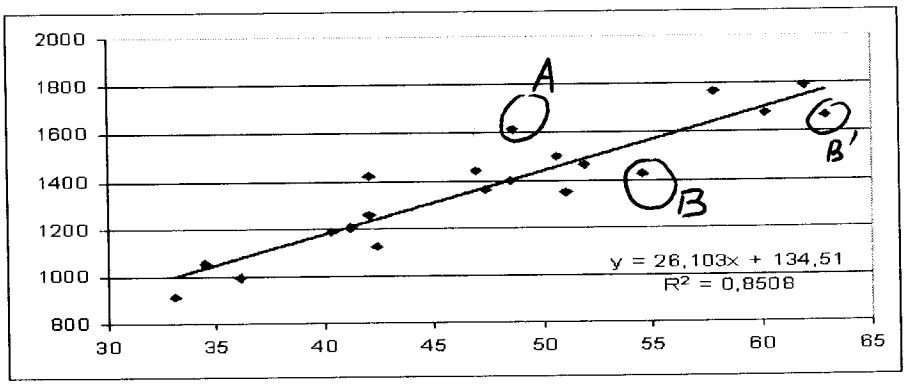
NOM : Date :
PRENOM : Groupe :

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

Exercice 1. : Etude des résidus d'une régression linéaire Pour étudier la relation qui existe entre la masse corporelle d'un individu (son poids diminué de ses réserves de graisse) et son métabolisme (le nombre moyen de calories brulées en 24 heures), des chercheurs ont recueilli les données suivantes portant sur 20 individus :

x_i =masse corporelle	60,2	62	62,9 (B')	36,1	54,6 (B)	48,5	42	47,4	50,6	42
y_i =métabolisme	1679	1792	1666	995	1425	1396	1418	1362	1502	1256
x_i =masse corporelle	48,7 (A)	40,3	33,1	51,9	42,4	34,5	51,1	41,2	57,9	46,9
y_i =métabolisme	1614	1189	913	1460	1124	1052	1347	1204	1767	1439

1. Ils effectuent une régression linéaire sur ces données et obtiennent le résultat présenté sur la figure suivante :



Quel point A du nuage présente le résidu ε_i positif le plus grand ? Indiquer ses coordonnées et calculer ce résidus ($\varepsilon_i = y_i - \hat{y}_i = y_i - (\hat{a}x_i + \hat{b})$).

C'est le point le plus éloigné verticalement de la droite de regression et situé au-dessus d'elle. Visiblement(1) ce point est le point $A(48.7, 1614)$. Pour ce point on a

$\varepsilon_i = 1614 - (26.103 \cdot 48.7 + 134.51) = 208.29$

2. Quel point B du nuage présente le résidu ε_i négatif le plus grand ? Indiquer ses coordonnées et calculer ce résidus.

C'est le point le plus éloigné verticalement de la droite de regression mais situé en-dessous d'elle cette fois. Il semble(2) que ce soit $B(54.6, 1425)$. Pour ce point on a

$\varepsilon_i = 1425 - (26.103 \cdot 54,6 + 134.51) = -134,71$

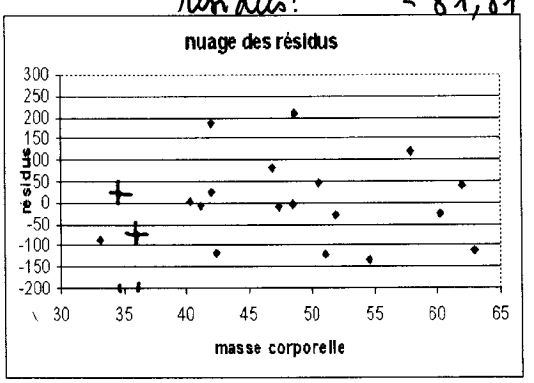
3. Quel est, selon ce modèle, le métabolisme d'un individu dont la masse corporelle vaut 50kg ? Expliquer.

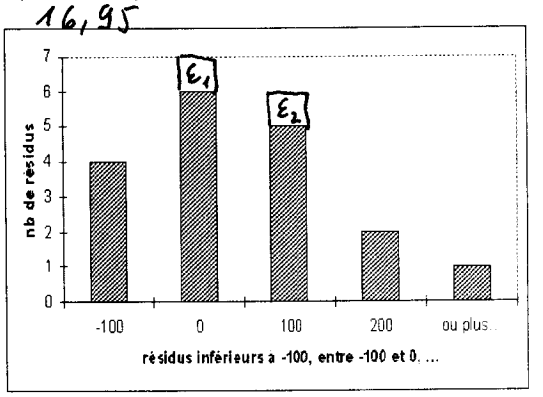
Le modèle calculé (par Excel) est $y = 26.103x + 134.51$. Donc pour $x = 50$ on a $y = 26.103 \cdot 50 + 134,51 = 1439.66$. Un individu de masse corporelle 50kg brûle donc, selon ce modèle, 1440 cal en moyenne par 24h.

(1) des calculs permettraient de vérifier cela mais je n'en vois pas le besoin
(2) cela pourrait aussi être le point $B'(62.9, 1666)$ mais je n'ai pas le temps de le vérifier

4. On a reproduit sur les deux graphiques suivants d'une part le nuage des résidus (x_i, ε_i), et d'autre part l'histogramme dès résidus (à noter que ce que Excel appelle un histogramme n'en est pas exactement un puisque sa surface totale ne vaut pas 1 mais plutôt un *diagramme en batons* des effectifs par classes). Les deux points (36,1,995) et (34,5,1052) ont été oubliés.

résidus : −81,81 16,95





Ajouter les deux points correspondants sur la figure de gauche. Puis modifier celle de droite (en rectifiant la hauteur des batons) de façon à ce qu'elle prenne en compte les deux points oubliés.

Exercice 2. Résidus non indépendants Pour être valide, un modèle de régression linéaire doit non seulement avoir un R^2 assez proche de 1, ainsi qu'un nuage ne présentant pas les abérations mentionnées mais il doit aussi posséder des résidus qui peuvent raisonnablement être assimilés à des *erreurs :* ils doivent être *indépendants* les uns des autres et présenter une distribution ayant plus ou moins une allure de cloche de Gauss. Voici un exemple où ce n'est pas le cas :

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
$\hat{y}_i$	5,1	5,4	5,7	5,9	6,2	6,5	6,7	7,0	7,3	7,6	7,8	8,1
ε_i	−0,85	−0,3	0,02	0,35	0,58	0,62	0,45	0,18	−0,08	−0,35	−0,32	−0,29

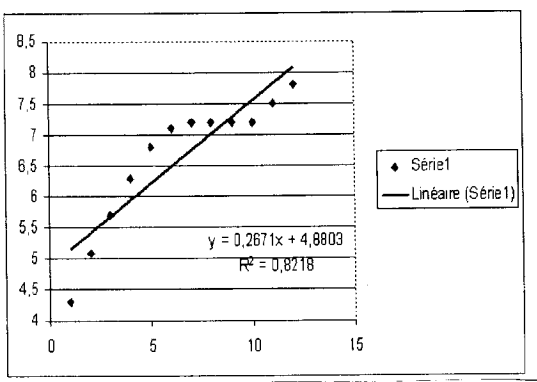
1. Un statisticien pressé a fait calculer par sa machine la droite des moindres carrés pour ces données et a trouvé la relation poids $= 4,88 + 0,267$ age. S'est-il trompé ?

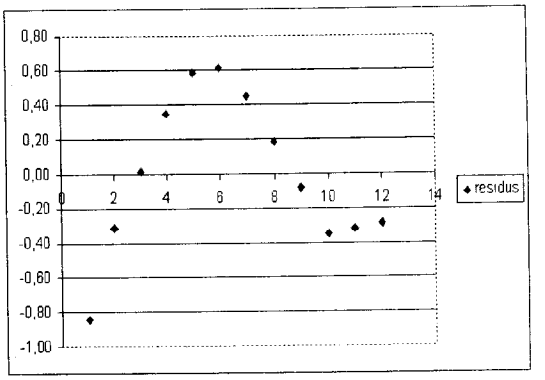
Nm

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

On calcule et trouve $R^2 = 0,82$ qui est un "bon" R^2
Le modèle pourrait donc bien être pertinent ; pourtant...

3. Calculer puis tracer les résidus en fonction de la classe d'age. Vous constaterez 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.





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.

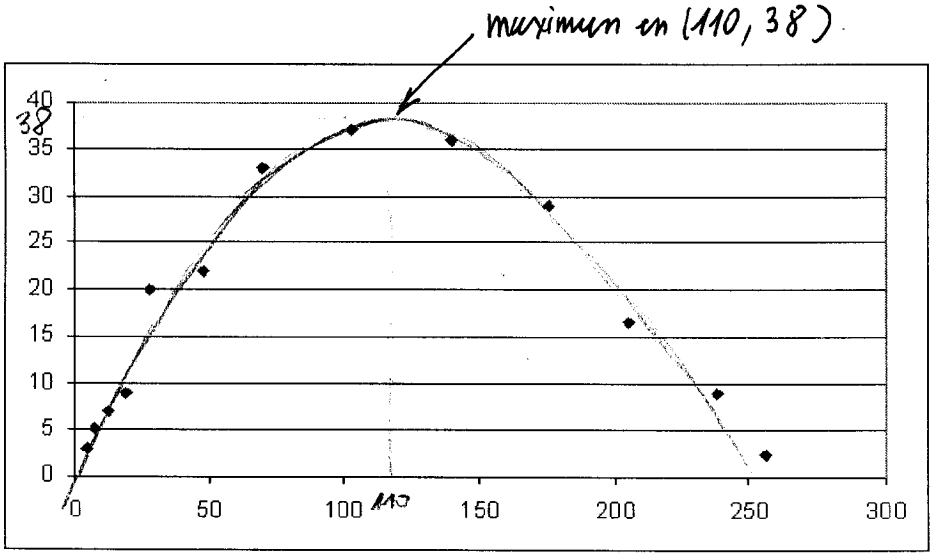
Une première approche peut être faite en comparant le *graphe théorique* de f et un *graphe empirique* que l'on peut construire à partir des observations recueillies sur la population que l'on étudie. Dans l'exemple d'un modèle logistique, le graphe théorique est celui de la parabole $f(N) = rN(1 - \frac{N}{K})$, 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.



On observe un "maximum" pour cette courbe situé auprès du point (103, 37), peut-être à (110, 38) approximativement.

Comme le maximum théorique est situé en $(\frac{K}{2}, \frac{2K}{4})$, on en déduit que $\begin{cases} \frac{K}{2} = 110 \\ \frac{2K}{4} = 38 \end{cases}$ et donc que

$K = 220$ et $r = \frac{4(38)}{220}$

d'où $r = 0,69$. On sent bien que ce choix est plutôt arbitraire compte tenu du petit nombre de points et de la forme irrégulière de la courbe.

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 $\tau_i = \frac{N_{i+1} - N_i}{(t_{i+1} - t_i)N_i}$.

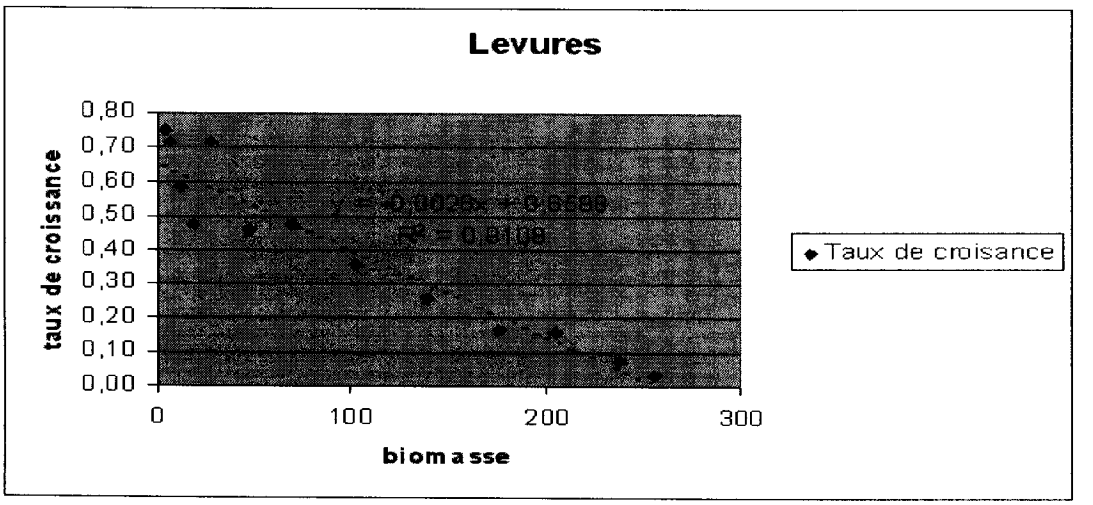
1. Expliquer pourquoi τ_i est une approximation de $\frac{N'}{N}$

$N'(t)$ est par définition la limite quand $h \to 0$ de $\frac{N(t+h) - N(t)}{(t+h) - t}$
Si l'on prend $h = t_{i+1} - t_i$ comme "pas de temps" et s'il est assez petit, le rapport $\frac{N_{i+1} - N_i}{t_{i+1} - t_i}$ sera proche de N_i' et donc $\frac{N_{i+1} - N_i}{(t_{i+1} - t_i)N_i}$ proche de $\frac{N_i'}{N_i}$.

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.

L'équation $\frac{N'}{N} = r - \frac{r}{K}N$ est de la forme $\frac{N'}{N} = aN + b$ avec $a = -\frac{r}{K}$ et $b = r$. Le tableau précédent indique que le taux de croissance est égal à $-0,0026$ fois la biomasse, plus $0,6588$.
Donc $r = 0,6588$ et $-\frac{r}{K} = -0,0026$. Donc $K = \frac{0,6588}{0,0026} \simeq 253$.