Histoire 1

Carl Freidrich Gauss (1777 - 1855), surnommé le *Prince des mathématiciens*, étudia des domaines très variés des mathématiques (arithmétique, analyse, probabilités, algèbre ...). Il fut aussi astronome et dirigea l'observatoire de Göttingen. En 1801, il introduisit la méthode des moindres carrés pour réaliser un ajustement de ses observations ce qui lui permit d'établir l'orbite de Cérès, découverte la même année par un astronome italien.

Karl Pearson (1857 - 1936) est considéré comme le fondateur de la statistique moderne avec la définition du coefficient de corrélation linéaire et la loi du χ^2 pour mesurer la qualité d'un ajustement.

1 Ajustement affine

1.1 Rappels sur les statistiques à une variable

Définition 1 ***Moyenne arithmétique et variance***

Soit une série statistique des valeurs d'un caractère quantitatif x mesurées sur une population de taille n. On note $(x_i)_{1\leqslant i\leqslant n}$ cette **série statistique à une variable**.

☞ La **moyenne arithmétique** de cette série se note $\overline{x}$ et elle est égale à :

$$\boxed{\overline{x}=\frac{1}{n}(x_1+x_2+\ldots+x_n)=\frac{1}{n}\sum_{i=1}^{n}x_i}$$

☞ La **variance** $\mathrm{V}(x)$ de cette série est la *moyenne des carrés des écarts à la moyenne*, c'est un nombre positif :

$$\mathrm{V}(x)=\frac{\left(x_1-\overline{x}\right)^2+\left(x_2-\overline{x}\right)^2+\cdots+\left(x_p-\overline{x}\right)^2}{n}$$

Avec le symbole de sommation $\sum$, on peut écrire :

$$\boxed{\mathrm{V}(x)=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}$$

☞ **L'écart-type** σ est la racine carré de la variance : $\sigma_x=\sqrt{\mathrm{V}(x)}$.

L'écart-type est homogène à la moyenne : si la moyenne est en mètres, l'écart-type est en mètres alors que la variance est en mètres carrés.

Pour une série statistique à une variable :

- la moyenne arithmétique est un **indicateur de tendance centrale**;
- l'écart-type est un **indicateur de dispersion**.

Capacité 1 *Calculer la moyenne et la variance d'une série statistique à 1 variable*

On considère les séries de notes de deux groupes d'élèves :

Groupe A

Note	5	8	9	11	14	15
Effectif	2	2	1	1	4	2

Groupe B

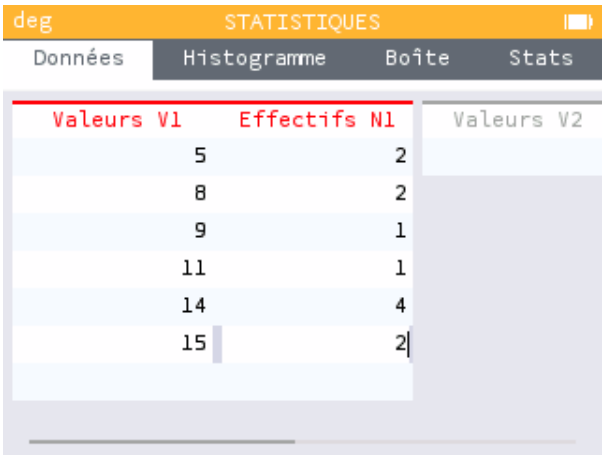
9 - 9 -10-10-11-11-
11-11-12-12-13-13-

1. À l'aide du module *Statistiques* de la calculatrice, calculer la moyenne, la variance et l'écart-type de la série de notes du groupe A.

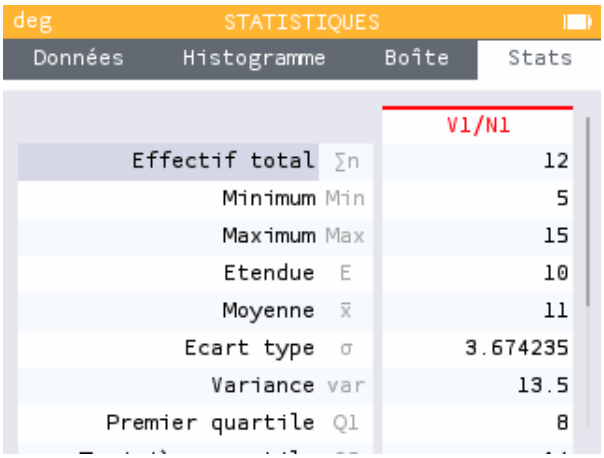
⚠ Les valeurs identiques ont été regroupées d'où l'apparition d'un paramètre d'effectif n_i pour une valeur x_i.

On donne ci-dessous les deux étapes avec une calculatrice Numworks, un tutoriel pour calculatrice TI est téléchargeable sur la page 36 élèves 36 calculatrices.

Saisie des données



Affichage des indicateurs



2. Calculer la moyenne, la variance et l'écart-type de la série de notes du groupe B.

Moyenne : 11
Écart-type : ≈ 1,29

Comparer les deux groupes à partir des couples (moyenne, écart-type).

Les deux groupes ont la même moyenne (même tendance centrale) mais l'écart-type du groupe B est plus petit ce qui traduit une plus grande homogénéité.

Algorithmique 1

1. Compléter le code de la fonction `moyenne(liste_notes)` ci-dessous pour qu'elle retourne la moyenne de la liste de notes passée en argument, en supposant que celle-ci est non vide.

```
def moyenne(liste_notes):
    somme = 0
```

```
    effectif = len(liste_notes)
    for k in range(effectif):
        somme = somme + liste_notes[k]
    return somme/effectif

#test sur le groupe B
groupe_B = [9,9,10, 10, 11, 11, 11, 11, 12, 12, 13, 13]
moyenne(groupe_B) == 11.0
```

2. Compléter les codes Python des fonctions `variance(liste_notes)` et `ecart_type(liste_notes)` ci-dessous pour qu'elles retournent respectivement la variance et l'écart-type de la liste de nombres `liste_notes`.

```
from math import sqrt

def variance(liste_notes):
    m = moyenne(liste_notes)
    effectif = len(liste_notes)
    somme = 0
    for note in liste_notes:
        somme = somme + (note-m)**2
    return somme / effectif

def ecart_type(liste_notes):
    return sqrt(variance(liste_notes))

assert round(variance(groupe_B), 2) == 1.67
```

1.2 Série statistique à deux variables

Définition 2

Sur une population de taille n, on étudie deux caractères statistiques quantitatifs x et y.
Pour chacun des n individus de la population, on note x_i et y_i l es valeurs respectives des caractères x et y.
On obtient alors une **série statistique à deux variables quantitatives** ou **série statistique double** qui peut être représentée par :

- une liste de couples $\left((x_i\,;y_i)\right)_{1\leqslant i\leqslant n}$;
- ou un tableau.

x_i	x_1	x_2	$\ldots$	x_n
y_i	y_1	y_2	$\ldots$	y_n

Contrairement aux séries statistiques à une variable, on n'associe pas d'effectifs aux valeurs x_i ou y_i.
En général les couples $(x_i;y_i)$ sont ordonnés par valeurs croissantes des x_i.

Exemple 1

Dans une grande surface, on considère une **série statistique à deux variables quantitatives** constituée de six valeurs du caractère x des *Dépenses publicitaires mensuelles* et du caractère y du *Chiffre d'affaires mensuel*.

Dépense publicitaire x_i	Chiffre d'affaires y_i
5	400
20	460
30	870
42	1070
50	980
60	1170

1.3 Nuage de point et point moyen

Définition 3

Soit une série statistique à deux variables x et y représentée par le tableau :

x_i	x_1	x_2	$\ldots$	x_n
y_i	y_1	y_2	$\ldots$	y_n

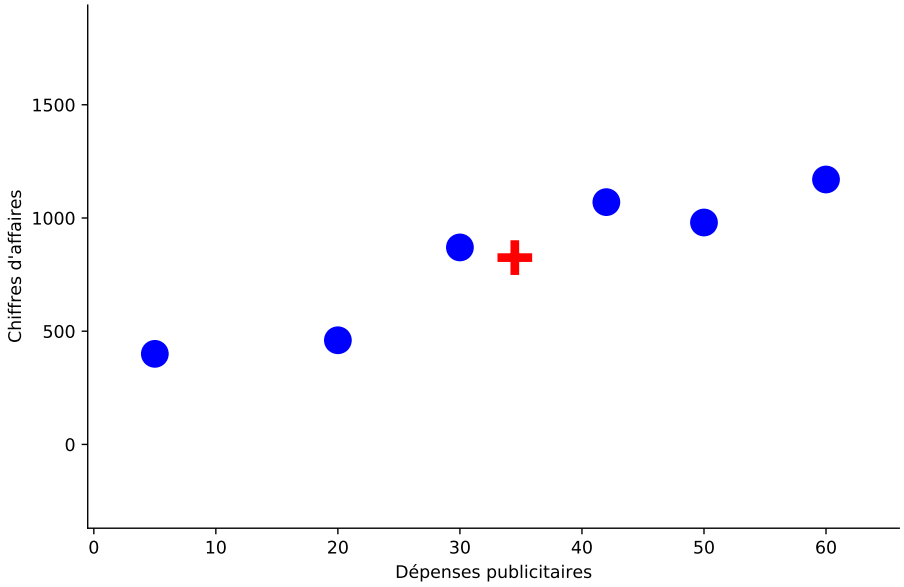
Dans un repère orthogonal du plan, on choisit de représenter le caractère x en abscisse et le caractère y en ordonnée.

- Le **nuage de points** représentant la série statistique est l'ensemble des points $\left(M_i\left(x_i\,;y_i\right)\right)_{1\leqslant i\leqslant n}$.

- Le **point moyen** du nuage est le point de coordonnées $(\overline{x}\,;\overline{y})$ où $\overline{x}$ est la moyenne arithmétique de la série $(x_i)_{1\leqslant i\leqslant n}$ et $\overline{y}$ est la moyenne de la série $(y_i)_{1\leqslant i\leqslant n}$.

Exemple 2

On a représenté ci-dessous le nuage de points (des ronds) et le point moyen (une croix) de la série statistique double de l'exemple 1.



Capacité 2 *Représenter un nuage de points, calculer les coordonnées d'un point moyen*

Le tableau ci-dessous donne le nombre des unions civiles, PACS* ou mariages, enregistrées en France entre 2005 et 2016. (*PACS*. Pacte Civil de Solidarité*)

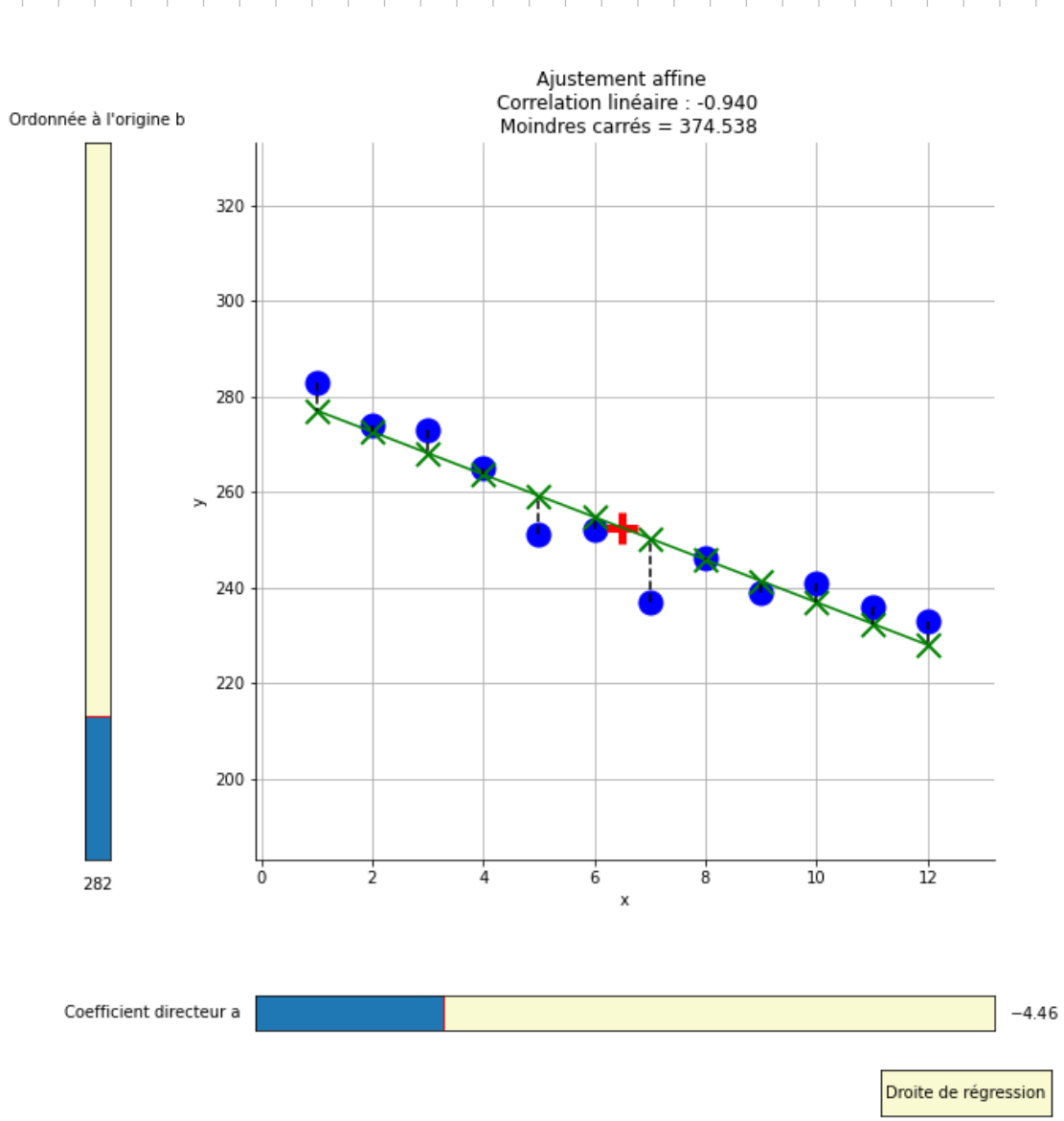
Année	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Rang de l'année	1	2	3	4	5	6	7	8	9	10	11	12
Nombre de mariages (en milliers)	283	274	273	265	251	252	237	246	239	241	236	233
Nombre de PACS (en milliers)	60	77	102	146	174	205	152	160	169	174	189	192

(d'après INSEE *Mariages et PACS en* 2017)

1. Sur le graphique ci-dessous, représenter le nuage de points de coordonnées $(x_i\,;\,y_i)$ où x_i désigne le rang de l'année et y_i le nombre de mariages.

2. Calculer les coordonnées du point moyen G. Placer G dans le repère.

G(6,5 ; 252,5)



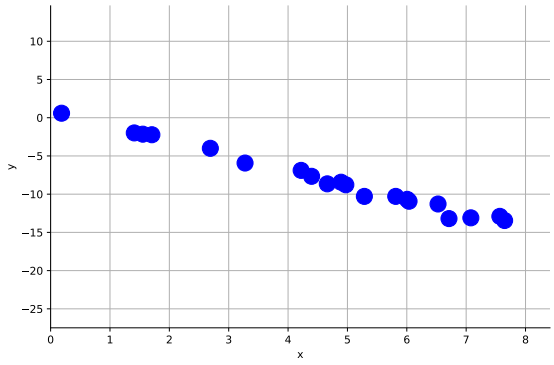
deg REGRESSION

Data | Graph | Stats

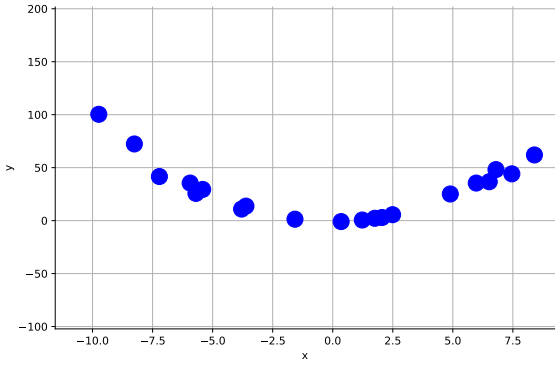
	X1	Y1
Mean	6.5	252.5
Sum	78	3030
Sum of squares	650	768296
dard deviation	3.452053	16.38343
Variance	11.91667	268.4167
nber of points		12

- un **ajustement affine** d'équation $y = ax + b$ si les points semblent alignés le long d'une droite ;
- un **ajustement parabolique** d'équation $y = ax^2 + bx + c$ si les points semblent alignés le long d'une parabole ;
- un **ajustement exponentiel** d'équation $y = ke^{ax+b}$ si les points semblent alignés le long d'une courbe d'exponentielle ...

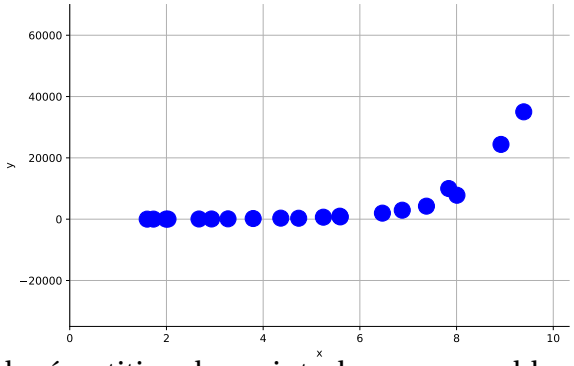
Ajustement affine



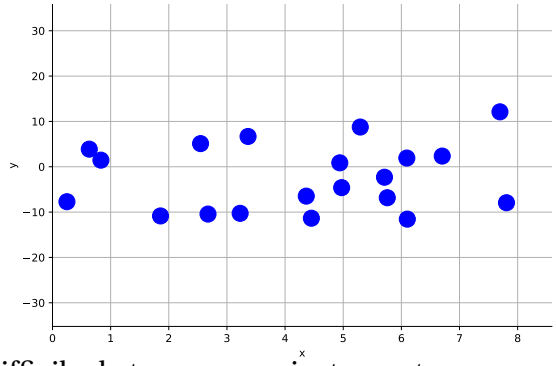
Ajustement parabolique



Ajustement exponentiel



Nuage aléatoire



Si la répartition des points du nuage semble aléatoire, il est difficile de trouver un ajustement.
Nous étudierons dans ce cours les **ajustements affines**, les ajustements exponentiel et parabolique pouvant s'y ramener par changement de variable.

Définition 5

Réaliser un **ajustement affine** d'une série statistique à deux variables $\left((x_i ; y_i)\right)_{1\leqslant i\leqslant n}$ consiste à déterminer des coefficients réels a et b tels que la droite d'équation $y = ax + b$ passe *au plus près* de l'ensemble des points $\left(M_i\left(x_i ; y_i\right)\right)_{1\leqslant i\leqslant n}$ du nuage.

La mesure de cette distance entre la droite d'ajustement et le nuage sera précisée par la méthode des moindres carrés.

Capacité 3 *Réaliser un ajustement affine, voir exo 1 p. 257*

On reprend l'énoncé de la capacité 2.

1. On réalise un ajustement affine du nuage de points de la série statistique double $((x_i\,;y_i))_{1\leqslant i\leqslant 12}$ à l'aide de la droite (d_1) d'équation : $y=-4,4x+281,1$.

Tracer la droite (d_1) sur le graphique en indiquant les coordonnées des points utilisés.

Voir ci-dessous

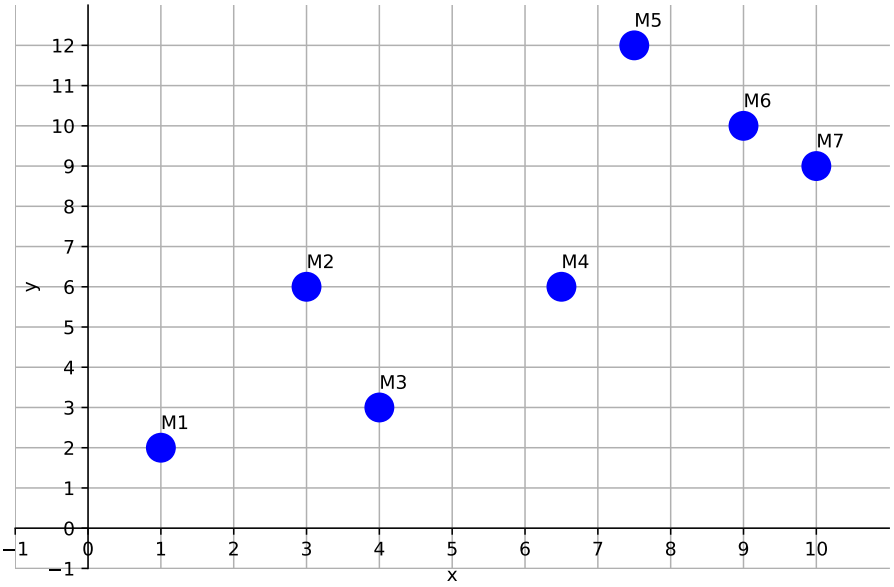
2. On suppose que ce modèle d'ajustement reste valable jusqu'en 2025. Déterminer par extrapolation le nombre de mariages prévisibles en 2020. Préciser la démarche utilisée.

x = 1 pour 2005 donc x = 16 pour 2020

Par extrapolation, on peut estimer en 2020 le nombre de mariages avec l'ajustement affine : -4,4 * 16 + 281,1 = 210,7 milliers

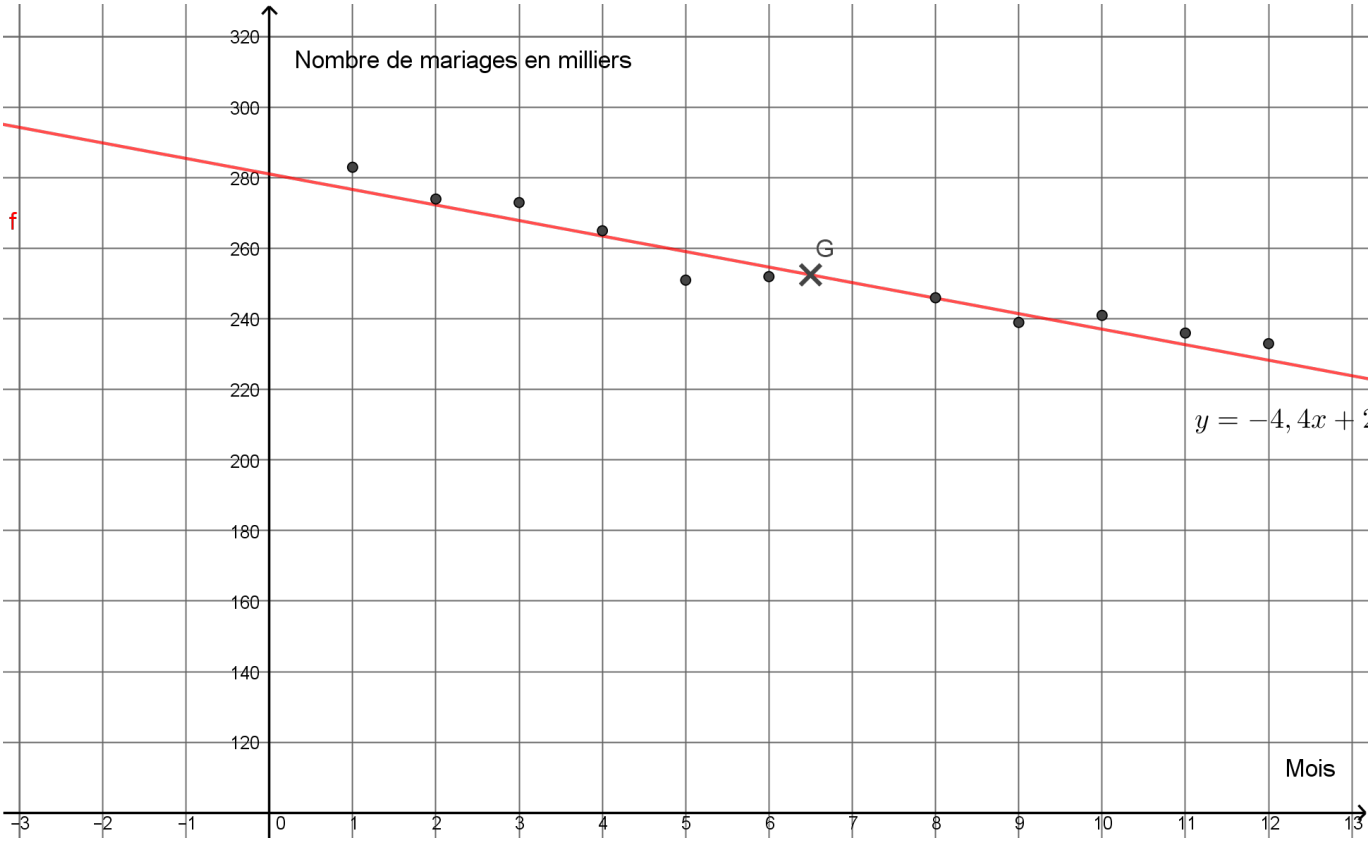
Capacité 4 *Réaliser un ajustement affine*

On considère une série statistique à deux variables quantitatives dont on donne le nuage de points ci-dessous.

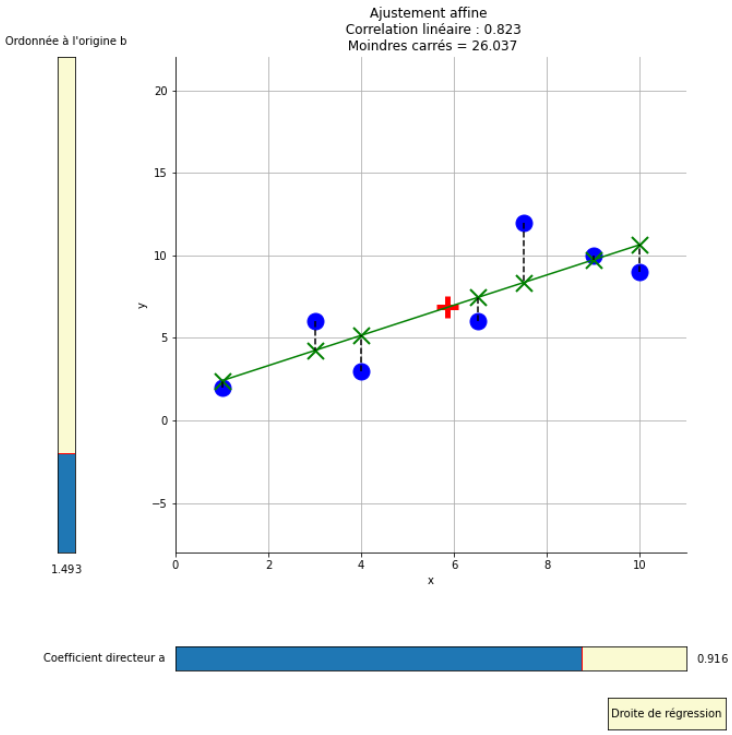


1. Calculer les coordonnées du point moyen G du nuage et le placer dans le repère.

G a pour coordonnées approchées à 0,01 près (5;86 ; 6;86)



Capacité 2 question 1)



Capacité 4

2. Quelle droite passant par deux points distincts du nuage semble réaliser le meilleur ajustement affine possible du nuage ?

On peut conjecturer qu'une droite de meilleur ajustement passe par le point moyen en minimisant la somme des distances aux autres points

3. Déterminer une équation réduite de la droite d'ajustement choisie.

En approchant les coefficients à 0,001 près :
$y = 0{,}916\,x + 1{,}493$

2 Méthode des moindres carrés

2.1 Principe de la méthode

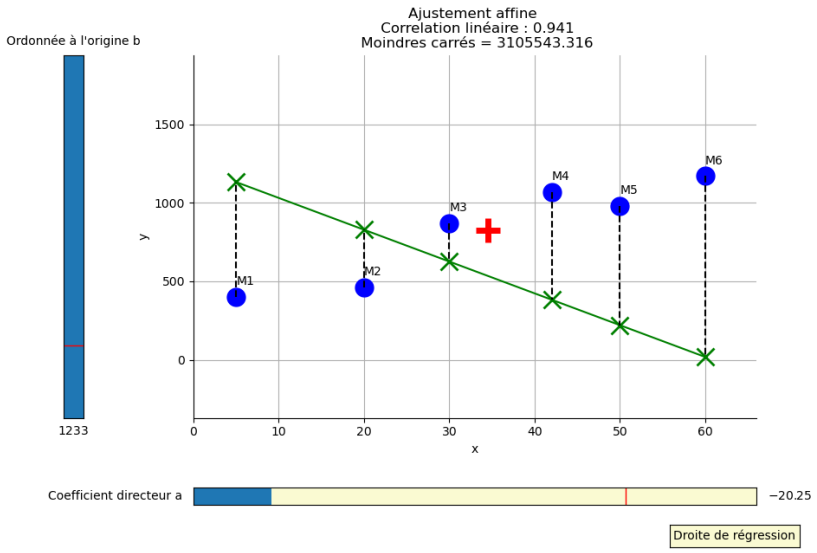
Méthode

Dans un repère orthogonal du plan, étant donné le nuage de points $\left(M_i\left(x_i\,;\,y_i\right)\right)_{1\leqslant i\leqslant n}$ d'une série statistique à deux variables et une droite $\mathscr{D}$ d'équation $y = ax + b$, on associe à chaque point M_i le point $A_i\left(x_i\,;\,ax_i + b\right)$ de $\mathscr{D}$ qui a même abscisse.
On calcule ensuite la somme S des carrés des distances :

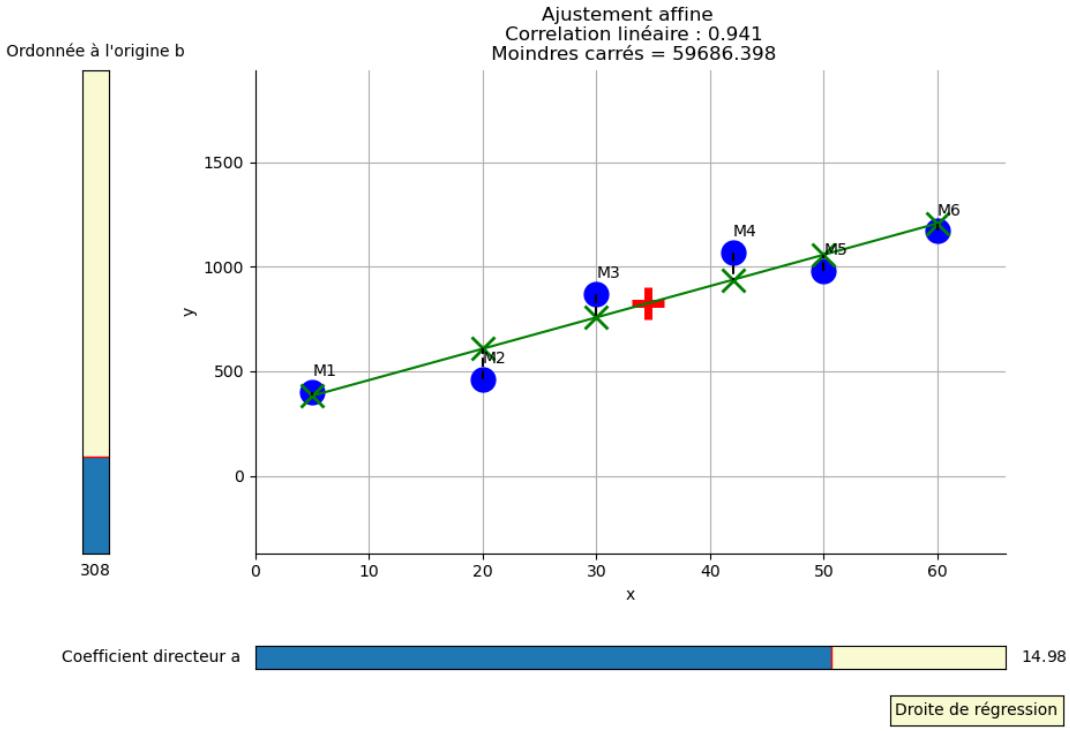
$$S = M_1A_1^2 + M_2A_2^2 + \ldots + M_nA_n^2 = \sum_{i=1}^{n} M_iA_i^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

Cette somme S représente la distance entre la droite $\mathscr{D}$ et l'ensemble des points du nuage, c'est pourquoi la **méthode des moindres carrés** choisit comme droite d'ajustement affine du nuage la droite $\mathscr{D}$ d'équation $y = ax + b$ telle que le couple $(a;b)$ rend minimale la somme S.

Mauvaise droite d'ajustement



Meilleure droite d'ajustement



2.2 Droite d'ajustement par la méthode des moindres carrés

Définition 6

La **covariance** d'une série statistique à deux variables quantitatives $\left((x_i\,;\,y_i)\right)_{1\leqslant i\leqslant n}$ se note σ_{xy} ou $\text{cov}(x,y)$, elle est égale à :

$$\boxed{\sigma_{xy}=\text{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}$$

$\overline{x}$ et $\overline{y}$ sont les moyennes arithmétiques des séries statistiques à une variable $(x_i)_{1\leqslant i\leqslant n}$ et $\left(y_i\right)_{1\leqslant i\leqslant n}$.

Corollaire

Soit $(x_i)_{1\leqslant i\leqslant n}$ une série statistique à une variable, $\text{cov}(x,x)$ est la variance de la série x notée $V(x)$.

Propriété 1 *Droite des moindres carrés*

Soit une série statistique à deux variables quantitatives dont le nuage de points est $\left(M_i\left(x_i\,;\,y_i\right)\right)_{1\leqslant i\leqslant n}$ dans un repère orthogonal du plan.
La **méthode des moindres carrés** donne une unique droite d'ajustement qui passe par le point moyen du nuage $\boxed{G\left(\overline{x};\overline{y}\right)}$ et a pour équation $\boxed{y=ax+b}$ avec :

$$\boxed{a = \frac{\text{cov}(x,y)}{V(x)} \text{ et } b = \overline{y} - a\overline{x}}$$

La droite d"ajustement d'un nuage de points obtenue avec la méthode des moindres carrés s'appelle une **droite de régression**.

Démonstration *Au programme*

Dans un repère orthogonal du plan, soit une série statistique à deux variables quantitatives dont le nuage de points est $\left(M_i\left(x_i\,;\,y_i\right)\right)_{1\leqslant i\leqslant n}$ et une droite d'équation $y = ax + b$.
D'après la méthode des moindres carrés, on recherche une condition sur a et b pour que la somme ci-dessous soit minimale :

$$S = \sum_{i=1}^{n} M_i A_i^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

Étape 1 : on fixe a

On développe et on isole les facteurs b et b^2 :

$$S = \sum_{i=1}^{n} ((y_i - ax_i) - b)^2 = \sum_{i=1}^{n} (y_i - ax_i)^2 - 2b(y_i - ax_i) + b^2$$
$$S = \sum_{i=1}^{n} (y_i - ax_i)^2 - 2b\sum_{i=1}^{n} (y_i - ax_i) + b^2 \sum_{i=1}^{n} 1$$
$$S = nb^2 - 2b\sum_{i=1}^{n} (y_i - ax_i) + \sum_{i=1}^{n} (y_i - ax_i)^2$$

On peut ainsi exprimer S comme une fonction polynôme du second degré en b de coefficients :
$S = ub^2 + vb + w$ avec $u = n$, $v = -2\sum_{i=1}^{n}(y_i - ax_i)$ et $w = sum_{i=1}^{n}(y_i - ax_i)^2$. .....

Pour a fixé S atteint un minimum pour : $b = -\frac{v}{2u} = \frac{\sum_{i=1}^{n}(y_i - ax_i)}{n} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right) - a\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \overline{y} - a\overline{x}$ .....
On en déduit que pour a fixé S est minimale si et seulement si $b = \overline{y} - a\overline{x}$.
.....

On en déduit que pour a fixé S est minimale si et seulement si $b = \overline{y} - a\overline{x}$.

Étape 2 : on remplace b

D'après l'étape 1 S est minimale pour (a, b) si et seulement si S est minimale pour $(a, b) = (a, \overline{y} - a\overline{x})$.
On peut donc rechercher le minimum de S en remplaçant b par $\overline{y} - a\overline{x}$ dans son expression puis en développant et en isolant les termes en a et en a^2 pour obtenir une fonction polynôme du second degré en a.

$$S = \sum_{i=1}^{n} ((y_i - (ax_i + \overline{y} - a\overline{x}))^2 = \sum_{i=1}^{n} ((y_i - \overline{y}) - a(x_i - \overline{x}))^2$$

$$S = \sum_{i=1}^{n} (y_i - \overline{y})^2 - 2a(y_i - \overline{y})(x_i - \overline{x}) + a^2(x_i - \overline{x})^2$$

$$S = a^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 - 2a \sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x}) + \sum_{i=1}^{n} (y_i - \overline{y})^2$$

On reconnaît les expressions de V(x), cov(x,y) et V(y)

$$S = a^2 n\mathrm{V}(x) - 2an\mathrm{cov}(x, y) + n\mathrm{V}(y)$$

On peut ainsi exprimer S comme une fonction polynôme du second degré en a de coefficients :
$S(a) = ua^2 + av + w$ avec $u = n\mathrm{V}(x)$, $v = -2n\mathrm{cov}(x, y)$ et $w = n\mathrm{V}(y)$.
Ainsi S atteint un minimum pour : $a = -\dfrac{-2n\mathrm{cov}(x, y)}{2n\mathrm{V}(x)} = \dfrac{\mathrm{cov}(x, y)}{\mathrm{V}(x)}$
Finalement S atteint un minimum si et seulement si $a = \dfrac{\mathrm{cov}(x, y)}{\mathrm{V}(x)}$ et $b = \overline{y} - a\overline{x}$.
Ce sont les coefficients de la droite d'ajustement déterminée par la méthode des moindres carrés.

Finalement S atteint un minimum si et seulement si $a = \dfrac{\mathrm{cov}(x, y)}{\mathrm{V}(x)}$ $b = \overline{y} - a\overline{x}$.
Ce sont les coefficients de la droite d'ajustement déterminée par la méthode des moindres carrés.

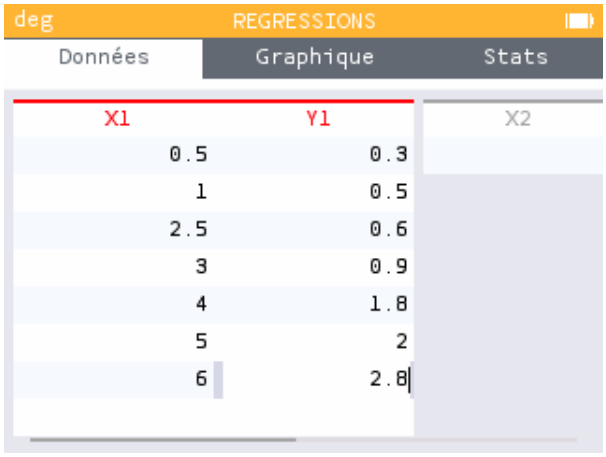
Méthode *Droite de régression avec la calculatrice*

Le webmestre d'un site de streaming s'intéresse à la corrélation entre la durée de téléchargement d'une vidéo et le nombre de clients connectés.
On considère une série statistique à deux variables quantitatives représentée par le tableau ci-dessous :

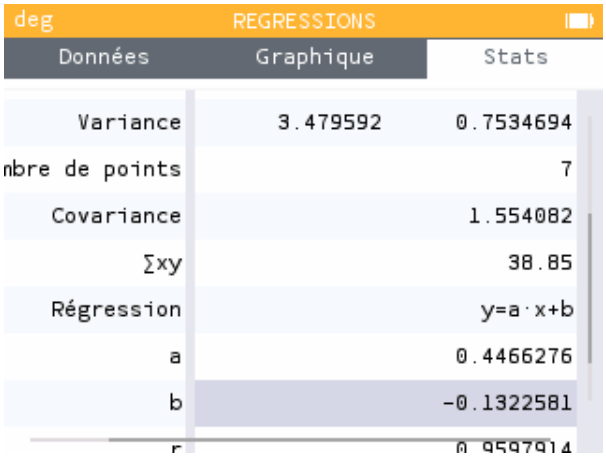
Nombre de clients en milliers x_i	0,5	1	2,5	3	4	5	6
Durée de téléchargement en secondes y_i	0,3	0,5	0,6	0,9	1,8	2	2,8

Avec le mode *Régressions* de la calculatrice, on peut saisir les données, calculer les indicateurs (V(x), cov(x, y)) et coefficients de la droite de régression et obtenir une représentation graphique.
On donne ci-dessous les principales étapes pour notre exemple sur une calculatrice Numworks.
Un tutoriel pour calculatrice TI est disponible sur :
https://www.lestutosmaths.fr/fr/statistiques/stats2variables

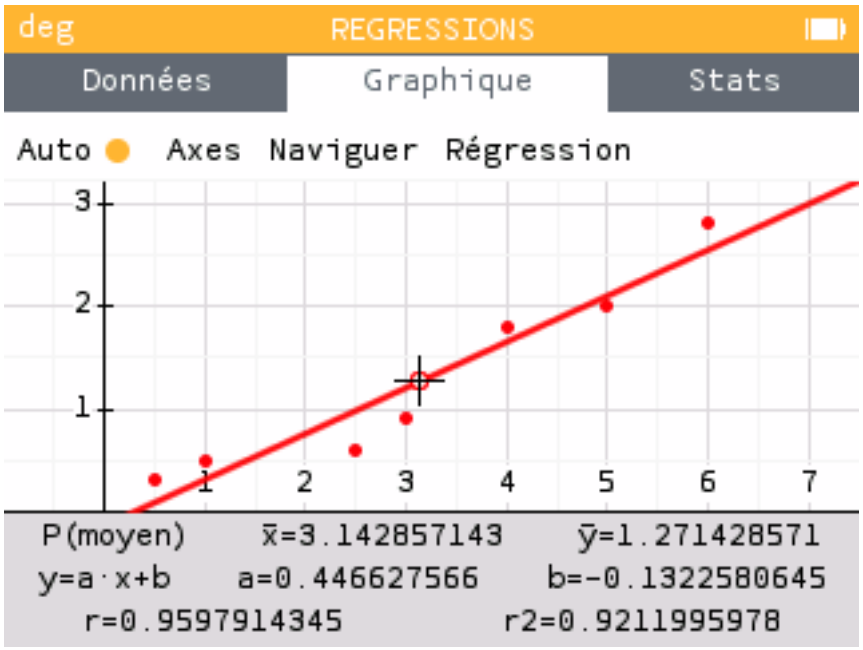
Étape 1 : saisie



Étape 2 : calculs



Étape 3 : graphique



Capacité 5 *Déterminer une droite de régression, voir exo 4 p.259*

Le glacier d'Aletsch, classé à l'UNESCO, est le plus grand glacier des Alpes, situé dans le sud de la Suisse, il alimente la vallée du Rhône.

Pour étudier le recul de ce glacier au fil des années, une première mesure a été effectuée en 1900 : ce glacier mesurait alors 25,6 km.

Des relevés ont ensuite été effectués tous les 20 ans : le recul du glacier est mesuré par rapport à la position où se trouvait initialement le pied du glacier en 1900.

Les mesures successives ont été relevées dans le tableau ci-dessous. On note x la durée, en années, écoulée depuis 1900, et y le recul correspondant, mesuré en kilomètres.

Année de mesure :	1900	1920	1940	1960	1980	2000
Durée x écoulée (depuis 1900) :	0	20	40	60	80	100
Recul y (en km) :	0	0,3	0,6	1	1,6	2,3

Mesures déduites de : The Swiss Glaciers, Yearbooks of the Glaciological Commission of the Swiss

Par exemple, en 1940 ($x = 40$), le recul du glacier par rapport à 1900 a été de 0,6 km : la longueur du glacier était donc de $25,6 - 0,6 = 25$km.

Dans cet exercice, les résultats seront arrondis, si nécessaire, à 10^{-3} près.

1. Tracer le nuage de points dans le repère donné en fin d'exercice (Durée x en abscisse, distance y en ordonnée).

2. Compléter le tableau ci-dessous.

x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$\times$
0	0	...	...	...	...	$\times$
20	0,3	...	...	...	...	$\times$
40	0,6	...	...	...	...	$\times$
60	1	...	...	...	...	$\times$
80	1,6	...	...	...	...	$\times$
100	2,3	...	...	...	...	$\times$
$\times$	$\times$	$\times$	$\times$	$\times$	...	Total

Déterminer les moyennes $\overline{x}$ et $\overline{y}$ des séries des caractères durée et recul, la variance $V(x)$ de la série des durées et la covariance $\text{cov}(x, y)$ de la série statistique à deux variables $\left((x_i\,;\,y_i)\right)_{1\leqslant i\leqslant 6}$.

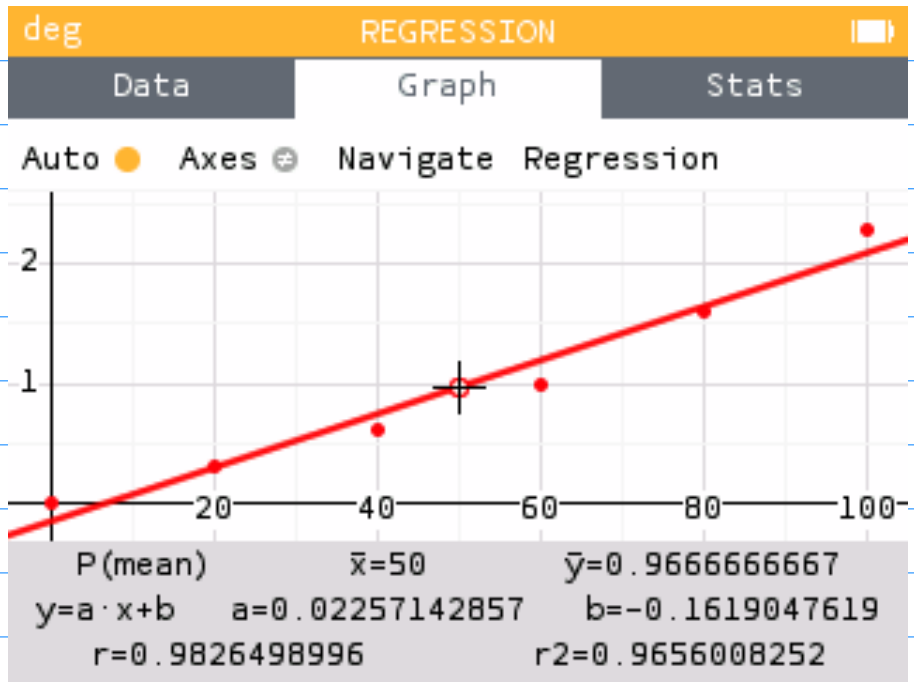
$\overline{x} = 50$ et $\overline{y} = 997$

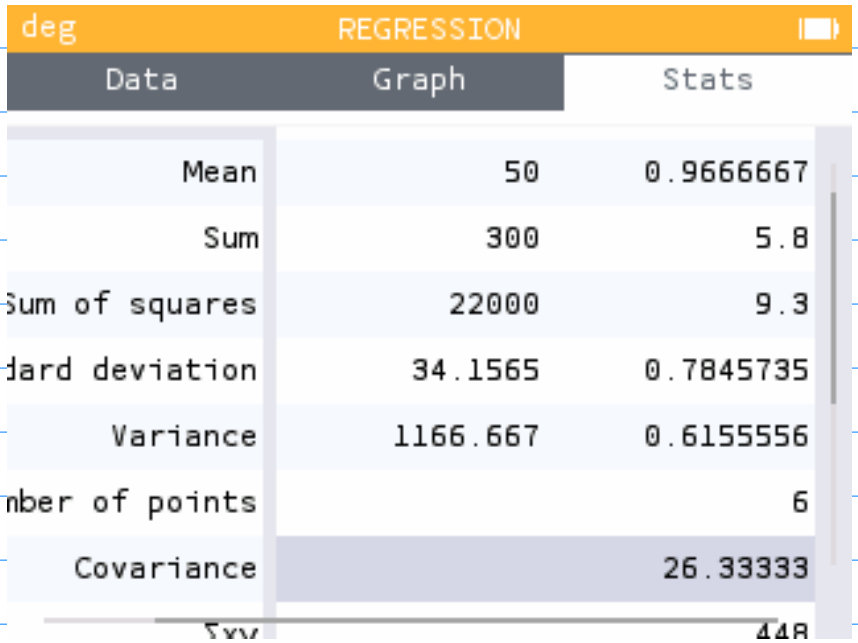
En déduire les coefficients de la droite d'ajustement affine par la méthode des moindres carrés de y en fonction de x.

$y = ax + b$ avec à 0;0001 près $b = -0;1619$ et $a = 0;0226$

3. Retrouver les résultats précédent à l'aide du mode *Régression* de la calculatrice.

	A	B	C	D	E	F	G
1	xi	yi	xi - xm...	yi - ymoy	(xi - xmoy)*(yi - y...	(xi-xmoy)²	
2	0	0	-50	-0.97	48.33	2500	
3	20	0.3	-30	-0.67	20	900	
4	40	0.6	-10	-0.37	3.67	100	
5	60	1	10	0.03	0.33	100	
6	80	1.6	30	0.63	19	900	
7	100	2.3	50	1.33	66.67	2500	
8							
9	xmoy	50	ymoy	0.97			
10							





Donner une équation de la droite d'ajustement affine par la méthode des moindres carrés de y en fonction de x et tracer cette droite dans le repère en précisant les coordonnées des points choisis.

Coefficients arrondis à 10^{-4} près

$y \approx 0{,}0226x - 0{,}1619$

4. À partir du modèle affine obtenu précédemment, estimer par le calcul :

a. Le recul puis la longueur du glacier en 1990.

1990 $\longrightarrow x = 90$

estimation par ajustement affine :

$y \approx 0{,}0226 \times 90 - 0{,}1619 \approx 1{,}8721$

b. Le recul puis la longueur du glacier en 2011.

2011 $\longrightarrow x = 111$

$y \approx 2{,}3467$

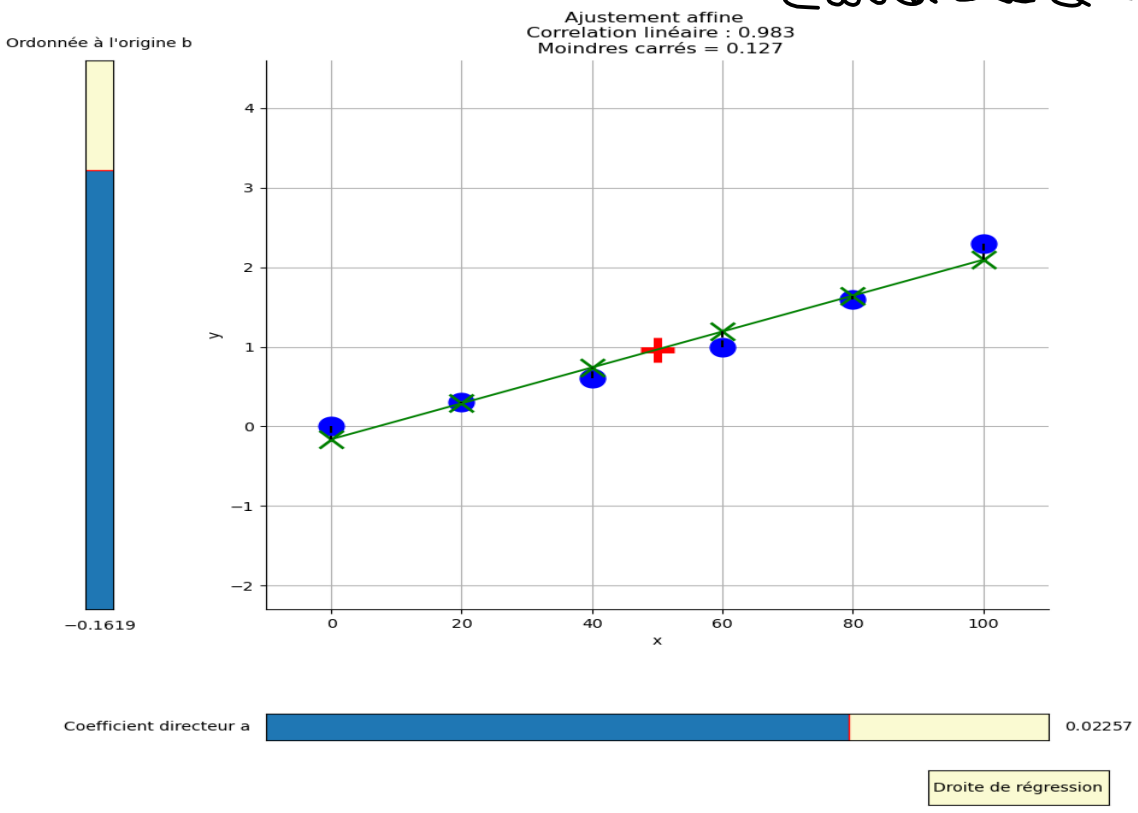
c. L'année de disparition du glacier (arrondir à l'unité).

On résout l'équation :

$25{,}6 = 0{,}0226x - 0{,}1619$

$\iff x = \dfrac{25{,}6 + 0{,}1619}{0{,}0226} \approx 1140$ ans

Arrondi à l'unité



3 Corrélation linéaire et changement de variable

3.1 Coefficient de corrélation linéaire

Définition 7

Soit une série statistique à deux variables quantitatives x et y :

x_i	x_1	x_2	$\ldots$	x_n
y_i	y_1	y_2	$\ldots$	y_n

Le **coefficient de corrélation linéaire** de la série statistique à deux variables x et y, se note r et il est égal à :

$$\boxed{r = \frac{\text{cov}(x,y)}{\sigma_x \times \sigma_y}}$$

- $\text{cov}(x,y)$ est la covariance de la série statistique à deux variables x et y
- $\sigma_x = \sqrt{\text{V}(x)}$ est l'écart-type de la série à une variable x
- $\sigma_y = \sqrt{\text{V}(y)}$ est l'écart-type de la série à une variable y

On peut calculer le **coefficient de corrélation linéaire** avec le mode *Régressions* de la calculatrice.

Propriété 2

Soit r le **coefficient de corrélation linéaire** d'une série statistique à deux variables.

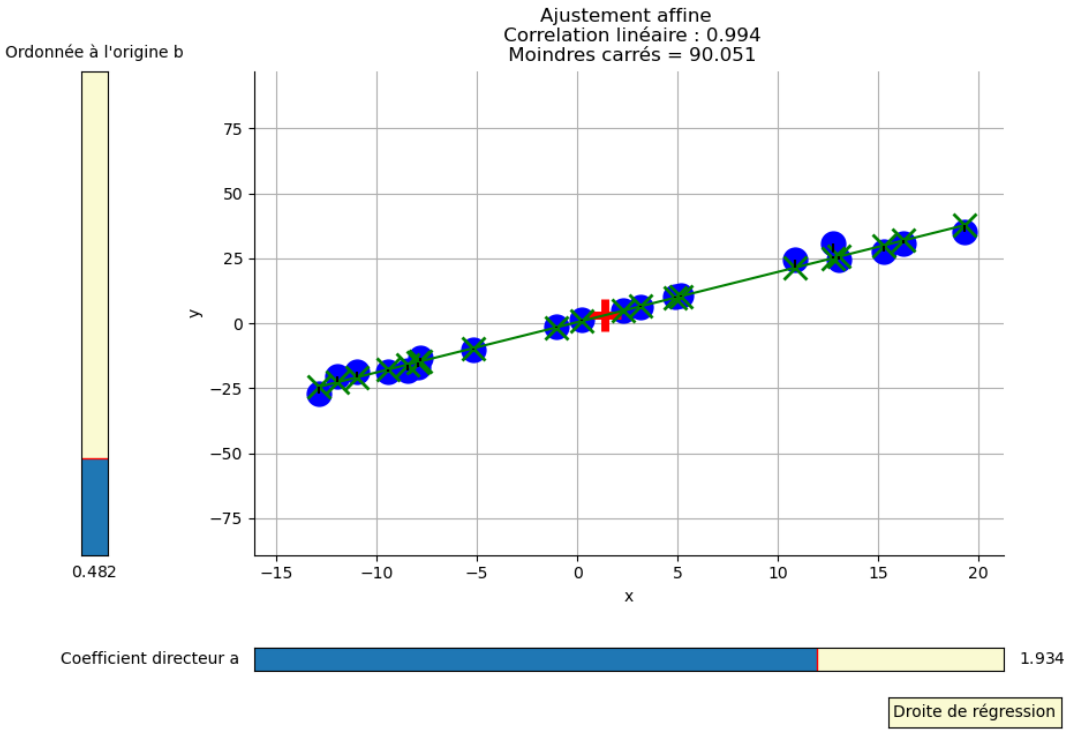
☞ r est un réel compris entre -1 et 1, $\boxed{-1 \leqslant r \leqslant 1}$;

☞ la valeur de r permet de mesure la qualité d'un ajustement affine obtenu par la méthode des moindres carrés :

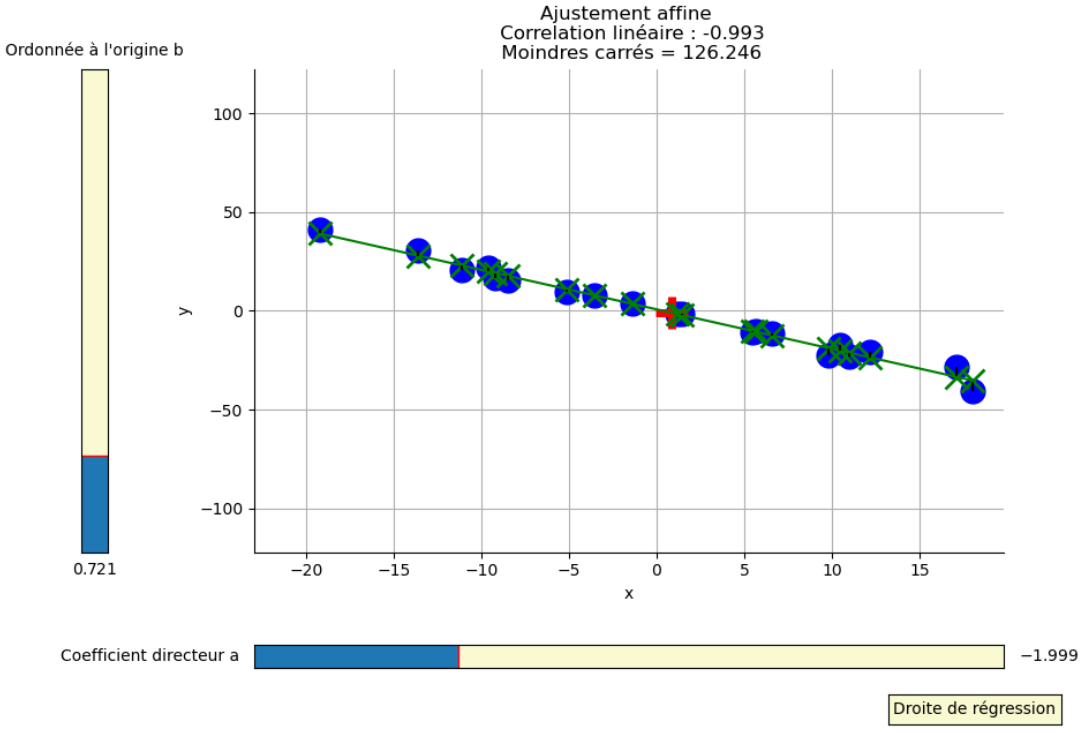
- plus la valeur absolue de r est proche de 0, moins la relation affine entre les variables x et y est forte ;
- si la valeur de r est proche de -1, les variables x et y ont une relation affine forte et varient dans des sens opposés ;
- si la valeur de r est proche de 1, les variables x et y ont une relation affine forte et varient dans le même sens.

On a représenté ci-dessous 4 nuages de points avec la valeur du coefficient de corrélation linéaire et la droite d'ajustement par la méthode des moindres carrés.

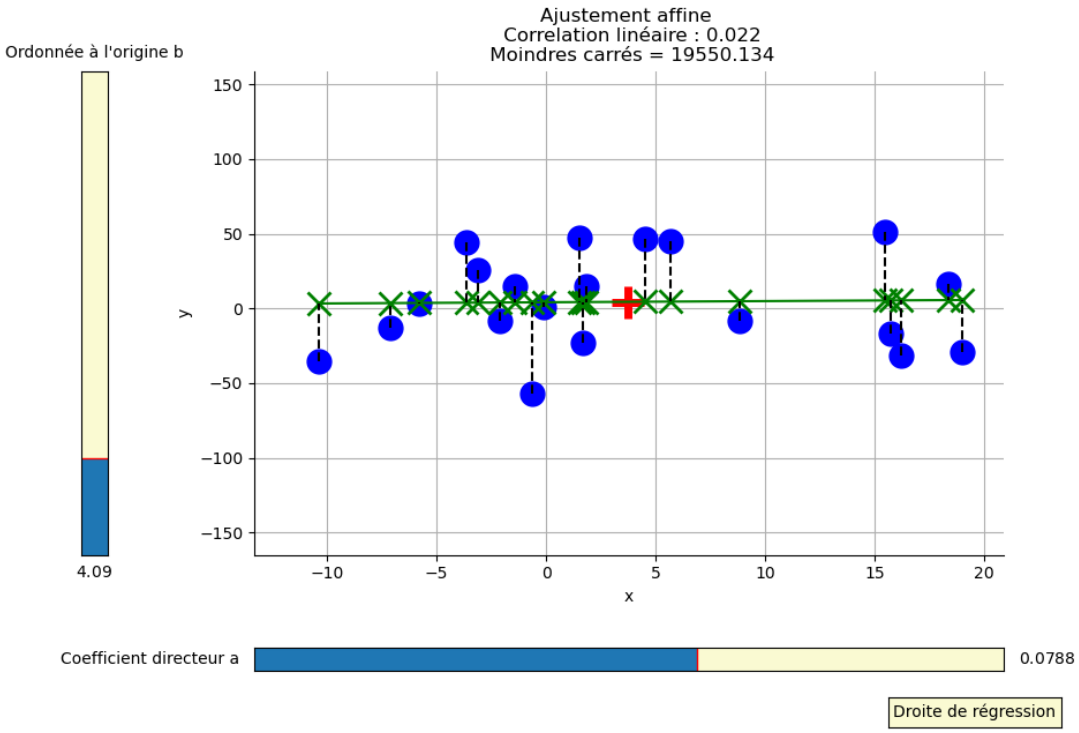
$\boxed{r \textbf{ proche de } 1}$



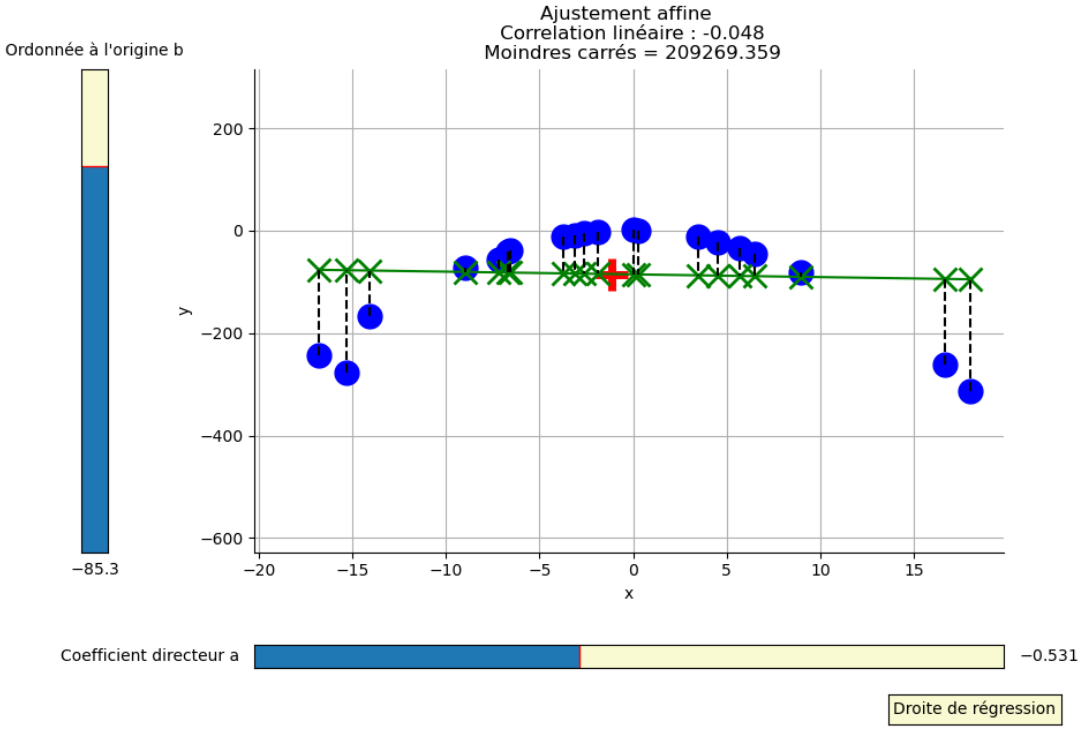
r proche de -1



$r > 0$ proche de 0



$r < 0$ **proche de** 0



Capacité 6 *Déterminer et utiliser le coefficient de corrélation linéaire, voir exo 27 p.265*

Le tableau ci-dessous donne le montant du SMIC horaire, en euros, entre 2013 et 2019.

Année	2013	2014	2015	2016	2017	2018	2019
Rang x_i	1	2	3	4	5	6	7
SMIC horaire y_i	9,43	9,53	9,61	9,67	9,76	9,88	10,03

Calculer et interpréter le coefficient de corrélation linéaire de cette série statistique à deux variables quantitatives.

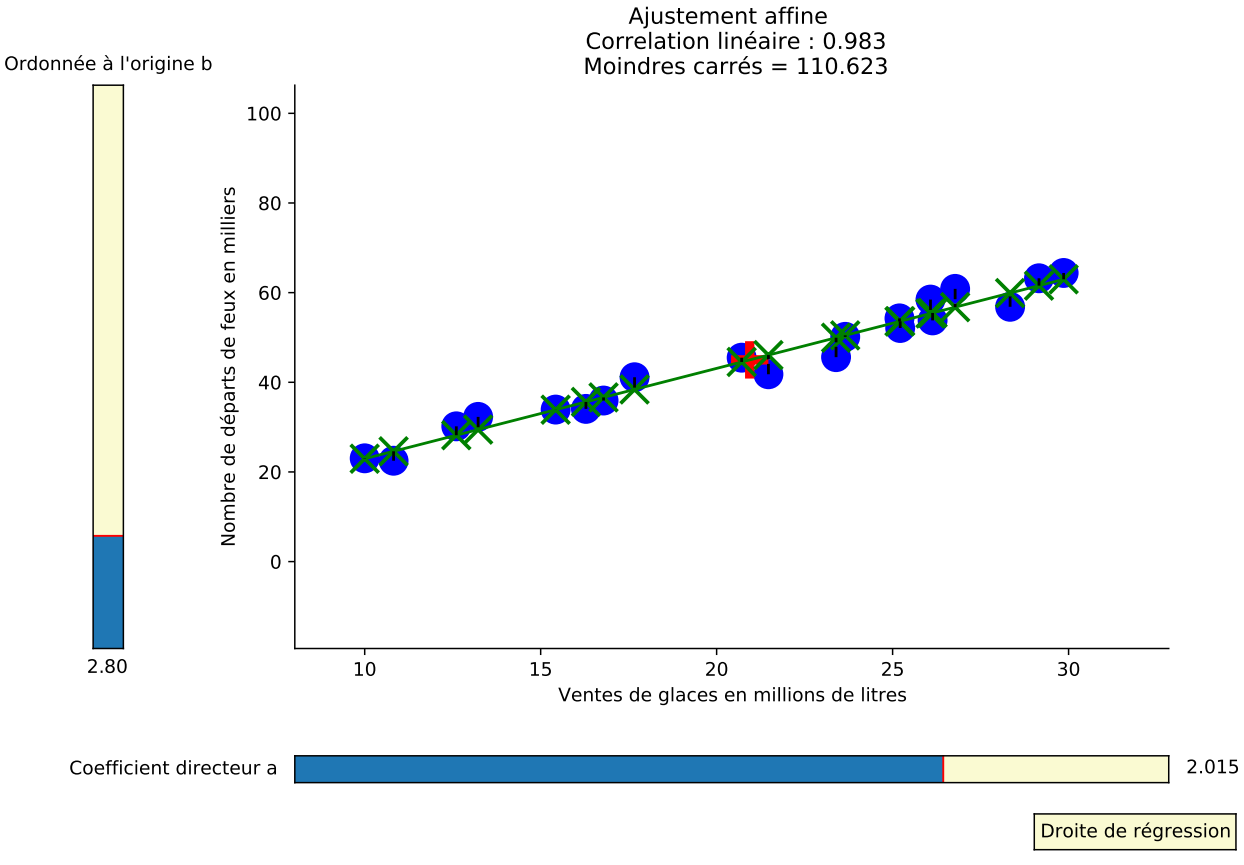
.....

.....

.....

Capacité 7 *Esprit critique*

On a mesuré sur 20 années consécutives dans un certain pays les ventes de glaces vendues en millions de litres (caractère x) et le nombre de départs de feux de forêts en milliers (caractère y) dans un certain pays.
La série statistique à deux variables ainsi obtenue est représentée par le nuage de points ci-dessous avec sa droite de régression et le coefficient de corrélation linéaire

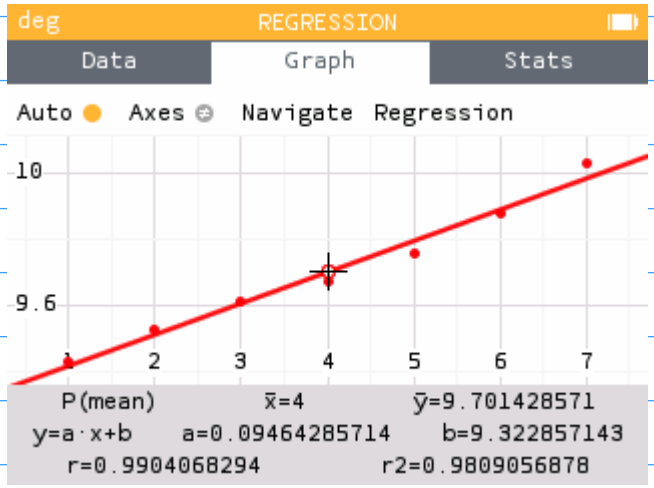


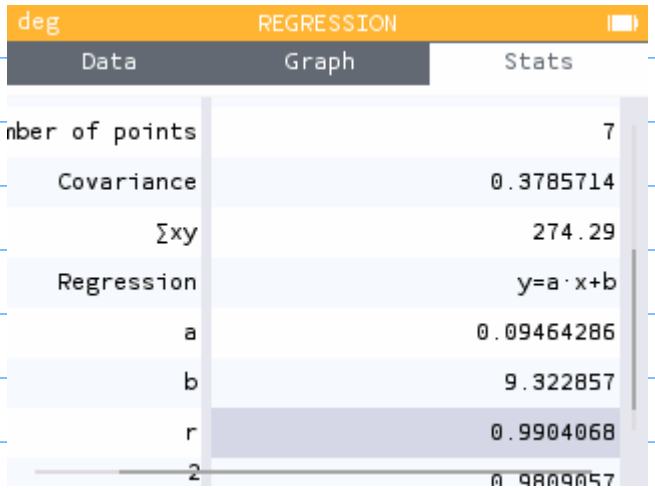
Peut-on établir une relation de causalité entre le nombre de glaces vendues et le nombre de départs de feux?

.....

Il est évident qu'il n'existe pas de lien de causalité entre les ventes de glaces et les incendies. On peut conjecturer que chacune de ces grandeurs est corrélée à une autre grandeur cachée (température, sécheresse ...).

Capacité 6





Le coefficient de corrélation linéaire est proche de 1
ce qui indique une forte corrélation entre x et y
qui varient dans le même sens

3.2 Ajustement affine après changement de variable

Méthode

Si le nuage de points associé à une série statistique à deux variables x et y ne se prête pas à un ajustement affine, on peut rechercher, selon la forme du nuage, un ajustement par une autre courbe (voir Définition 4) :

- d'équation $y = ax^2 + bx + c$ pour un **ajustement parabolique** ;
- d'équation $y = \mathrm{e}^{ax+b}$ pour un **ajustement exponentiel** ;
- d'équation $y = \ln(ax+b)$ pour un **ajustement logarithmique**.

Une méthode consiste à se ramener à un ajustement affine par **changement de variable**.

☞ **Étape 1 :** On choisit une fonction f telle que le nuage de points de coordonnées $\left(\left(x_i\,;\,f(y_i)\right)\right)_{1\leqslant i\leqslant n}$ permette un ajustement affine. Le coefficient de corrélation linéaire et la forme du nuage permettent d'affiner le choix du **changement de variable** f.

- $f(y) = \sqrt{y-c}$ pour un ajustement parabolique ;
- $f(y) = \ln(y)$ pour un ajustement exponentiel ;
- $f(y) = \mathrm{e}^y$ pour un ajustement logarithmique.

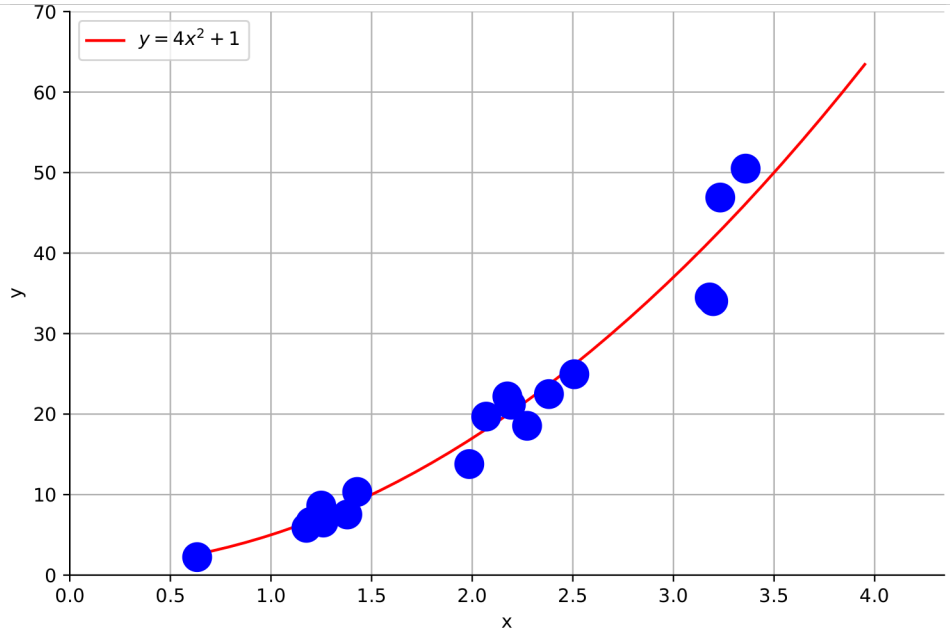
☞ **Étape 2 :** On fait le changement de variable $z = f(y)$ et pour le nuage de points $\left(\left(x_i\,;\,z_i = f(y_i)\right)\right)_{1\leqslant i\leqslant n}$, on détermine la droite de régression de $z = f(y)$ par rapport à x : $\boxed{z = ax+b}$.

☞ **Étape 3 :** On exprime y en fonction de x à l'aide de la fonction réciproque de f :

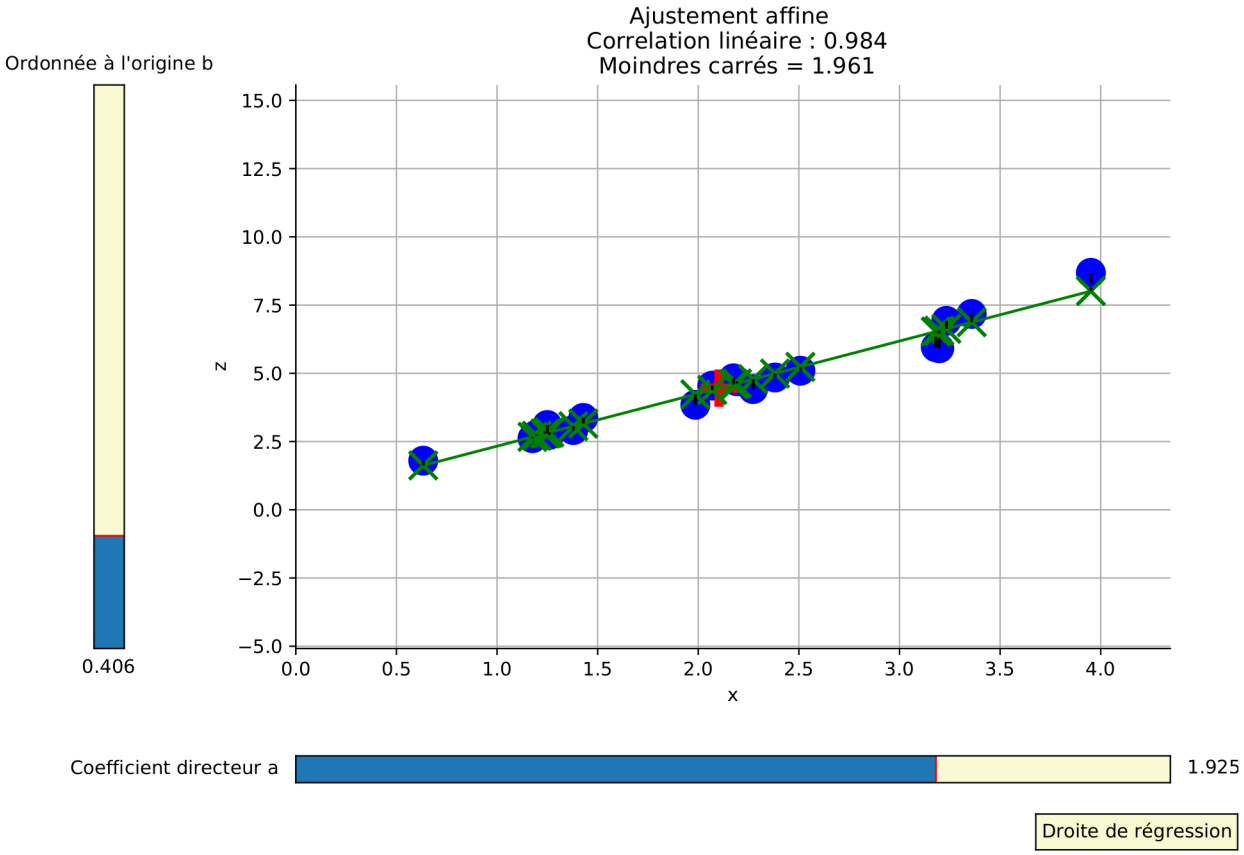
- pour un ajustement parabolique : $\boxed{\sqrt{y-c} = ax+b \Leftrightarrow y = (ax+b)^2 + c}$;
- pour un ajustement exponentiel : $\boxed{\ln(y) = ax+b \Leftrightarrow y = \mathrm{e}^{ax+b}}$;
- pour un ajustement logarithmique : $\boxed{\mathrm{e}^y = ax+b \Leftrightarrow y = \ln(ax+b)}$.

Exemple 3

On considère le nuage de points d'une série statistique à deux variables x et y, ci-dessous pour lequel un ajustement parabolique semble mieux adapté qu'un ajustement affine :



En effectuant le changement de variable $z = \sqrt{y-1}$, le nuage de points de la série statistique à deux variables x et z présente un coefficient de corrélation linéaire proche de 1 permettant un ajustement affine par la droite de régression d'équation $z = 1,925x + 0,406$.



On en déduit un ajustement de y par une fonction trinôme proche de celle conjecturée sur le premier graphique :

$$\begin{cases} z = 1,925x + 0,406 \\ z = \sqrt{y-1} \end{cases} \iff \begin{cases} y = (1,925x + 0,406)^2 + 1 \\ z^2 = y - 1 \end{cases}$$

Thème 1 *Corrélation et causalité*

Le tableau suivant donne la valeur de revente d'une machine outil au bout de t années d'utilisation (les prix sont donnés en centaines d'euros). On veut faire une estimation de son prix de revente au-delà de 6 ans.

Temps écoulé depuis l'achat t_i $0 \leqslant i \leqslant 6$	0	1	2	3	4	5	6
Valeur de revente y_i en centaines d'euros $0 \leqslant i \leqslant 6$	90	73,8	60	49,5	40,5	33	27

1. Déterminer le pourcentage de baisse du prix de revente de la machine au bout de six ans d'utilisation (de t_0 à t_6).

.....

.....

2. Étude d'un modèle affine

a. Représenter graphiquement le nuage de points $M_i\left(t_i\,;\,y_i\right)$ pour $0 \leqslant i \leqslant 6$ dans un repère orthogonal, en prenant comme unités graphiques : 1 cm pour une unité sur l'axe des abscisses ; 1 cm pour 10 unités sur l'axe des ordonnées.

b. Déterminer, à l'aide de la calculatrice, une équation de la droite de régression de y en t par la méthode des moindres carrés (les coefficients seront arrondis au centième). Tracer cette droite dans le repère précédent.

.....

.....

c. On sait qu'au bout de 10 ans la valeur de revente est de 1 000 euros. Le modèle vous semble-t-il adapté pour des calculs à plus long terme ?

.....

.....

3. Étude d'un modèle exponentiel

a. Pour $0 \leqslant i \leqslant 6$, on pose $z_i = \ln\left(y_i\right)$. Compléter le tableau suivant (en arrondissant les nombres au dixième) :

Temps écoulé depuis l'achat t_i $0 \leqslant i \leqslant 6$	0	1	2	3	4	5	6
$z_i = \ln\left(y_i\right)$							

b. Déterminer, à l'aide de la calculatrice, une équation de la droite d'ajustement de z en t par la méthode des moindres carrés (les coefficients seront arrondis au dixième).

.....

.....

c. En déduire que $y = \mathrm{e}^{-0,2t+4,5}$ est un ajustement exponentiel possible.

.....

.....

.....

d. Déterminer à l'aide de ce modèle une estimation de la valeur de revente au bout de 10 ans d'utilisation. Ce modèle vous semble-t-il mieux adapté que celui de l'ajustement affine ? Justifier la réponse.

.....

.....

Table des matières

31 Nathan est un papa comblé : son fils est en excellente santé. Dans son carnet de santé, il a noté son poids lors de chacun de ses anniversaires et il a obtenu le tableau ci-dessous.

Age x_i (en année)	7	8	9	10	11	12
Poids y_i (en kg)	22	24	28	34	42	51

Il souhaite avoir une idée de l'évolution du poids de son fils.

1. a) À l'aide de la calculatrice, représenter graphiquement le nuage de points $M_i(x_i ; y_i)$ avec $1 \leqslant i \leqslant 6$, associé à cette série.

b) Déterminer le coefficient de corrélation linéaire r entre x et y.

2. a) On pose $z_i = \sqrt{y_i}$.

Recopier et compléter le tableau suivant :

x_i	7	8	9	10	11	12
z_i						

b) À l'aide de la calculatrice, déterminer le coefficient de corrélation linéaire r' entre x et z.

Comparer r et r'.

c) Donner l'équation de la droite de régression de z en x.

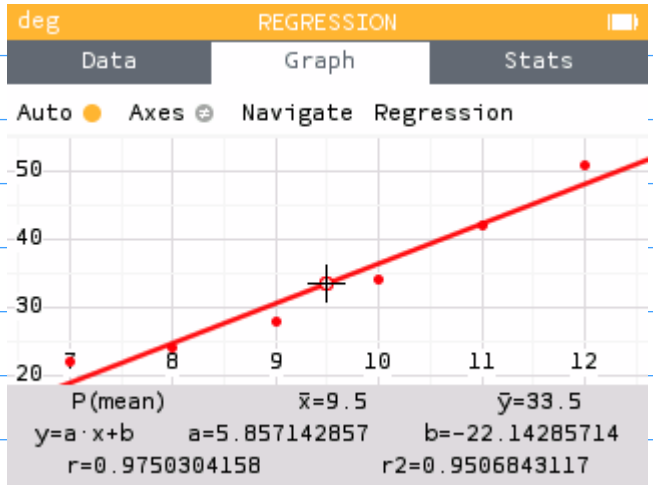
Arrondir au centième.

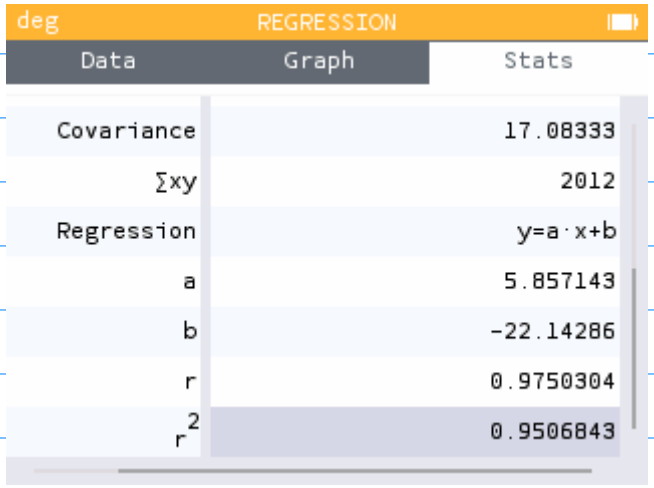
d) Estimer alors une relation entre y et x.

e) Utiliser cette relation pour estimer le poids que pourrait avoir son fils à l'âge de 15 ans.

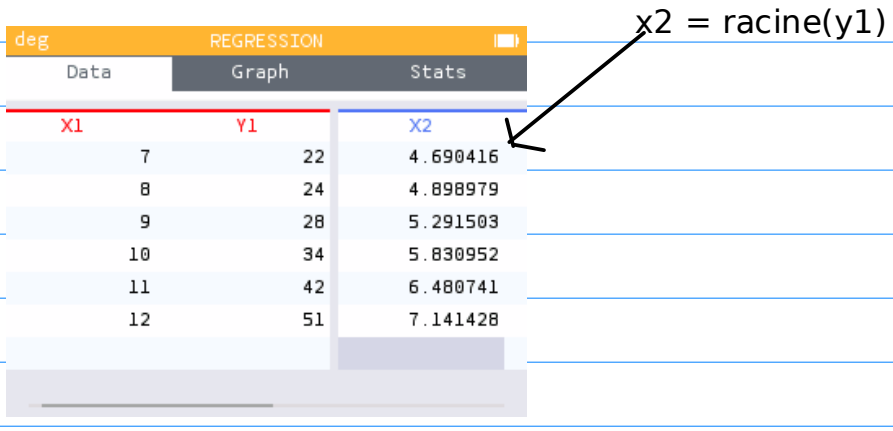
Que faut-il penser de ce résultat ?

Question 1)

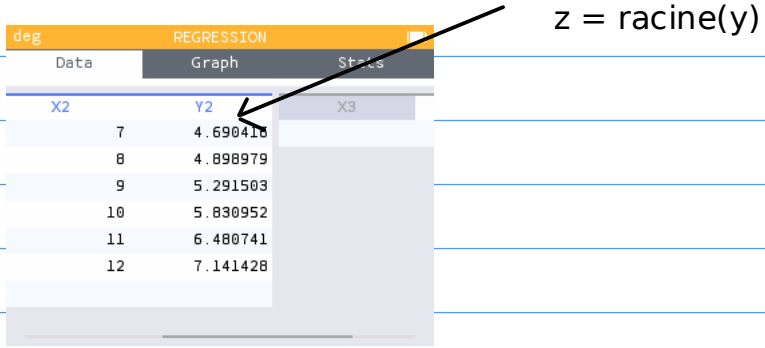


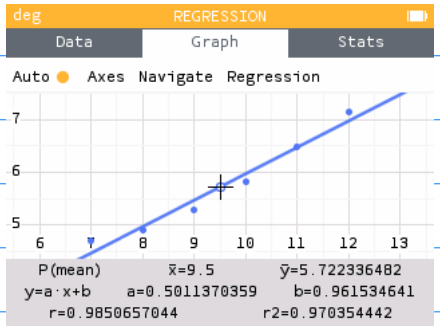


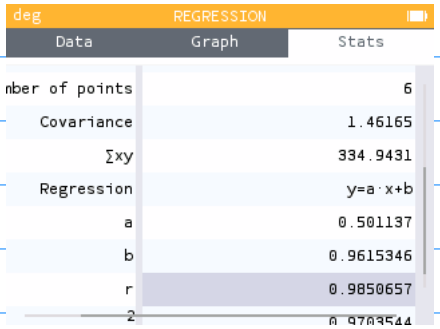
Question 2 :



Question 3)







Le coefficient de corrélation linéaire de z = racine(y) par rapport à x
est plus élevé que le coefficient de corrélation linéaire de y par rapport à x

La droite d'ajustement de z par rapport à x a pour équation :
z = 0,501137 x + 0,9615346
Or z = racine(y)
On en déduit que y = (0,501137 x + 0,9615346)^2

Pour x = 15 on en déduirait une estimation du poids de :
(0,501137 * 15 + 0,9615346)^2 soit environ 71,9 kg à 0,1 près

Que faut-il en penser ? Rien, cela dépend de la taille du fils, l'IMC se calcule avec la formule Masse / (Taille^2)

24 Ce tableau donne, pour les années 2005 à 2018 (sauf 2008), l'âge moyen a_i des femmes ayant accouché en France métropolitaine.

Année	Rang x_i	a_i
2005	1	29,9
2006	2	30
2007	3	30
2009	5	30,1
2010	6	30,2
2011	7	30,2

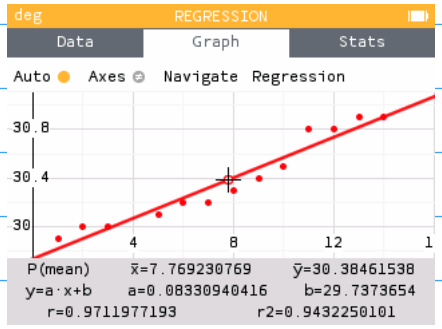
Année	Rang x_i	a_i
2012	8	30,3
2013	9	30,4
2014	10	30,5
2016	12	30,8
2017	13	30,9
2018	14	30,9

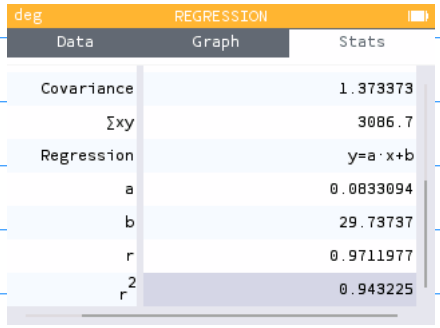
Source : Insee État Civil

a) À l'aide de la calculatrice, déterminer l'équation de la droite de régression de a en x. *Arrondir au centième.*

b) Utiliser cette droite pour estimer :

- en quelle année l'âge moyen des femmes ayant accouché en France métropolitaine pourrait dépasser 40 ans ;
- l'âge moyen des femmes ayant accouché en France en 2008.





Droite d'ajustement de y en x d'équation :
y = 0,0833094 x + 29,73737
donc pour x = 4 on obtient une estimation
de l'âge moyen des femmes ayant accouché
en 2008 par interpolation :
y = 0,0833094 * 4 + 29,73737
On obtient 30,07 ans à 0,01 près

En supposant que cet ajustement reste
vrai dans les années à venir, on peut déterminer
l'année à partir de laquelle l'âge moyen
des femmes ayant accouché dépassera 40 ans
en résolvant l'inéquation
0,0833094 x + 29,73737 > 40
<=> x > (40 - 29,73737) / 0,0833094
<= > x > 123,19

Si cet ajustement reste vrai, l'âge moyen
des femmes ayant accouché dépassera 40 ans
en 2004 + 124 = 2128

Thème 1 *Corrélation et causalité*

Le tableau suivant donne la valeur de revente d'une machine outil au bout de t années d'utilisation (les prix sont donnés en centaines d'euros). On veut faire une estimation de son prix de revente au-delà de 6 ans.

Temps écoulé depuis l'achat t_i $0 \leqslant i \leqslant 6$	0	1	2	3	4	5	6
Valeur de revente y_i en centaines d'euros $0 \leqslant i \leqslant 6$	90	73,8	60	49,5	40,5	33	27

1. Déterminer le pourcentage de baisse du prix de revente de la machine au bout de six ans d'utilisation (de t_0 à t_6).

$\frac{27-90}{90} = \frac{-63}{90} = \frac{-7}{10} = -0,7 = -70\%$

2. Étude d'un modèle affine

a. Représenter graphiquement le nuage de points $M_i(t_i\,;\,y_i)$ pour $0 \leqslant i \leqslant 6$ dans un repère orthogonal, en prenant comme unités graphiques : 1 cm pour une unité sur l'axe des abscisses ; 1 cm pour 10 unités sur l'axe des ordonnées.

b. Déterminer, à l'aide de la calculatrice, une équation de la droite de régression de y en t par la méthode des moindres carrés (les coefficients seront arrondis au centième). Tracer cette droite dans le repère précédent.

$y = ax + b$

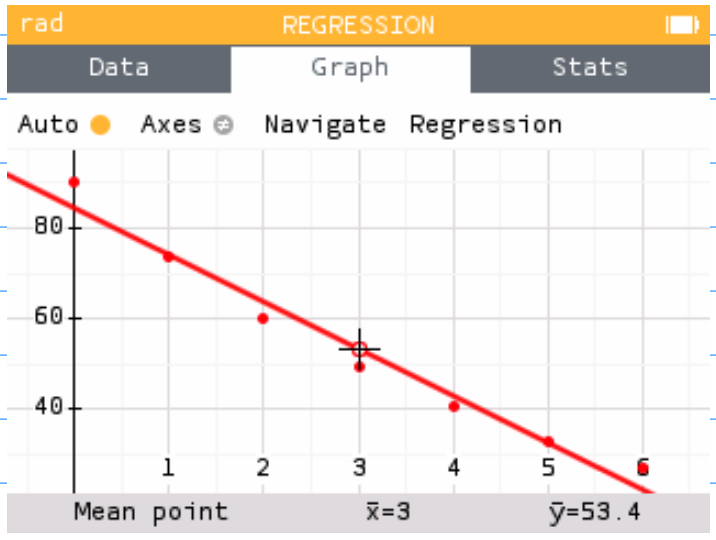
avec $a \approx -10,36071$ $b \approx 84,48214$

c. On sait qu'au bout de 10 ans la valeur de revente est de 1 000 euros. Le modèle vous semble-t-il adapté pour des calculs à plus long terme ?

La valeur prédite par cet ajustement au bout de 10 ans est :

$\approx -10,36071 \times 10 + 84,48214$

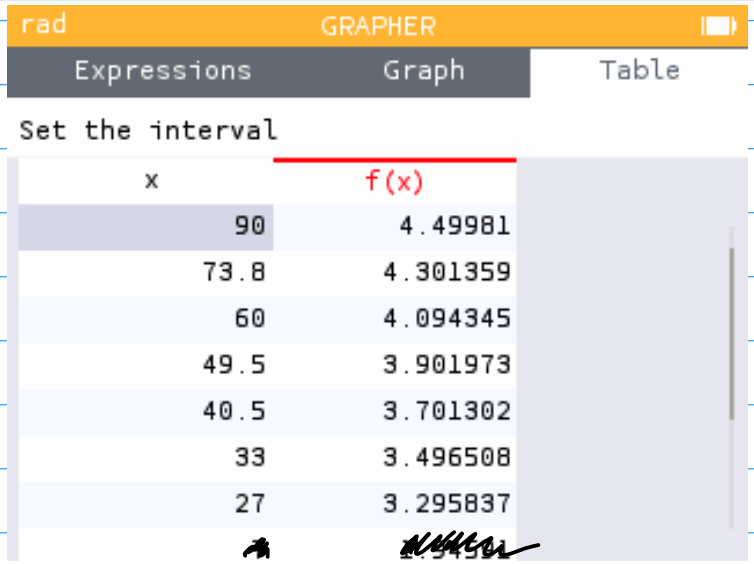
$\approx -19,12$ une valeur négative qui n'est pas réaliste.



3. Étude d'un modèle exponentiel

a. Pour $0 \leqslant i \leqslant 6$, on pose $z_i = \ln(y_i)$. Compléter le tableau suivant (en arrondissant les nombres au dixième) :

Temps écoulé depuis l'achat t_i $0 \leqslant i \leqslant 6$	0	1	2	3	4	5	6
$z_i = \ln(y_i)$							



b. Déterminer, à l'aide de la calculatrice, une équation de la droite d'ajustement de z en t par la méthode des moindres carrés (les coefficients seront arrondis au dixième).

$z \approx -0,2t + 4,5$

c. En déduire que $y = e^{-0,2t+4,5}$ est un ajustement exponentiel possible.

$\ln(y) = -0,2x + 4,5$ équivaut à $y = e^{-0,2x+4,5}$

d. Déterminer à l'aide de ce modèle une estimation de la valeur de revente au bout de 10 ans d'utilisation. Ce modèle vous semble-t-il mieux adapté que celui de l'ajustement affine? Justifier la réponse.

D'après l'ajustement exponentiel $y = e^{-0,2x+4,5}$ on peut prévoir au bout de 10 ans une valeur de revente de $y = e^{-0,2\times 10+4,5} = e^{2,5} \approx 12,18$ €

Cette valeur est plus réaliste que celle fournie par l'ajustement affine.