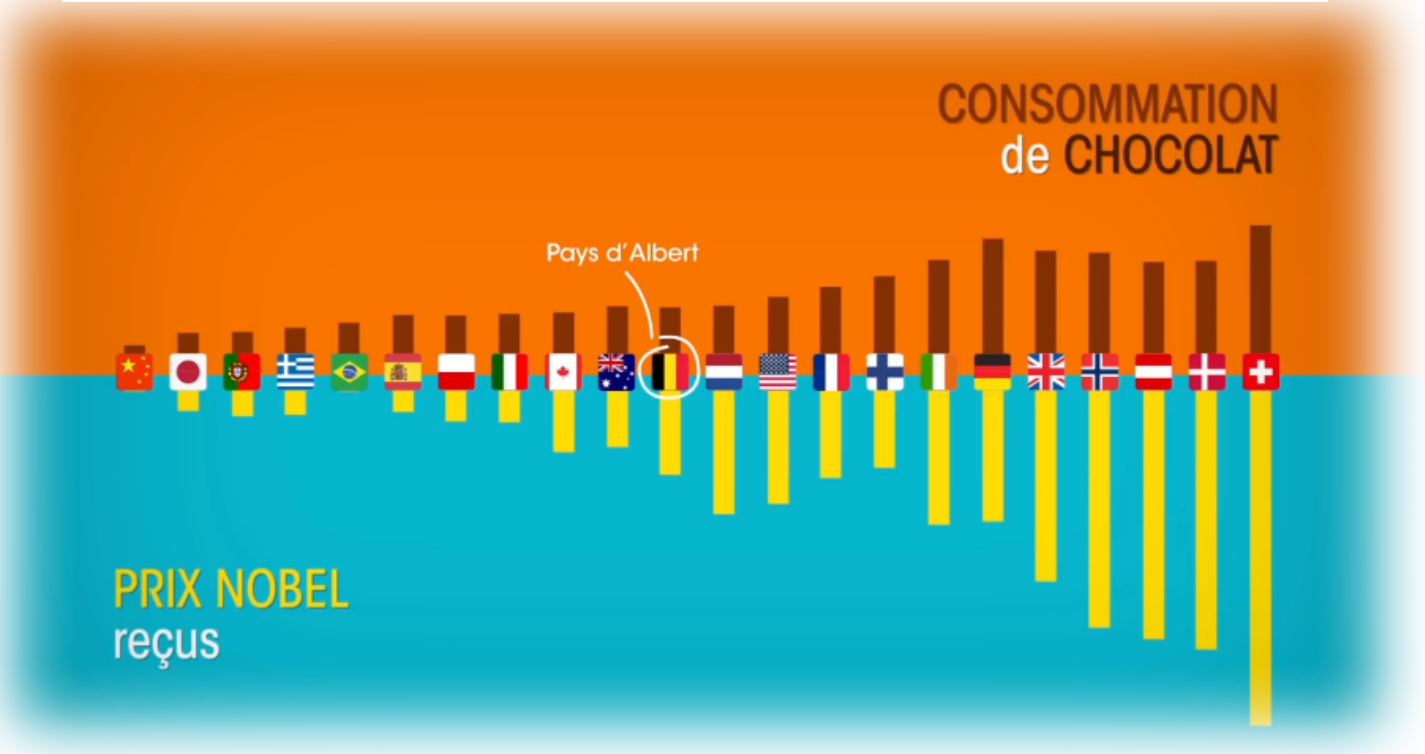
Cours 3

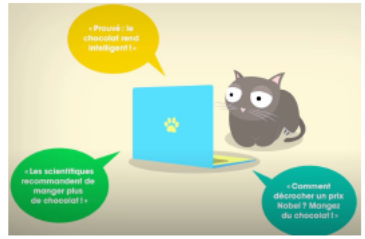
Corrélation entre 2 variables statistiques

Ressource R2.08 - Statistique Descriptive

2023-2024



Plan du cours



1. Notion de corrélation

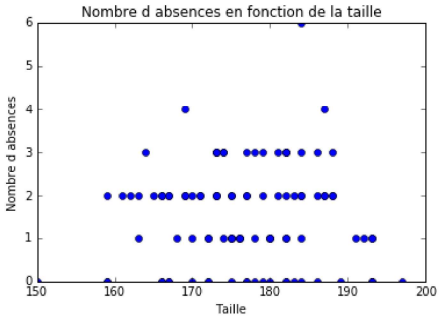
Selon Larousse :

Corrélation : Relation existant entre deux notions dont l'une ne peut être pensée sans l'autre, entre deux faits liés par une dépendance nécessaire.

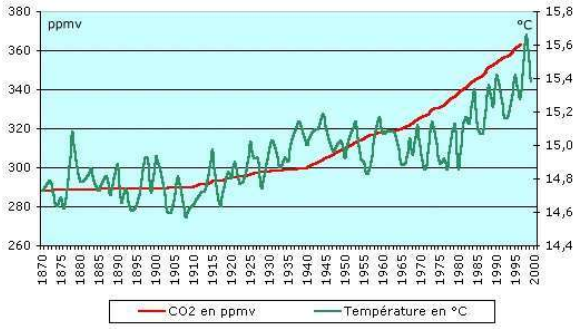
En probabilités et en statistiques, la corrélation entre plusieurs variables aléatoires ou statistiques est une notion de liaison qui contredit leur indépendance.

• Exemples

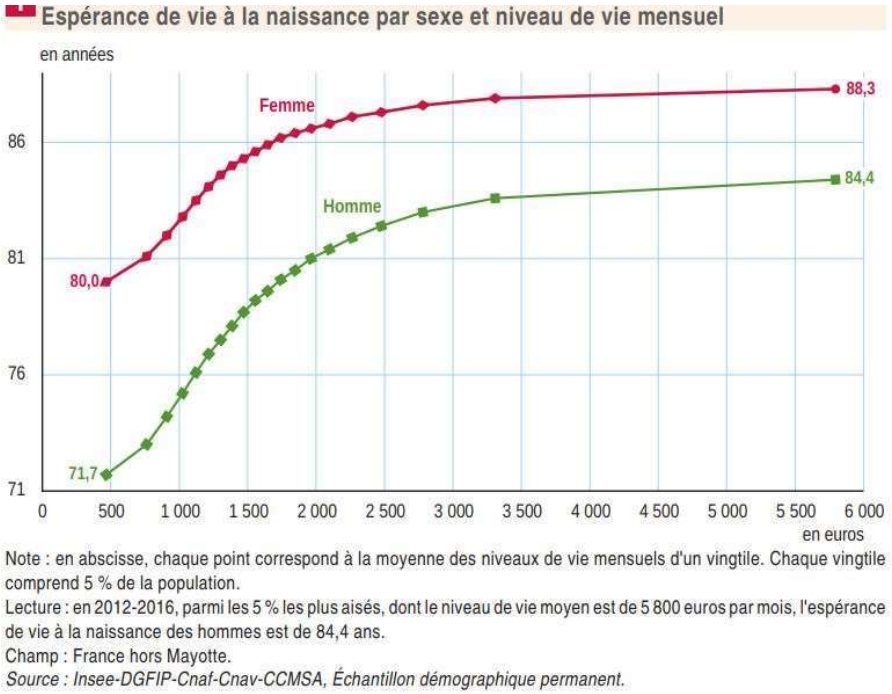
Quand on a 2 variables statistiques, on peut se demander si elles sont corrélées :



- Le nombre d'absences ne semble pas corrélé à la taille des étudiant.e.s

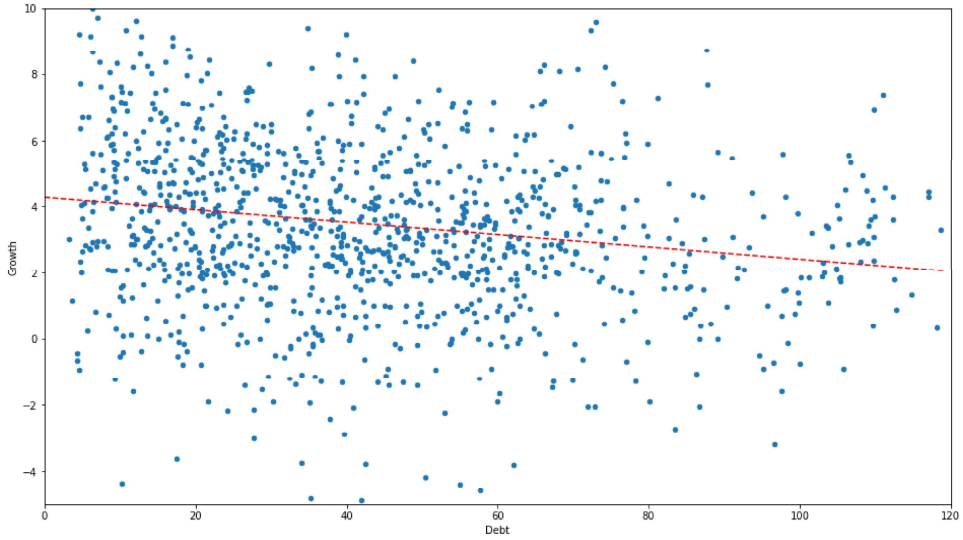


- Le réchauffement climatique semble être relié à l'effet de serre.



Note : en abscisse, chaque point correspond à la moyenne des niveaux de vie mensuels d'un vingtile. Chaque vingtile comprend 5 % de la population.
Lecture : en 2012-2016, parmi les 5 % les plus aisés, dont le niveau de vie moyen est de 5 800 euros par mois, l'espérance de vie à la naissance des hommes est de 84,4 ans.
Champ : France hors Mayotte.
Source : Insee-DGFIP-Cnaf-Cnav-CCMSA, Échantillon démographique permanent.

Croissance du PIB d'un pays en fonction de son niveau d'endettement

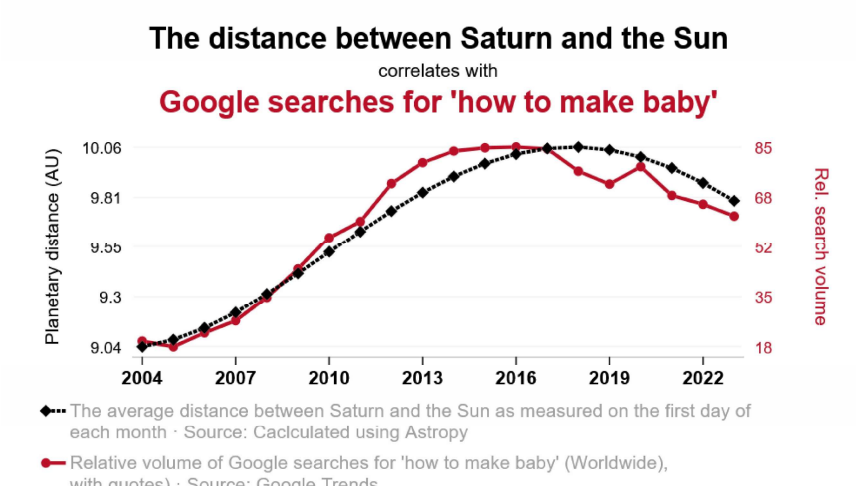


Dans ce graphique, chaque point représente un couple (pays, année): par exemple il y a un point correspondant à l'Australie en 1970, et un autre à l'Australie en 1980.

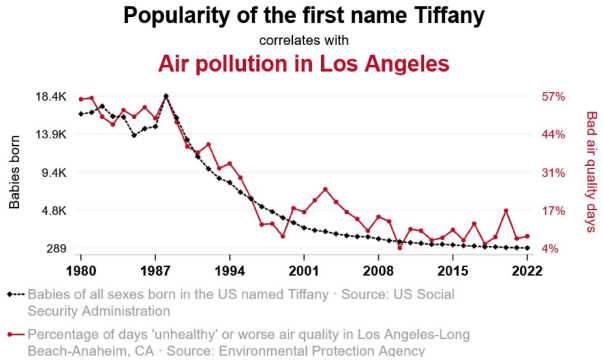
⟶ Exercice du TP3

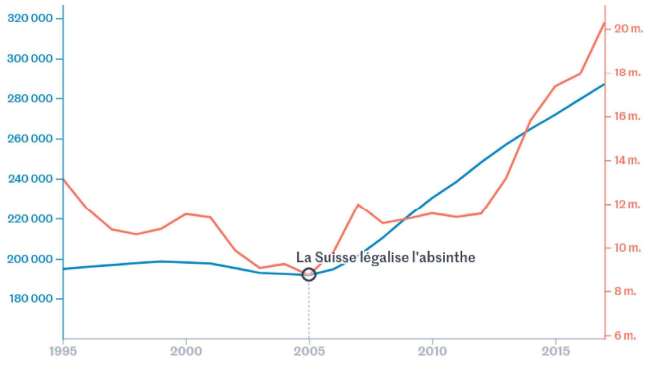
• Corrélation *vs* causalité : le hasard

Une corrélation peut arriver par hasard, surtout avec des données peu nombreuses.



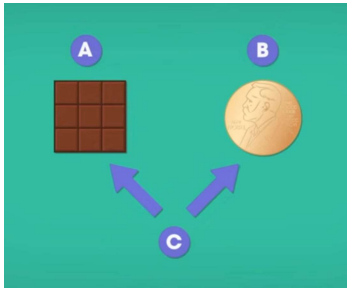
… le hasard peut être aidé par une IA cherchant aléatoirement des corrélations, et des internautes facétieux qui publient les plus ”pertinentes” …





• Corrélation *vs* causalité : l’effet cigogne

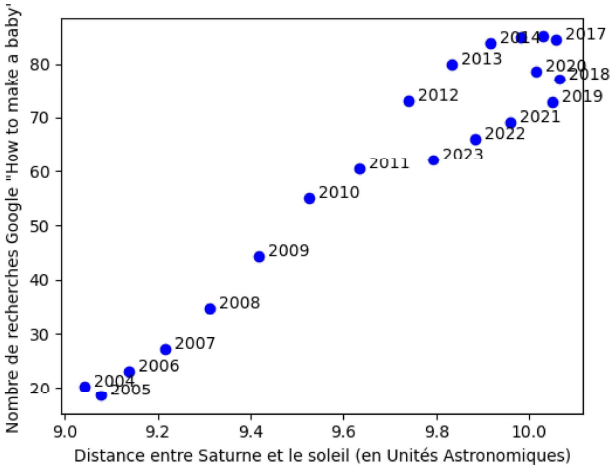
Clarification avec Albert le chat



3. Étude graphique

• Tracer un nuage de points

Pour observer si deux variables statistiques X et Y sont corrélées, on peut commencer par observer la forme du nuage de points les représentant :



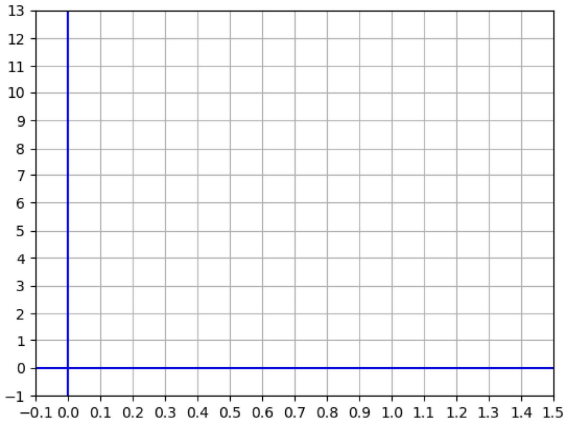
- on indique la variable X en abscisse (ici la distance Saturne-Soleil)
- on indique la variable Y en ordonnée (ici le nbre de recherches "How to make a baby")
- chaque point représente ces 2 chiffres pour une année donnée.

• Exemple 1 : âge d'un enfant en fonction de sa taille

On étudie l'âge d'un enfant en fonction de sa taille.
On a les données suivantes sur un échantillon de quatre enfants :

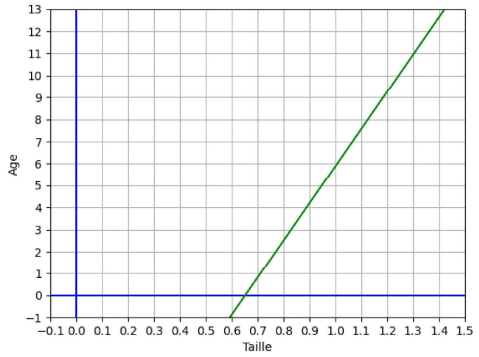
	Enfant A	Enfant B	Enfant C	Enfant D
Taille	0.78	0.91	1.05	1.26
Âge	2	4	8	10

Représenter le nuage de points de ces données sur le graphique ci-contre, un enfant étant représenté par un point :



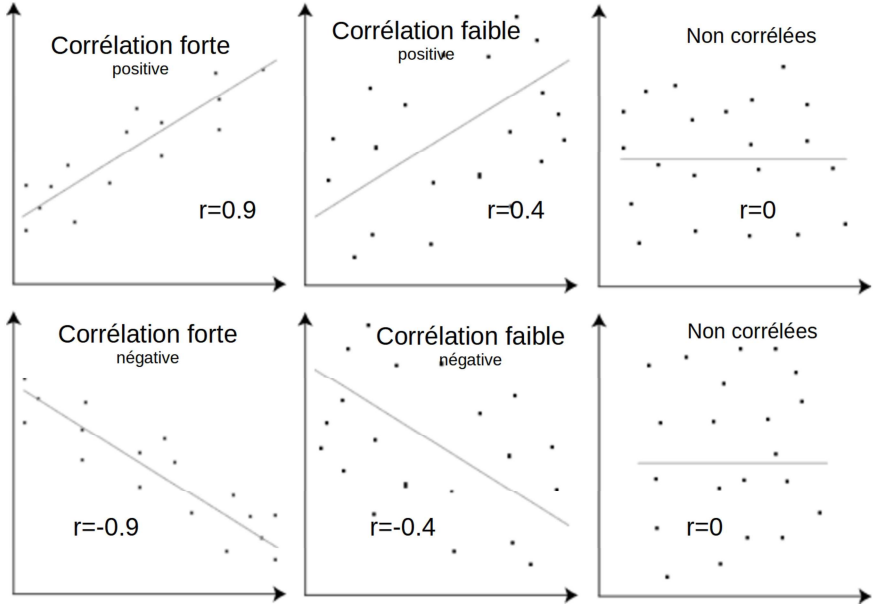
• Nuage de points et corrélation

Quand le nuage de points est proche d'une droite, on dit que les données X et Y sont corrélées linéairement.



Certaines données peuvent être corrélées autrement que *linéairement*, leur nuage de points prend des formes autres que des droites :

Dans ce cours, on va étudier uniquement les **corrélations linéaires**.
Pour cela, on va introduire le **coefficient de corrélation linéaire**, noté r ou $Cor(X, Y)$.



4. Covariance

On définit la *covariance* de 2 séries X et Y, qui va nous servir à définir le *coefficient de corrélation* de X et Y.

Définition

Soient X et Y des séries statistiques sur une même population P d'effectif N.
Alors la **covariance** de X et Y est la valeur

$$Cov(X,Y) = \frac{1}{N}\sum_{k\in P}(X(k)-\overline{X})\times(Y(k)-\overline{Y})$$

Proposition

Soient X et Y des séries statistiques sur une même population P d'effectif N.
Alors on peut calculer leur covariance avec la **2e formule de Koenig** :

$$Cov(X,Y) = \left(\frac{1}{N}\sum_{k\in P}X(k)\times Y(k)\right) - \overline{X}\,\overline{Y} = \overline{XY} - \overline{X}\,\overline{Y}$$

Remarque : en faisant la covariance d'une variable avec elle-même, on retrouve la variance : $Cov(X,X) = Var(X)$.

• **Exemple 1 : âge d'un enfant en fonction de sa taille**

	Enfant A	Enfant B	Enfant C	Enfant D
Taille (X)	0.78	0.91	1.05	1.26
Âge (Y)	2	4	8	10

Calculer d'abord les moyennes de X et de Y :

- $\overline{X} = \ldots\ldots$
- $\overline{Y} = \ldots\ldots$

Covariance de X et Y par la formule de la définition :

- $Cov(X,Y) = \ldots\ldots$

Covariance de X et Y par la formule de Koenig :

- $\overline{XY} = \ldots\ldots$
- $Cov(X,Y) = \ldots\ldots$

5. Coefficient de corrélation

Définition

Soient 2 séries statistiques X et Y définies sur une même population.
Le **coefficient de corrélation** de X et Y est défini par :

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

Remarque: $Cor(X,Y)$ est aussi souvent noté r.

Proposition

Pour 2 séries statistiques X et Y, on aura toujours $-1 \leq Cor(X,Y) \leq 1$.

Définition

Plus $|Cor(X,Y)|$ est proche de 1, plus les variables sont dites **corrélées**.
En particulier :

- si $Cor(X,Y) = 0$, les variables sont **indépendantes**.
- si $|Cor(X,Y)| \leq \frac{\sqrt{2}}{3} \approx 0.866$, on dit qu'il y a une **corrélation faible**.
- si $|Cor(X,Y)| > \frac{\sqrt{2}}{3} \approx 0.866$, on dit qu'il y a une **corrélation forte**.

• **Exemple 1 : âge d'un enfant en fonction de sa taille**

	Enfant A	Enfant B	Enfant C	Enfant D
Taille (X)	0.78	0.91	1.05	1.26
Âge (Y)	2	4	8	10

Voici des valeurs approchées des écart-types de X et de Y :

- $\sigma_X \approx 0.18$
- $\sigma_Y \approx 3.16$

Calculer le coefficient de corrélation de X et Y :

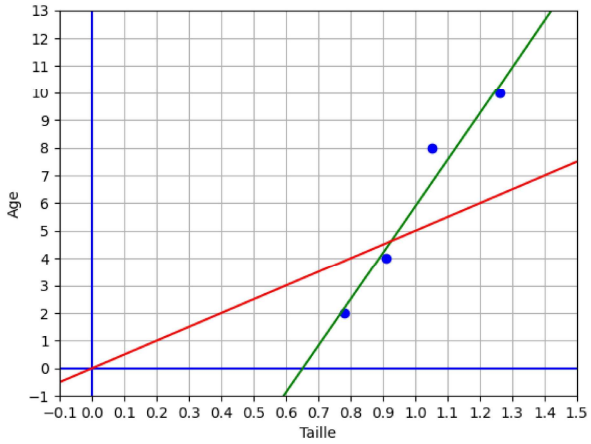
- $Cor(X,Y) = \ldots\ldots$

La corrélation est-elle forte ? $\ldots\ldots\ldots$

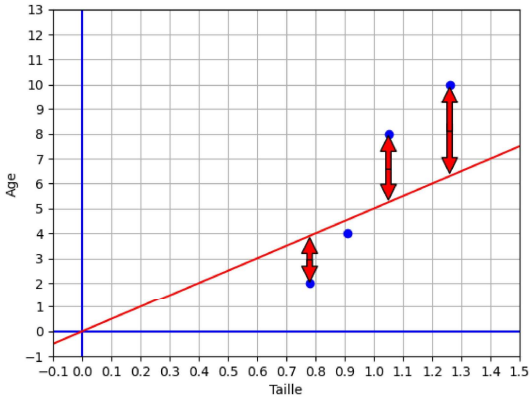
6. Droite d'ajustement linéaire

- **Exemple 1 : âge d'un enfant en fonction de sa taille**

	Enfant A	Enfant B	Enfant C	Enfant D
Taille (X)	0.78	0.91	1.05	1.26
Âge (Y)	2	4	8	10



- **Définition de la droite d'ajustement linéaire**



Objectif : Trouver la "meilleure droite" d'équation $y = ax + b$ passant par notre nuage de points

Définition

Soient X et Y deux variables statistiques sur une même population.
La *méthode des moindres carrés* consiste à chercher la droite $y = ax + b$ qui minimise la somme des carrés des écarts entre

- les points $(X(k), Y(k))$ du nuage de points
- et les points $(X(k), aX(k) + b)$ correspondant sur la droite.

Cette "meilleure droite" est appelée **droite d'ajustement linéaire**.

- **Equation de la droite d'ajustement linéaire**

Proposition

Soient X et Y deux variables statistiques sur une même population.
La droite d'ajustement linéaire de Y en X (c'est-à-dire avec X en abscisse et Y en ordonnée) a pour équation $y = ax + b$ avec :

$$a = \frac{Cov(X,Y)}{Var(X)}, \qquad b = \overline{Y} - a\overline{X}.$$

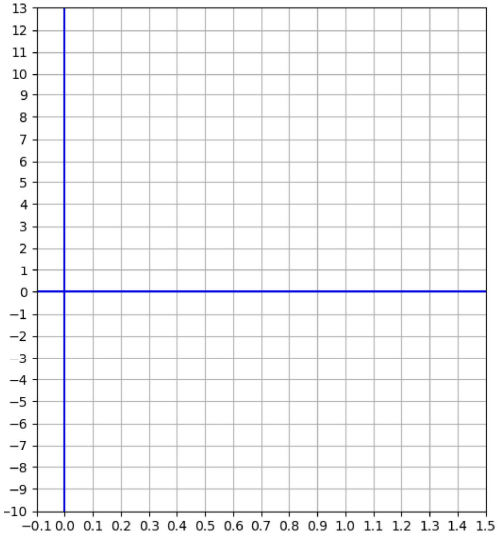
Exemple 1 : âge d'un enfant en fonction de sa taille

Calculer a et b et tracer la droite d'ajustement linéaire de Y en X sur le graphique ci-dessous.

$a = \ldots\ldots$

$b = \ldots\ldots$

Remarque : Attention, en général on n'obtient pas la même droite si on échange les variables X et Y.



Proposition

Plus le coefficient de corrélation $Cor(X,Y)$ est proche de 1, plus le nuage de points (X,Y) est proche de la droite d'ajustement linéaire de Y en X.

Remarque : Plus précisément, on peut démontrer que la somme des carrés des distances des points à la droite d'ajustement vaut $N\ Var(Y)\ Cor(X,Y)$.