fréquences ;
- un histogramme est une suite de rectangles dont les aires correspondent aux différentes fréquences (pertinent pour des données rangées par classes).

Caractéristiques de position et de dispersion

1 ▹ On appelle ***mode*** d'une série statistique une modalité de x dont l'effectif est maximal.

Sur un diagramme en bâtons, cela correspond au bâton le plus haut (ou à l'un d'entre eux si plusieurs sont à la même hauteur).

Dans le cas de valeurs regroupées en classes, une ***classe modale*** est une classe dont l'effectif divisé par l'amplitude de la classe est maximal.

Sur un histogramme, cela correspond au rectangle de plus grande aire (ou à l'un d'entre eux).

2 ▹ La ***moyenne*** de la série statistique $x = (x_1, \ldots, x_p)$, notée $\overline{x}$, est définie par :

$$\overline{x} = \frac{n_1 x_1 + \cdots + n_p x_p}{n_1 + \cdots + n_p} = \frac{1}{\mathrm{N}} \sum_{k=1}^{p} n_k x_k = \sum_{k=1}^{p} f_k x_k.$$

Dans le cas d'une séries de valeurs regroupées par classes, on remplace x_k par le milieu de la classe.

3 ▹ La ***médiane*** d'une série statistique, notée Q_2, est une valeur qui partage la population en deux groupes de même effectif :

- si N est impair alors on appelle ***individu médian*** le $\frac{\mathrm{N}+1}{2}$-ième individu (en les classant par ordre croissant) et la médiane est la modalité associée à cet individu ;
- si N est pair alors les individus médians sont les $\frac{\mathrm{N}}{2}$-ième et $\frac{\mathrm{N}+2}{2}$-ième individus et la médiane est la moyenne des modalités associées à ces deux individus.

4 ▹ De façon générale, on appelle *quartiles* et on note $\mathrm{Q}_1, \mathrm{Q}_2, \mathrm{Q}_3$ les valeurs qui partagent la population en quatre groupes de même effectif.

On définit de façon analogue les déciles $\mathrm{D}_1, \ldots, \mathrm{D}_9$ et les percentiles.

5 ▹ L'***étendue*** d'une série statistique est la différence entre les modalités la plus grande et la plus petite pour lesquelles l'effectif n'est pas nul (dans le cas d'un regroupement par classe, on considère la borne supérieure de la plus grande classe et la borne inférieure de la plus petite).

6 ▹ La ***variance*** de x est le nombre réel noté s_x^2 défini par :

$$s_x^2 = \frac{1}{\mathrm{N}} \sum_{k=1}^{p} n_k (x_k - \overline{x})^2 = \sum_{k=1}^{p} f_k (x_k - \overline{x})^2.$$

Dans le cas d'un regroupement par classes, on remplace x_k par le milieu de la classe.

Puisque $s_x^2 \geqslant 0$, on peut considérer sa racine carrée (positive) s_x que l'on appelle ***écart-type*** de x.

On a : $s_x^2 = \overline{x^2} - \overline{x}^2$ (formule de Koenig-Huygens).

⮊ Notion de statistique bivariée

On considère une population d'effectif total N sur laquelle on observe un couple (x, y) de caractères quantitatifs, ce qui se traduit par un N-uplet de couples :

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_{\mathrm{N}}, y_{\mathrm{N}}).$$

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

Afin d'étudier la dispersion des points de coordonnées (x_k, y_k) autour du point moyen, on introduit :

– la ***covariance*** de la série double, définie par :

$$s_{xy} = \frac{1}{\mathrm{N}} \sum_{k=1}^{\mathrm{N}} (x_k - \overline{x})(y_k - \overline{y}),$$

– et le ***coefficient de corrélation*** de la série double, défini par :

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

On a : $s_{xy} = \overline{xy} - \overline{x}\,\overline{y}$.

On vérifie que $\left|r_{xy}\right| \leqslant 1$ et que $\left|r_{xy}\right| = 1$ si et seulement s'il existe des réels a et b tels que $y = ax+b$, ce qui signifie que les points (x_k, y_k) sont alignés. C'est pourquoi lorsque le coefficient de corrélation est proche de ± 1, les points (x_k, y_k) sont « presque » alignés et on dit alors que x et y sont ***bien corrélés***.

⮊ Ajustement affine

Considérons le cas de deux caractères $x = (x_1, \ldots, x_{\mathrm{N}})$ et $y = (y_1, \ldots, y_{\mathrm{N}})$ tels que les N points de coordonnées (x_k, y_k) soient tous distincts et ne soient jamais à la verticale l'un de l'autre.

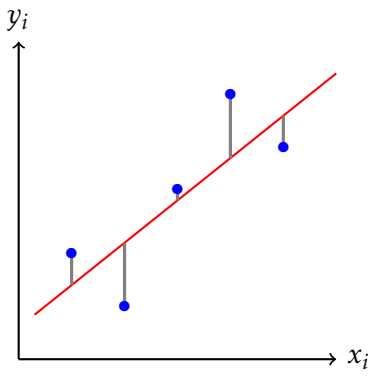
On appelle ***droite de régression*** (de y par rapport à x) du nuage de points de la série statistique double, la droite passant par le point moyen $\mathrm{G}(\overline{x}, \overline{y})$ et de coefficient directeur $a = \frac{s_{xy}}{s_x^2}$ c'est-à-dire la droite d'équation :

$$\mathrm{Y} = \frac{s_{xy}}{s_x^2}\mathrm{X} + \left(\overline{y} - \frac{s_{xy}}{s_x^2}\overline{x}\right).$$

Cette droite est l'unique droite d'équation $Y = aX + b$ telle que a et b minimisent la somme :

$$S = \sum_{k=1}^{N} \Big(y_k - (ax_k + b) \Big)^2$$

des carrés des distances verticales entre la droite et chacun des points du nuage.



B - Savoir-faire élémentaires

■ Comment représenter des données statistiques ?

SF1 Utiliser un diagramme en bâtons ou un histogramme

Cela consiste à représenter en abscisse les différentes modalités d'un caractère et à tracer un rectangle vertical fin dont la hauteur correspond à la fréquence de la modalité (ou à son effectif).

Exemple

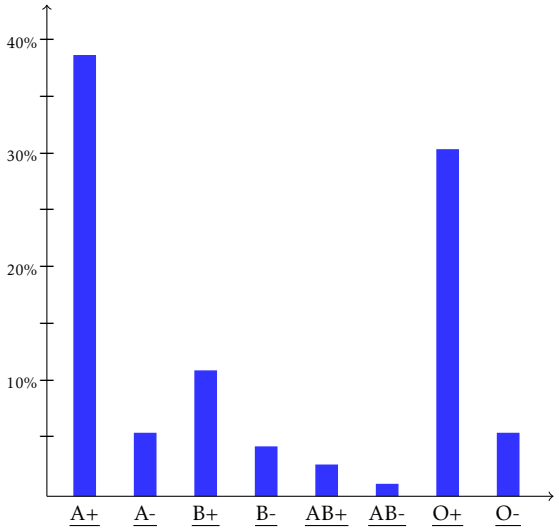
Le relevé des groupes sanguins (avec rhésus) d'une population étant donné par le tableau ci-dessous, établissons le tableau des fréquences puis les représenter graphiquement :

A+	A-	B+	B-	AB+	AB-	O+	O-
70	10	20	7	5	3	55	10

L'effectif total étant de 180 individus, le tableau des fréquences (en %) est :

	A+	A-	B+	B-	AB+	AB-	O+	O-
effectif	70	10	20	8	5	2	55	10
fréquence	38,9	5,6	11,1	4,4	2,8	1,1	30,6	5,6

Graphiquement, on a donc :



Dans le cas où le caractère représente des données continues regroupées en classes, on peut utiliser un histogramme : on représente en abscisse les différentes modalités en faisant notamment apparaître les seuils entre les classes et on trace des rectangles dont les aires sont proportionnelles aux fréquences des différentes modalités (ou aux effectifs).

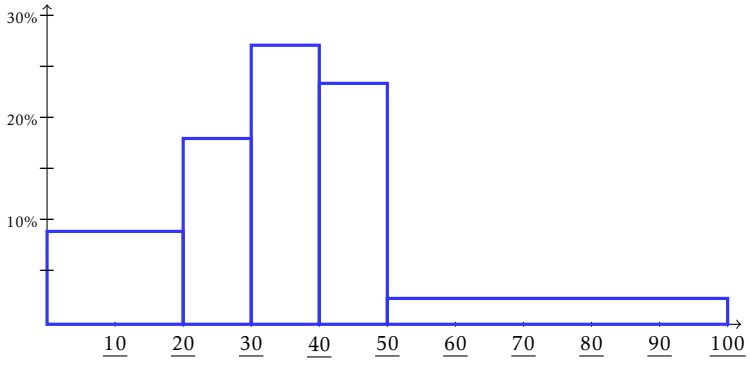
Exemple

Un commerçant relève le nombre de ventes effectuées en les classant en fonction de leur prix comme indiqué dans le tableau ci-dessous. Représentons ces données à l'aide d'un histogramme.

prix	$]0,20]$	$]20,30]$	$]30,40]$	$]40,50]$	$]50,100]$
effectif	20	20	30	26	14

Indiquons les fréquences (en %) puis traçons l'histogramme en tenant bien compte de l'amplitude de chaque classe :

prix	$]0,20]$	$]20,30]$	$]30,40]$	$]40,50]$	$]50,100]$
fréquence	18,2	18,2	27,3	23,6	12,7



SF2 Utiliser un nuage de points

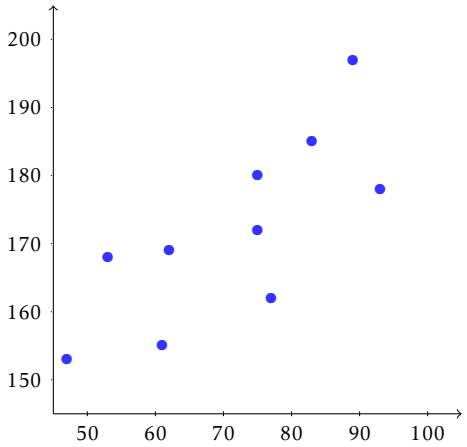
Dans le cas d'une série statistique double constituée de deux caractères x et y, on représente les modalités de x en abscisse et celles de y en ordonnée puis on place les points de coordonnées (x_k, y_k). On peut éventuellement jouer sur la taille des points si des couples ont plusieurs occurrences.

Exemple

Lors d'une visite médicale, on relève le poids et la taille de 10 patients comme indiqué dans le tableau ci-dessous. Représentons cette série statistique.

patients	1	2	3	4	5	6	7	8	9	10
poids (en kg)	53	62	89	75	83	77	75	61	47	93
taille (en cm)	168	169	197	172	185	162	180	155	153	178

On représente le poids en abscisse et la taille en ordonnée.



■ Comment décrire des données statistiques ?

SF3 Calculer et exploiter les caractéristiques de position et celles de dispersion

Il faut faire la distinction entre ce qui relève de la position des données (mode, moyenne, médiane, point moyen pour une série double) et ce qui relève de leur dispersion (variance et écart-type, quartiles, déciles).

Il peut être pertinent d'utiliser également le *coefficient de variation* qui est le quotient de l'écart-type par la moyenne. Il mesure en quelque sorte la dispersion relative. De plus, il n'a pas d'unité et permet donc de comparer des séries statistiques exprimées dans des unités différentes.

Exemple

Le tableau ci-dessous donne les débits mensuels moyens en $m^3.s^{-1}$ de certains fleuves français. Pour chacune des séries statistiques, calculons la moyenne et l'écart-type.

mois	**Loire** (Saint-Nazaire)	**Rhône** (Beaucaire)	**Garonne** (Mas-d'Agenais)
Janvier	1830	1960	926
Février	1740	2010	1030
Mars	1630	2010	951
Avril	1180	1930	905
Mai	998	1860	832
Juin	554	1760	564
Juillet	328	1340	299
Août	242	1080	190
Septembre	294	1140	212
Octobre	443	1410	321
Novembre	787	1880	514
Décembre	1210	1900	849

Notons x, y et z ces trois séries statistiques.

Pour la Loire, on trouve : $\overline{x} \approx 936\ m^3.s^{-1}$ et $s_x \approx 558\ m^3.s^{-1}$.

Pour le Rhône : $\overline{y} \approx 1690\ m^3.s^{-1}$ et $s_y \approx 332\ m^3.s^{-1}$.

Pour la Garonne : $\overline{z} \approx 633\ m^3.s^{-1}$ et $s_z \approx 304\ m^3.s^{-1}$.

Comparons maintenant les coefficients de variation.

On trouve(approximativement) : $0,60$ pour la Loire, $0,20$ pour le Rhône et $0,48$ pour la Garonne. On peut remarquer que le débit du Rhône est beaucoup moins sensible aux variations saisonnières, notamment en raison des réserves d'eau constituées par la glace et la neige en montagne (apport beaucoup plus faible pour la Garonne et presque inexistant pour la Loire).

SF4 Établir un ajustement affine

Dans le cas d'une série statistique double, le point moyen renseigne sur la position des valeurs, la covariance et le coefficient de corrélation renseignent sur leur dispersion autour du point moyen.

Exemple

Calculons la covariance et le coefficient de corrélation pour l'exemple des relevés de poids et de taille vu précédemment.

En notant x le poids et y la taille, on obtient :

$$\overline{x} \approx 71,5\,kg\ ,\ s_x \approx 14,49\,kg\ ,\ \overline{y} \approx 171,9\,cm \text{ et } s_y \approx 12,92\,kg.$$

On trouve également : $\overline{xy} \approx 12434,4$ kg.cm, d'où :

$$s_{xy} = \overline{xy} - \overline{x}\,\overline{y} \approx 143,6.$$

Le fait que la covariance soit positive indique que x et y ont tendance à varier « dans le même sens ».

D'autre part, le coefficient de corrélation est :

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \approx 0,7.$$

Cela s'interprète en disant que le poids et la taille de ces 10 individus semblent peu corrélés.

Le cas d'un coefficient de corrélation proche de ±1 signifie qu'il y a presque une relation affine entre les deux caractères x et y. Dans ce cas, il est pertinent de chercher une droite donnant l'allure générale du nuage de points. La droite de régression (de y par rapport à x) est la droite minimisant les écarts-verticaux aux points du nuage.

Exemple

On relève les notes en mathématiques (x) et en physique (y) de 10 étudiants dans le tableau ci-dessous. Étudions la corrélation puis déterminons la droite de régression.

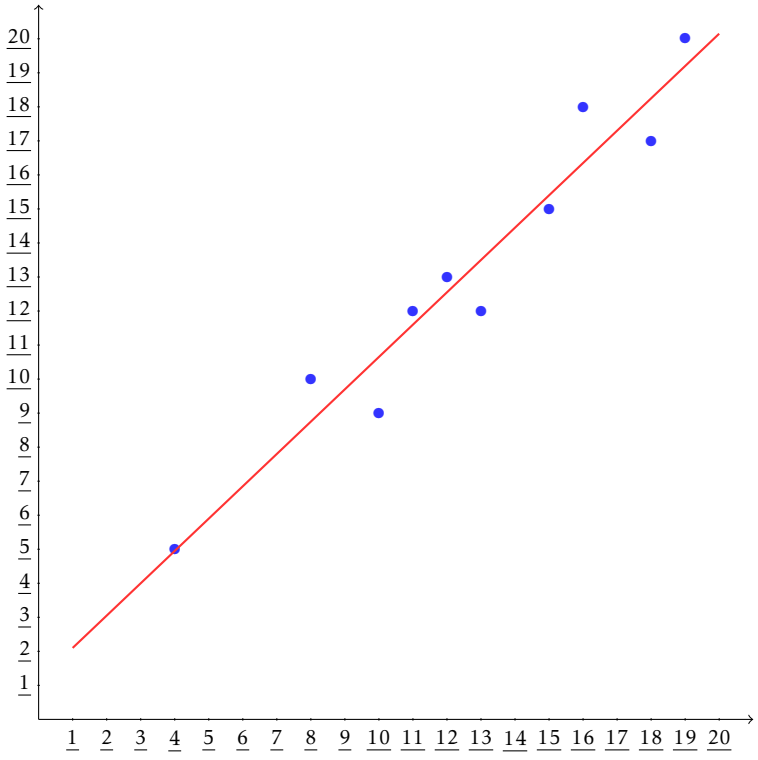
x	19	16	13	18	8	12	10	15	4	11
y	20	18	12	17	10	13	9	15	5	12

On trouve : $\overline{x} \approx 12,6$, $\overline{y} \approx 13,1$, $\overline{xy} \approx 183,3$, $s_{xy} \approx 18,2$, $s_x \approx 4,4$, $s_y \approx 4,3$, $r_{xy} \approx 0,967$.

La corrélation est donc bonne. Déterminons maintenant le droite de régression :

$$\frac{s_{xy}}{s_x^2} \approx 0,95 \text{ et } \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x} \approx 1,15.$$

La droite a pour équation (aux valeurs approchées près) : $Y = 0,95X + 1,15$.



C - Lien entre régression et projection orthogonale

Considérons n points de $\mathbb{R}^2$, $(x_1,y_1),\dots,(x_n,y_n)$ non alignés verticalement.

On cherche la droite qui «approche» au mieux ces n points. Si on note $y=ax+b$ l'équation d'une droite, on cherche à minimiser l'erreur :

$$\mathrm{E}_r=\sum_{i=1}^{n}d_i^2=\sum_{i=1}^{n}(ax_i+b-y_i)^2.$$

Traduisons matriciellement le problème. Posons :

$$\mathrm{X}=\begin{pmatrix}a\\b\end{pmatrix},\quad \mathrm{A}=\begin{pmatrix}x_1&1\\x_2&1\\\vdots&\vdots\\x_n&1\end{pmatrix}\quad\text{et}\quad \mathrm{B}=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}\quad\text{de sorte que}\quad \mathrm{AX}-\mathrm{B}=\begin{pmatrix}ax_1+b-y_1\\ax_2+b-y_2\\\vdots\\ax_n+b-y_n\end{pmatrix}.$$

Si l'on considère le produit scalaire canonique de $\mathscr{M}_{n,1}(\mathbb{R})$ et la norme associée, alors :

$$\mathrm{E}_r=\|\mathrm{AX}-\mathrm{B}\|^2.$$

Les deux colonnes de la matrice A forment une famille libre (puisque les points ne sont pas alignés verticalement) donc la matrice A est de rang 2.

Il existe donc un unique vecteur X_0 minimisant $\|\mathrm{AX}-\mathrm{B}\|$. Pour le calculer, on écrit :

$${}^t\mathrm{AA}=\begin{pmatrix}\sum_{i=1}^{n}{x_i}^2&\sum_{i=1}^{n}x_i\\\sum_{i=1}^{n}x_i&n\end{pmatrix}\in\mathscr{M}_2(\mathbb{R}).$$

Justifions l'inversibilité de ${}^t\mathrm{AA}$:

$$\det\left({}^t\mathrm{AA}\right)=n\sum_{i=1}^{n}{x_i}^2-\left(\sum_{i=1}^{n}x_i\right)^2.$$

D'après l'inégalité de Cauchy-Schwarz de $\mathbb{R}^n$ avec les vecteurs **non colinéaires** $x=(x_1,\dots,x_n)$ et $u_0=(1,\dots,1)$:

$$\langle u_0,x\rangle^2=\left(\sum_{i=1}^{n}x_i\right)^2<n\sum_{i=1}^{n}x_i^2\ \text{ donc }\ \det\left({}^t\mathrm{AA}\right)\neq 0$$

donc la matrice ^{t} AA est inversible et son inverse est donné par :

$$\left({}^t\mathrm{AA}\right)^{-1}=\frac{1}{n\sum_{i=1}^{n}{x_i}^2-\left(\sum_{i=1}^{n}x_i\right)^2}\begin{pmatrix}n&-\sum_{i=1}^{n}x_i\\-\sum_{i=1}^{n}x_i&\sum_{i=1}^{n}{x_i}^2\end{pmatrix}.$$

De plus ${}^t\mathrm{AB}=\begin{pmatrix}\sum_{i=1}^{n}x_iy_i\\\sum_{i=1}^{n}y_i\end{pmatrix}$ ce qui permet d'expliciter le vecteur X_0 :

$$a=\frac{n\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n\sum_{i=1}^{n}{x_i}^2-\left(\sum_{i=1}^{n}x_i\right)^2}=\frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)}{\frac{1}{n}\sum_{i=1}^{n}{x_i}^2-\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2}.$$

Si $\bar{x}$ (resp. $\bar{y}$) et σ_x (resp. σ_y) désignent la moyenne et l'écart-type empirique de la série statistique $\{x_i \mid i \in [\![1,n]\!]\}$ (resp. $\{y_i \mid i \in [\![1,n]\!]\}$), $\mathrm{Cov}(x,y)$ désigne la covariance empirique de x et y et $\rho_{x,y}$ désigne le coefficient de corrélation empirique, alors la droite de régression linéaire de y en x a pour équation :

$$y - \bar{y} = \rho_{x,y}\frac{\sigma_y}{\sigma_x}(x-\bar{x}) = \frac{\mathrm{Cov}(x,y)}{\sigma_x^2}(x-\bar{x}).$$

D - Programmation en python

```
import numpy as np
import matplotlib.pyplot as plt

def regression(listepoints):
    X=np.array([listepoints[i][0] for i in range(len(listepoints))])
    Y=np.array([listepoints[i][1] for i in range(len(listepoints))])
    mx , my , mxy , mx2 = X.mean() , Y.mean() , (X*Y).mean() , (X**2).mean()
    a = (mxy-mx*my)/(mx2-mx**2)
    b = (my*mx2-mx*mxy)/(mx2-mx**2)
    plt.figure()
    plt.plot(X,Y,'gs',label="données")
    Xd = np.linspace(X.min(),X.max())
    plt.plot(Xd,a*Xd+b,'r--',label="droite de régression")
    plt.legend(loc="upper left")
    plt.show()
```

On peut tester cette fonction avec un nuage aléatoire :

```
from random import randrange

test = [ ( i*0.5 , (i+1)*(0.80+randrange(201)/500) ) for i in range(11) ]

regression(test)
```

On obtient :