NOM :
PRENOM : Corrigé

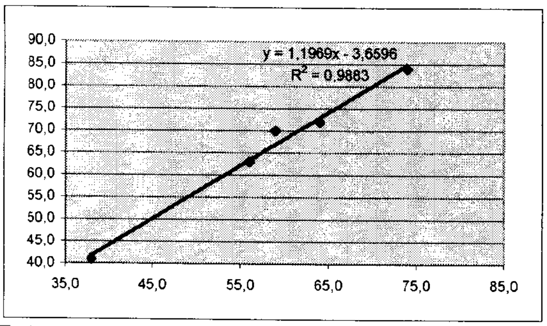
Date :
Groupe :

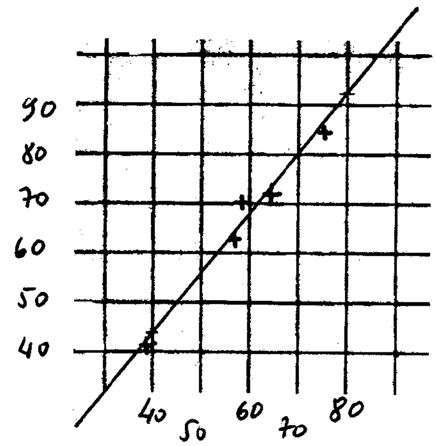
Mathématiques pour la Biologie : Feuille-réponses du TD 9
Régression linéaire

Exercice 1. : On possède 6 spécimens fossiles d'un animal disparu et ces spécimens sont de tailles différentes. On estime que si ces animaux appartiennent à la même espèce il doit exister une relation linéaire entre la longueur de deux de leurs os, le fémur et l'humérus. Voici les données de ces longueurs en cm pour les 5 spécimens possédant ces deux os intacts :

X=fémur	38	56	59	64	74
Y=humérus	41	63	70	72	84

1. Tracer le nuage de point correspondant à ces données. Pensez-vous que les 5 spécimens peuvent appartenir à la même espèce et ne différer en taille que parce que certains sont plus jeunes que d'autres ?





2. Calculer à l'aide des fonctions statistiques de votre calculatrice les moyennes, variances et covariance de ces deux séries. Indiquer brillèvement quelles touches, quelles fonctions, ...vous avez utilisé et le modèle de votre calculatrice.

$\mu(X) =$ 58.2 $\quad \mu(Y) =$ 66.0 $\quad \mathrm{Var}(X) =$ 139.4 $\quad \mathrm{Var}(Y) =$ 202.0 $\quad \mathrm{Cov}(X, Y) =$ 166.8

Sous Excel μ se calcule au moyen de f_x-Statistiques-MOYENNE et Var(X) au moyen de f_x-Statistiques-VAR.P, et Cov(X,Y) au moyen de f_x-Statistiques-COVARIANCE

A noter que f_x-Statistiques-VAR donne la valeur de $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu(x))^2$ qui est un "estimateur sans biais" à partir d'un "échantillon" de la Population

3. En déduire l'équation $y = \hat{a}x + \hat{b}$ de la droite des moindres carrés. Contrôler vos calculs en ajustant son graphe au nuage de points.

On a $\hat{a} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \frac{166.8}{139.4} = 1.1969..$ et $\hat{b} = \mu(Y) - \hat{a}\,\mu(X) = -3.66..$

Pour tracer la droite il suffit de calculer deux de ses points, par exemple pour X=40 et X=80. On trouve $f(40) = \hat{a}\cdot 40 + \hat{b} = 44.2$ et $f(80) = 92.1$

4. Calculer le coefficient de corrélation linéaire ρ et le R^2. Qu'en concluez-vous ?

$\rho = \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = 0{,}994...$ et $R^2 = \rho^2 = 0{,}988..$

R^2 est très voisin de 1 et on considère dans ce cas que la regression linéaire est très satisfaisante ce qui revient à conclure que les 5 spécimens correspondent à la même espèce.(*)

(*) On note toutefois qu'un seul point est au-dessus de la droite de regression : le spécimen correspondant à ce point mériterait à être réexaminé avec soin !

5. Reprenez les 2 questions précédentes en effectuant directement la régression linéaire au moyen des fonctions statistiques de votre calculatrice si celle-ci le permet. Vérifier que vos résultats sont identiques.

Voir la figure imprimée au moyen d'Excel, obtenue au moyen de Assistant Graphique - Nuage de points - Nuage de points, puis "click-droit" sur un point - Ajouter une courbe de tendance... - Option cocher les cases "Afficher l'équation..." et "Afficher le coefficient..(R^2).."

Exercice 2. :

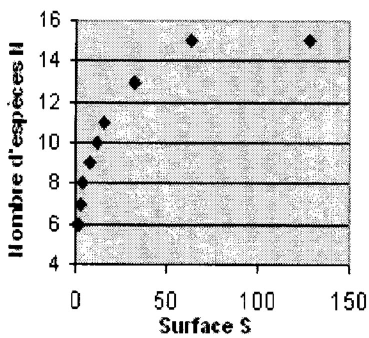
L'une des rares lois que l'on a pu mettre en évidence en Ecologie est la relation existant entre le nombre N d'espèces présentes dans un habitat donné (bien délimité) et la surface S de cet habitat. On considère généralement que cette relation est de la forme

$$N = AS^B \tag{1}$$

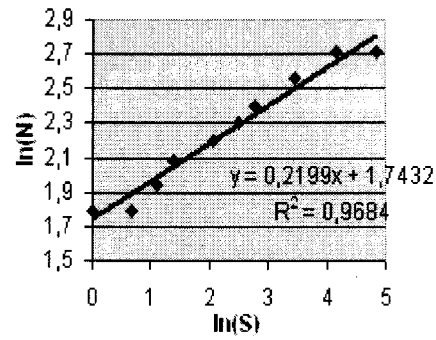
où A et B sont deux constantes. Afin de vérifier cette relation pour les plantes présentes dans une prairie (pissenlit, paquerettes, orties, boutons d'or, ...), on a effectué les mesures indiquées dans le premier tableau ci-dessous. On a représenté sur la première figure ci-dessous les valeurs de N en fonction de celles de S et sur la deuxième les valeurs de $\tilde{N} = \ln(N)$ en fonction de celles de $\tilde{S} = \ln(S)$. On voit que la regression linéaire de $\tilde{N}$ sur $\tilde{S}$ a donné :

$$\tilde{N} = 0,2199\tilde{S} + 1,7432 \text{ avec } R^2 = 0,9684 \tag{2}$$

S	N
1	6
2	6
3	7
4	8
8	9
12	10
16	11
32	13
64	15
128	15



ln(S)	ln(N)
0,00	1,79
0,69	1,79
1,10	1,95
1,39	2,08
2,08	2,20
2,48	2,30
2,77	2,40
3,47	2,56
4,16	2,71
4,85	2,71



1. Pourquoi n'a-t-on pas effectué directement une régression linéaire de N sur S ? Expliquez l'intéret de cette transformation des données.

La fonction $f(S) = AS^B$ n'est pas de la forme $aS+b$; en revanche $\tilde{N} = ln(N) = ln(AS^B) = ln(A) + B\,ln(S) = b + a\tilde{S}$ est bien une fonction linéaire-affine de $\tilde{S} = ln(S)$, avec $b = ln(A)$ et $a = B$

2. Que représente R^2 et que peut-on déduire de sa valeur ?

R^2 est le carré de la corrélation linéaire entre $\tilde{N}$ et $\tilde{S}$. On a $R^2 = \frac{DR}{DT}$ où DT est la dispersion totale de $\tilde{N}$ et DR est la dispersion expliquée par la regression. On voit qu'avec $R^2 = 0.9684.. \approx 97\%$ la dispersion totale est très bien expliquée par la regression linéaire.

3. A partir de la régression linéaire (2), calculer les constantes A et B de la relation (1).

On lit sur l'equation de la droite de regression imprimée sur le graphe que $a = 0.2199$ et $b = 1.7432$. Or $B = a = 0.2199$ et $A = e^b$ (puisque $b = ln(A)$) et donc $A = exp(1.7432) = 5.7156..$ et $B = 0.2199$

4. Quelle valeur $\tilde{N}$ ce modèle linéaire prédit-il pour $\tilde{S} = ln(128)$? En comparant avec la valeur de $\tilde{S}$ observée, calculer le résidu ε en ce point.

La valeur prédite $\tilde{N}$ est $a\tilde{S} + b$, et donc $\tilde{N} = 0.2199\,ln(128) + 1.7432 = 2.810..$

Le résidu ε en ce point vaut $\varepsilon = 2.71.. - 2.81.. = -0.10..$

5. Quelle valeur $\tilde{N}$ ce modèle linéaire prédit-il pour $\tilde{S} = ln(100)$? En déduire le nombre d'espèces pouvant coexister dans un habitat de surface $S = 100$, selon ce modèle.

Pour $\tilde{S} = ln(100)$ le modèle prédit $\hat{N} = a\tilde{S} + b = 0.2199 \cdot 4.60.. + 1.74..$ donc $\hat{N} = 2.75..$ et donc $N = exp(\hat{N}) = e^{2.75..} = 15.73..$ c'est-à-dire 15 à 16 espèces différentes sur un terrain de surface 100.