NOM : Corrigé
PRENOM :

Date :
Groupe :

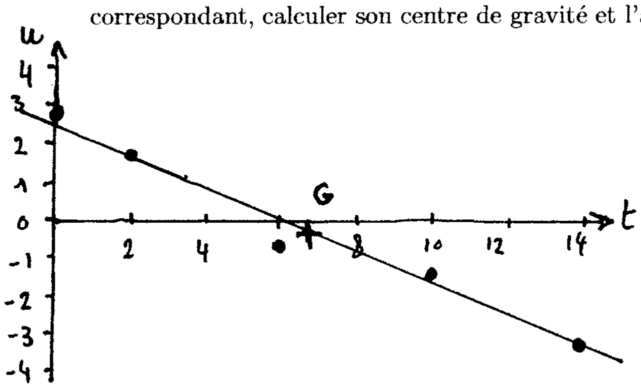
Mathématiques pour la Biologie : Feuille-réponses du TD 10
Régression linéaire : suite

Exercice 1. : On étudie la croissance d'une plante à partir d'un instant considéré comme instant initial. On effectue des mesures du diamètre de la tige principale et on obtient les résultats ci-dessous.

1. Pour tout $t \geq 0$, on pose $u(t) = \ln\left(\frac{8}{d(t)} - 1\right)$. Compléter les valeurs manquantes de u_i.

Temps t_i, en semaines	0	2	6	10	14
Diamètre d_i, en centimètres	0,4	1,2	5,4	6,4	7,8
$u_i = \ln\left(\frac{8}{d_i} - 1\right)$	2,94	1,73	-0,73	-1,39	-3,66

2. On considère les données (t_i, u_i) comme une série de 5 points. Représenter le nuage de points correspondant, calculer son centre de gravité et l'ajouter au dessin. Centre de gravité G



$x_G = (0+2+6+10+14)/5 = 6,4$

$y_G = (2,94+1,73-0,73-1,39-3,66)/5 \simeq -0,22$

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

i	t i	u i	(t i)^2	(u i)^2	(t i)*(u i)
1	0	2,94	0	8,67	0,00
2	2	1,73	4	3,01	3,47
3	6	-0,73	36	0,53	-4,39
4	10	-1,39	100	1,92	-13,86
5	14	-3,66	196	13,42	-51,29
total:	32	-1,10	336	27,56	-66,07
moyennes	6,4	-0,22	67,2	5,512	-13,214

$Var(t) = 67,2 - 6,4^2 = 26,24$

$Var(u) = 5,512 - (-0,22)^2 \simeq 5,46$

$Cov(t,u) = -13,214 - (6,4)(-0,22) \simeq -11,806$

$\boxed{Var(t) = 26,24 \qquad Var(u) = 5,46 \qquad cov(t,u) = -11,806}$

4. Déterminer, par la méthode des moindres carrés ordinaires une équation de la droite de régression de u en t. $\boxed{u = -0,4498t + 2,6586}$

5. Tracer cette droite sur le graphique de la deuxième question.

6. Calculer le coefficient de corrélation linéaire ; commenter. $\boxed{\rho(t,u) = -0,986}$

7. En déduire que, pour cette plante, le diamètre de sa tige principale est donné par une relation de la forme $d(t) = \frac{8}{1+Ce^{-at}}$ dans laquelle on précisera C et a.

Comme $\rho = -0,986$ proche de -1 on peut penser que la droite $u(t) = -0,4498t + 2,6586$ représente bien les données.

Mais $u(t) = \ln(8/d(t) - 1)$ donne $d(t) = \frac{8}{1+e^{u(t)}}$ où

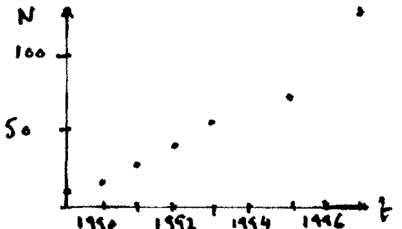
$e^{u(t)} = e^{2,6586} e^{-0,4498t}$, donc $a = 0,4498$ et

$C = e^{2,6586} \simeq 14,276$

Exercice 2. : *Posidonia oceanica* est une plante à fleurs poussant sur les fonds de faibles profondeurs (inférieures à 40 mètres) en mer méditerranée. En 1989, des individus de posidonies provenant de différentes prairies méditerranéennes ont été transplantées dans une baie du parc national de Port-Cros (Var) pour étudier la croissance des transplants. Un suivi a été mis en place : régulièrement, le nombre de plantes vivantes a été décompté par des plongeurs et on a obtenu les données suivantes :

Année (t)	1989	1990	1991	1992	1993	1995	1997
Nb de plantes (N_t)	12	17	27	36	54	71	129

1. Le tracé du nuage de points correspondant (t, N_t) suggère un relation linéaire de $y(t) = \ln \frac{N_t}{N_{1989}}$ sur t. Expliquer pourquoi.



Un tracé rapide du nuage de points suggère une croissance exponentielle de la population à partir de la population de départ en 1989, donc $\frac{N_t}{N_{1989}}$ pourrait être de la forme $e^{y(t)}$ avec $y(t) = at+b$

2. Dans le tableau ci-dessous, on a procédé à la régression linéaire de $y(t) = \ln(N_t/N_{1989})$ sur t selon la méthode MCO. Compléter les 8 cellules vides puis justifier vos réponses.

	Année (t)	y=ln (Nt / No)	t^2	y^2	t*y
	1989	0	3956121	0,00	0,00
	1990	0,35	3960100	0,12	693,13
	1991	0,81	3964081	0,66	1614,56
	1992	1,10	3968064	1,21	2188,44
	1993	1,50	3972049	2,26	2997,63
	1995	1,78	3980025	3,16	3546,66
	1997	2,37	3988009	5,64	4742,69
Somme	13947	8	27788449	13,05	15783
Moyenne	1992,43	1,13	3969778,4	1,86	2254,7

Var(t)	Var(y)	Cov(t,y)	a chapeau	b chapeau	R^2
6,82	0,59	1,97	0,2889	-574,48	0,9645

$\hat{a} = Cov(t,y)/Var(t) \simeq 0{,}2889$

$\hat{b} = \mu(y) - \hat{a}\mu(t) = 1{,}13 - 0{,}2889 \times 1992{,}43 \simeq -574{,}48$

$$R^2 = \rho^2 = \frac{Cov(t,y)^2}{Var(t)\,Var(y)} = \frac{1{,}97^2}{6{,}82 \times 0{,}59} \simeq 0{,}9645$$

3. Quelle est l'équation de la droite de régression de $y(t)$ sur t ?

$y(t) = 0{,}2889\,t - 574{,}48$

4. Selon vous, le modèle linéaire obtenu est-il justifié ?

$R^2 \simeq 0{,}9645$ donc plus de 96 % de la variation des données y s'explique par la droite de régression $y(t) = 0{,}2889t - 574{,}48$

C'est un très bon résultat.

Exercice 3. : Calibration d'un modèle par MCO Lorqu'on choisit un modèle dynamique, par exemple un modèle logistique, pour étudier la dynamique d'une population donnée, l'une des premières difficultés est de déterminer les constantes (dans le cas du modèle logistique il y en a deux r et K) qui correspondent le mieux à la population particulière que l'on étudie. Choisir ces constantes s'appelle la *calibration* (en anglais *fitting*) du modèle.

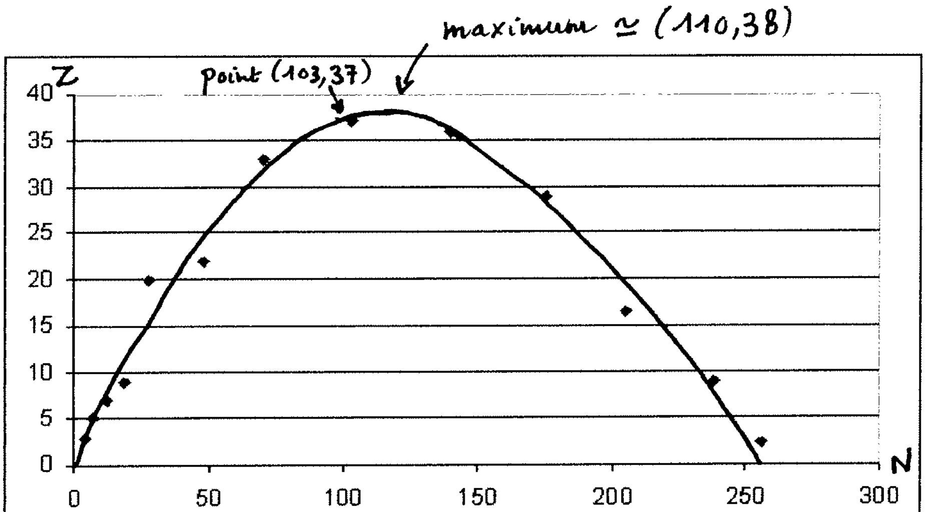
Une première approche peut être faite en comparant le *graphe théorique* de f et un *graphe empirique* que l'on peut construire à partir des observations recueillies sur la population que l'on étudie. Dans l'exemple d'un modèle logistique, le graphe théorique est celui de la parabole $f(N) = rN(1 - \frac{N}{K})$, de sommet $(\frac{K}{2}, \frac{rK}{4})$ et le graphe empirique est obtenu en traçant en abscisse la suite des effectifs N_1, N_2, ..., N_n mesurés en des instants successifs t_1, t_2, ..., t_n et en ordonnée il conviendrait de tracer la dérivée N_i' par rapport au temps pour chaque valeur N_i de l'effectif (puisqu'on a la relation $N' = f(N)$). En pratique, on remplace cette dérivée qui ne peut être mesurée par les taux de variations Z de ces effectifs par unité de temps $Z_1 = \frac{N_2 - N_1}{t_2 - t_1}$, $Z_2 = \frac{N_3 - N_2}{t_3 - t_2}$, ..., $Z_{n-1} = \frac{N_n - N_{n-1}}{t_n - t_{n-1}}$. S'il est raisonnable de supposer que la population a bien un comportement de type logistique, ce graphe empirique de la fonction f doit avoir l'allure d'une parabole et les coordonnées de son sommet doivent être égales à $(\frac{K}{2}, \frac{rK}{4})$. Ceci permet d'estimer les deux constantes r et K approximativement à partir des coordonnées observées du sommet.

Voici un exemple d'application :

En 1927, Pearl a étudié la dynamique d'une culture de cellules de levure et il a obtenu les mesures suivantes (la taille de la levure est exprimée en biomasse ($mg\ 100ml^{-1}$) :

t=Heures	0	1	2	3	4	5	6	7	8	9	10	12	14	18
N=Biomasse	4	7	12	19	28	48	70	103	140	176	205	238	256	265
Z=taux	3	5	7	9	20	22	33	37	36	29	16,5	9	2,25	—

Calculer les taux de variation (compléter le tableau) et tracer le graphe empirique de f correspondant (en utilisant votre calculatrice pour tracer le graphe empirique de f, si vous avez une calculatrice graphique, ou approximativement à *la main*) Puis utiliser ce graphe pour calibrer un modèle logistique aux données de Pearl (en proposant des valeurs de K et r). Expliquez vos choix.



Le graphe empirique donne un sommet peu après le point (103,37) en approximativement (110,38). On peut alors considérer que le sommet théorique $\left(\frac{K}{2}, \frac{rK}{4}\right) \simeq (110,38)$. De $\frac{K}{2} \simeq 110$ on déduit facilement $\boxed{K \simeq 220}$. Alors de $\frac{rK}{4} \simeq 38$ on tire $\boxed{r \simeq \frac{4 \times 38}{220} \simeq 0,69}$

Le tracé assez peu précis de la courbe rend ces valeurs très approximatives.

La méthode précédente est en fait assez grossière. En pratique, une bonne calibration relève plutôt de méthodes statistiques, la plus simple de ces méthodes statistiques étant la *méthode des moindres carrés* que nous venons d'introduire. Voici comment on procède.

On note que si $N(t)$ désigne la taille de la population à l'instant t l'équation différentielle logistique $N' = rN(1 - \frac{N}{K})$ peut se ré-écrire $\frac{N'}{N} = r - \frac{r}{K}N$. On en déduit que le *taux de biomasse*, $\frac{N'}{N}$, est en réalité une *fonction linéaire* (ou plutôt *affine*) de la biomasse. Donc si l'on peut calculer ce taux à partir des données, une simple régression linéaire pourra permettre de calculer les deux constantes $\frac{r}{K}$ (pente de la droite de régression) et r (ordonnée à l'origine de la droite de régression). En pratique on approxime ce taux de biomasse par le rapport $\tau_i = \frac{N_{i+1}-N_i}{(t_{i+1}-t_i)N_i}$.

1. Expliquer pourquoi τ_i est une approximation de $\frac{N'}{N}$

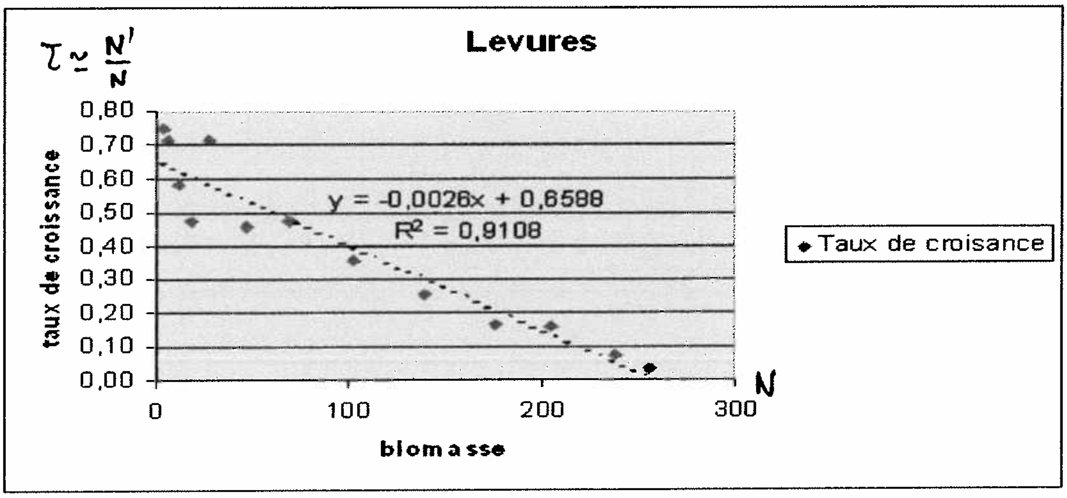
$\frac{N_{i+1}-N_i}{t_{i+1}-t_i}$ est la pente de la droite qui joint les deux points consécutifs (t_i, N_i) et (t_{i+1}, N_{i+1}) du graphe de $N(t)$. Si ces points sont assez rapprochés cette pente est proche de la dérivée en t_i

$\tau_i \simeq N'(t_i)/N(t_i)$

2. Compléter le tableau suivant :

t=Heures	0	1	2	3	4	5	6	7	8	9	10	12	14	18
N=Biomasse	4	7	12	19	28	48	70	103	140	176	205	238	256	265
τ	0,75	0,71	0,58	0,47	0,71	0,46	0,47	0,36	0,26	0,16	0,08	0,04	0,01	—

3. Voici le resultat de la régression linéaire de τ sur N obtenu sous Excel. Calculer les valeurs de deux constantes r et K en expliquant vos calculs.



En déduire les valeurs de K et r et le modèle logistique suivi par la biomasse de levure $N(t)$. Comparer avec les résultats précédents et commenter.

Comme $R^2 = 0,9108$ on peut raisonnablement considérer que

$\frac{N'}{N} = r - \frac{r}{K}N$ est correctement représenté par $-0,0026N + 0,6588$

D'où l'on tire $\boxed{r = 0,6588}$ et $-\frac{r}{K} = -0,0026$ donc $\boxed{K = \frac{0,6588}{0,0026} \simeq 253}$

Ces valeurs sont proches de celles trouvées précédemment mais plus précises.