NOM : CORRIGÉ

PRENOM :

Date : 28 Novembre au 2 Décembre 2011

Groupe :

Mathématiques pour la Biologie : Feuille-réponses du 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 de leur fémur et de leur humérus.

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_iy_i
1	38	41	1444	1681	1558
2	56	61	3136	3721	3416
3	59	70	3481	4900	4130
4	64	72	4096	5184	4608
5	75	84	5625	7056	6300
Moyenne :	58,4	65,6	3556,4	4508,4	4002,4

$Var(x) = 3556,4 - (58,4)^2 = \frac{1}{n}\sum x_i^2 - \mu_x^2$

$Var(y) = 4508,4 - (65,6)^2 = \frac{1}{n}\sum y_i^2 - \mu_y^2$

$Cov(x,y) = 4002,4 - 58,4 \times 65,6$
$= \frac{1}{n}\sum x_i y_i - \mu_x \mu_y$

$Var(x) = 145,84$ $\quad Var(y) = 205,04$ $\quad cov(x,y) = 171,36$

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

$$\hat{a} = \frac{Cov(x,y)}{Var(x)} = \frac{171,36}{145,84} \simeq 1,175$$

$$\hat{b} = \mu(y) - \hat{a}\mu(x) = 65,6 - 1,175 \times 58,4 \simeq -3,019$$

$y = 1,175x - 3,019$

3. Cette droite passe-t-elle par G ? Expliquez pourquoi.

Le centre de gravité G a pour coordonnées $(\mu(x) ; \mu(y))$. On a

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

Donc G est sur la droite de régression.

4. Calculer le coefficient de corrélation linéaire. Commenter.

$$\rho(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} = \frac{171,36}{\sqrt{145,84 \times 205,04}} \simeq 0,991$$

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

$\rho(x,y) = 0,991$

5. Calculer la longueur prédite par ce modèle pour l'humérus d'un spécimen dont le fémur mesurerait $50cm$.

longueur de l'humérus $= 1,175 \times 50 \text{ cm} - 3,019$
$\simeq 55,73 \text{ cm}$

Exercice 2. :

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

classe d'age	1	2	3	4	5	6	7	8	9	10	11	12
poids moyen	4,3	5,1	5,7	6,3	6,8	7,1	7,2	7,2	7,2	7,2	7,5	7,8
$\hat{y}_i$	5,147	5,414	5,681	5,948	6,215	6,482	6,749	7,016	7,283	7,550	7,817	8,084
ε_i	0,847	0,314	-0,019	-0,352	-0,585	-0,618	-0,451	-0,184	0,083	0,350	0,317	0,284

1. Un statisticien pressé a fait calculer par sa machine la droite des moindres carrés pour ces données et a trouvé la relation poids = 4,88 + 0,267 age. S'est-il trompé ? Expliquer.

On a $\mu(x) = 6,5$ $\mu(y) \simeq 6,617$ $\mu(xy) \simeq 46,192$ $\mu(x^2) \simeq 54,167$

D'où $Var(x) = \mu(x^2) - \mu(x)^2 \simeq 11,917$

et $Cov(x,y) = \mu(xy) - \mu(x)\mu(y) \simeq 3,183$

et donc $\hat{a} = Cov(x,y) / Var(x) \simeq 0,267$

et $\hat{b} = \mu(y) - \hat{a}\,\mu(x) \simeq 4,880$

Les calculs sont justes

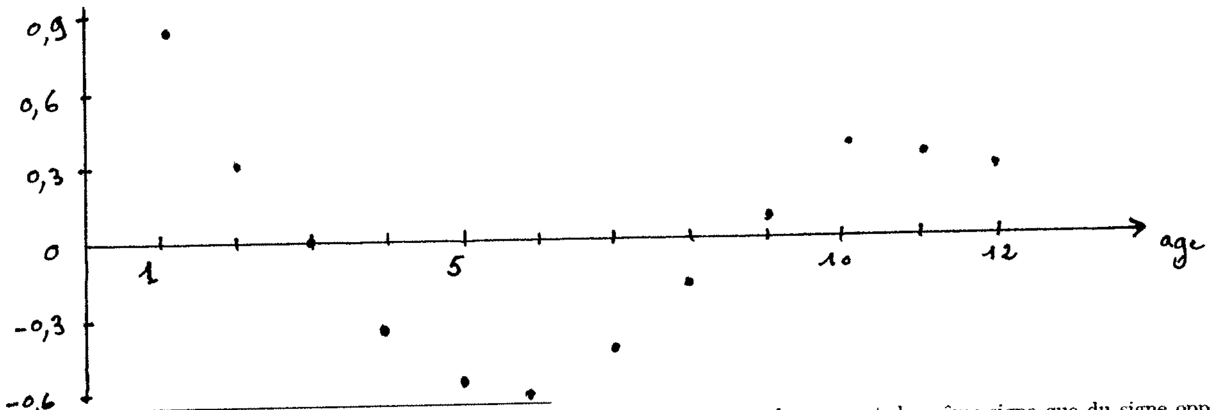
2. A votre avis, quelle est la pertinence de son modèle ?

Le coefficient de corrélation linéaire $\rho(x,y) = \dfrac{Cov(x,y)}{\sqrt{Var(x)\,Var(y)}} \simeq 0,907$

ce qui est assez bon mais on peut néanmoins douter de la pertinence d'un modèle linéaire en examinant les résidus qui sont plusieurs fois de l'ordre de 10% du poids mesuré (cf 4.)

3. Compléter le tableau en calculant les poids prédits par le modèle $\hat{y}_i$.

4. Calculer puis tracer les résidus en fonction de la classe d'age[1].



[1]Vous pouvez constater que deux résidus successifs sont beaucoup plus souvent du même signe que du signe opposé. Ceci n'est pas compatible avec le fait qu'ils soient supposés indépendants. On dit que les résidus sont *autocorrélés*. C'est une raison de rejeter le modèle et d'en rechercher un meilleur.