Statistique Descriptive
Chapitre 02 : Statistique bivariée

La problématique

Introduction

- Dans le chapitre précédent, les notions de base de la statistique ont été vues.
- Dans la réalité, une étude statistique porte sur plus d'une variable.

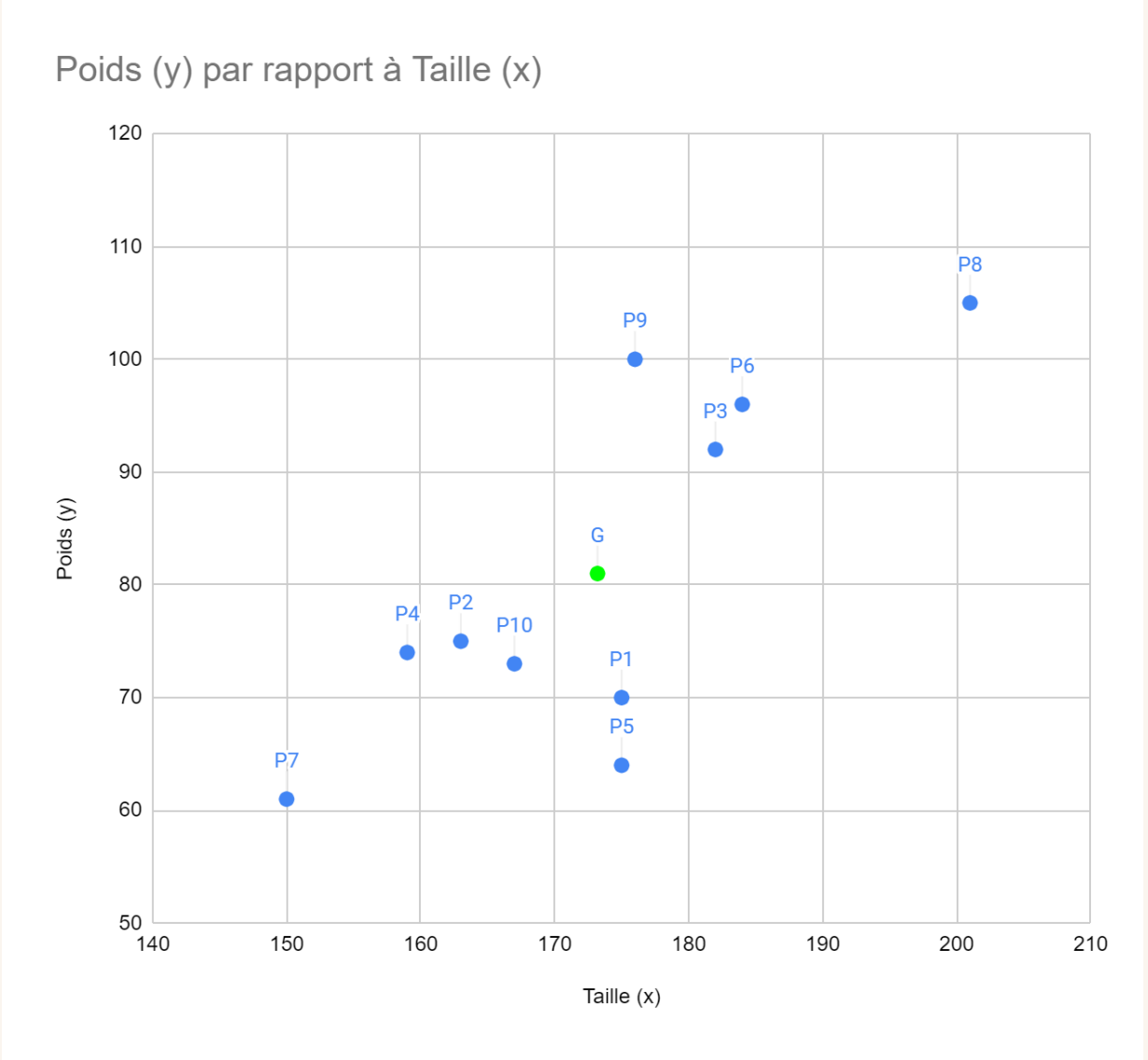
Exemple : relevé statistique

- Voici les relevés de la taille x_i (cm) et le poids y_i (kg) d'une population adulte d'un hameau.
- Avec, en complément, les relevés de l'âge z_i (ans) et de la pratique sportive w_i (qualitatif).

Individu	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
X : Taille (cm)	175	163	182	159	175	184	150	201	176	167
Y : Poids (kg)	70	75	92	74	64	96	61	105	100	73
Z : Âge (ans)	25	28	31	34	38	42	60	55	47	63
W : Pratique sportive	Régul.	Aucune	Régul.	Aucune	Occas.	Régul.	Aucune	Occas.	Aucune	Occas.

Exemple : nuage de points

Avec son centre de gravité en vert



Point moyen : g

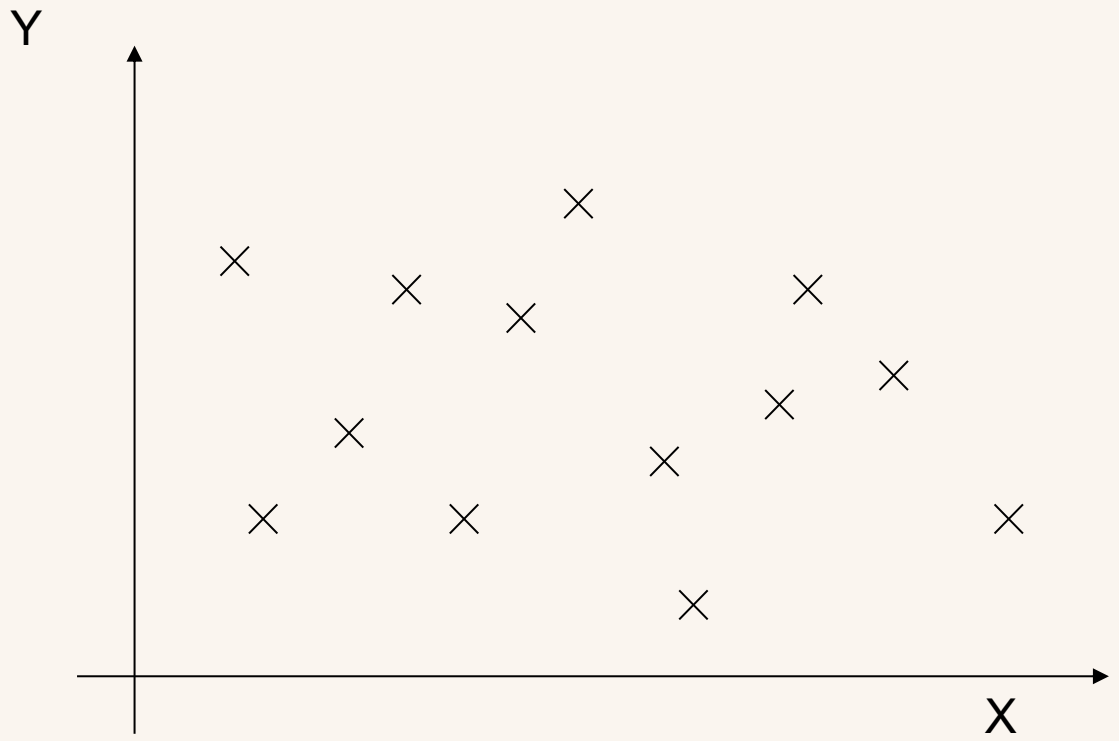
- Analogie avec la physique et le centre de gravité "g"
- Il est possible de calculer le point moyen g avec les formules suivantes

$$g = (\overline{X}; \overline{Y}) = (\frac{\sum_{i=1}^{n} x_i}{n}; \frac{\sum_{i=1}^{n} y_i}{n}) = (\ 173{,}2\ ;\ 81{,}0\ )$$

La corrélation entre les deux variables

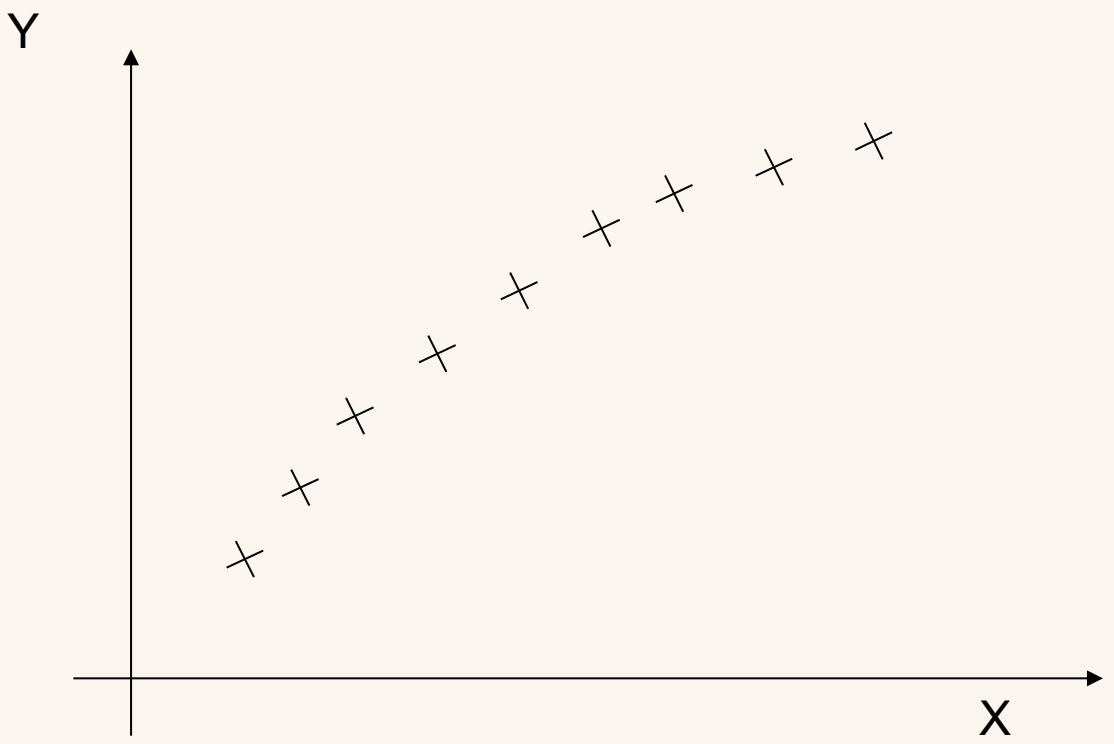
Indépendance

- Deux variables peuvent être indépendantes si le nuage de points couvre tout le graphe.
- Corrélation proche de 0 ou 0%.
- Exemple :
  - X : gain au Lotto d'une personne
  - Y : température extérieure le jour où la personne apprend son gain



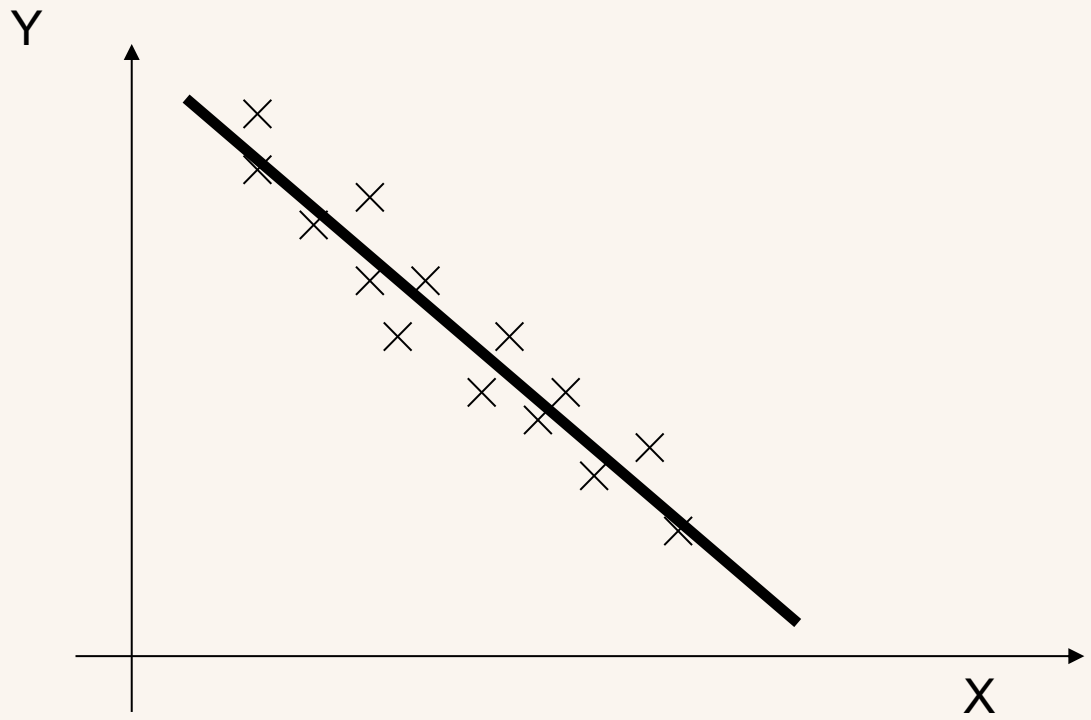
Dépendance fonctionnelle

- Deux variables peuvent avoir une dépendance fonctionnelle entre elles.
  - Une variable est fonction de l'autre.
- Corrélation proche de 1 ou 100%.
- Exemple :
  - La quantité de déchets produits par un ménage est proportionnelle au nombre de personnes dans le ménage.



Dépendance statistique ou corrélation

- Il se peut que les deux variables se répartissent autour d'une courbe.
  - Pas *sur* la courbe, mais plus ou moins proche



Rappel : Mathématiques

Équation d'une droite

Sens de l'équation

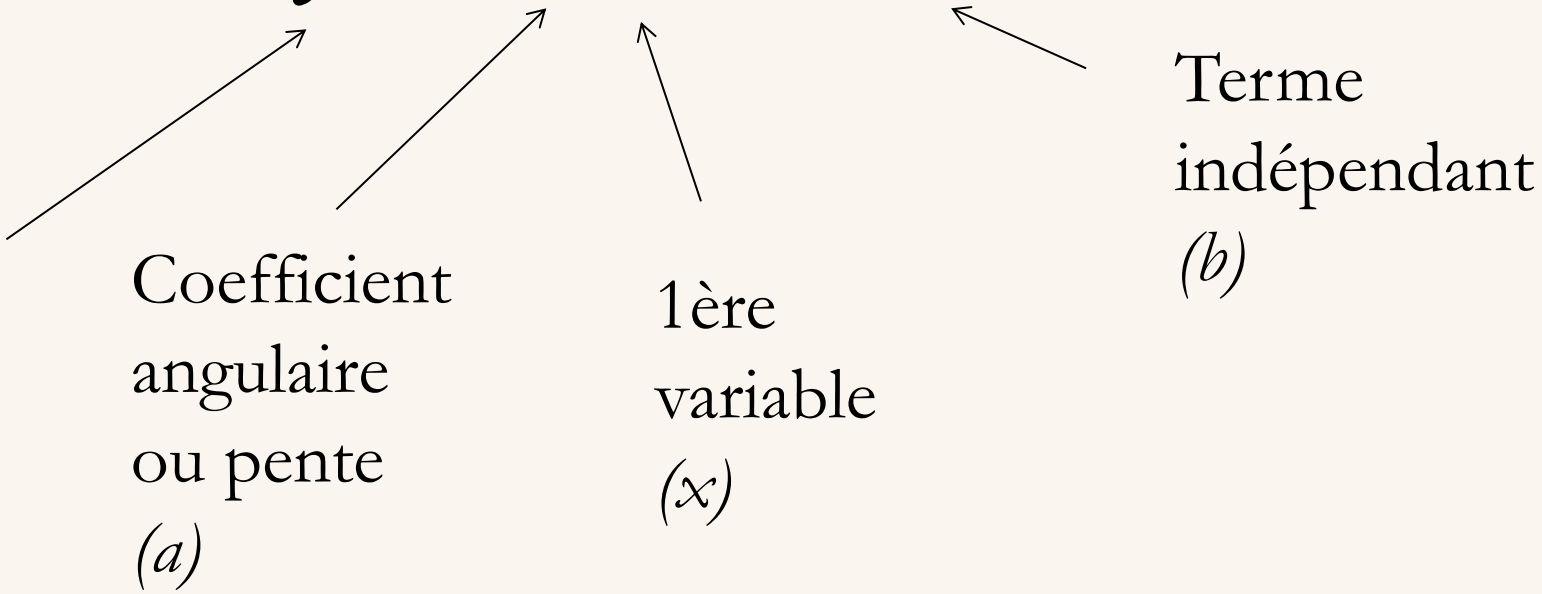
Coefficient angulaire

Terme indépendant

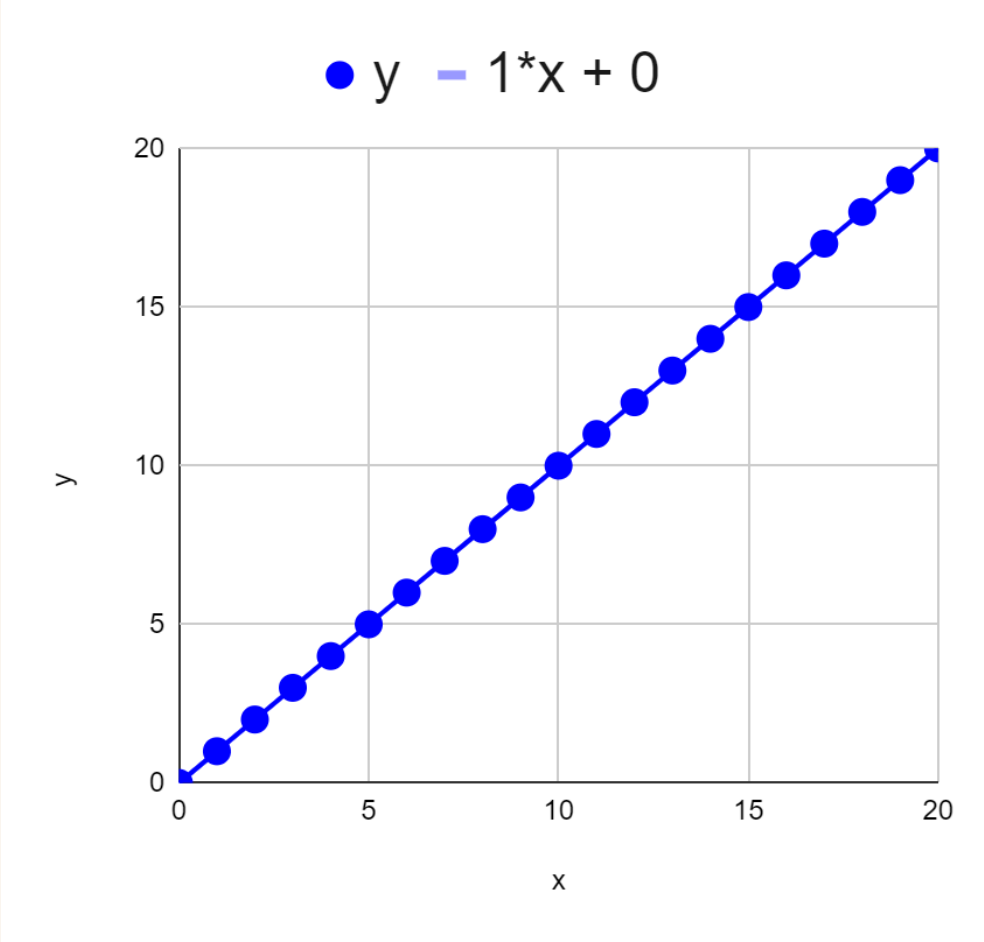
Équation d'une droite passant par un point donné

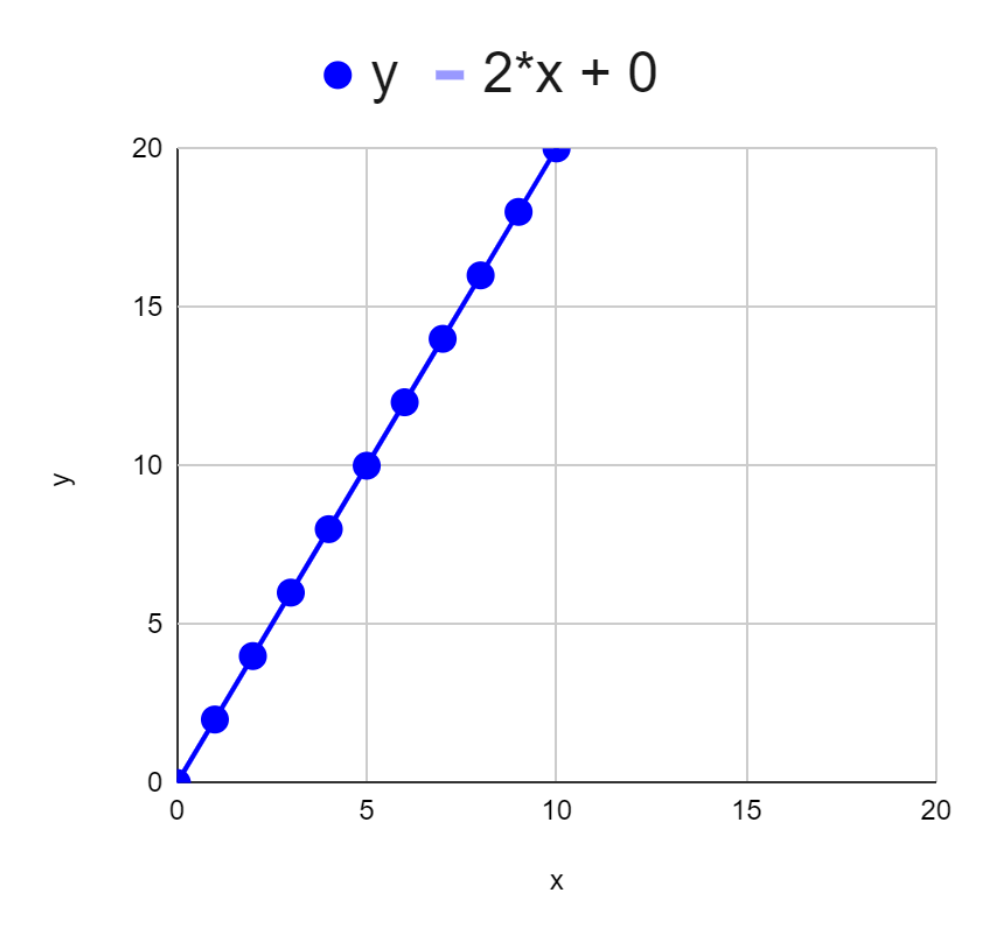
Rappel : Mathématiques : la droite

$$D \equiv y = ax + b$$

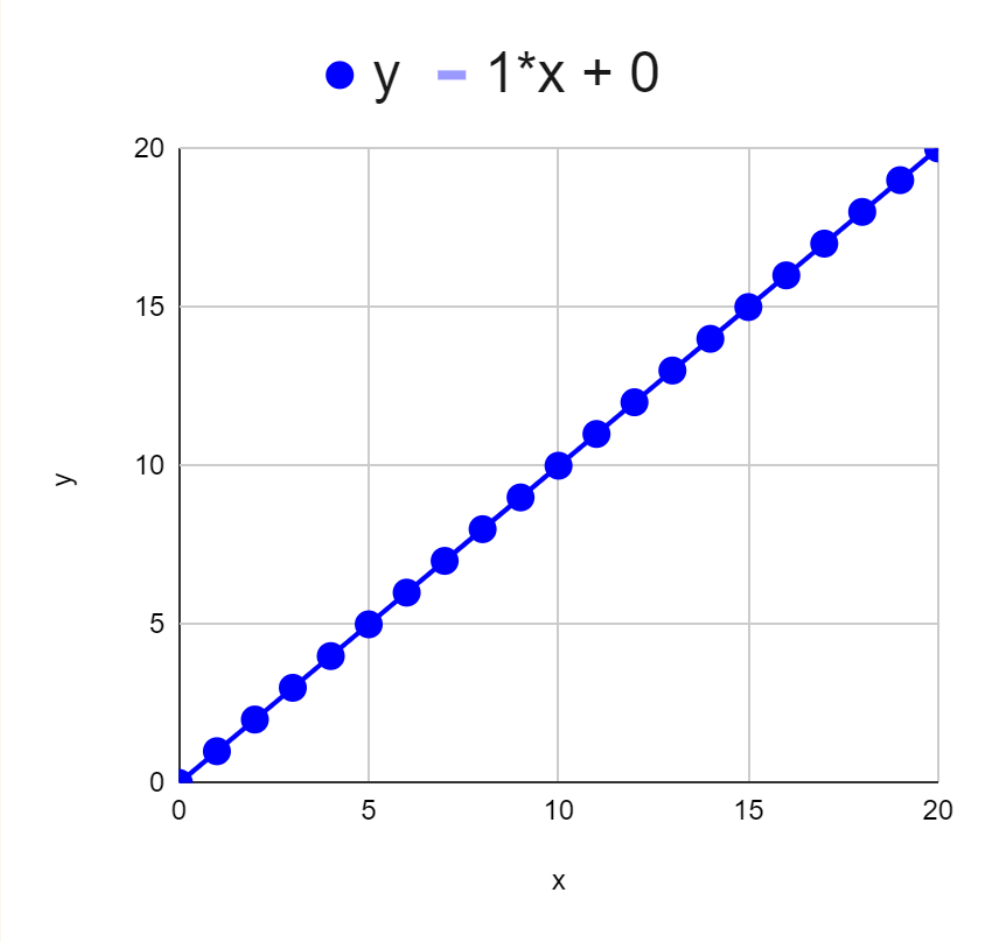


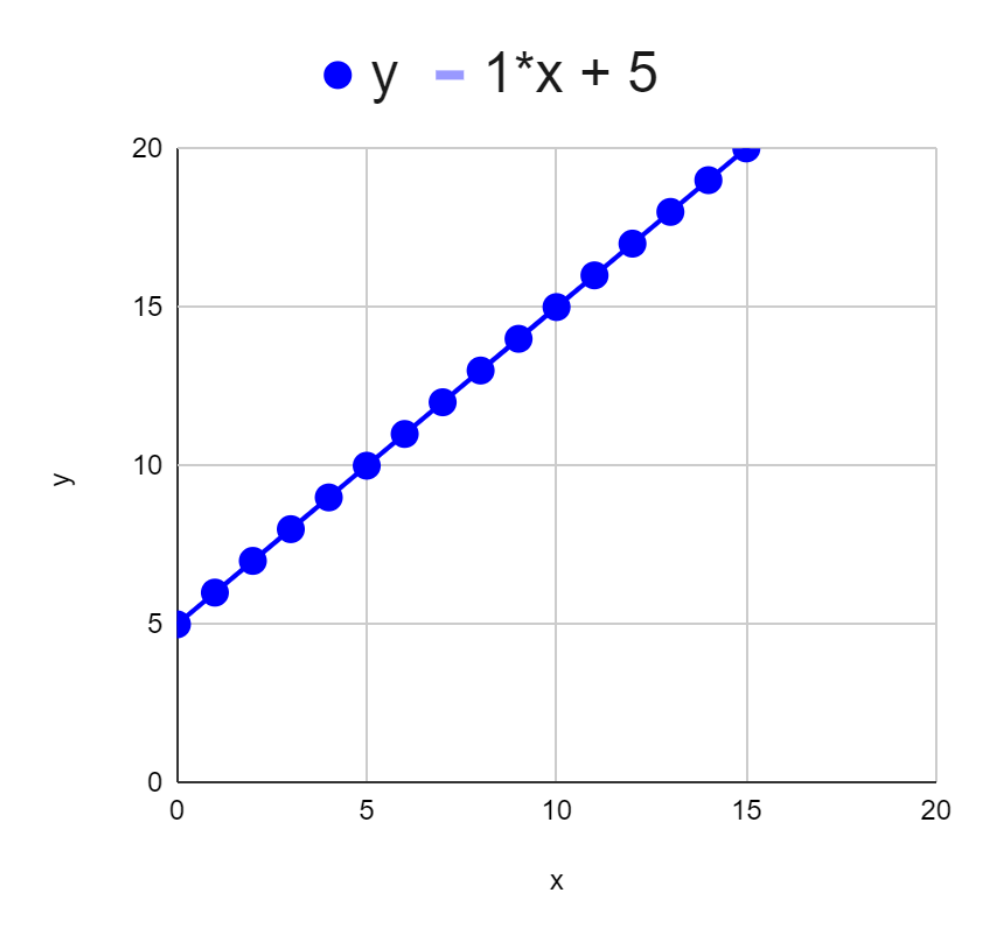
Rappel : Mathématiques : la droite



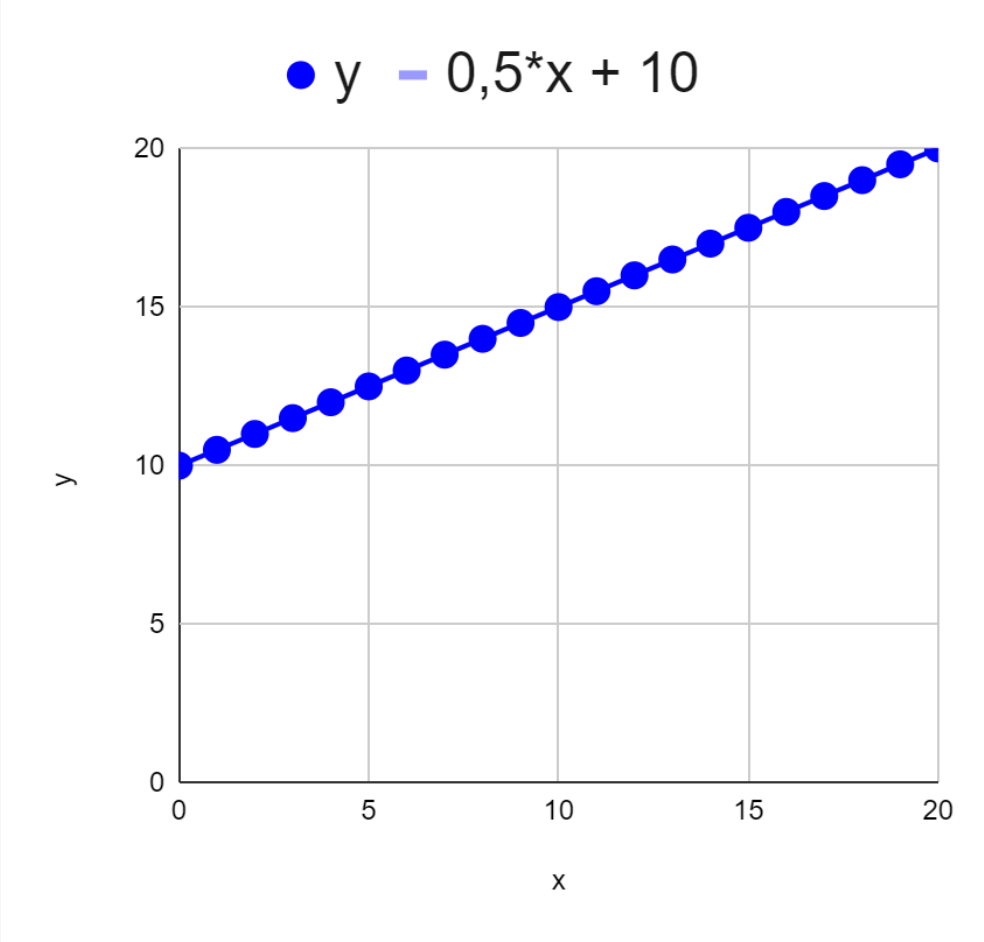


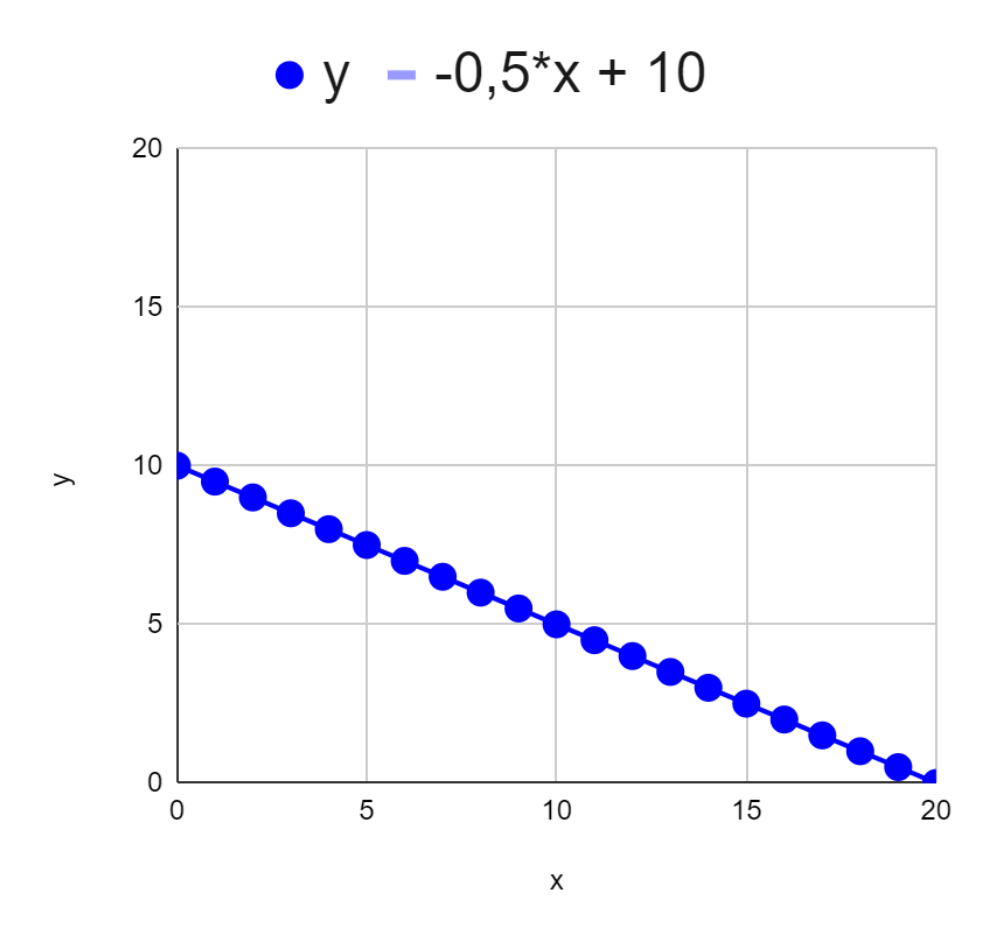
Rappel : Mathématiques : la droite





Rappel : Mathématiques : la droite





Rappel : Mathématiques : la droite passant par un point *(X,Y)*

- *X*, *Y* et *a* sont connus. Que vaut *b* ?
- L'équation de la droite de pente *a* passant par un point donné *(X,Y)* est :

$$y - Y = a\,(x - X)$$

$$y - Y = ax - aX$$

$$y = ax + (Y - aX)$$

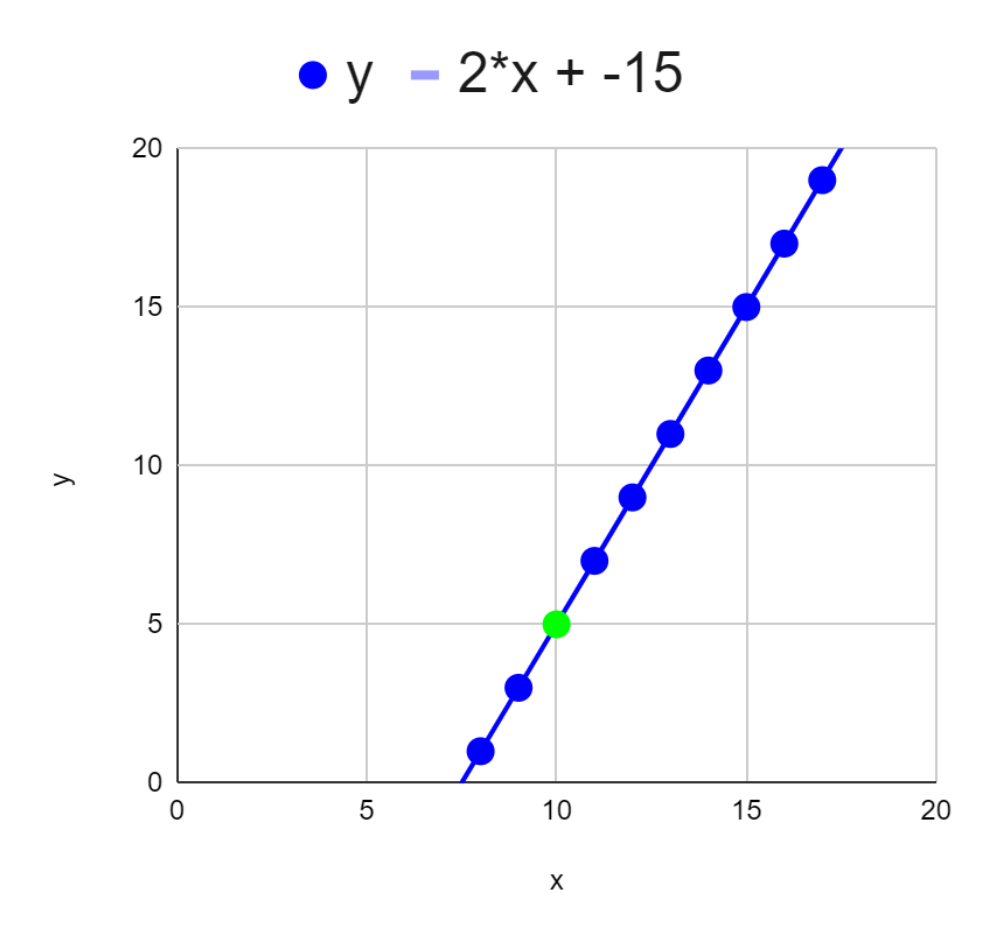
- Le terme indépendant *b* vaut donc

$$b = Y - aX$$

Rappel : Mathématiques : la droite passant par (X,Y)

- Exemple 1
- La droite de pente *2* passant par le point *(10, 5)* s'écrit donc

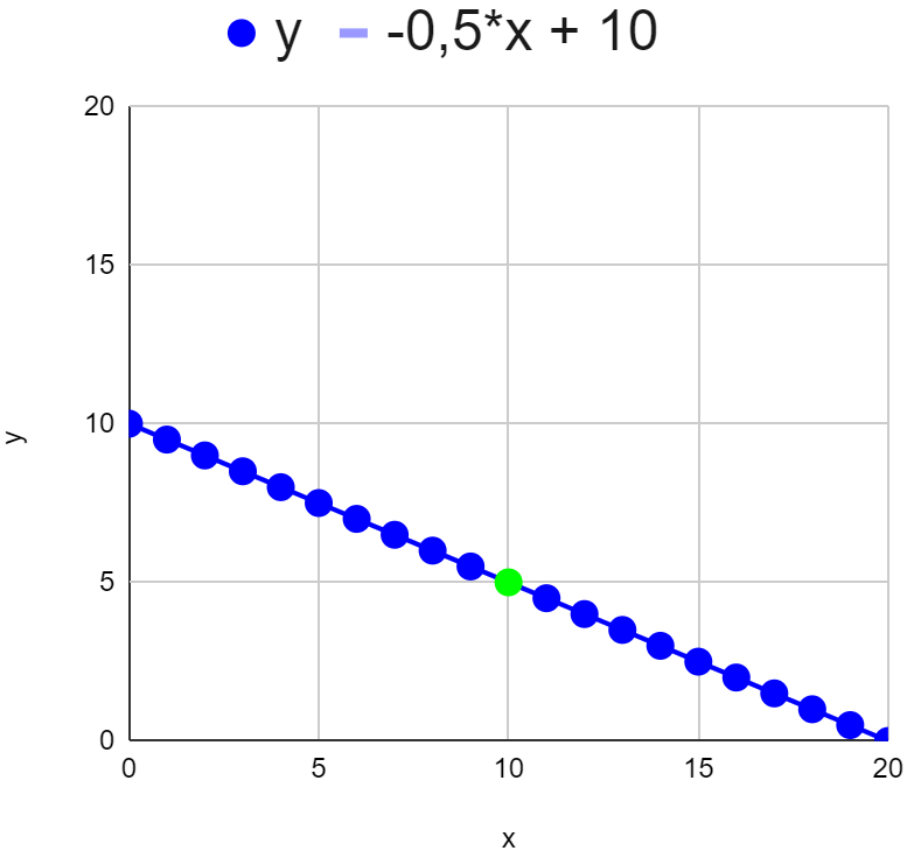
$$a = 2$$
$$b = 5 - 2 * 10 = -15$$
$$y = 2x - 15$$



Rappel : Mathématiques : la droite passant par (X,Y)

- Exemple 2
- La droite de pente *-0,5* passant par le point *(10, 5)* s'écrit donc

$$a = -0{,}5$$
$$b = 5 - (-0{,}5) * 10$$
$$= 10$$
$$y = -0{,}5\,x + 10$$



Ajustement linéaire

Corrélation entre deux variables

coefficient de corrélation "r"

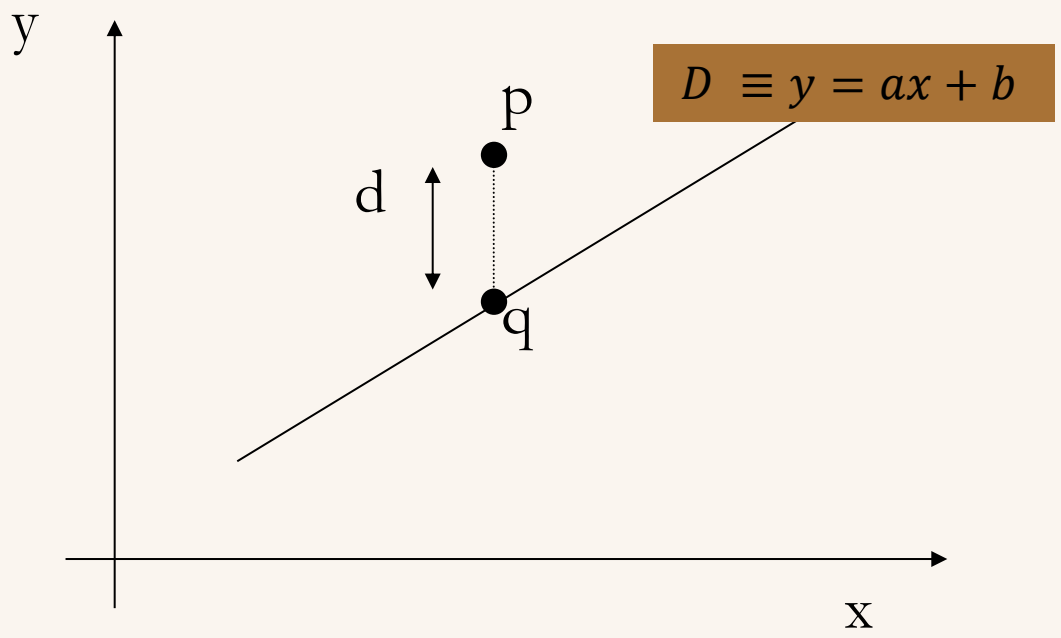
Ajustement linéaire

- Lorsqu'on s'aperçoit qu'il y a corrélation entre les deux variables, on peut effectuer un ajustement linéaire suivant la droite.
- On veut trouver la droite qui passe "au milieu des points".

$$D \equiv y = ax + b$$
$$a : \text{à } déterminer$$
$$b : \text{à } déterminer$$

- En fait, on va calculer deux droites,
  - l'une qui ajuste y par rapport à x : D
  - l'autre qui ajuste x par rapport à y : D'
- Remarque : Dans la réalité, le plus souvent
  - on cherche avant tout comment y dépend de x
  - dans ce cas, seul D compte

Droite de régression (1/2)



- D est la droite de régression de y en fonction de x
- *a* minimise le carré de la distance verticale entre p et D

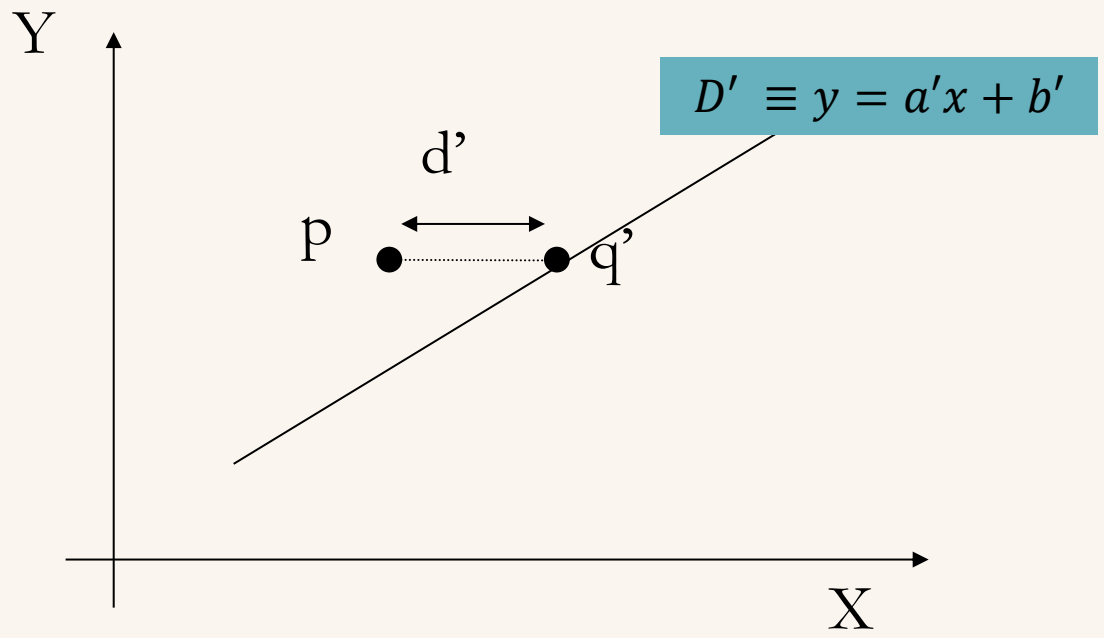
$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

⇩

$$y - \bar{Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}.(x - \bar{X})$$

$$\rightarrow b = \bar{Y} - a\bar{X}$$

Droite de régression (2/2)



- D' est la droite de régression de X en fonction de Y
- a' minimise le carré de la distance horizontale entre p et D

$$a' = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}$$

⇩

$$y - \bar{Y} = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}.(x - \bar{X})$$

$$\rightarrow b' = \bar{Y} - a'\bar{X}$$

Exemple (1/3)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1732}{10} = 173{,}2$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{810}{10} = 81$$

x_i	y_i	$x_i - \bar{X}$	$y_i - \bar{Y}$	$(x_i - \bar{X})^2$	$(y_i - \bar{Y})^2$	$(x_i - \bar{X})(y_i - \bar{Y})$
175	70	1,8	-11	3,24	121	-19,8
163	75	-10,2	-6	104,04	36	61,2
182	92	8,8	11	77,44	121	96,8
159	74	-14,2	-7	201,64	49	99,4
175	64	1,8	-17	3,24	289	-30,6
184	96	10,8	15	116,64	225	162
150	61	-23,2	-20	538,24	400	464
201	105	27,8	24	772,84	576	667,2
176	100	2,8	19	7,84	361	53,2
167	73	-6,2	-8	38,44	64	49,6
1732	**810**	**0**	**0**	**1863,6**	**2242**	**1603**

Exemple (2/3)

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{X}).(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

$$= \frac{1603}{1863,6} = 0,8602$$

$$b = \bar{Y} - a\bar{X}$$

$$= 81 - 0,8602 * 173,2$$

$$= -67,952$$

Droite de régression D donnant le poids en fonction de la taille :

$$D \equiv y = 0,86x - 67,952$$

$$a' = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{\sum_{i=1}^{n}(x_i - \bar{X}).(y_i - \bar{Y})}$$

$$= \frac{2242}{1603} = 1,3986$$

$$b' = \bar{Y} - a'\bar{X}$$

$$= 81 - 1,3986 * 173,2$$

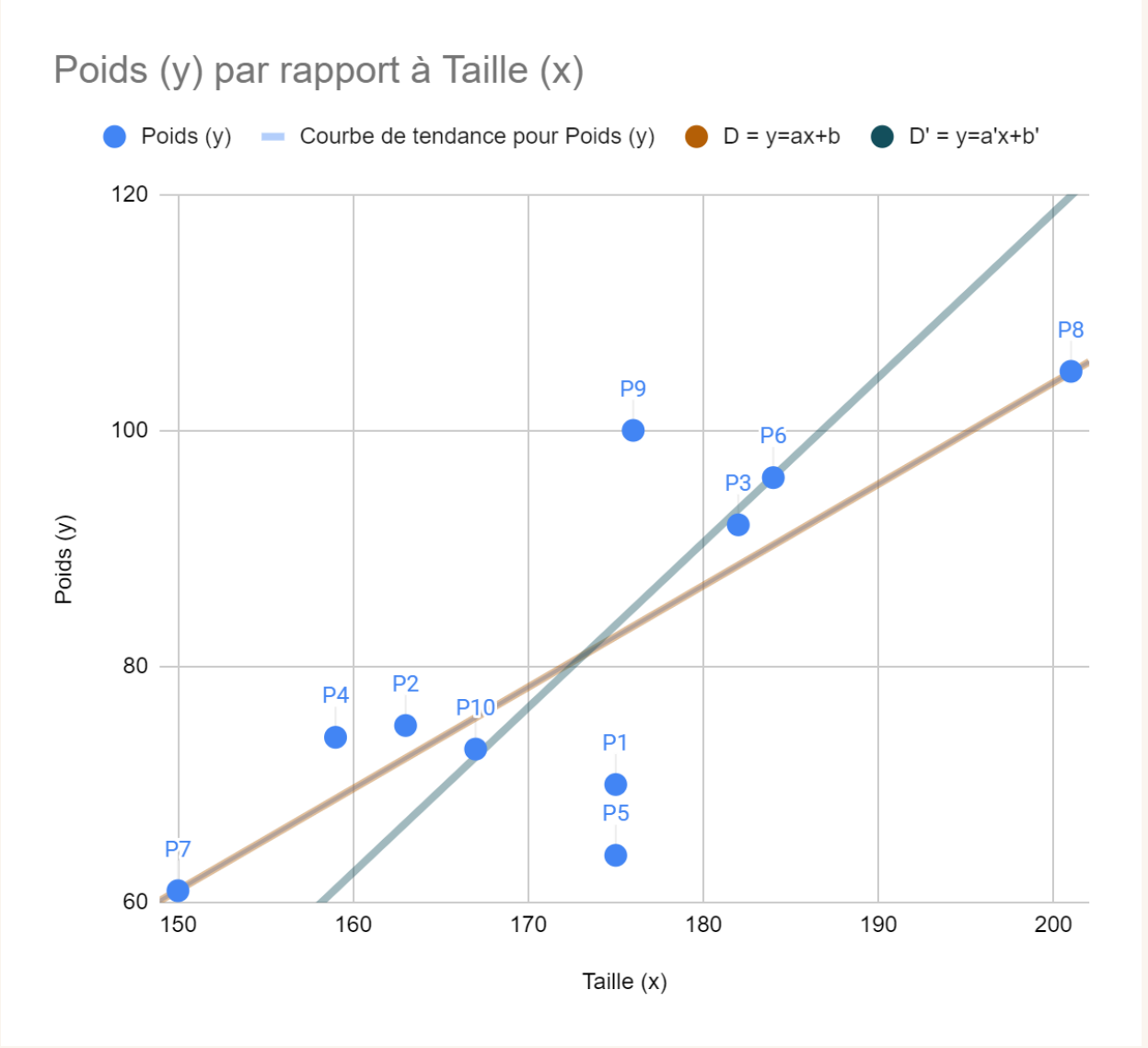
$$= -161,48$$

Droite de régression D' donnant la taille en fonction du poids :

$$D' \equiv y = 1,4x - 161,48$$

Exemple (3/3)

- a (ou a') est positif
  - corrélation positive
- a (ou a') est négatif
  - corrélation négative



Coefficient de corrélation linéaire

- Si on estime qu'il y a une corrélation linéaire entre les variables
  on voudra « mesurer » l'intensité de cette corrélation.
  →coefficient de corrélation linéaire "r" (Pearson)
- "r" a une valeur comprise entre
  - **0** : corrélation faible
  - **1** : corrélation forte

$$r^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{X}).(y_i - \bar{Y})\right)^2}{\left(\sum_{i=1}^{n}(x_i - \bar{X})^2\right).\left(\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2\right)} = \frac{a}{a'}$$

$$r^2 = \frac{0{,}8601}{1{,}3986} = 0{,}614972$$

$\longrightarrow$

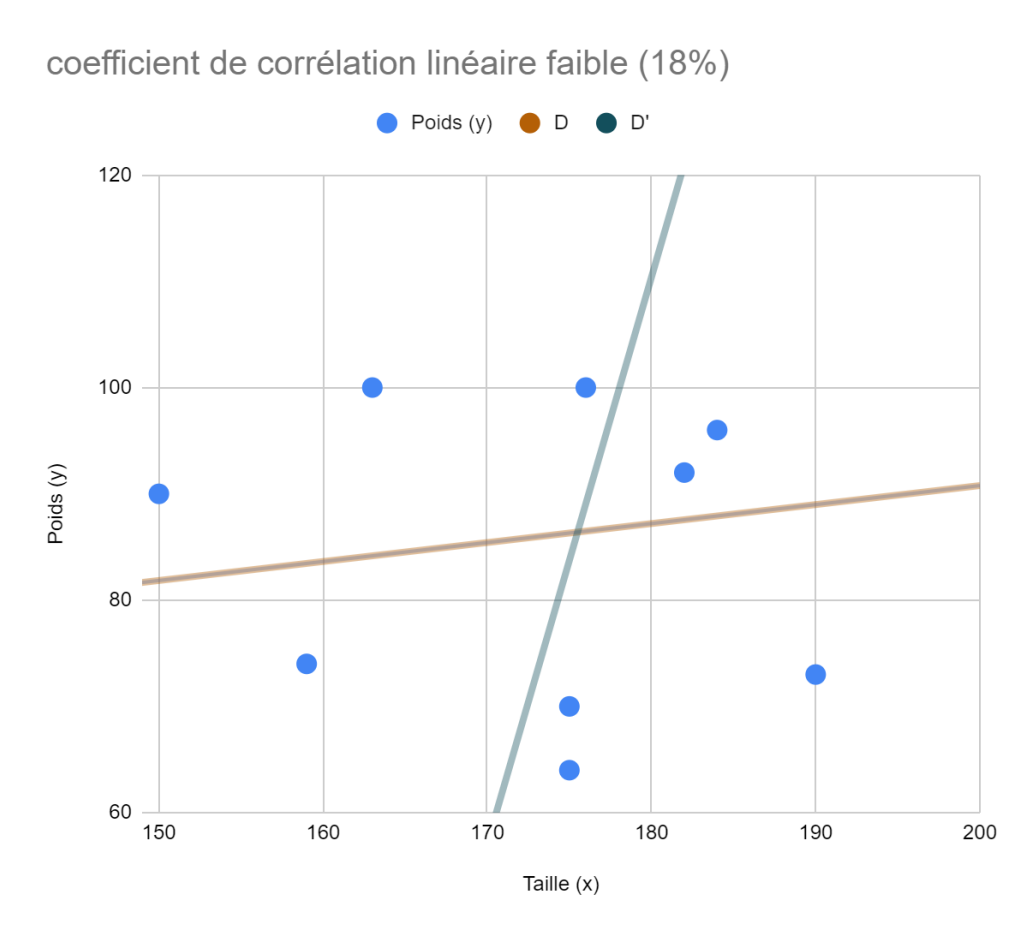
$$r = \sqrt{0{,}614972} = \mathbf{0{,}7842}$$

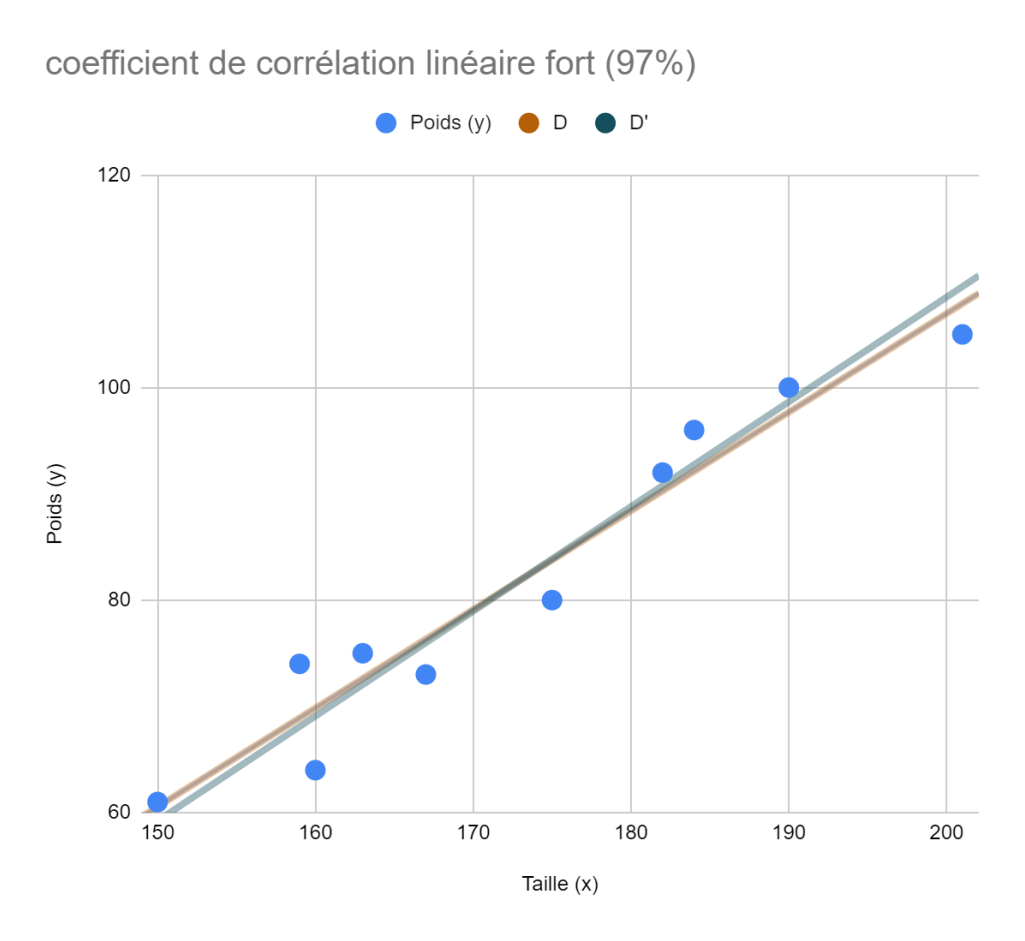
Coefficient de corrélation linéaire - autre approche

- Calcul alternatif, équivalent au précédent à un détail près.
  - Intéressant si on connait les variances : σ_x, σ_y , σ_{xy}
- "r" a une valeur comprise entre
  - **-1** : corrélation négative forte
  - **+1** : corrélation positive forte
  - Passant par **0** : corrélation faible

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{X}).(y_i - \bar{Y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{X})^2\right)} . \sqrt{.\left(\sum_{i=1}^{n}(y_i - \bar{Y})^2\right)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad r = \frac{1603}{\sqrt{1863,6} . \sqrt{2242,0}} = \mathbf{0,7842}$$

Coefficient de corrélation linéaire : exemples





Détermination

prédiction d'une valeur

qualité de la prédiction d'une valeur

coefficient de détermination R^2

modèle et équation de régression multivariables

Le problème de départ

- Comment prédire le poids d'un individu, sans aucune autre information ?
  - Meilleure stratégie : annoncer la moyenne
    - `poids = 78 kg`
  - Peu précis
- Comment prédire son poids, connaissant sa taille ?
  - Mieux que la moyenne, utiliser la droite de régression
    - `poids = a * taille + b`
  - Une partie de l'erreur a disparu
- On prédira encore mieux son poids avec d'autres informations, peut-être avec l'âge ?
  - Régression multiple
    - $poids = a_1 \times taille + a_2 \times âge + b$
  - Mais l'âge joue-t-il un vrai rôle ?

Coefficient de détermination R^2

- R^2 mesure la part de variance expliquée par le modèle, par rapport à la variance totale.
- $R^2 = \frac{\text{variance expliquée}}{\text{variance totale}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$

  y_i : valeur observée

  $\hat{y}_i$: valeur prédite par la droite

  $\bar{y}$: moyenne

Application à notre exemple

- $R^2 = 0{,}615$
- La taille explique 62 % de la variation du poids.
- Les 38 % restants sont non expliqués par la taille
  - mais par l'alimentation, sport, morphologie individuelle, …
  - illustration de la dispersion des points autour de la droite.
- Généralisation
  - R^2 se réfère à un modèle de régression, faisant intervenir plusieurs variables quantitatives
  - Ici, 2 mais on peut impliquer d'autres

Application à notre exemple

Élément	Valeur
a_1 (kg par cm de taille)	0,863
a_2 (kg par année d'âge)	0,120
b (intercept)	−73,6 kg
R^2 simple (taille seule)	0,615
R^2 multiple (taille + âge)	0,625
Gain apporté par l'âge	seulement +0,010
r (taille -> poids)	0,773
r (âge -> taille)	-0,029
r (âge -> poids)	0,124

- Le R^2 multiple (0,625) mobilise deux colonnes à la fois.
  - Aucun des trois r par paires ne pourrait le résumer.
- Le gain de l'âge est minuscule (+0,01 point de R^2).

Lien élégant : $R^2 = r^2$

- régression linéaire simple :
  - coefficient de détermination = carré de r de Pearson
- $r = 0{,}784$
- $r^2 = 0{,}784^2 = 0{,}615$ ✓
- Deux regards sur la même information.



Où R^2 est-il vraiment utile ?

Régression simple
2 variables

- $R^2 = r^2$
  - redondant avec le r déjà calculé
- Intérêt
  - lecture immédiate en « pourcentage de variance expliquée »

Régression multiple
3+ variables

- « r » multiples
  - R^2 unique devient irremplaçable
- Intérêt
  - comparaison de modèles : taille seule vs taille + âge
- Si R^2 passe de 0,62 à 0,65 : l'âge n'apporte presque rien
  - non-corrélation

Limites

- R^2 ne prouve pas la causalité — juste une association
- R^2 augmente mécaniquement avec le nombre de variables
  - même celles ne jouant aucun vrai rôle dans la régression
  - d'où le R^2 ajusté en régression multiple
- R^2 élevé ≠ bon modèle
  - un nuage incurvé peut tromper
- Petits échantillons : avec $n = 10$, R^2 est très instable et peu significatif.

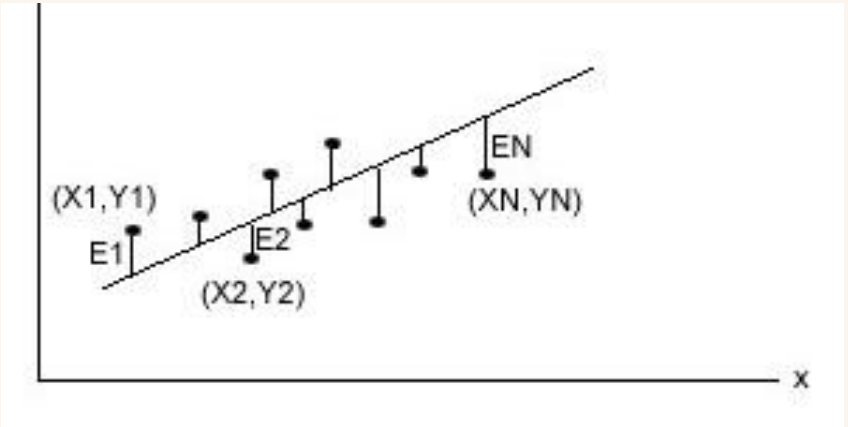
Annexe

Détails théoriques de la méthode des moindres carrés

Cette partie du cours n'est pas à connaître pour l'examen. Ce n'est pas une raison pour ne pas la lire. ☺

Méthode des moindres carrés (1/7)

- Pour une valeur quelconque de X, par exemple X1, il y aura une différence ou erreur entre la valeur de Y1 et la valeur qui lui correspond sur la courbe. Cette erreur est indiquée par E sur la figure ci-dessus. Les valeurs que peuvent prendre celles-ci peuvent être positives, négatives ou simplement, nulles.



Méthode des moindres carrés (2/7)

- On peut mesurer l'efficacité d'un ajustement en effectuant la somme des carrés des erreurs de l'ensemble des données. Si cette somme est petite, l'ajustement est considéré comme étant bon. À l'inverse, si elle est grande, l'ajustement est mauvais. On dit donc que parmi toutes les courbes possible qui s'approche d'un ensemble de données, celle révélant le meilleur ajustement est celle qui correspond à la propriété suivante:

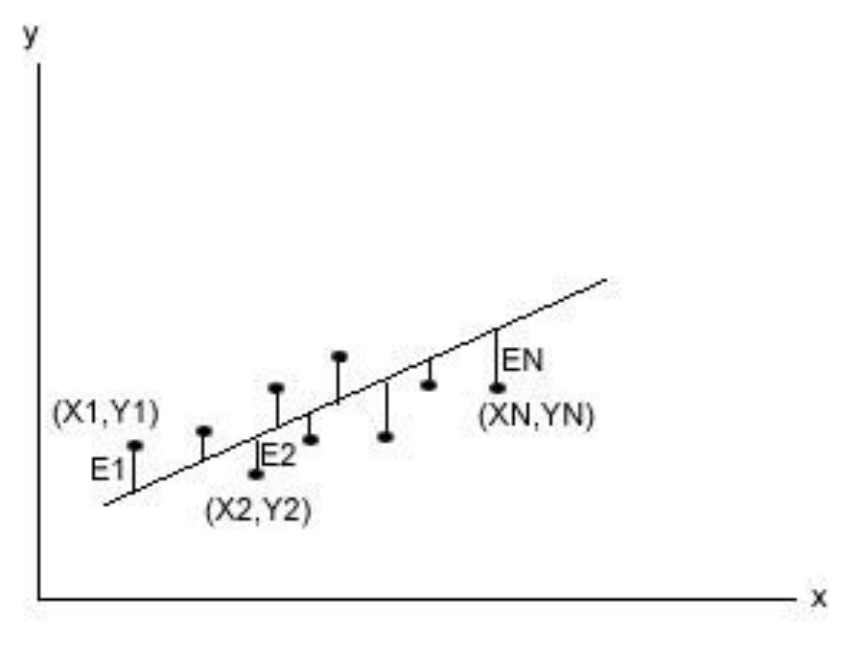
$$\text{Minimum} \left(E_1^2 + E_2^2 + \ldots + E_N^2 \right)$$

- On appelle courbe des moindres carrées une courbe qui vérifie cette propriété.

Méthode des moindres carrés (3/7)

- La droite des moindres carrés ajustant les points (X_1,Y_1), (X_2,Y_2) … (X_N,Y_N) possède l'équation suivante :

$$D \equiv y = ax + b$$



Si on prend le point E_1 et qu'on calcule le point à sa verticale qui "tombe" sur la droite D, on obtient le point D_1 : (X_1, aX_1+b).

La distance entre E_1 et D_1 est $Y_1 - aX_1 - b$.

Le but de la méthode est de minimiser la somme des écarts des points avec la droite :

$$\sum_{i=1}^{N} (y_i - ax_i - b)^2$$

Méthode des moindres carrés (4/7)

$$f_a(b) = \sum_{i=1}^{N} ((y_i - ax_i) - b)^2$$

⇩ Définition du carré parfait de la différence de 2 nombres

$$f_a(b) = \sum_{i=1}^{N} (y_i - ax_i)^2 - 2b.\sum_{i=1}^{N} (y_i - ax_i) + Nb^2$$

La fonction a son minimum en b lorsque la dérivée de f par rapport à b vaut 0

$$f_a'(b) = 2Nb - 2.\sum_{i=1}^{N} (y_i - ax_i) = 0$$

$$b = \frac{1}{N}\sum_{i=1}^{N} (y_i - ax_i) = \overline{Y} - a\overline{X}$$

Méthode des moindres carrés (5/7)

$$f_a(b) = \sum_{i=1}^{N} (y_i - ax_i - b)^2 \qquad b = \overline{Y} - a\overline{X}$$

$$\Longrightarrow \quad S(a) = \sum_{i=1}^{N} \left(y_i - ax_i - \left(\overline{Y} - a\overline{X}\right)\right)^2$$

$$\Longrightarrow \quad S(a) = \sum_{i=1}^{N} \left((y_i - \overline{Y}) - a\left(x_i - \overline{X}\right)\right)^2$$

⇩ Définition du carré parfait de la différence de 2 nombres

$$S(a) = \left(\sum_{i=1}^{N} \left(x_i - \overline{X}\right)^2\right) a^2 - 2\left(\sum_{i=1}^{N} \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)\right) a + \sum_{i=1}^{N} \left(y_i - \overline{Y}\right)^2$$

Méthode des moindres carrés (6/7)

La fonction a son minimum en a lorsque la dérivée de S par rapport à a vaut 0

$$S'(a) = 2\left(\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2\right)a - 2\left(\sum_{i=1}^{N}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)\right) = 0$$

$$\left(\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2\right)a = \sum_{i=1}^{N}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)$$

$$a = \frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}$$

Méthode des moindres carrés (7/7)

$$D \equiv y = ax + b$$

$$a = \frac{\sum_{i=1}^{N} \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{N} \left(x_i - \overline{X}\right)^2}$$

$$b = \overline{Y} - a\overline{X}$$

$$D \equiv y = ax + \overline{Y} - a\overline{X}$$

$$D \equiv y - \overline{Y} = a\left(x - \overline{X}\right)$$

$$D \equiv y - \overline{Y} = \frac{\sum_{i=1}^{N} \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{N} \left(x_i - \overline{X}\right)^2}\left(x - \overline{X}\right)$$