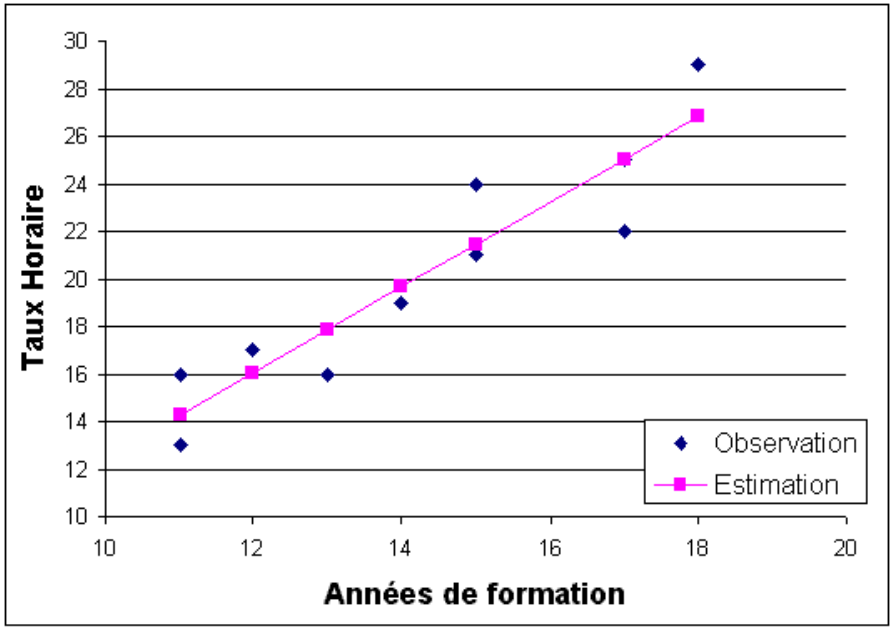
Corrélation et régression linéaire

1. Concept de corrélation
2. Analyse de régression linéaire
3. Différences entre valeurs prédites et observées d’une variable

2. Analyse de régression

Après avoir calculé ***r*** l'intensité du lien linéaire entre les variables x et y et testé la validité de ce lien on passe à l'étape suivante :

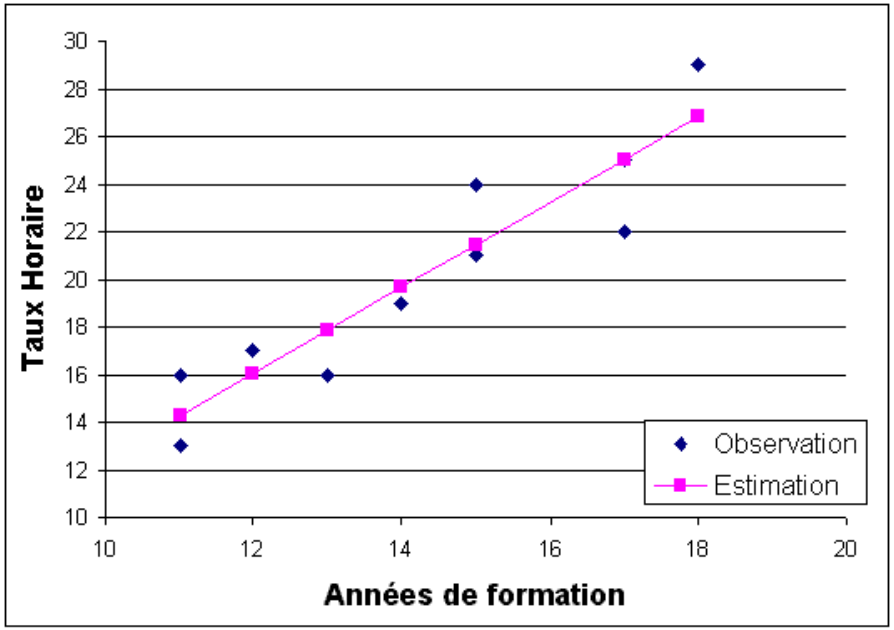
la construction de la droite de régression linéaire par la méthode des moindres carrés ordinaires



On va calculer l'équation d'une droite de la forme :
$\hat{y} = a\,x + b$. On calcule **a** et **b**

Le tracé de cette droite sur le même graphique que le nuage de points est celui qui s'ajuste le mieux au nuage de points

Régression linéaire (droite d'estimation)



Objectif de l'étude

La méthode de la régression a pour but de **décrire la relation** entre une ***variable aléatoire dépendante ou variable réponse Y*** et un ensemble de ***variables indépendantes* ou *explicatives* ou *prédictives X***,

Y est la variable dont on veut expliquer les valeurs

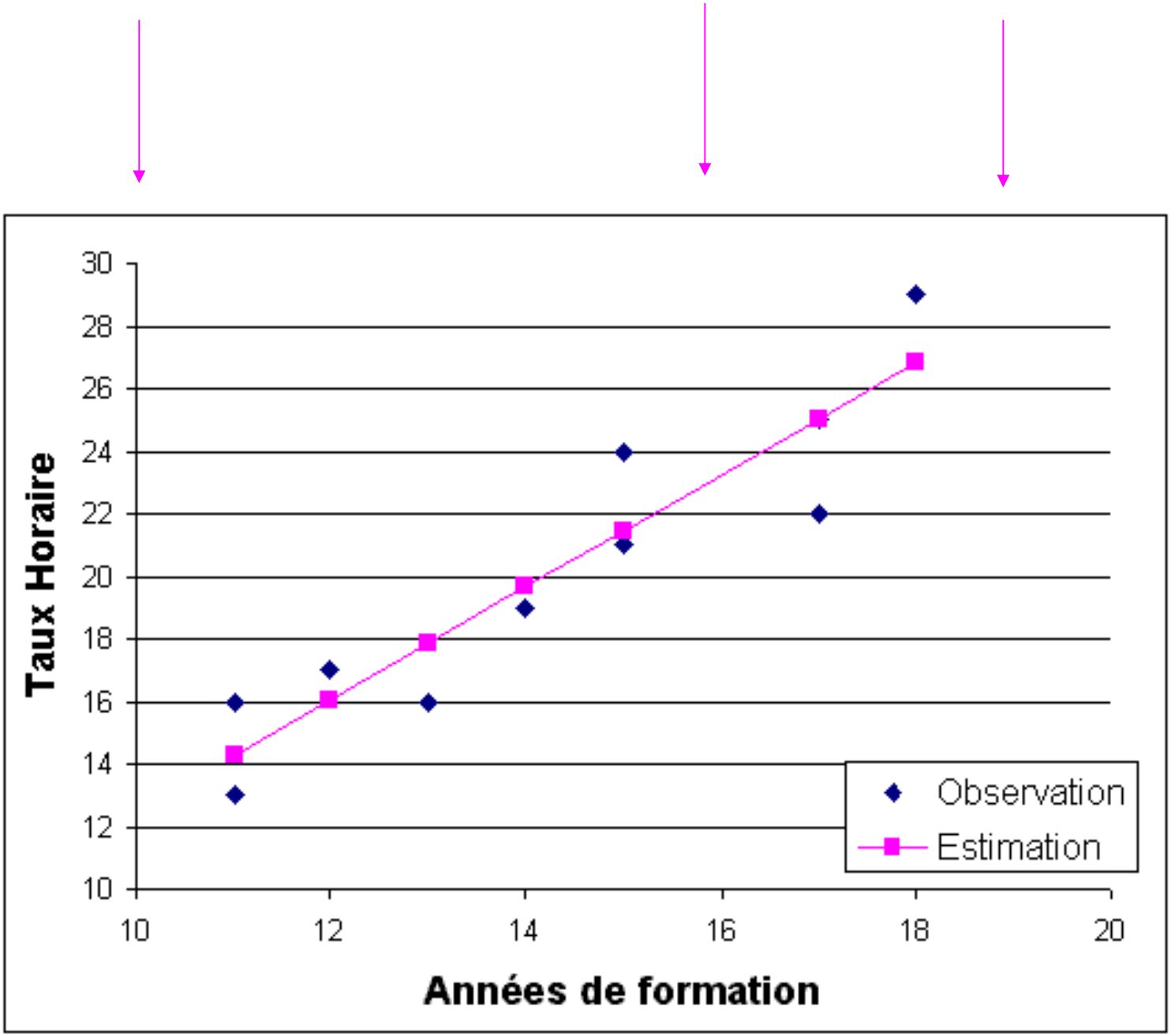
X est la variable que l'on veut utiliser pour expliquer Y

régression = modélisation
plusieurs objectifs possibles

1.Description : trouver le meilleur modèle liant les variables Y et X

2. Inférence : tester des hypothèses précises se rapportant aux paramètres du modèle dans la population comme la pente A et l'ordonnée à l'origine B de la droite Y = AX + B.
A et B sont estimés par a et b

3. Prédiction : Prévoir et prédire les valeurs de la variable dépendante Y pour de nouvelles valeurs de la variable X


Taux Horaire
30
28
26
24
22
20
18
16
14
12
10
10
12
14
16
18
20
Années de formation
Observation
Estimation

• Si les variables ***x*** sont **contrôlées ou indépendantes** on parle de régression de modèle I : *droite des moindres carrés*.
• Si les variables ***x*** sont **aléatoires**, on parle de régression de modèle II : *droite des moindres rectangles*

$$Equation:\ \hat{y} = ax + b$$

est l'équation d'une *ligne droite :*
- traversant le nuage de points
- permettant de calculer une valeur estimée pour chaque point
- d'axe des *x* correspondant à la variable prédictive.

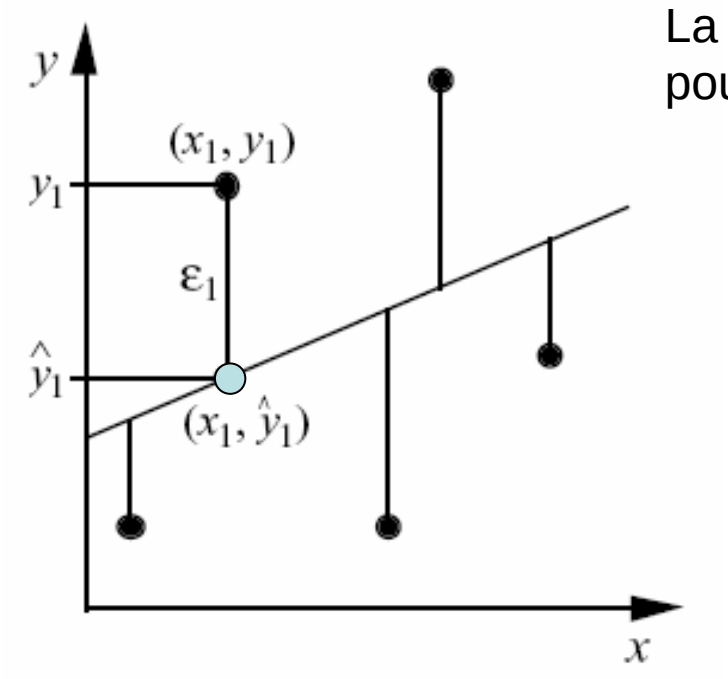
ŷ = ***droite d'estimation*** **ou** ***droite de régression de y en x.***
C'est celle qui est obtenue avec la calculatrice autorisée

x et y ne jouent plus des rôles symétriques

Principe des moindres carrés

$$\hat{y} = ax + b$$

Faire passer la droite d'estimation, à travers le nuage de points, de façon à ce que les différences (y – ŷ) soient les plus faibles possible pour l'ensemble des points.



La différence $\boldsymbol{\varepsilon_i = e_i = (y_i - \hat{y}_i)}$ porte le nom de ***résidu*** pour l'observation *i*.

On minimise $E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b - ax_i)^2$

en calculant $\frac{\partial E}{\partial a} = 0$ et $\frac{\partial E}{\partial b} = 0$

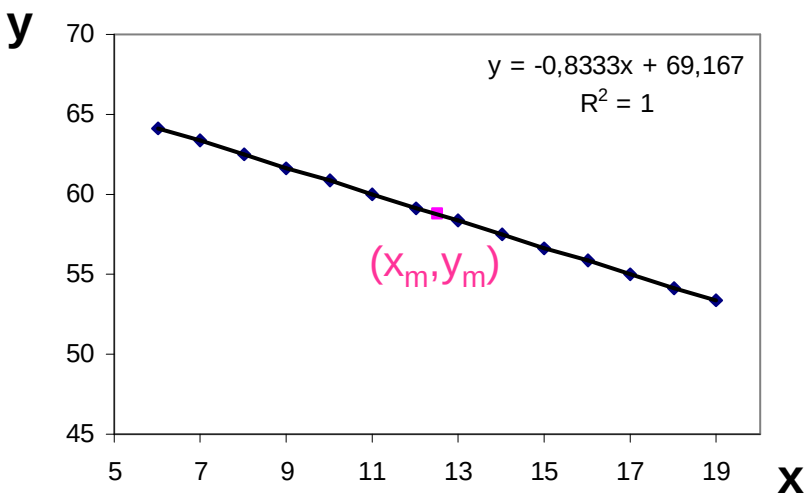
$$a = \frac{cov(x,y)}{var(x)} = \frac{cov(x,y)}{S^2(x)}$$

et

$$b = \bar{y} - a\bar{x}$$

???

Régression de y en x



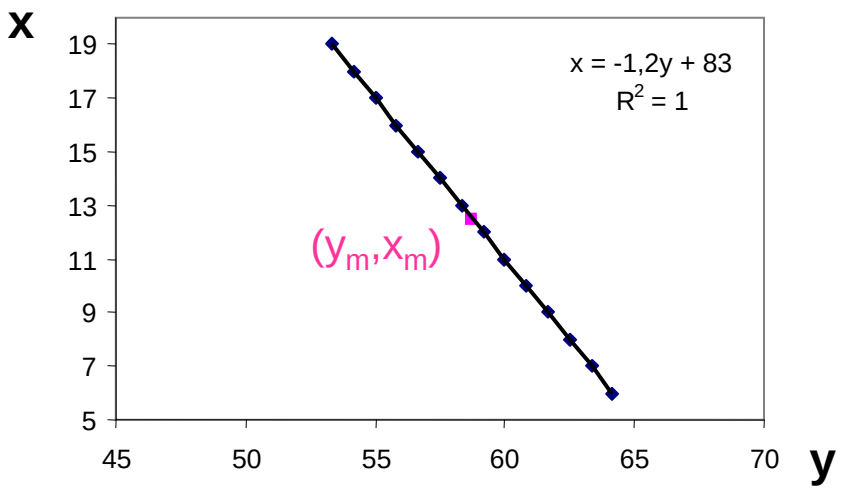
$$\hat{y} = ax + b$$

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

$$b = \bar{y} - a\bar{x}$$

$$r^2 = aa'$$

Régression de x en y



$$\hat{x} = a'y + b'$$

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

$$b' = \bar{x} - a'\bar{y}$$

Sachant que $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \times \sqrt{\sum (y_i - \bar{y})^2}}$

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

$\Rightarrow$

$$a = r \frac{s_y}{s_x}$$

ou

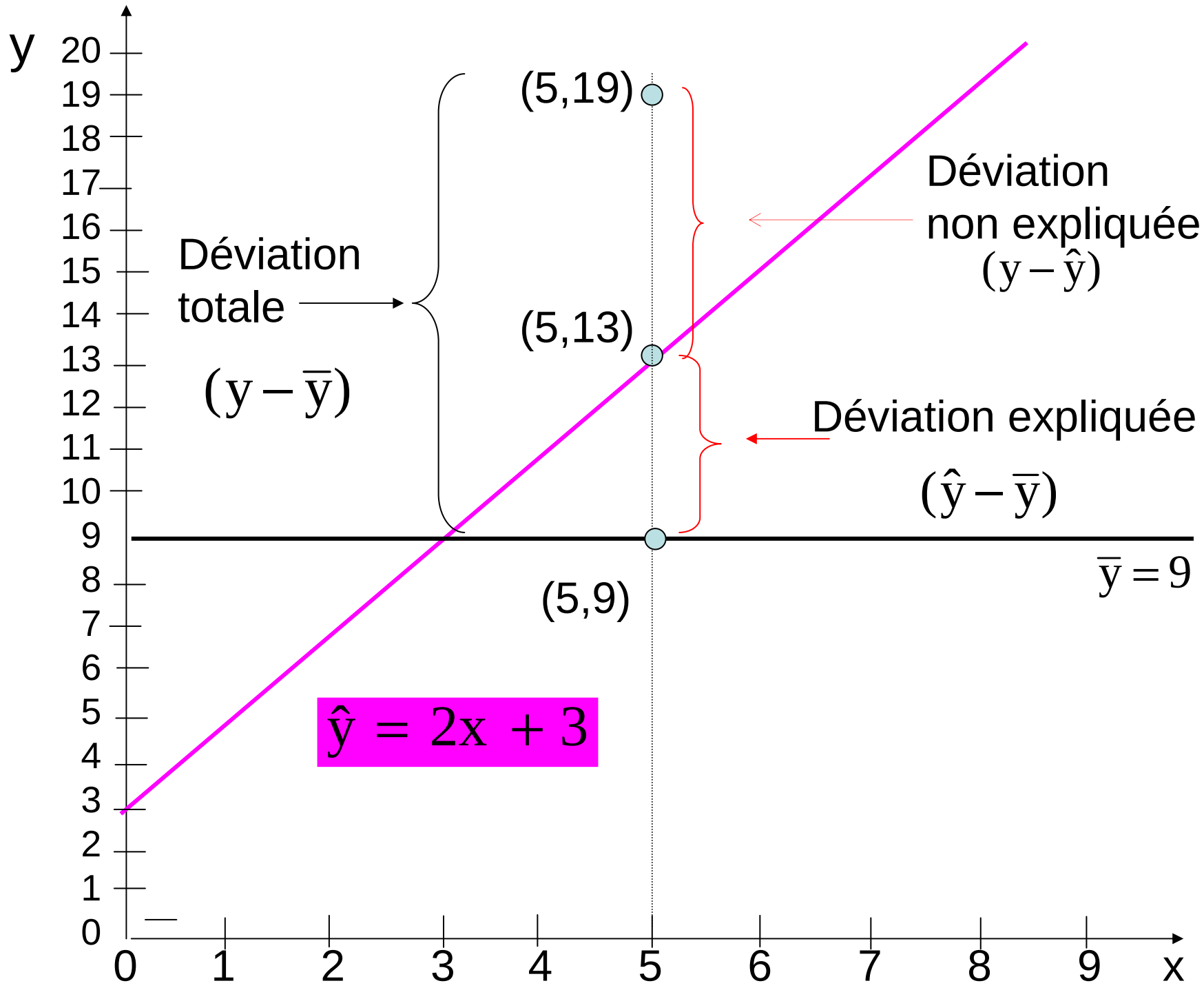
$$r = a \frac{s_x}{s_y}$$

s_y est l'écart type estimé de Y, s_x est l'écart type estimé de X

Le coefficient de détermination évalue la quantité d'information

$$r^2 = \frac{\text{variance expliquée}}{\text{variance totale}}$$

apportée par la droite de régression c'est-à-dire la part de variabilité totale de y expliquée par la droite de régression


y
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
x
(5,19)
(5,13)
(5,9)
Déviation totale
$(y - \bar{y})$
Déviation non expliquée
$(y - \hat{y})$
Déviation expliquée
$(\hat{y} - \bar{y})$
$\bar{y} = 9$
$\hat{y} = 2x + 3$

Point expérimental (5,19)

$\bar{y} = 9$ Sur la verticale x = 5 on a

$$(y - \bar{y}) = (y - \hat{y}) + (\hat{y} - \bar{y})$$

19 – 9 = (19 -13) + (13 – 9)

Déviation totale = déviation résiduelle non expliquée par la régression linéaire + déviation expliquée par droite de régression

Déviation ou écart

En élevant les écarts au carré et en sommant sur tous les y_i expérimentaux et tous les $\hat{y}_i$ on montre que l'on a toujours :

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

SCE_T = SCE_R + SCE_E

$$\sum(y-\bar{y})^2=\sum(y-\hat{y})^2+\sum(\hat{y}-\bar{y})^2$$

SCE_T = SCE_R + SCE_E

T pour « total »
R pour « résiduel »
E pour « expliqué par la régression linéaire »

$$\frac{SCE_T}{SCE_T}=1=\frac{SCE_R}{SCE_T}+\frac{SCE_E}{SCE_T}$$

On montre que $\frac{SCE_E}{SCE_T}=r^2=\frac{\text{variance expliquée}}{\text{variance totale}}$

r^2 est appelé le coefficient de détermination

Exercice : on étudie la toxicité d'un médicament chez la souris en mesurant la survie des souris en fonction de la dose administrée. Chaque dose est donnée à cinq souris, on étudie 5 doses. On sacrifie toutes les souris à la 21ème semaine

Dose mg/kg	1	2	3	4	5
Survie (jours)	19	17	11	9	6
	19	17	11	9	7
	20	18	12	12	7
	20	18	13	13	8
	21	20	14	13	9

a. établir l'équation de la droite de régression qui exprime la survie en fonction de la dose administrée

b. Calculer le coefficient de corrélation linéaire r entre la dose et la survie

c. Ce coefficient r a - t'il un sens ?

d. l'équation $\hat{y} = ax + b$ est elle pertinente ?

Les moyennes observées des variables X et Y sont

$$\bar{x} = 3{,}00 \qquad \bar{y} = 13{,}72$$

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

$$a = \frac{-158}{50} = -3{,}16 \qquad \Rightarrow b = \bar{y} - a\bar{x} = 13{,}72 + 3{,}16 * 3 = 23{,}2$$

vérification

L'équation de la droite de régression est : $\hat{y} = -3{,}16\,x + 23{,}2$

b. Calculer le coefficient de corrélation linéaire r entre la dose et la survie

$$a = r\frac{s_y}{s_x} \quad \Rightarrow \quad r = a\frac{s_x}{s_y} \qquad \text{ou}$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \times \sqrt{\sum (y_i - \bar{y})^2}} = \frac{\sum x_i y_i - n\,\bar{x}\,\bar{y}}{\sqrt{\left(\sum (x_i^2) - n\bar{x}^2\right) \times \left(\sum (y_i^2) - n\bar{y}^2\right)}}$$

S_x=1,4433 S_y=4,8176

r = -3,16×1,4433 / 4,8176 = - 0,9467 : forte corrélation

$r^2 = (- 0,9467)^2 = 0,896$: le modèle linéaire permet d'expliquer 89,6 % de la variabilité totale de le survie des souris : ce qui est important

c. Ce coefficient *r* a - t'il un sens ?

La statistique de test est : $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ avec n = 25

$t = -14{,}095$

$H_0: \rho = 0$ $H_1 : \rho \neq 0$ ddl = 23

Valeurs critiques pour $\alpha = 0{,}05$: $t_\alpha = \pm 2{,}069$

t est dans la zone critique $\leftrightarrow$ $|t| > t_\alpha$ donc H_0 rejetée

La valeur r calculée a un sens : elle est différente de 0 de façon significative.

Il existe une corrélation linéaire significative entre les variables X et Y dans la population

3. Différences entre valeurs prédites et observées d’une variable

- test de validité de la pente a

- Intervalle de confiance de la pente

Test de la validité de a

L'équation $\hat{y} = ax + b$ est elle pertinente ?

Le modèle linéaire Y = AX + B est-il adapté pour décrire Y en fonction de X dans la population ?

On teste $H_0 : A = 0$ contre $H_1 : A \neq 0$

$a \Rightarrow t = \frac{a - A}{S(a)}$ qui suit une loi de Student à n -2 d.d.l.

avec $S^2(a)$ = variance (a) = estimation de var(A) dans la population

On montre que $S^2(a) = \frac{\frac{s_y^2}{s_x^2} - a^2}{n - 2}$

suite de l'exercice sur la survie des souris

$s^2(a) = (4{,}8176^2 / 1{,}4433^2 - 3{,}16^2) / 23 = 1{,}1560 / 23 = 0{,}0502$

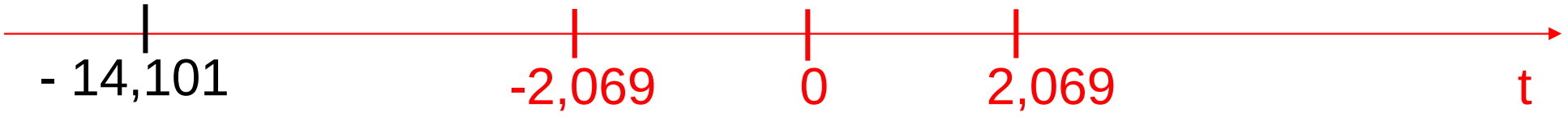
$s(a) = 0{,}2241$

La statistique de test est $t_{calculé} = \dfrac{a}{s(a)} = -14{,}101$

Au seuil de rejet de 5% et pour ddl = 23 les valeurs critiques de la table de Student sont $t_\alpha = \pm 2{,}069$

t est dans la zone critique : on rejette H_0

on peut conclure à l'existence d'une pente significativement différente de zéro



Intervalle de confiance de la pente

Le coefficient ***A*** de la droite de régression dans la population est estimé par ***a***, à partir des couples de valeurs (x,y) observés sur un échantillon

avec $a = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ ou $\dfrac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum (x_i^2) - n \bar{x}^2}$

On sait que la variance de a : $S^2(a) = \dfrac{\frac{s_y^2}{s_x^2} - a^2}{n - 2}$

L'intervalle de confiance de la pente de la droite de régression est donné par : $a \pm t_{n-2,\alpha} \sqrt{s^2(a)}$

cet intervalle est centré sur a

Suite de l'exemple précédent :

La pente A de la droite de régression exprimant le temps de survie en fonction de la dose du médicament dans la population est estimée par le calcul à partir de l'échantillon est a = - 3,16

Sa variance est $s^2(a) = (4{,}8176^2 / 1{,}4433^2 - 3{,}16^2) / 23 = 0{,}0502$

$s(a) = 0{,}2241$

Au seuil de rejet de 5% et pour ddl = n – 2 = 23
la valeur critique de la table de Student est $t_{n-2,\alpha} = 2{,}069$

$a - t_\alpha \times s(a) < A < a + t_\alpha \times s(a) \Leftrightarrow t_\alpha \times s(a) = 0{,}46$ = rayon de I.C.

-3,16 - 0,46 < A < -3,16 + 0,46

- 3,62 < A < - 2,70

APPORTER

SA CALCULATRICE

en TD