Statistique à deux variables

1. Généralités

On dit qu'on a affaire à des statistiques à **deux variable *X* et *Y*** lorsque deux variables statistiques **varient en même temps**.

On appellera **série chronologique** une série statistique à deux variables dont la variable X correspond à des dates.

On note généralement x_i les valeurs prises par la première variable et y_i celles prises par la seconde.

Les points du plan rapporté à un repère orthogonal constituent un **nuage de points**.

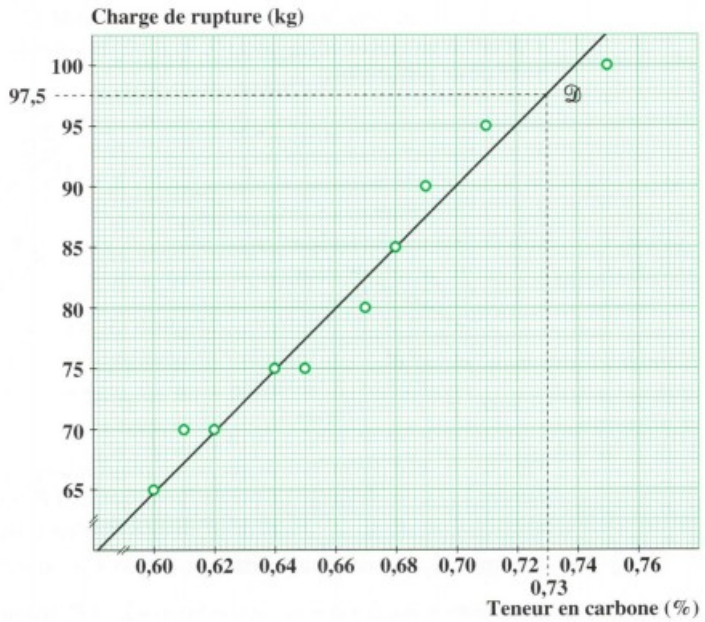
On appelle **point moyen** du nuage le point dont l'abscisse est la moyenne des abscisse des points et dont l'ordonnée est la moyenne des ordonnées des points du nuage. On le note souvent G.

2. Ajustement affine

Faire un **ajustement affine** consiste à trouver une droite qui passe le plus près possible des points du nuage.

Exemple : Ajustement graphique

L'ajustement graphique est une méthode empirique qui consiste à faire passer une droite dans le nuage de points pour rendre compte, au mieux de la tendance observée. Le graphique ci-dessous peut permettre d'estimer la charge de rupture pour une teneur en carbone de 0,73 %, on obtient alors environ 97,5 kg.



• Calcul du **point moyen** (on note G ce point)

$$x_G = \frac{0{,}60 + 0{,}61 + \ldots + 0{,}75}{10} = 0{,}662 \ ; \ y_G = \frac{65 + 70 + \ldots + 100}{10} = 80{,}5$$

$G(0{,}662 ; 80{,}5)$.

2a. Ajustement par la méthode des moindres carrés

a) Définition de la covariance

Définition

Considérons une série statistique double (x_i, y_i), on appelle covariance et on note $\text{cov}(x, y)$ le réel :

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

À savoir

On montre que : $\text{cov}(x, y) = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \overline{x}\,\overline{y}$.

On peut vérifier que pour l'exemple précédent, on a :

Teneur en carbone (en %) x_i	0,60	0,61	0,62	0,64	0,65	0,67	0,68	0,69	0,71	0,75
Charge de rupture (en kg) y_i	65	70	70	75	75	80	85	90	95	100

$$cov(x, y) = \frac{537,8}{10} - 0,662 \times 80,5 = 0,489$$

Équation de la droite $D_{Y/x}$:

À savoir

On montre que cette droite a une équation du type :

$$y = ax + b \quad \text{avec} \quad a = \frac{\text{cov}(x, y)}{V(x)} \quad \text{et} \quad b = \overline{y} - a\overline{x} \qquad (G(\overline{x}, \overline{y}) \in \mathcal{D}_{Y/x}).$$

x est appelé la variable explicative et y la variable à expliquer.

La variance *V(x)* et le coefficient *a* sont donnés par :

$$V(x) = \frac{\sum_{i=1}^{N} x_i^2}{N} - \overline{x}^2 \qquad \text{et} \qquad a = \frac{\frac{\sum_{i=1}^{N} x_i y_i}{N} - \overline{x}\,\overline{y}}{\frac{\sum_{i=1}^{N} x_i^2}{N} - \overline{x}^2} = \frac{\sum_{i=1}^{N} x_i y_i - N\overline{x}\,\overline{y}}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2}.$$

Donc $a = \frac{537,8 - 10 \times 0,662 \times 80,5}{4,4026 - 10 \times 0,662^2} \approx 242,56$ et $b \approx -80,07$.

L'équation de la droite est : $y \approx 242,56\,x - 80,07$.

Pour une teneur en carbone (x) de 0,73 %, la charge de rupture estimée (y) est de 97 kg.

Équation de la droite $D_{X/y}$:

À savoir

On montre qu'une équation de la droite $\mathcal{D}_{X/y}$ est :

$$x = \alpha y + \beta \quad \text{avec} \quad \alpha = \frac{\text{cov}(x, y)}{\text{V}(y)} \quad \text{et} \quad \overline{x} = \alpha \overline{y} + \beta \quad (G(\overline{x}, \overline{y}) \in \mathcal{D}_{X/y}).$$

Dans ce cas y est la variable explicative et x la variable à expliquer.

Exemple

En reprenant toujours le même exemple, on obtient en utilisant $\sum y_i^2 = 66\,025$.

$$\alpha = \frac{537{,}8 - 10 \times 0{,}662 \times 80{,}5}{66\,025 - 10 \times 80{,}5^2} = 0{,}004.$$

On a : $\overline{x} = \alpha \overline{y} + \beta$ donc $\beta = 0{,}662 - 0{,}004 \times 80{,}5$; $\beta = 0{,}34$.

Une équation de la droite de régression $\mathcal{D}_{X/y}$ est donc :

$$x = 0{,}004\,y + 0{,}34.$$

Cette équation permet de donner une estimation de la variable x pour une valeur donnée de y.

Si on cherche la teneur en carbone permettant d'obtenir une charge de rupture de 105 kg, on utilise cette équation de droite,

d'où $x \approx 0{,}004 \times 105 + 0{,}34$; $x \approx 0{,}76$.

La teneur en carbone estimée est 0,76 %.

À savoir

- **Si on cherche une estimation de y connaissant la variable x on utilise $\mathcal{D}_{Y/x}$.**
- **Si on cherche une estimation de x connaissant la variable y on utilise $\mathcal{D}_{X/y}$.**

2b. Coefficient de corrélation linéaire

À savoir

On appelle coefficient de corrélation linéaire le réel r défini par :

$$r = \frac{\sum_{i=1}^{N} x_i y_i - N\overline{x}\,\overline{y}}{\sqrt{\left(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\right)\left(\sum_{i=1}^{N} y_i^2 - N\overline{y}^2\right)}} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}.$$

Remarques

- $r^2 = a\alpha$.
- Si r est très voisin de -1 ou de 1, on dira qu'il y a une très bonne corrélation entre les deux séries étudiées.
- Quelles que soient les séries statistiques étudiées, on aura :

$$\boxed{-1 \leqslant r \leqslant 1.}$$

(r est en fait le cosinus d'un angle). Si r est proche de zéro, la corrélation est mauvaise, il est alors inutile de calculer l'équation d'une droite de régression.

N.B. : Un coefficient de corrélation proche de 1 ou de -1 ne prouve rien hors de son contexte.

Pour l'exemple précédent, on a : $r = \dfrac{537{,}8 - 10 \times 0{,}662 \times 80{,}5}{\sqrt{(4{,}4026 - 10 \times 0{,}662^2) \times (66025 - 10 \times 80{,}5^2)}} \approx 0{,}985$

Le coefficient de corrélation linéaire est très proche de 1, donc il y a une très bonne corrélation.

3. Changement de variable en statistique

Si on remplace toutes les valeurs x_i du caractère d'une série statistique par $X_i=x_i-a$ (a réel) alors :

$$\bar{X}=\bar{x}-a \qquad \text{et} \qquad \sigma_X=\sigma_x \quad .$$

Exemple :

Appliquer ce résultat à la série statistique suivante :

x_i	746	746,5	747	747,5	748	748,5	749	749,5	750
n_i	1	2	4	10	29	32	15	5	2

en prenant comme changement de variable $X_i = x_i - 748$.

Le nouveau tableau statistique obtenu est :

X_i	– 2	– 1,5	– 1	– 0,5	0	0,5	1	1,5	2
n_i	1	2	4	10	29	32	15	5	2

L'avantage de cette méthode est, lorsqu'on ne dispose pas de matériel de calcul performant, de permettre des calculs moins fastidieux qu'avec la série statistique d'origine. On obtient :

$\overline{X} = 0{,}285$ donc $\bar{x} = 748 + 0{,}285 = 748{,}285$ et $\sigma \approx 0{,}70$.

4. Régression exponentielle

On donne le tableau suivant indiquant pour sept années consécutives le montant des ventes mondiales d'un produit récent sur le marché en milliers de dollars. Pour faciliter le calcul les années sont repérées par leur rang (ceci correspond en fait à un changement de variable du type $X_i = x_i - a$, où a est l'entier donnant la première année – 1).

Rang de l'année x_i	1	2	3	4	5	6	7
Montant des ventes m_i	37 000	45 000	74 000	99 000	148 000	297 000	601 000

On pose $y_i = \ln m_i$. Calculer le coefficient de corrélation entre x_i et y_i. Donner l'équation de la droite de régression de Y en x. En déduire le lien entre m_i et x_i.

Les calculs faits avec une calculatrice donnent :
$r \approx 0{,}984$ (ce qui prouve que la corrélation est très bonne).
$a \approx 0{,}458$; $b \approx 9{,}847$.
L'équation de la droite de régression est donc :

$$y \approx 0{,}458x + 9{,}847.$$

Si on remplace y par $\ln m$, on obtient :
$\ln m \approx 0{,}458x + 9{,}847$
soit $m \approx e^{0{,}458x + 9{,}847}$
donc $m \approx e^{9{,}847} \times e^{0{,}458x}$
$m \approx 18\,902 \times e^{0{,}458x}$ ou $m \approx 18\,902 \times 1{,}58^x$.

En pratique

• Lorsque le nuage a une forme du type de la figure 1 :
(Phénomène qui tend à s'accélérer),
on pose $Y_i = \ln y_i$ (régression exponentielle).

• Lorsque le nuage a une forme du type de la figure 2 :
(phénomène qui continue à progresser mais de moins en moins vite),
on pose $Y_i = \exp y_i$ (régression logarithmique).

Pour savoir si on a bien choisi le changement de variable, il est préférable de calculer en premier lieu le coefficient de corrélation linéaire obtenu avec la nouvelle série statistique.

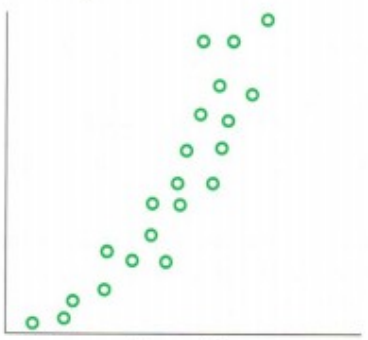


Figure 1

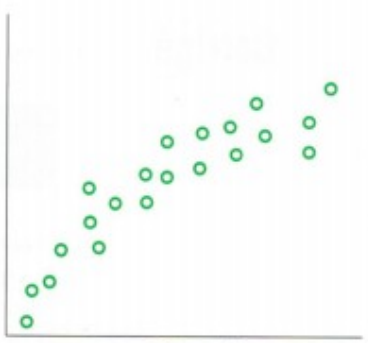


Figure 2

5. Représenter une série statistique à deux variables et déterminer le point moyen avec une calculatrice graphique

Exemple.

Une entreprise a relevé le coût total de production mensuel (en k€), noté y, en fonction de la production x (en tonnes).

Production x (en tonnes)	1	2	4	6	8	10
Coût total y (en k€)	36,3	38,5	44,6	48,4	51,1	54,2

1. Représenter le nuage de points $M_i(x_i ; y_i)$ dans un repère orthogonal.

2. Déterminer les coordonnées du point moyen G de la série.

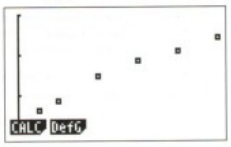
Avec une calculatrice Casio Graph 35+

• *Dans* MENU STATS EXE, *on sélectionne* ▶ *en tapant* F6 . DEL-A *avec* F4 , YES *avec* EXE .

On entre les valeurs x_i dans List 1, les valeurs y_i dans List 2.

• *On sélectionne* ▶ *par* F6 *puis* GRAPH *par* F1 , *puis on choisit* GPH1 *avec* F1 .

On obtient le nuage de points.



• *Pour obtenir les coordonnées de G taper* EXIT *2 fois.*

Sélectionner CALC *avec* F2 *puis* Set *avec* F6 .

Sur 2VarX-List , *on sélectionne List1 ;*

sur 2VarY-List , *on sélectionne List2 ;* EXE .

F2 *pour* 2VAR .

On lit $\bar{x} = 5{,}166\ldots$, *puis* $\bar{y} = 45{,}516\ldots$.

Avec une calculatrice TI 82 stats.fr ou 83 Plus

• *On tape* Stats *puis on sélectionne* 4:Effliste .

On tape 2nde 1 , 2nde 2 *(pour L1, L2) puis* entrer

(on a effacé les listes précédentes.)

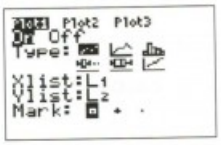
• *On tape* Stats *puis on sélectionne* 1:Edit entrer .

On entre les valeurs x_i dans L1, les valeurs y_i dans L2.

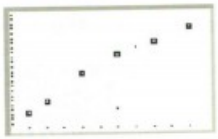
• *On tape* 2nde f(x) *pour graph stats puis* 1 .

• *On sélectionne sur l'écran ci-dessous :*

- On ,
- *le type de représentation,*
- *L1 pour ListeX et L2 pour ListeY,*
- *le type de marque.*



• *On tape* ZOOM 9 *et on obtient le nuage de points.*



• *Pour obtenir les coordonnées de G :*

on tape Stats , *on sélectionne* CALC *puis* 2: Stats2Var

et on tape 2nde 1 , 2nde 2 entrer .

On lit $\bar{x} = 5{,}166\ldots$, $\bar{y} = 45{,}516\ldots$

6. Réaliser un ajustement affine par la méthode des moindres carrés avec une calculatrice graphique

1. Déterminer le coefficient de corrélation de la série.
2. Donner une équation de la droite D, droite de régression de y en x obtenue par la méthode des moindres carrés.
3. Estimer à l'aide de cette droite le coût total correspondant à une production de 12 tonnes.

Avec une calculatrice Casio Graph 35+

- On tape MENU STAT EXE ,
- On efface les listes précédentes en sélectionnant ▶ par F6 , DEL-A par F4 et YES par F1 .
- On entre les x_i dans List 1, les y_i dans List 2.
- On sélectionne ▶ par F6 puis CALC par F2 et Set par F6 .

Sur la ligne 2VarXList , on sélectionne List1 ;
sur la ligne 2VarYList , on sélectionne List2 ;
sur la ligne 2VarFreq, on sélectionne 1 puis EXE .

- On obtient les résultats en sélectionnant REG par F3 puis X par F1 et ax+b par F1 :

```
LinearReg(ax+b)
   a =2.00356164
   b =35.1649315
   r =0.98851473
   r²=0.97716137
 MSe=1.42689041
y=ax+b
```

Remarque :
Pour obtenir une équation de la droite de régression de x en y,
sur la ligne 2VarXList , on remplace List1 par List2 ;
et sur la ligne 2VarYList , on remplace List2 par List1 ;
et on obtient a' et b' de l'équation $x = a'y + b'$. ▶ ▶ ▶

Avec une calculatrice TI 82 stats.fr ou 83 Plus

- On tape stats et on sélectionne 4 : Effliste 2nde 1 , 2nde 2 entrer pour supprimer les listes précédentes.

On tape stats , on sélectionne 1 : Edite entrer .

- On entre les x_i dans L1, les y_i dans L2.

On tape stats , on sélectionne CALC . On sélectionne 4 : Reglin (ax + b) .

- On entre L1, L2 avec 2nde 1 , 2nde 2 entrer .

Cela donne le résultat :

```
LinReg
 y=ax+b
 a=2.003561644
 b=35.16493151
```

si r^2 et r ne sont pas affichés, on tape 2nde 0 et on sélectionne CorrelAff dans la liste.

Remarque :
Pour obtenir une équation de la droite de régression de x en y,
Sélectionner 4 : Reglin (ax + b) , taper 2nde 2 , 2nde 1 pour L2, L1.
On obtient les coefficients a' et b' de l'équation $x = a'y + b'$.

De cette manière, on trouve les résultats suivants :

1. Le coefficient de corrélation est $r = 0{,}988515$.
r est proche de 1, l'ajustement affine est justifié.

2. La droite D de régression de y en x a pour équation :

$$y = 2x + 35{,}16.$$

3. Pour une production de 12 tonnes, on peut estimer que :

$$y = 2 \times 12 + 35{,}16 \quad \text{soit} \quad y = 59{,}16.$$

Exercices :

18 C Dans cet exercice, les calculs seront effectués à 10^{-3} près.
L'étude du coût de maintenance annuel d'une installation de chauffage dans un immeuble de bureaux, en fonction de l'âge de l'installation, a donné les résultats suivants.

Âge x_i (en années)	1	2	3	4	5	6
Coût y_i (en k€)	7,55	9,24	10,74	12,84	15,66	18,45

1. Représenter le nuage de points $M_i(x_i ; y_i)$ dans un repère orthogonal (unités graphiques : 2 cm en abscisse, 1 cm en ordonnée).
Peut-on envisager un ajustement affine de ce nuage ?

2. a) Déterminer le coefficient de corrélation linéaire de la série statistique double $(x_i ; y_i)$.
Le résultat obtenu confirme-t-il l'observation faite à la question **1.** ?

b) Déterminer, par la méthode des moindres carrés, une équation de la droite de régression D de y en x.

Tracer D dans le même repère que celui de la question **1.**

c) En admettant que l'évolution du coût constaté pendant six ans se poursuive les années suivantes, donner une estimation du coût de maintenance de l'installation lorsqu'elle aura huit ans.

19 Le nombre d'internautes en France est donné (en millions) dans le tableau suivant :

Année	2001	2003	2005	2007	2009	2011
x : rang de l'année	1	3	5	7	9	11
y : nombre d'internautes (en millions)	12,86	20,67	25,07	29,55	33,64	39,36

1. Donner le coefficient de corrélation linéaire entre les séries x et y. Arrondir le résultat au centième.

2. On envisage un ajustement affine. Donner une équation de la droite de régression de y en x obtenue par la méthode des moindres carrés. Arrondir les coefficients au centième.

3. En utilisant l'équation précédente, estimer le nombre d'internautes en 2015, en arrondissant le résultat au demi-million.

20 En 2013, une caisse de retraite propose à ses adhérents un barème de rachat d'un trimestre de cotisation des années antérieures selon le tableau suivant.

Âge de l'adhérent (en années)	56	57	58	59	60
Rang x_i	0	1	2	3	4
Montant y_i du rachat d'un trimestre de cotisation (en euros)	3 906	3 994	4 081	4 167	4 251

1. Donner une équation de la droite de régression D de y en x, obtenue par la méthode des moindres carrés.

2. Quel serait avec cet ajustement affine le montant du rachat d'un trimestre pour un salarié âgé de 62 ans ?

21 C Une machine fabrique en grande série des billes d'acier.
La moyenne des diamètres des billes produites en une semaine varie au cours du temps. La fabrication est jugée valable tant que cette moyenne reste dans l'intervalle [3,25 ; 3,32].
La semaine numérotée 0 correspond à celle du réglage initial. Des contrôles hebdomadaires effectués lors des quatre premières semaines de fonctionnement ont donné les résultats suivants.

Semaine s	0	1	2	3	4
Moyenne m	3,32	3,32	3,31	3,29	3,27

1. a) Calculer le coefficient de corrélation de la série statistique.

b) Déterminer une équation de la droite d'ajustement de s en m par la méthode des moindres carrés.

2. En déduire un pronostic pour la valeur maximale du temps séparant deux réglages successifs.

22 Afin de mesurer l'évolution de l'utilisation du vélo, une communauté urbaine organise le comptage régulier des vélos en plusieurs points de l'agglomération. Le tableau ci-dessous indique le nombre moyen, sur un mois, de vélos comptés par jour.

Mois	Mars 2005	Juin 2005	Déc. 2005	Juin 2006	Déc. 2006	Juin 2007
Rang x_i du mois	0	3	9	15	21	27
Nombre moyen y_i de vélos comptés par jour (en milliers)	3,9	4,4	5,1	6,4	7,1	7,6

1. Représenter le nuage de points $M_i(x_i ; y_i)$ dans un repère orthogonal.
On prendra pour unités graphiques : en abscisse, 1 centimètre pour représenter 3 mois et en ordonnées, 1 centimètre pour représenter 1 millier.

2. Déterminer les coordonnées du point moyen G et le placer sur la représentation graphique.

3. Déterminer, en utilisant la calculatrice, l'équation de la droite d'ajustement affine de y en x obtenue par la méthode des moindres carrés. On arrondira les coefficients obtenus à 10^{-2} près.
Tracer la droite d'ajustement sur la représentation graphique.

4. À l'aide de l'ajustement réalisé, déterminer une estimation du nombre moyen de vélos que l'on devait prévoir par jour au mois de décembre 2007 (on arrondira le résultat à 10^{-1}).

5. On sait qu'en décembre 2007, le nombre moyen de vélos observés a été en fait de 7600.
Déterminer, en pourcentage, l'erreur commise dans l'estimation précédente.

23 **C** Sur un parcours donné, la consommation y d'une voiture est donnée en fonction de sa vitesse moyenne x par le tableau suivant :

x (en km/h)	80	90	100	110	120
y (en L/100 km)	4	4,8	6,3	8	10

1. La consommation est-elle proportionnelle à la vitesse moyenne ? Justifier la réponse.

2. a) Représenter le nuage de points correspondant à la série statistique $(x_i ; y_i)$ dans un repère orthogonal du plan (on prendra 2 cm pour 10 km/h sur l'axe des abscisses et 1 cm pour 1 L sur l'axe des ordonnées).

b) Déterminer les coordonnées du point moyen G du nuage et le placer sur le graphique.

c) À l'aide d'une calculatrice, donner une équation, sous la forme $y = ax + b$, de la droite d'ajustement affine de y en x par la méthode des moindres carrés et tracer cette droite (on arrondira a au millième et b au centième).

d) En utilisant cet ajustement, estimer la consommation aux 100 km (arrondie au dixième) de la voiture pour une vitesse de 130 km/h.

3. La forme du nuage permet d'envisager un ajustement exponentiel.

On pose $z = \ln(y)$ et on admet que la droite d'ajustement obtenue pour les cinq points $(x ; z)$ du nuage par la méthode des moindres carrés, a pour équation :

$$z = 0{,}023\,4x - 0{,}508\,0.$$

a) Écrire y sous la forme $y = A\,e^{Bx}$ (donner A et B arrondis à 10^{-4}).

b) Tracer, sur le même graphique, la courbe d'équation $y = A\,e^{Bx}$ pour x élément de l'intervalle $[80 ; 120]$.

c) En utilisant cet ajustement, estimer la consommation aux 100 km (arrondie au dixième) de la voiture, pour une vitesse de 130 km/h.

4. Des deux valeurs obtenues dans les questions **2. d)** et **3. c)**, pour la consommation à une vitesse de 130 km/h, laquelle vous semble la plus proche de la consommation réelle ?
Expliquer votre choix.

24 On a relevé mois après mois, le coût d'amortissement d'une pompe hydraulique.

Mois x	1	2	3	4	5	6
Coût C (en €)	400	300	270	220	180	150

1. L'allure du nuage des points de la série $(x ; C)$ conduit à poser $y = \ln C$.

a) Dresser le tableau de la série statistique $(x ; y)$ en prenant des valeurs décimales arrondies à 10^{-3} près.

b) Calculer le coefficient de corrélation linéaire de cette série. (On en donnera une valeur décimale arrondie à 10^{-3} près.)

c) Justifier la pertinence d'un ajustement affine.

2. Déterminer une équation de la forme :

$$y = ax + b,$$

où a et b désignent des nombres réels, de la droite de régression de y en x.
(On prendra pour valeurs de a et b leurs valeurs décimales arrondies à 10^{-3} près.)

3. À partir du résultat de la question **2.**, déterminer l'expression de C, en fonction de x sous la forme $C = \alpha\,\beta^x$, où α et β sont des réels dont on donnera des valeurs approchées à 10^{-2} près.

4. Utiliser l'expression précédente pour évaluer le coût d'amortissement de la pompe au mois numéro 7, à 10^{-2} (€) près.