

CORRIGÉ

TD 9 : Régression linéaire

Exercice 1. : On reprend l'exemple des 5 spécimens fossiles d'un animal disparu pour lesquels on possède les mesures de la longueur en cm de leur humérus x et de leur fémur y .

1. Compléter le tableau suivant et en déduire les valeurs des variances et covariance :

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	44	40	1936	1600	1760
2	65	60	4225	3600	3900
3	71	59	5041	3481	4189
4	75	65	5625	4225	4875
5	87	77	7569	5929	6699
μ	68,4	60,2	4879,2	3767	4284,6

Calculs effectués pour variances et covariance :

$$\text{Var}(x) = \mu(x^2) - \mu(x)^2 = 4879,2 - 68,4^2 = 200,64$$

$$\text{Var}(y) = \mu(y^2) - \mu(y)^2 = 3767 - 60,2^2 = 142,96$$

$$\begin{aligned}\text{Cov}(x, y) &= \mu(xy) - \mu(x)\mu(y) = 4284,6 - 68,4 \cdot 60,2 \\ \text{Cov}(x, y) &= 166,92\end{aligned}$$

$\text{Var}(x) = 200,64$	$\text{Var}(y) = 142,96$	$\text{Cov}(x, y) = 166,92$
--------------------------	--------------------------	-----------------------------

2. Déterminer, par la méthode des moindres carrés ordinaires, l'équation de la droite de régression de y en x .

$$\hat{a} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} = \frac{166,92}{200,64} \simeq 0,83194 \quad \text{et} \quad \hat{b} = \mu(y) - \hat{a}\mu(x) \simeq 60,2 - 0,83194 \times 68,4 \simeq 3,2955$$

$y = 0,83194x + 3,2955$

3. Passe-t-elle par le centre de gravité G ? Justifier par un calcul.

Le centre de gravité G a pour coordonnées $(\mu(x), \mu(y)) = (68,4, 60,2)$.

On vérifie que $60,2 \simeq 0,83194 \times 68,4 + 3,2955$.

Plus généralement, plus abstraitement et plus exactement, vu la définition de $\hat{b}$ on a

$$\hat{a}\mu(x) + \hat{b} = \hat{a}\mu(x) + \mu(y) - \hat{a}\mu(x) = \mu(y)$$

donc la droite de régression passe par le centre de gravité G .

4. Calculer le coefficient de corrélation linéaire. Commenter.

$$\rho(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} = \frac{166,92}{\sqrt{200,64 \times 142,96}} \simeq 0,9856$$

$\rho(x, y) = 0,9856$

$\rho(x, y)$ est très proche de 1, l'approximation du nuage de points par la droite de régression est donc très bonne.

5. Calculer la longueur, selon ce modèle, du fémur d'un spécimen dont l'humérus mesurerait 55 cm.

D'après ce modèle la longueur du fémur serait :

$$0,83194 \times 55 + 3,2955 \simeq 49,05 \text{ cm.}$$

Exercice 2. : Pour étudier les problèmes de malnutrition dans un pays pauvre, on a calculé le poids moyen par âge d'un échantillon de 2400 enfants répartis uniformément en 12 classes d'âge. On a obtenu le tableau suivant :

x_i =classe d'âge	1	2	3	4	5	6	7	8	9	10	11	12
y_i =poids moyen	3,5	3,5	3,3	4,4	4,4	4,2	5,1	5,3	5,5	5,5	6,2	5,7
y_i^2	12,25	12,25	10,89	19,36	19,36	17,64	26,01	28,09	30,25	30,25	38,44	32,49
$x_i y_i$	3,5	7,0	9,9	17,6	22,0	25,2	35,7	42,4	49,5	55,0	68,2	68,4

1. Compléter ce tableau. Déterminer la droite des moindres carrés.

On calcule $\mu(x) = 6,5$ et $\mu(x^2) \simeq 54,167$ puis $\mu(y) \simeq 4,716667$ et $\mu(y^2) \simeq 23,1067$ enfin $\mu(xy) = 33,7$. D'où $\text{Var}(x) = \mu(x^2) - \mu(x)^2 \simeq 11,917$ et $\text{Cov}(x, y) = \mu(xy) - \mu(x)\mu(y) \simeq 3,0417$. Donc finalement :

$$\hat{a} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \simeq \frac{3,0417}{11,917} \simeq 0,25524 \quad \text{et} \quad \hat{b} = \mu(y) - \hat{a}\mu(x) \simeq 4,716667 - 0,25524 \times 6,5 \simeq 3,0576$$

2. Calculer le coefficient de corrélation linéaire. Commenter.

On calcule $\text{Var}(y) \simeq \mu(y^2) - \mu(y)^2 \simeq 0,8597$ d'où le coefficient de corrélation linéaire

$$\rho(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} \simeq \frac{3,0417}{\sqrt{11,917 \times 0,8597}} \simeq 0,9503$$

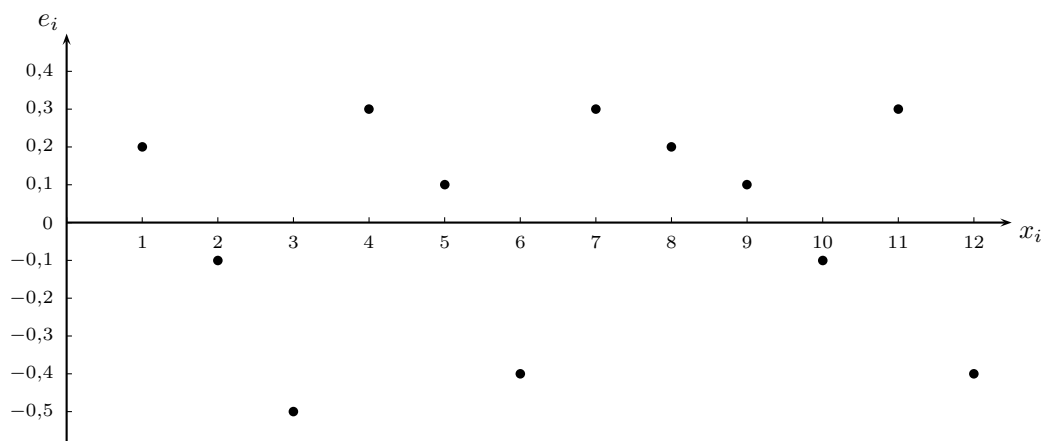
ce qui est proche de 1 : le nuage est proche de la droite.

3. Compléter le tableau

x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	3,5	3,5	3,3	4,4	4,4	4,2	5,1	5,3	5,5	5,5	6,2	5,7
$\hat{y}_i$	3,3	3,6	3,8	4,1	4,3	4,6	4,8	5,1	5,4	5,6	5,9	6,1
e_i	0,2	-0,1	-0,5	0,3	0,1	-0,4	0,3	0,2	0,1	-0,1	0,3	-0,4

où $\hat{y}_i$ est la valeur prévue par le modèle par classe d'âge et $e_i = y_i - \hat{y}_i$ le résidu.

4. Tracer les résidus e_i en fonction des classes d'âge et commenter.



Les résidus sont distribués au hasard ce qui confirme que la droite de régression représente bien l'évolution du poids en fonction de la classe d'âge.