

 **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  $\chi^2$  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  $(x_i)_{1 \leq i \leq n}$  cette **série statistique à une variable**.

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

$$\bar{x} = \frac{1}{n} (x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i$$

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

$$V(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_p - \bar{x})^2}{n}$$

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

$$V(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

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

*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-

- 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

Valeurs V1	Effectifs N1	Valeurs V2
5	2	
8	2	
9	1	
11	1	
14	4	
15	2	

Affichage des indicateurs

	V1/N1
Effectif total $\Sigma n$	12
Minimum Min	5
Maximum Max	15
Etendue E	10
Moyenne $\bar{x}$	11
Ecart type $\sigma$	3.674235
Variance var	13.5
Premier quartile Q1	8

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

.....

.....

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

.....

.....

## Algorithmique 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 = .....

return .....

#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 = .....
    return somme / effectif

def ecart_type(liste_notes):
    return .....

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  $y_i$  les 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  $((x_i; y_i))_{1 \leq i \leq n}$ ;
- ou un tableau.

$x_i$	$x_1$	$x_2$	...	$x_n$
$y_i$	$y_1$	$y_2$	...	$y_n$

 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 :

$x_i$	$x_1$	$x_2$	...	$x_n$
$y_i$	$y_1$	$y_2$	...	$y_n$

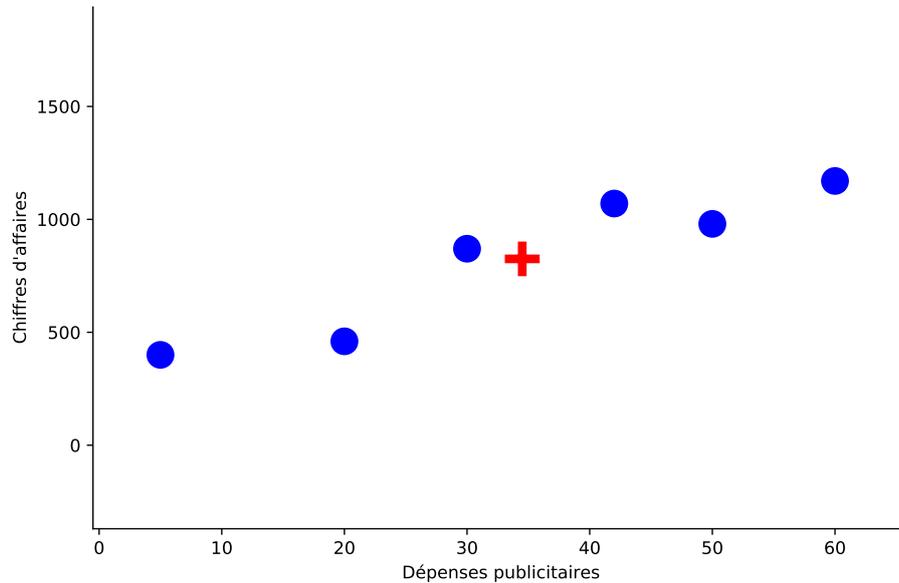
Dans un repère orthogonal du plan, on choisit de représenter le caractère  $x$  en abscisse et le caractère  $y$  en ordonnée.

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

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

## 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

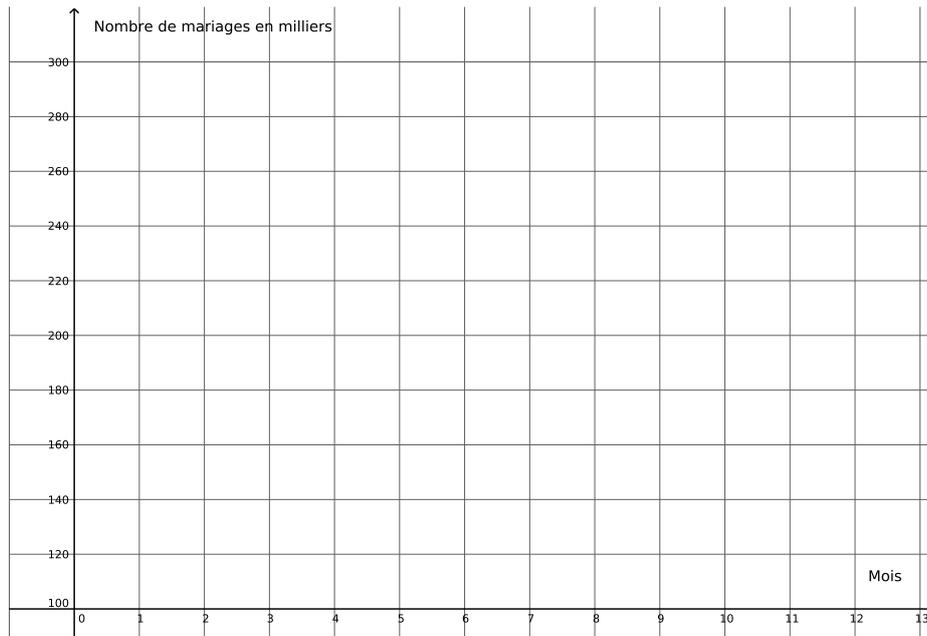
(d'après INSEE *Mariages en 2017*)

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

.....

.....

.....



## 1.4 Ajustement affine, interpolation, extrapolation



### Définition 4

Étant donné une **série statistique à deux variables quantitatives**  $x$  et  $y$ , on va se demander si en fonction d'une nouvelle valeur du caractère  $x$  (par exemple une *dépense publicitaire*) on peut estimer la valeur du caractère  $y$  (par exemple le *chiffre d'affaires*). On recherche une **corrélacion** entre les caractères  $x$  et  $y$ .

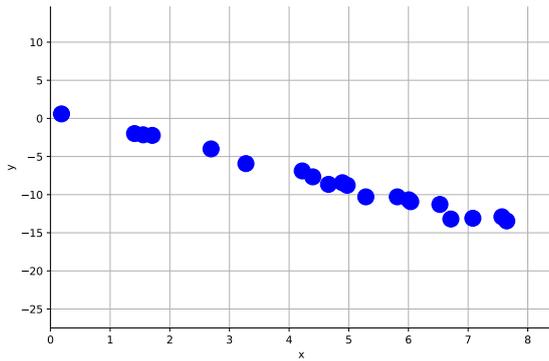
- ☞ l'estimation d'une valeur de  $y$  pour une valeur de  $x$  comprise entre deux valeurs  $x_p$  et  $x_q$  de la série s'appelle une **interpolation**;
- ☞ sinon l'estimation d'une valeur de  $y$  pour une valeur de  $x$  en dehors de la plage de valeurs des  $x_i$  de la série s'appelle une **extrapolation**.

Une fonction permettant de réaliser de telles estimations s'appelle un **ajustement**.

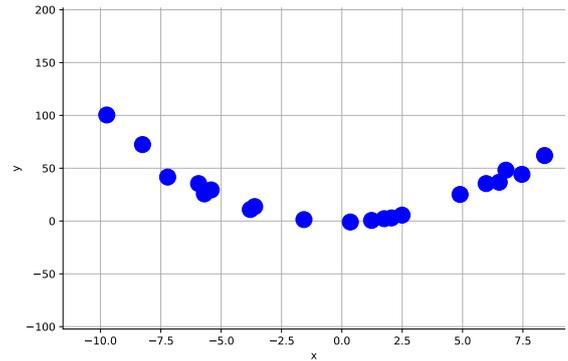
Selon la forme du nuage de points de la série double, on peut rechercher :

- ☞ 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 ...

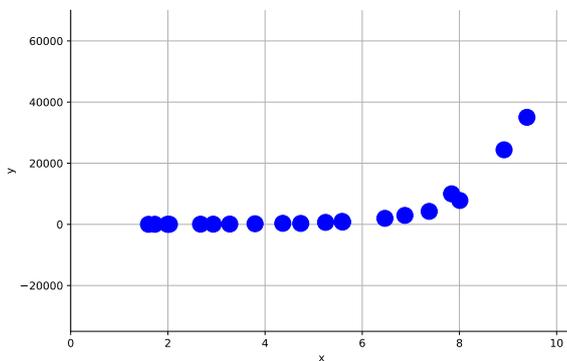
Ajustement affine



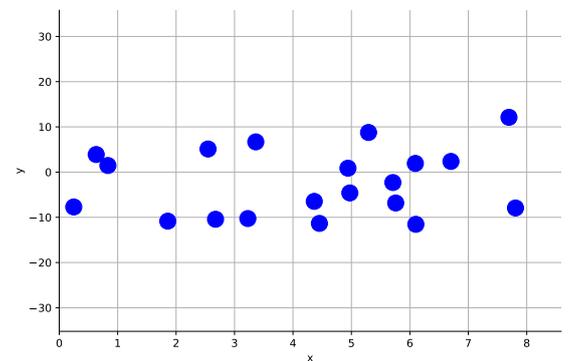
Ajustement parabolique



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  $((x_i; y_i))_{1 \leq i \leq n}$  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  $(M_i(x_i; y_i))_{1 \leq i \leq n}$  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.

- On réalise un ajustement affine du nuage de points de la série statistique double  $((x_i; y_i))_{1 \leq i \leq 12}$  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.

.....

.....

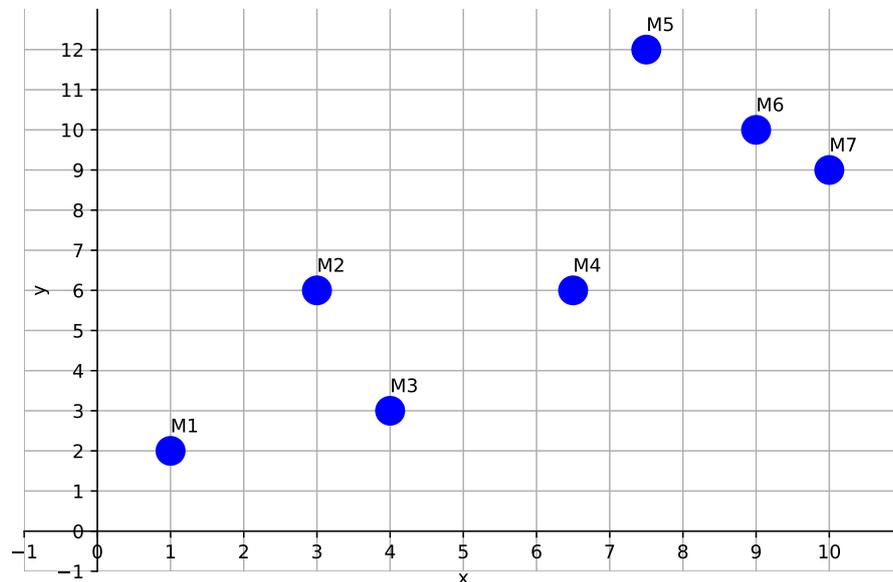
.....

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.

.....  
.....  
.....

**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.

.....  
.....

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

.....  
.....

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

.....  
.....

## 2 Méthode des moindres carrés

### 2.1 Principe de la méthode

#### Méthode

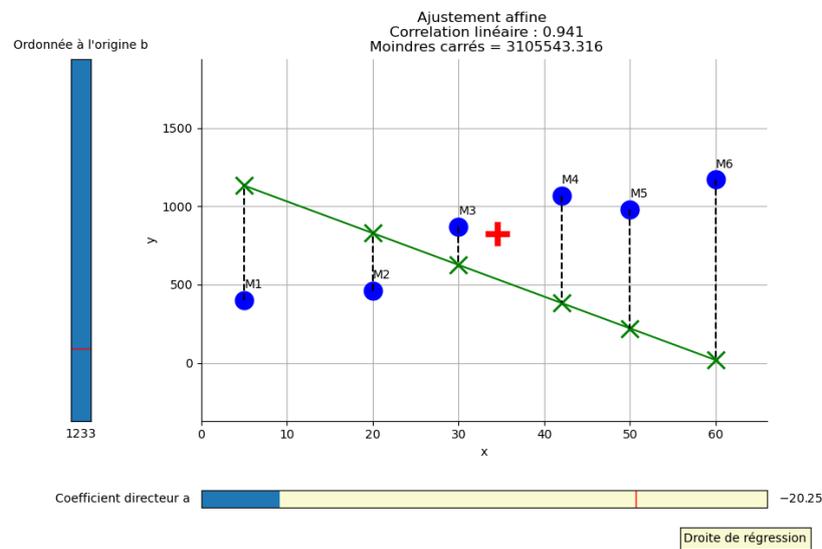
Dans un repère orthogonal du plan, étant donné le nuage de points  $(M_i(x_i; y_i))_{1 \leq i \leq n}$  d'une série statistique à deux variables et une droite  $\mathcal{D}$  d'équation  $y = ax + b$ , on associe à chaque point  $M_i$  le point  $A_i(x_i; ax_i + b)$  de  $\mathcal{D}$  qui a même abscisse.

On calcule ensuite la somme  $S$  des carrés des distances :

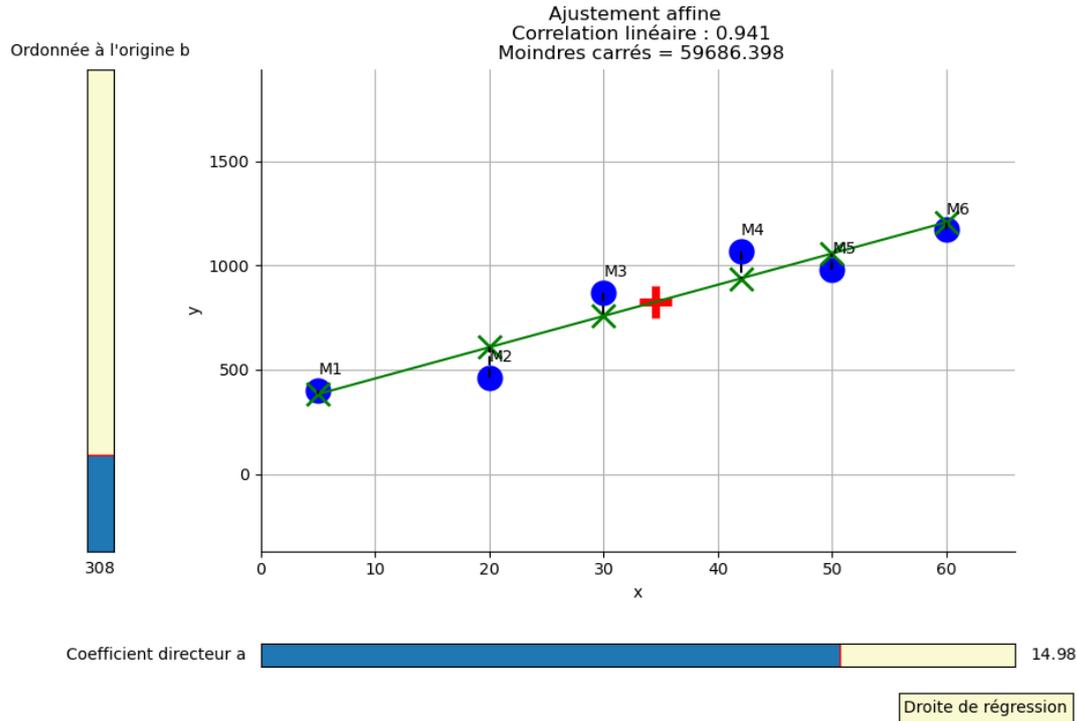
$$S = M_1 A_1^2 + M_2 A_2^2 + \dots + M_n A_n^2 = \sum_{i=1}^n M_i A_i^2 = \sum_{i=1}^n (y_i - (ax_i + b))^2$$

Cette somme  $S$  représente la distance entre la droite  $\mathcal{D}$  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  $\mathcal{D}$  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  $((x_i; y_i))_{1 \leq i \leq n}$  se note  $\sigma_{xy}$  ou  $\text{cov}(x, y)$ , elle est égale à :

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

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

### Corollaire

Soit  $(x_i)_{1 \leq i \leq n}$  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  $(M_i(x_i; y_i))_{1 \leq i \leq n}$  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  $G(\bar{x}; \bar{y})$  et a pour équation  $y = ax + b$  avec :

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

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  $(M_i(x_i; y_i))_{1 \leq i \leq n}$  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 :

$$S = ub^2 + vb + w \text{ avec } u = n, v = -2 \sum_{i=1}^n (y_i - ax_i) \text{ et } w = \sum_{i=1}^n (y_i - ax_i)^2.$$

$$\text{Pour } a \text{ fixé } S \text{ atteint un minimum pour : } b = -\frac{v}{2u} = \frac{\sum_{i=1}^n (y_i - ax_i)}{n} = \frac{1}{n} \left( \sum_{i=1}^n y_i \right) - a \frac{1}{n} \left( \sum_{i=1}^n x_i \right) = \bar{y} - a\bar{x}$$

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

#### Étape 2 : on remplace $b$

D'après l'étape 1  $S$  est minimale pour  $(a, b)$  si et seulement si  $S$  est minimale pour  $(a, b) = (a, \bar{y} - a\bar{x})$ .

On peut donc rechercher le minimum de  $S$  en remplaçant  $b$  par  $\bar{y} - a\bar{x}$  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  $a$ .

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

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

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

On reconnaît les expressions de  $V(x)$ ,  $\text{cov}(x, y)$  et  $V(y)$

$$S = a^2 nV(x) - 2a \text{ncov}(x, y) + nV(y)$$

On peut ainsi exprimer  $S$  comme une fonction polynôme du second degré en  $a$  de coefficients :

$$S(a) = ua^2 + av + w \text{ avec } u = nV(x), v = -2ncov(x, y) \text{ et } w = nV(y).$$

$$\text{Ainsi } S \text{ atteint un minimum pour : } a = -\frac{-2ncov(x, y)}{2nV(x)} = \frac{cov(x, y)}{V(x)}$$

$$\text{Finalement } S \text{ atteint un minimum si et seulement si } a = \frac{cov(x, y)}{V(x)} \text{ et } b = \bar{y} - a\bar{x}.$$

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>

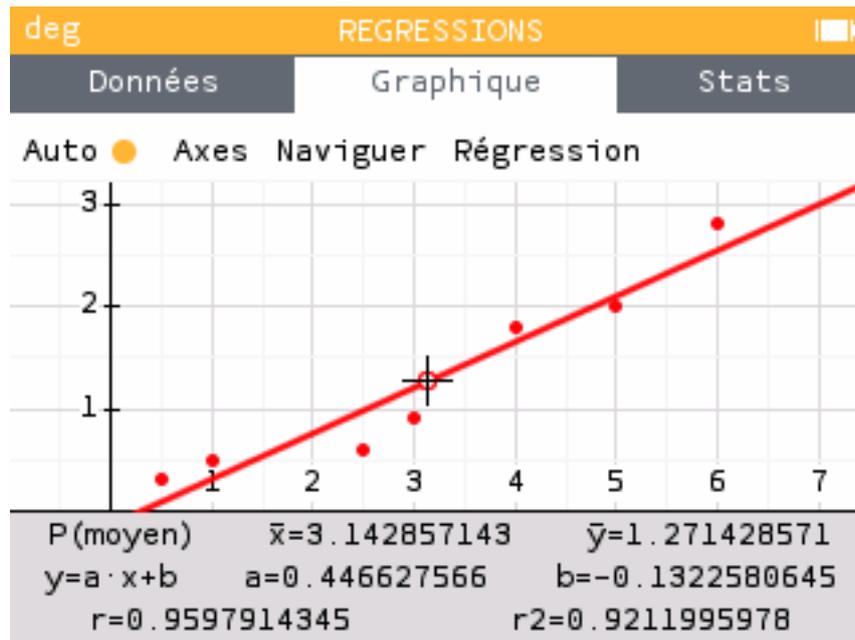
### Étape 1 : saisie

deg REGRESSIONS		
Données	Graphique	Stats
X1	Y1	X2
0.5	0.3	
1	0.5	
2.5	0.6	
3	0.9	
4	1.8	
5	2	
6	2.8	

### Étape 2 : calculs

deg REGRESSIONS		
Données	Graphique	Stats
Variance	3.479592	0.7534694
nombre de points		7
Covariance		1.554082
$\sum xy$		38.85
Régression		$y = a \cdot x + b$
a		0.4466276
b		-0.1322581
r		0.9597914

### 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.

$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$\times$
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  $\bar{x}$  et  $\bar{y}$  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  $((x_i; y_i))_{1 \leq i \leq 6}$ .

.....

.....

.....

.....

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

.....

.....

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

.....

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.

.....

.....

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

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

.....

.....

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

.....

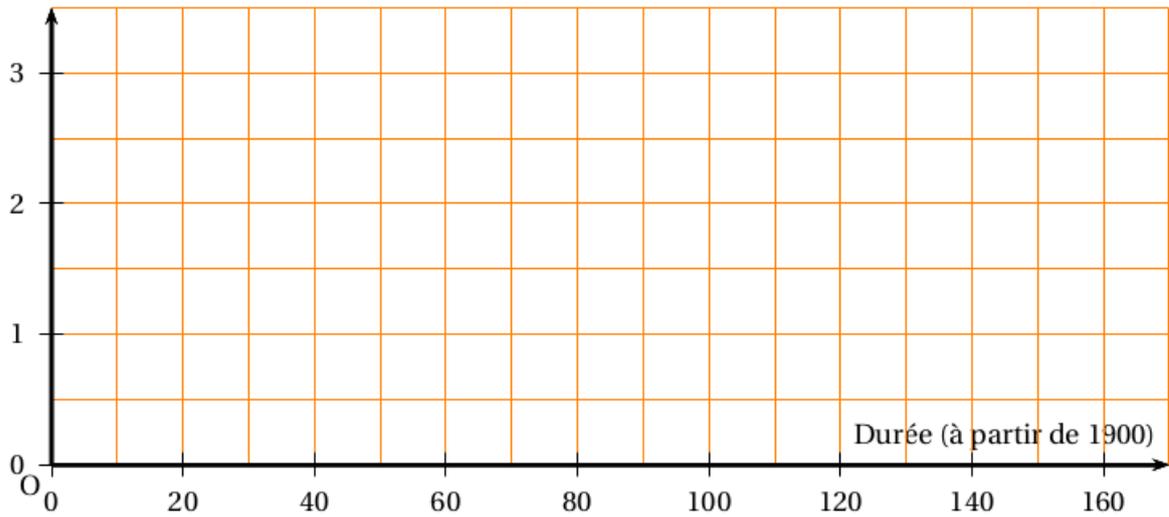
.....

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

.....

.....

Recul (mesuré en km)



## 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  $y$  :

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

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

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

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

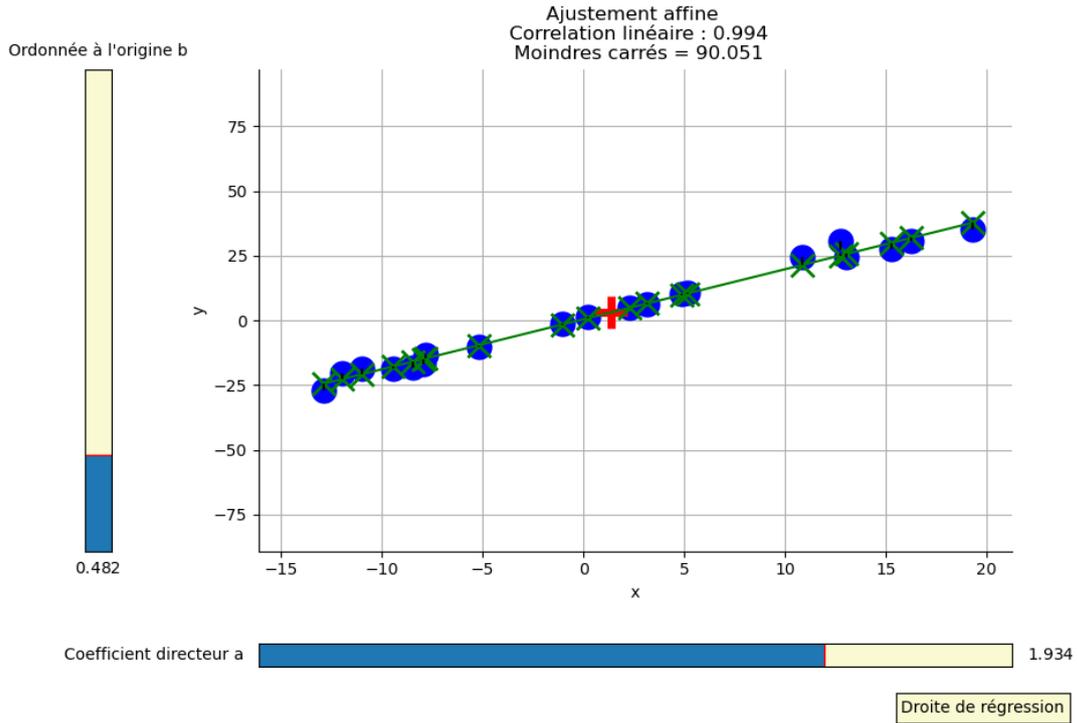
#### Propriété 2

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

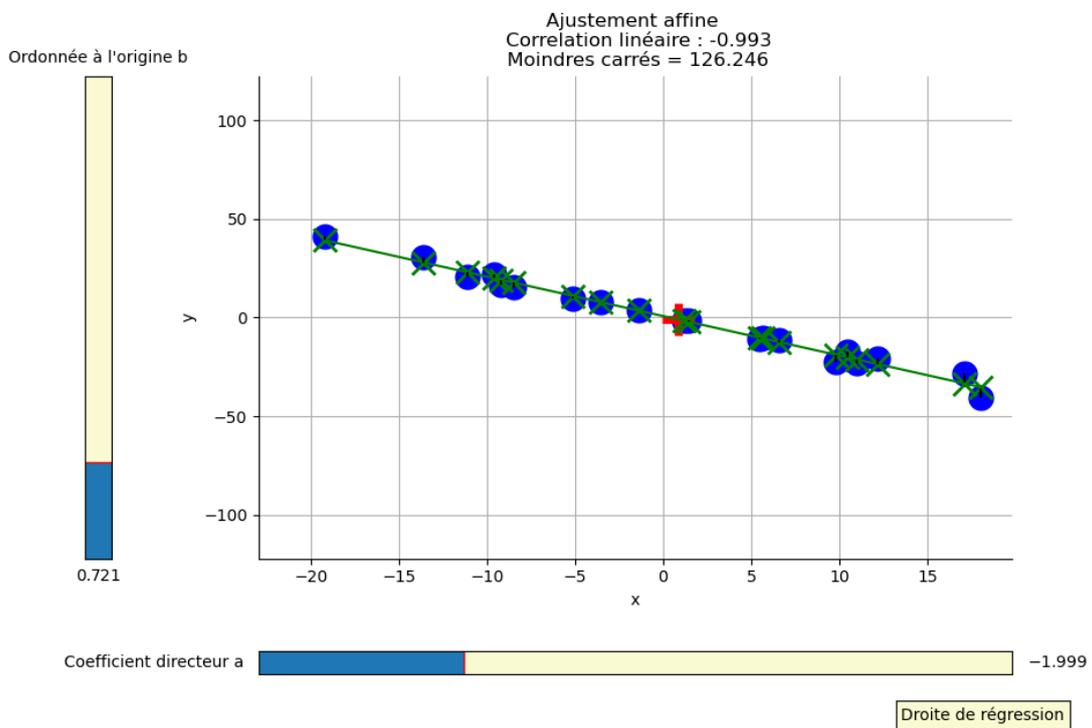
- ☞  $r$  est un réel compris entre  $-1$  et  $1$ ,  $-1 \leq r \leq 1$  ;
- ☞ la valeur de  $r$  permet de mesurer la qualité d'un ajustement affine obtenu par la méthode des moindres carrés :
  - plus la valeur absolue de  $r$  est proche de  $0$ , moins la relation affine entre les variables  $x$  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.

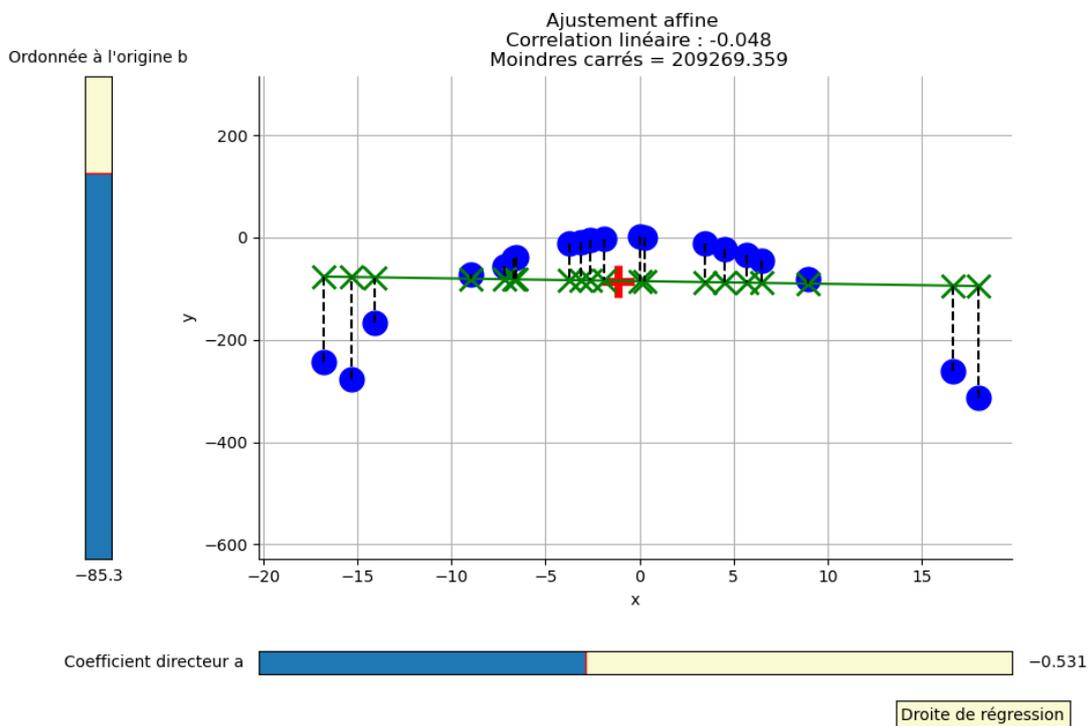
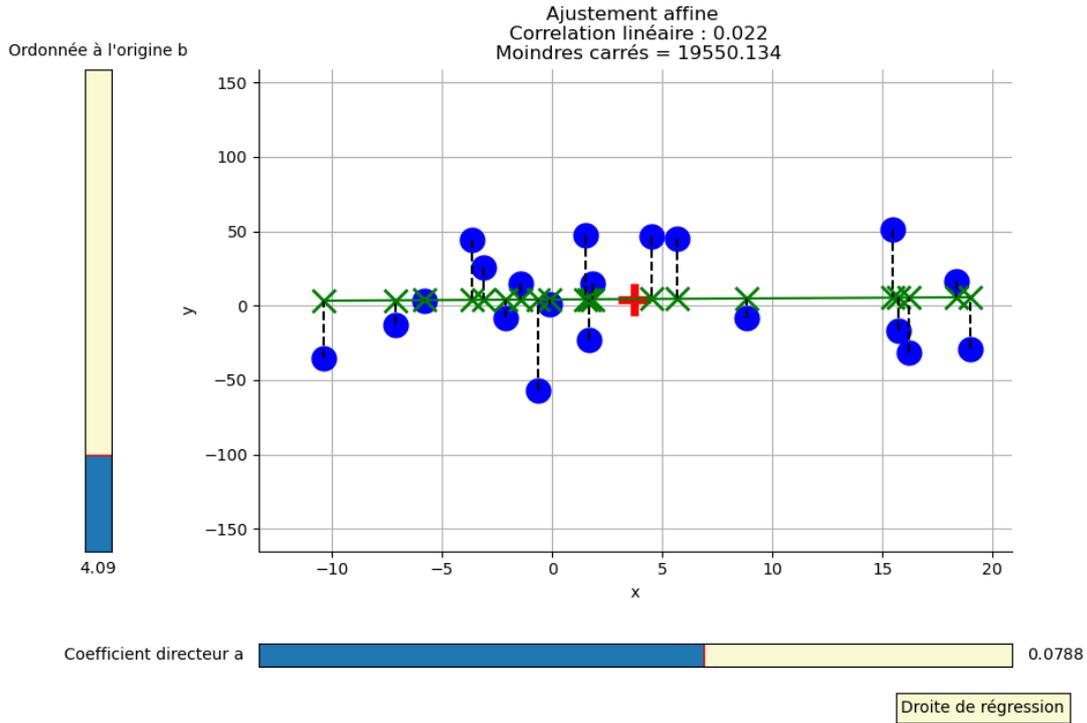
$r$  proche de  $1$



$r$  proche de  $-1$



$r > 0$  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.

.....

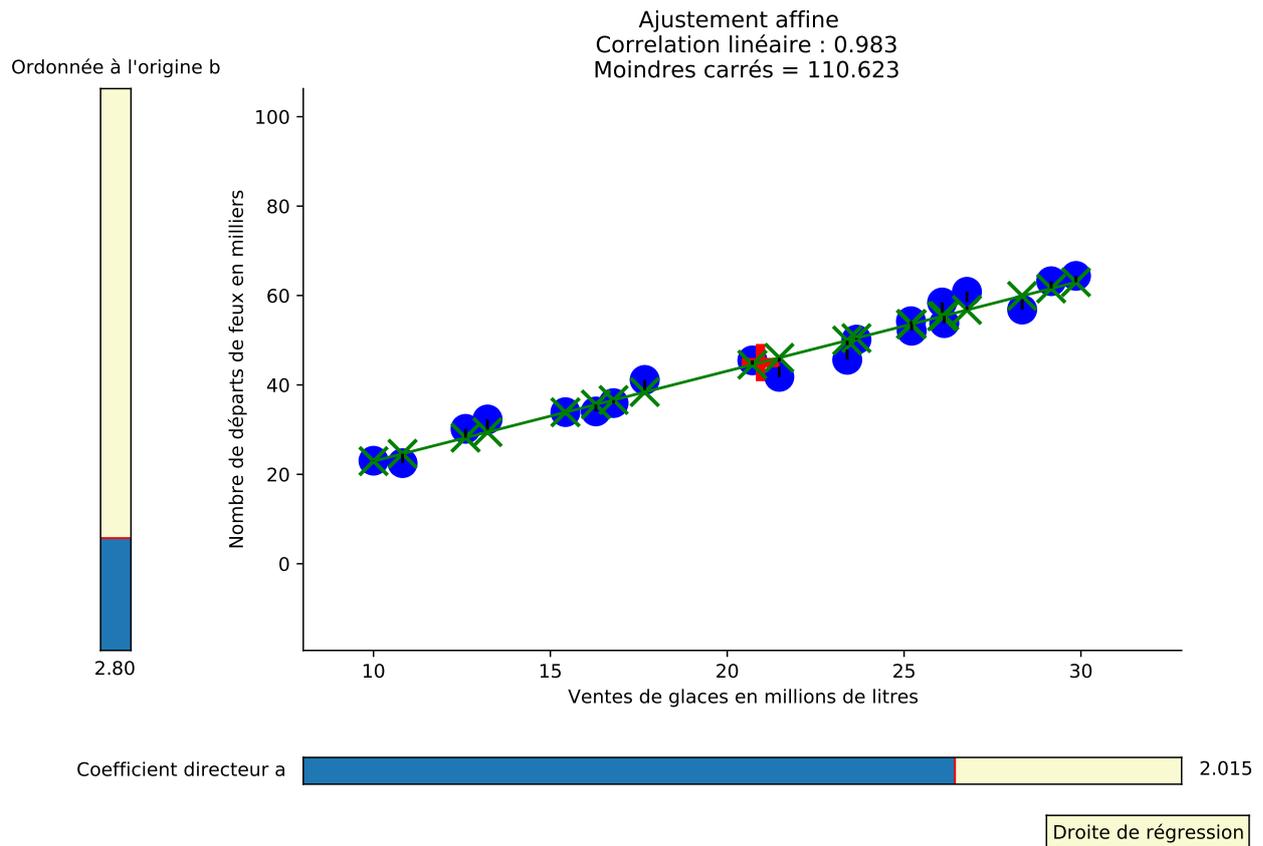
.....

.....

## 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?

.....

.....

D'autres exemples de corrélations sans causalité : <https://tylervigen.com/spurious-correlations>.

## 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  $y = e^{ax+b}$  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  $((x_i; f(y_i)))_{1 \leq i \leq n}$  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 ;
- $f(y) = e^y$  pour un ajustement logarithmique.

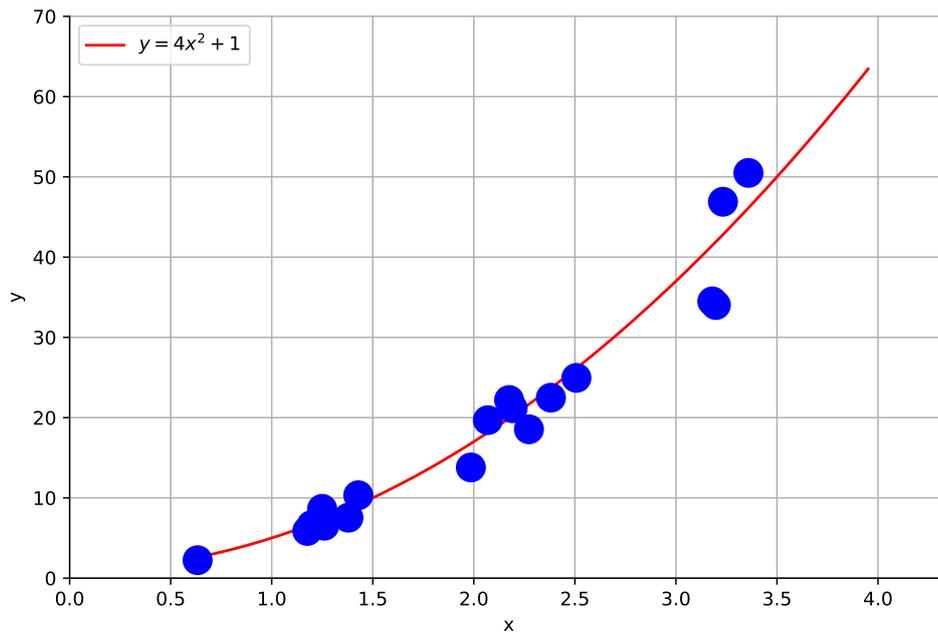
☞ **Étape 2** : On fait le changement de variable  $z = f(y)$  et pour le nuage de points  $((x_i; z_i = f(y_i)))_{1 \leq i \leq n}$ , on détermine la droite de régression de  $z = f(y)$  par rapport à  $x$  :  $z = ax + b$ .

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

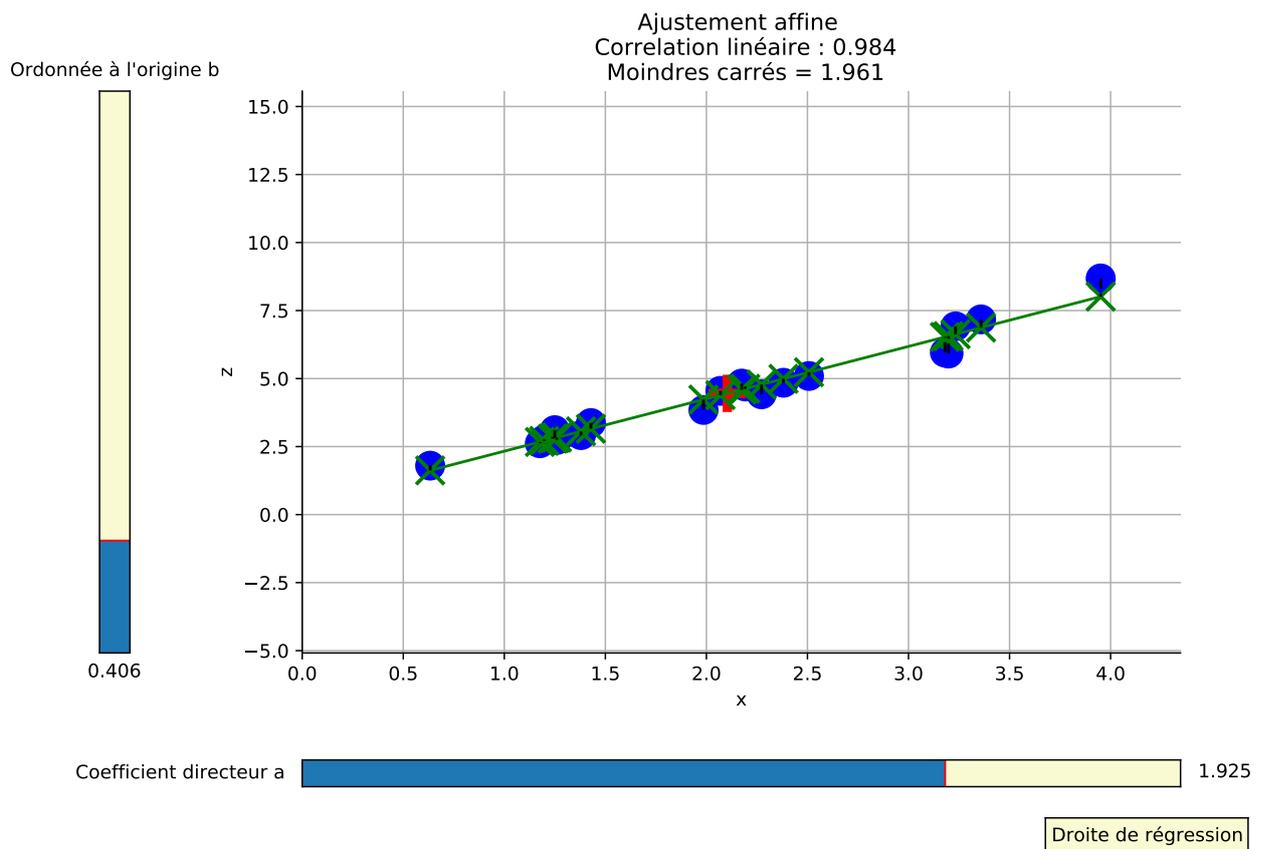
- pour un ajustement parabolique :  $\sqrt{y-c} = ax + b \Leftrightarrow y = (ax + b)^2 + c$  ;
- pour un ajustement exponentiel :  $\ln(y) = ax + b \Leftrightarrow y = e^{ax+b}$  ;
- pour un ajustement logarithmique :  $e^y = ax + b \Leftrightarrow y = \ln(ax + b)$ .

### 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 $t_i$ $0 \leq i \leq 6$	0	1	2	3	4	5	6
Valeur de revente $y_i$ en centaines d'euros $0 \leq i \leq 6$	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  $M_i(t_i ; y_i)$  pour  $0 \leq i \leq 6$  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  $0 \leq i \leq 6$ , on pose  $z_i = \ln(y_i)$ . Compléter le tableau suivant (en arrondissant les nombres au dixième) :

Temps écoulé depuis l'achat $t_i$ $0 \leq i \leq 6$	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).

.....  
.....

**c.** En déduire que  $y = e^{-0,2t+4,5}$  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

<b>1 Ajustement affine</b>	<b>1</b>
1.1 Rappels sur les statistiques à une variable . . . . .	1
1.2 Série statistique à deux variables . . . . .	4
1.3 Nuage de point et point moyen . . . . .	4
1.4 Ajustement affine, interpolation, extrapolation . . . . .	6
<b>2 Méthode des moindres carrés</b>	<b>9</b>
2.1 Principe de la méthode . . . . .	9
2.2 Droite d'ajustement par la méthode des moindres carrés . . . . .	10
<b>3 Corrélation linéaire et changement de variable</b>	<b>16</b>
3.1 Coefficient de corrélation linéaire . . . . .	16
3.2 Ajustement affine après changement de variable . . . . .	20