CORRIGÉ

NOM : Date : 5 au 9 décembre 2011

PRENOM : Groupe :

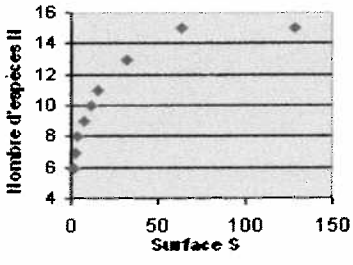
Mathématiques pour la Biologie : Feuille-réponses du TD 10
Régression linéaire : compléments

Exercice 1. : 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

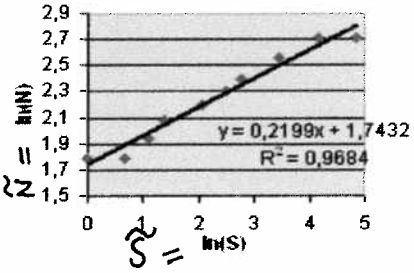
$$N = AS^B \tag{1}$$

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 régression linéaire de $\tilde{N}$ sur $\tilde{S}$ a donné l'équation $\tilde{N} = 0,2199\tilde{S} + 1,7432$, avec $R^2 = 0,9684$.

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 et a-t-on préféré transformer N en $\tilde{N}$ et S en $\tilde{S}$?

Pratiquement N en fonction de S (graphe de gauche) n'a pas l'air d'une droite. Si $N = AS^B$ ce n'est pas une droite si $B \neq 1$

Par contre $\tilde{N} = \ln(N) = \ln(AS^B) = B\ln(S) + \ln(A) = B\tilde{S} + \ln(A)$ est une droite de pente B et d'ordonnée à l'origine $\ln(A)$.

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

D'après l'équation trouvée $\boxed{B = 0,2199}$ et $\ln(A) = 1,7432$ donc $\boxed{A = \exp(1,7432) \simeq 5,7156}$

3. Quelle valeur $\tilde{N}$ ce modèle linéaire prédit-il pour $\tilde{S} = ln(64)$? En comparant avec la valeur de $\tilde{N}$ observée, calculer le résidu ε en ce point.

Pour $\tilde{S} = \ln(64)$ on prévoit $\tilde{N} = 0,2199 \times \ln(64) + 1,7432 \simeq \boxed{2,6577}$

Or on observe $\tilde{N} = 2,71$ on a donc un résidu $\varepsilon = 2,71 - 2,6577$

$\boxed{\varepsilon \simeq 0,052}$

4. 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 $\tilde{S} = \ln(100)$ on prévoit $\tilde{N} = 0,2199 \times \ln(100) + 1,7432 \simeq 2,756$

Donc le nombre d'espèces pouvant cohabiter sur une surface de 100 est $\boxed{N \simeq e^{2,756} \simeq 15,73 \simeq 16}$

Exercice 2. : Des écologistes se sont intéressés à la répartition de la végétation dans un site aride du sud de la France, la plaine de la Crau (13). Ils ont effectué 9 prélèvements de sol (S1, ..., S9) pour lesquel ils ont retenu, après analyse, 6 mesures (pH, C/N, Ca, Mg, K, P) et dans le même temps ils ont évalué dans chaque cas le pourcentage de recouvrement au sol par la végétation (%V). Ces données brutes sont regroupées dans le tableau suivant :

	pH	C/N	Ca	Mg	K	P	$\%V$
$S1$	5,5	30,75	0,55	0,01	0,42	0,01	15
$S2$	7	19,29	1,02	0,07	0,43	0,02	42
$S3$	6,8	31,47	1,02	0,05	0,45	0,01	26
$S4$	7,3	2,93	1,82	0,09	0,44	0,02	50
$S5$	5,62	32,45	0,42	0	0,41	0,01	25
$S6$	6,6	20,05	0,75	0,01	0,44	0,01	30
$S7$	7	3,35	1,33	0,07	0,46	0,02	50
$S8$	5,8	33,18	0,48	0,02	0,43	0,01	18
$S9$	7	5,32	1,35	0,07	0,48	0,005	38

On veut étudier de quelle façon les variables pH, C/N et P influent sur le pourcentage de recouvrement au sol par la végétation et on va pour cela chercher à expliquer par une régression linéaire la quantité %V en fonction de chacune de ces 3 variables explicatives.

1. Exprimer, par une régression linéaire, %V en fonction de pH et donner l'équation de la droite des moindres carrés ainsi que la valeur du coefficient de détermination R^2 (on pourra utiliser le fait que la variance de pH vaut 0,417, que la variance de %V vaut 150,44 et que leur covariance vaut 6,92).

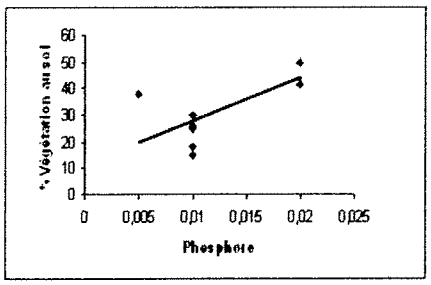
On a les moyennes $\mu(pH) \simeq 6{,}513$ et $\mu(\%V) \simeq 32{,}667$

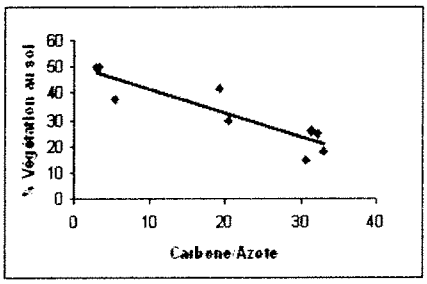
D'où $\hat{a} = Cov(pH, \%V)/Var(pH) \simeq 16{,}60$ et $\hat{b} = \mu(\%V) - \hat{a}\,\mu(pH) \simeq -75{,}43$

Donc l'équation $\boxed{\%V = 16{,}60\,pH - 75{,}43}$

et le coefficient $\boxed{R^2 = Cov(pH, \%V)^2/(Var(pH) \times Var(\%V)) \simeq 0{,}76}$

2. On a aussi cherché à expliquer de la même façon %V à l'aide de deux autres variables, C/N et P. On a obtenu les dessins des nuages et des droites de régression suivants :





Mais on a mélangé les résultats ; les équations $y = 1617{,}4x + 12$ et $y = -0{,}904x + 50{,}691$ et les coefficients $R^2 = 0{,}8178$ et $R^2 = 0{,}4937$. Sans faire de nouveaux calculs, indiquer quelle équation et quel coefficient correspond à quel dessin en justifiant vos réponses.

Dessin de gauche La droite est de pente >0 et son ordonnée à l'origine un peu au dessus de 10. C'est donc l'équation $\boxed{y = 1617{,}4\,x + 12}$

La dispersion autour de la droite est importante donc $\boxed{R^2 = 0{,}4937}$ petit

Dessin de droite La droite est de pente <0 et son ordonnée à l'origine proche de 50. C'est donc l'équation $\boxed{y = -0{,}904\,x + 50{,}691}$

La dispersion autour de la droite est plus faible donc R^2 est plus proche de 1 : $\boxed{R^2 = 0{,}8178}$

3. Parmi ces trois régressions, y-en-a-t-il à votre avis qui soient acceptables ? Justifier votre réponse.

Les régressions du couvert végétal en fonction du pH ou du rapport carbone/azote donnent des coefficients R^2 acceptables, 0,76 et 0,82 respectivement. Par contre la régression en fonction du phosphore donne un $R^2 = 0,49$ ce qui est faible.

4. Finalement, compte tenu de ces résultats, peut-on conclure que la chimie du sol peut expliquer le pourcentage de végétation ? On pourra se servir de la part de dispersion de $\%V$ expliquée par les régressions.

Les régressions de %V en fonction du pH ou du rapport C/N expliquent 76% et 82% de la dispersion de %V : ce sont des facteurs chimiques importants. Par contre la régression de %V en fonction du phosphore n'explique que 49% de la dispersion de %V.

Exercice 3. : Des biologistes ont introduit au printemps 1937 sur l'ile de la Protection (Etat de Washington, USA), 8 individus d'une population de faisans qu'ils ont ensuite recencé chaque printemps jusqu'à ce qu'en 1943 l'armée débarque sur l'ile, décimant cette population de faisans. Leurs observations ont été les suivantes :

Année	1937	1938	1939	1940	1941	1942
Effectifs	8	30	81	282	705	1325

On désigne par x le nombre d'années écoulées depuis 1937, par $N(x)$ la taille de la population de faisans à la date x et par $y(x) = \ln N(x)$ le logarithme de $N(x)$.

1. Calculer la droite de régression linéaire de y par rapport à x (on pourra utiliser que $Var(x) \simeq 2,92$, $Var(y) \simeq 3,17$ et $Cov(x,y) \simeq 3,02$).

On a les moyennes $\mu(x) = 2,5$ et $\mu(y) \simeq 4,88$

D'où $\hat{a} = Cov(x,y) / Var(x) \simeq 1,036$ et $\hat{b} = \mu(y) - \hat{a}\,\mu(x) \simeq 2,287$

et l'équation $\boxed{\hat{y} = 1,036\,x + 2,287}$

2. Calculer la valeur prédite $\hat{y}(6)$ par ce modèle pour $x = 6$ et en déduire l'effectif prédit pour l'année 1943.

Le modèle prévoit $\hat{y}(6) \simeq 1,036 \times 6 + 2,287 \simeq 8,504$

et donc $\boxed{\hat{N}(6) = e^{\hat{y}(6)} \simeq 4935}$ faisans en 1943

3. Cette régression linéaire est-elle valide selon vous ? Expliquer.

Le calcul du $R^2 = Cov(x,y)^2 / (Var(x) \times Var(y))$ donne $R^2 \simeq 0,989$ ce qui signifie une très bonne qualité de la régression linéaire sur l'intervalle de temps étudié.

4. A quelle expression pour $N(x)$ cette régression de $y(x)$ conduit-elle ? Comment s'appelle ce type de modèle ? Quel est son principal défaut ?

On aura $\hat{N}(x) = \exp(\hat{y}(x)) \simeq e^{1,036x + 2,287} = e^{2,287} \times e^{1,036x}$

donc $\boxed{\hat{N}(x) \simeq 9,84\, e^{1,036x}}$ ce qui correspond à un modèle malthusien qui n'est valable que pour décrire le début de la croissance d'une population partant de petits effectifs mais pas à long terme (et d'autant moins qu'on fait intervenir l'armée !)

5. On opte finalement pour un modèle logistique pour $N(x)$ avec pour constante $r = 1,3$ et $K = 2000$. On a donc

$$\frac{dN(x)}{dx} = (1,3)N(x)\left(1 - \frac{N(x)}{2000}\right).$$

On rappelle que, si $N(0)$ est l'effectif en $x = 0$, la solution exacte de cette équation est donnée par la formule $N(x) = \frac{N(0)Ke^{rx}}{K+N(0)(e^{rx}-1)}$.

Calculer $N(1)$ selon ce modèle, puis calculer une valeur approchée de $N(1)$ par la méthode d'Euler en utilisant un pas $h = 1$, arrondies à l'entier le plus proche. Comparer avec l'effectif observé après un an et commenter.

$N(1) = 8 \times 2000 \times e^{1,3\times 1} / (2000 + 8 \times (e^{1,3\times 1} - 1))$

$\boxed{N(1) \simeq 29}$

$N(1)$ approché $= N(0) + h\,N'(0) = 8 + 1 \times 1,3 \times 8 \times (1 - 8/2000)$

$\boxed{N(1) \text{ approché} \simeq 18}$

La valeur observée est $N(1) = 30$, la valeur donnée par le modèle est 29 ce qui est proche, la valeur approchée par la méthode d'Euler est 18 ce qui est éloigné mais le pas choisi $h = 1$ est beaucoup trop grand (voir la question suivante).

6. On a préféré finalement calculer les valeurs de la solution exacte du modèle (toujours pour $N(0) = 8$) et celles de la solution approchée, mais en prenant cette fois un pas $h = 0,1$. Le tableau suivant indique les premières valeurs, exactes et approchées, arrondies à l'entier le plus proche. Compléter les trois valeurs manquantes en expliquant quels calculs vous faites pour cela. A noter que ce qui est étudié ici est un modèle mathématique : il n'est pas supposé, bien entendu que les faisans font, brusquement, 10 pontes reproductives par an !

x	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
solution exacte	8	9	10	12	13	15	17	20	22	26	29
solution approchée	8	9	10	12	13	15	17	19	21	24	27

$N(0,8) = 8 \times 2000 \times e^{1,3\times 0,8} / (2000 + 8 \times (e^{1,3\times 0,8} - 1))$

$\boxed{N(0,8) \simeq 22}$

$N(0,1)$ approché $= N(0) + 0,1 \times N'(0) = 8 + 0,1 \times 1,3 \times 8 \times (1 - 8/2000)$

$\boxed{N(0,1) \text{ approché} \simeq 9}$

$N(0,5)$ approché $= N(0,4)$ approché $+ 0,1 \times N'(0,4) = 13 + 0,1 \times 1,3 \times 13 \times (1 - 13/2000)$

$\boxed{N(0,5) \text{ approché} \simeq 15}$

Avec un pas de 0,1 les valeurs approchées sont très proches des valeurs exactes, elles s'en éloignent un peu au bout d'un certain temps. C'est la dérive de la méthode d'Euler vers l'extérieur des rivages.