ÉTUDE D'UNE SÉRIE STATISTIQUE À DOUBLE VARIABLES

RÉGRESSION LINÉAIRE

Benchikh Tawfik

Faculté de Médecine, UDL, SBA

1ère année Médecine

18 Octobre 2023

PLAN DU COURS

INTRODUCTION

- Les statistiques à une variable s'intéressaient, pour une population donnée, à un caractère donné: le nombre d'enfants par famille, les salaires dans une entreprise, et ...
- Lorsque l'on s'intéresse à l'**étude simultanée de deux caractères** d'une même population, on fait ce que l'on appelle des **statistiques à deux variables**, en étudiant des **séries statistiques doubles**.

OBJECTIF GÉNÉRAL:

- On étudie une population à partir de **données** recueillies sur un **échantillon** d'individus tirés au sort dans la population.
- **Les données:** proviennent de deux variables qui sont mesurées simultanément sur les individus apportent une information partielle sur la population étudiée.
- On se pose des questions sur la population étudiée. Pour y répondre:
  - ⋆ 1 ère phase descriptive: analyses des données observées, à l'aide de méthodes descriptives adaptées.

Les méthodes descriptives pour séries statistiques doubles:

- **Visualiser** et **mesurer par des indices** les éventuelles **relations** existant entre ces variables pour des données échantillonnées:
  - ⋆ **les graphiques** et **indices statistiques** calculés apportent une information **partielle** sur la variable dans la population et préparent les **analyses inférentielles**.
- **Modélisation des relations** entre une variable quantitative que l'on cherche à **expliquer** par une autre variables quantitative appelée variable **explicatives**:
  - ⋆ Méthodes de **régression linéaire**.

Définition

- On considère
  - une population d'effectif n,
  - deux variables statistique X et Y (deux caractères).
- Chaque individu de cette population est désigné par un nombre compris entre 1 et n.
- À chaque individu i ($1 \leq i \leq n$) correspond un couple $(x_i; y_i)$, ou x_i est la modalité du caractère X et y_i est la modalité du caractère Y associé à l'individu i.
- L'ensemble des couples $(x_i; y_i)$ définit une série statistique à deux variables.

Exemple 1:

- Un médecin mesure sur 12 femmes d'âges (X) différents la pression sanguine systolique (Y). Il obtient les résultats suivant:

x (ans)	56	42	72	36	63	47
y (mm Hg)	147	125	160	118	149	128
x (ans)	55	49	38	42	68	60
y (mm Hg)	150	145	115	140	152	155

  - ⋆ **Population:** femme (d'une ville, d'un pays,...).
  - ⋆ **Caractère 1:** l'âge.
  - ⋆ **Caractère 2:** pression sanguine systolique.

Exemple 2:

- Le tableau ci-dessous permet de suivre l'évolution de l'espérance de vie à la naissance (en années) en algerie de 2000 à 2009 pour les hommes:

X (année)	2000	2001	2002	2003	2004
Y (Espérance de vie)	76.9	77.1	77.3	77.5	77.6
X (année)	2005	2006	2007	2008	2009
Y (Espérance de vie)	77.8	78	78.1	78.2	78.2

  - ⋆ **Population:** les hommes en algerie.
  - ⋆ **Caractère 1:** l'année.
  - ⋆ **Caractère 2:** l'espérance de vie.

Définition

- Lorsque l'un des deux caractères est une année, une date, on dit que la série statistique double est une **série chronologique ou série temporelle** (Exemple 2).

REMARQUE:

- De la donnée de la série statistique double, on peut déduire les séries statistiques simples décrivant séparément les caractères X et Y.

x (ans)	56	42	72	36	63	47	55	49	38	42	68	60

y (mm Hg)	147	125	160	118	149	128	150	145	115	140	152	155

NUAGE DE POINTS

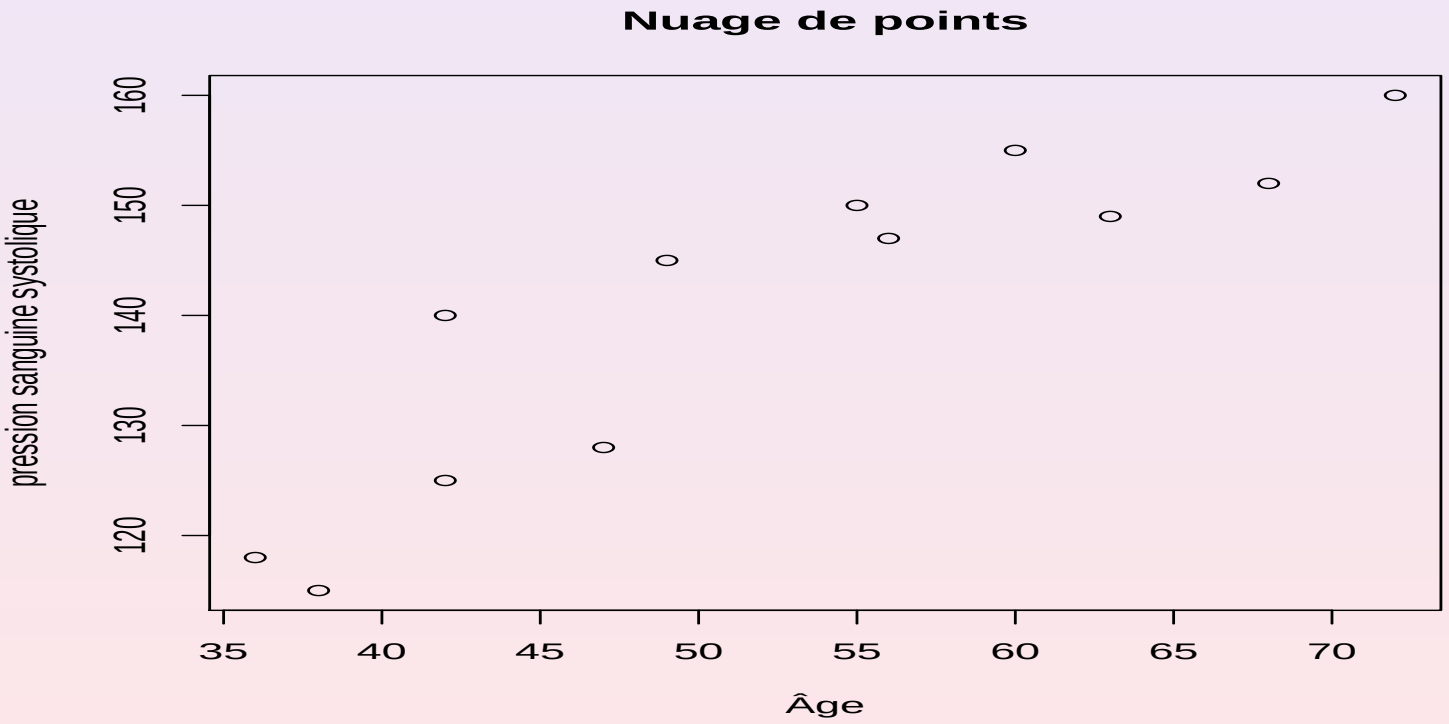
- Ces observations sont représentées sur un **diagramme de dispersion** appelé (**nuage de points**) dans lequel un point i a pour coordonnées:

$$x_i = \text{Âge}$$

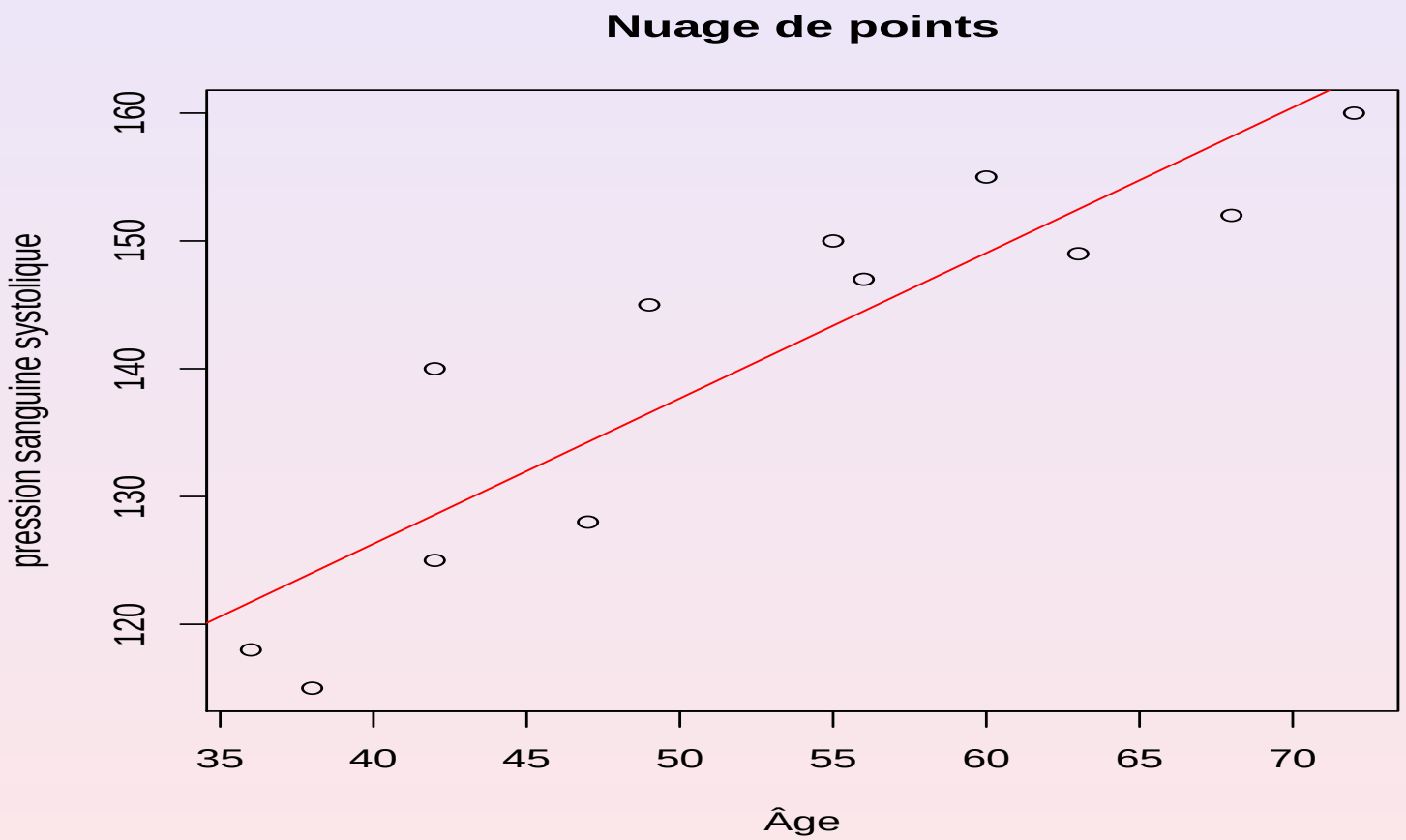
$$y_i = \text{pression sanguine systolique}$$

NUAGE DE POINTS

- Représentation de la pression sanguine systolique en fonction de l'âge:



NUAGE DE POINTS: DROITE DE REGRESSION



Point moyen du nuage

- On appelle point moyen $G(x;y)$ du nuage, le point dont les coordonnées sont les moyennes des variables X et Y de la série: $G = (\overline{X}, \overline{Y})$; où

$$m_X = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

et

$$m_Y = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} x_i \,.$$

La covaiance

- La covaiance de X et Y vaut

$$\begin{aligned} Cov(X,Y) = S_{X,Y} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \overline{X}\,\overline{Y}\,. \end{aligned}$$

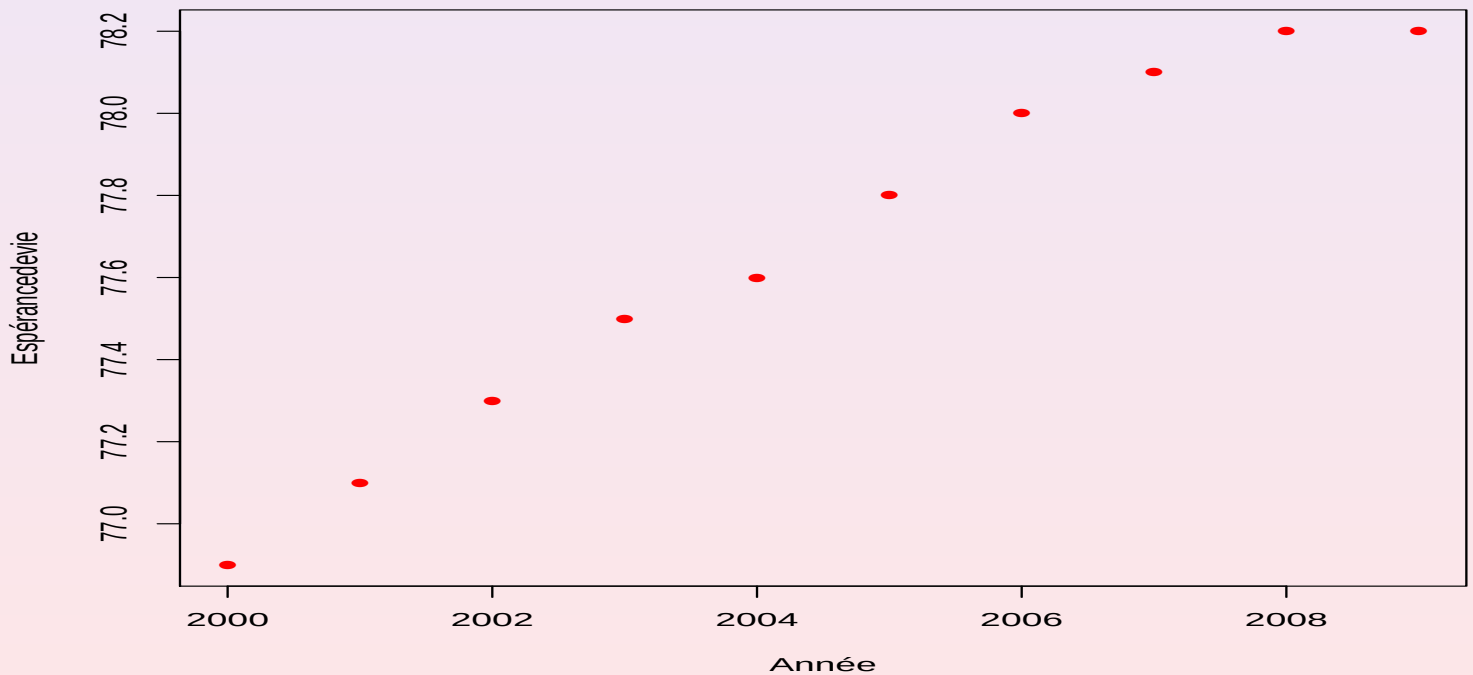
Cœfficient de corrélation linéaire

- Pour deux variables X et Y, le cœfficient de corrélation linéaire $r = \rho(X,Y) = cor(X,Y)$ vaut:

$$r = cor(X,Y) = \frac{cov(X,Y)}{S_X S_Y}$$

TYPES DES NUAGES DE POINTS

- Exemple1: l'évolution de l'espérance de vie à la naissance (en années) en algerie de 2000 à 2009 pour les hommes



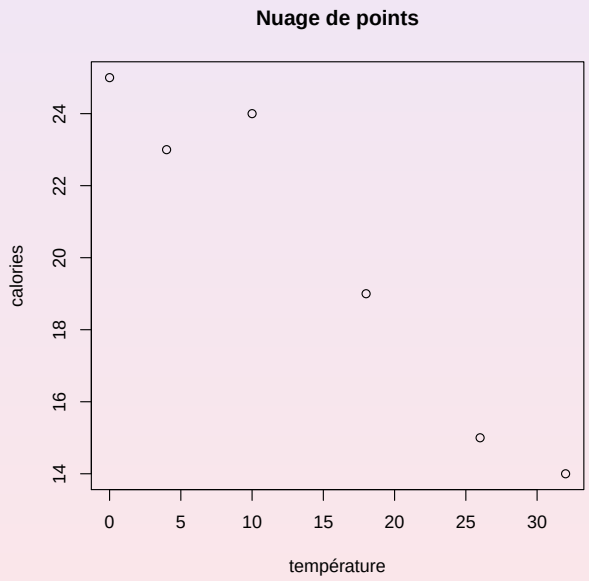
EXEMPLE 3

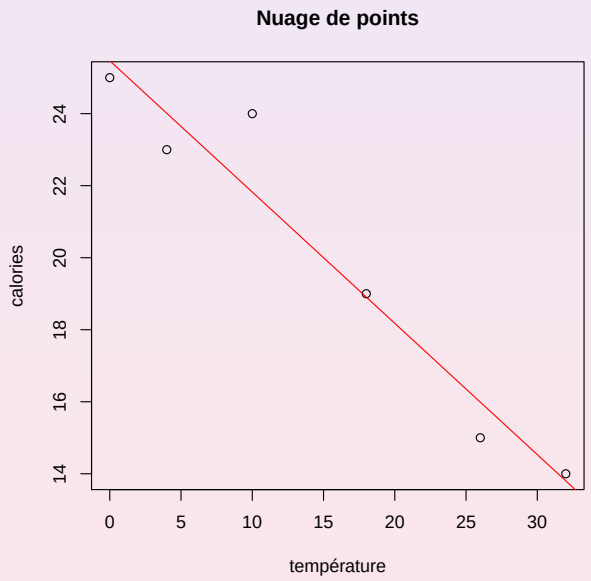
- On a mesuré la quantité d'énergie métabolisée en 10 heures (calories) par un moineau soumis à différentes températures $(^{\circ}C)$; Les résultats sont les suivants:

x = température:	0	4	10	18	26	32
y = calories:	25	23	24	19	15	14

NUAGE DE POINTS: EXEMPLE 3

- Représentation de la quantité d'énergie métabolisée en 10 heures (calories) en fonction de la températures:



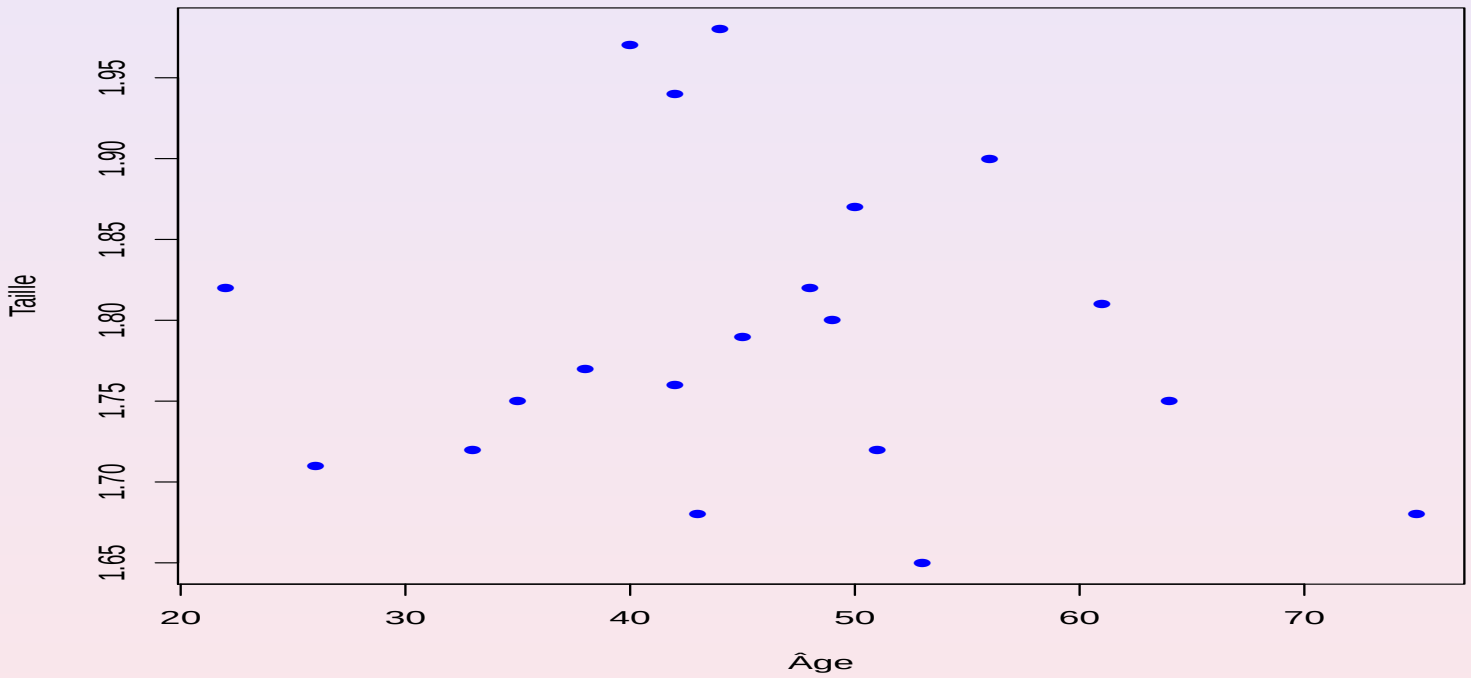


TYPES DES NUAGES DE POINTS: EXEMPLE 4

- Exemple 4: On dispose des données suivantes concernant un échantillon de 20 hommes: âges x_i en années, tailles y_i en mètres.

X (Âge)	22	26	33	35	38	40	42	42	43	44
Y (Taille)	1.82	1.71	1.72	1.75	1.77	1.97	1.94	1.76	1.68	1.98
X (Âge)	45	48	49	50	51	53	56	61	64	75
Y (Taille)	1.79	1.82	1.8	1.87	1.72	1.65	1.9	1.81	1.75	1.68

TYPES DES NUAGES DE POINTS: EXEMPLE 4



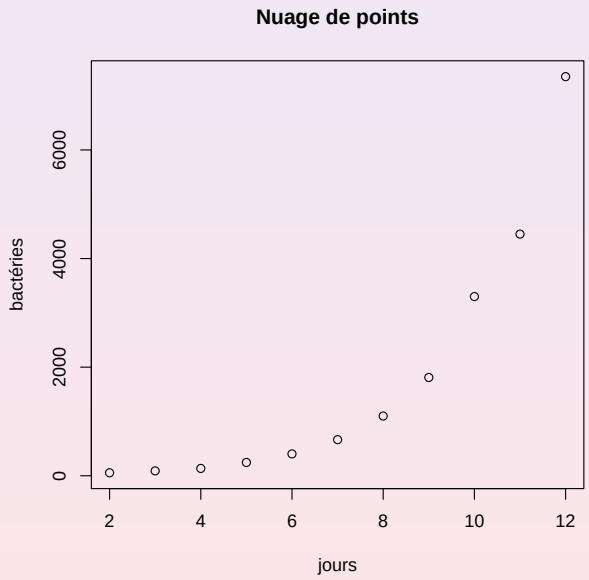
RÉGRESSION NON LINÉAIRE: EXEMPLE 5

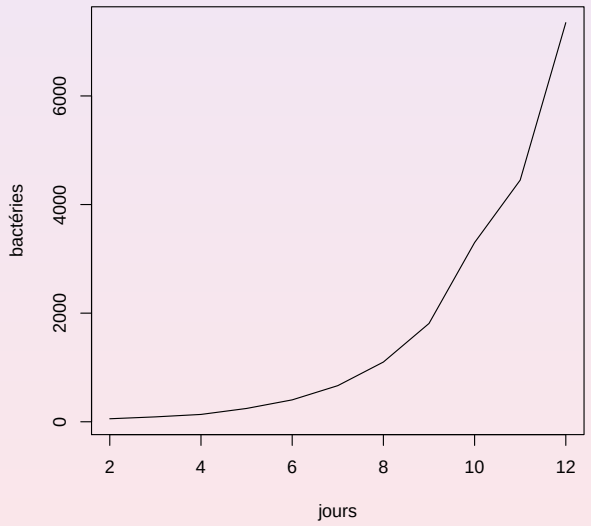
- En l'absence de mortalité, on souhaite décrire l'évolution dans le temps de la croissance d'une population de bactéries. Des numérations faites tous les jours à partir du 2ième donne les résultats suivants:

jours	bactéries
2	55
3	90
4	135
5	245
6	403
7	665
8	1100
9	1810
10	3300
11	4450
12	7350

NUAGE DE POINTS: EXEMPLE 5

- Représentation de la croissance de la population de bactéries en fonction des jours:





TYPES DES NUAGES DE POINTS: REMARQUES

- En regardant ces graphes, on constate que le type de fonction à ajuster sera différent selon les données:
  1. Pour le 1 et 2 exemples, les points sont alignées de façon croissante. On peut ajuter une fonction de type "une droite" (croissante)
  2. Pour le 3 exemples, les points sont alignées de façon décroissante. On peut ajuter une fonction de type "une droite" (décroissante)
  3. Pour l'exemple 4, les points sont disposé de façon quelconque.
  4. Pourl'exemple 5, le graphique: le nombre de bactéries croit de manière rapide (exponentiel).

Types des nuages de points: Généralisation

- Lorsque on a un nuage de points (x_i, y_i) différent situations peuvent se présenter
  - A) les points sont disposés de façon quelconque : on dira que les caractère Age et Taille sont indépendants
  - B) les points sont disposés autour d'une certaine courbe: on pourra faire un "ajustement graphique", c'est à dire tracer au mieux cette courbe.
    - La courbe la plus simple que l'on puisse obtenir est une droite.
    - Parfois il s'agira d'une parabole, d'une fonction du troixième degré, fonction puissance ou exponentielle, ...

RÉGRESSION LINÉAIRE

- **Regression linéaire:** modèle le plus simple:

$$Y = f(X) + \varepsilon = \alpha + \beta \times X + \varepsilon$$

  - Interprétation
  - Estimations des paramètres
  - Prédiction.

Régression linéaire

- α représente l'ordonnée à l'origine et β représente la pente de la droite.
- On utilise des lettres grecques pour représenter l'ordonnée à l'origine et la pente pour bien insister sur le fait que ce sont des paramètres inconnus.
- Leur valeur respective serait connue si on avait accès à toute la population, ce qui n'est jamais le cas en pratique. Il nous faudra donc les estimer.

- **Droite de régression:**
  - Résume le mieux le nuage de point
    - $\Rightarrow$ La plus proche de tous les points
    - $\Rightarrow$ Erreurs ε petits $+ + +$

PRINCIPE DES CALCULS

- Estimer α et β tel que ε petits +++

1. ε_i: écart entre la droite et le point i

$$y_i = \alpha + \beta \times x_i + \varepsilon_i$$

$$\Rightarrow \varepsilon_i = y_i - (\alpha + \beta \times X)$$

2. Somme des Carrés des Écarts

$$SCE = \sum_{i=1}^{n} (\varepsilon_i)^2$$

3. Estimer α et β tel que: SCE minimum

Calcul de la pente β

- La pente β est donnée par la formule suivante:

$$b = \frac{Cov(X,Y)}{Var(X)}$$

ESTIMATION DE α:

- La droite passe par m_Y et m_X:

$$m_Y = a + b m_X$$

où $m_Y = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ et $m_X = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, D'où:

$$a = m_Y - b m_X$$

EXEMPLE

- Covariance de la pression sanguine et de l'âge:

$$Cov(pressionsanguinesystolique, Age) = Cov(X, Y) = S_{X,Y} = 160.4242$$

- Variance de l'âge: $S^2(Age) = S^2(X) = 140.9697$
- Estimation de β

$$b = Cov(pression, Age)/S^2(Age) = \frac{S_{X,Y}}{S^2(X)} = 1.138005$$

- Estimation de α:

$$a = M_{pressionsanguinesystolique} - b \times M_{Age} = m_Y - bm_X = 80.77773$$

L'équation s'écrit donc:

$$\text{Pression sanguine systolique} = 80.778 + 1.138 \times \text{âge} + \varepsilon$$

où

REMARQUE

Une fois les paramètres a et b calculés, on en déduit les valeurs ajustées $\hat{y}_i = a + bx_i$ puis les résidus estimés $\varepsilon_i = y_i - \hat{y}_i$.

Coefficient de détermination

- Pourcentage de variance expliquée:

$$R^2 = \frac{\text{Part de variance expliquée par la régression}}{\text{Variance totale}}$$

  - Donc R^2 est la proportion de la dispersion des Y qui est expliquée par la dispersion des X (ou par le modèle).
  - $R^2 \simeq 1 \Rightarrow$, le modèle est meilleur (parfait): la connaissance des valeurs de X permet de deviner avec précision celle de Y.
  - Lorsque $R^2 \simeq 0 \Rightarrow$ que X n'apporte pas d'informations utiles (intéressantes) sur Y , la connaissance des valeurs de X ne nous dit rien sur celles de Y .
  - **Remarque:** Dans le cas de la régression simple, $R = r$ (estimation du coefficient de corrélation entre X et Y).

Exemple 1:

- Cœfficient de corrélation estimé entre X et Y

$$\begin{aligned} r &= cor(pressionsanguinesystolique, Age) \\ &= \frac{cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = 0.8961394 \end{aligned}$$

- Estimation de R^2:

$$r * r = 0.8030658$$

ce qui indique que la relation entre l'âge et la pression sanguine systolique est très forte .

REMARQUES: LES FAUSSES CORRÉLATIONS

- Qu'est-ce qu'une corrélation ? C'est une relation positive ou négative entre deux phénomènes, mais elle n'est pas absolue.
  - Exemple: il y a une corrélation positive entre la taille et le poids des hommes : ceux qui mesurent un mètre quatre-vingt pèsent en général plus lourd que ceux dont la taille ne dépasse pas un mètre soixante. Mais il y a des petits gros et des grands maigres.

REMARQUES: LES FAUSSES CORRÉLATIONS

- Souvent, une corrélation est le signe d'une relation de cause à effet. Le plus souvent, on sait ce qui est la cause et ce qui est l'effet :
  - c'est la consommation de tabac qui provoque le cancer du poumon et non la prédisposition à ce cancer qui donne envie de fumer. Mais dans certains cas, les choses sont beaucoup moins évidentes. Et il peut arriver aussi que chacun des deux phénomènes soit à la fois cause et effet.

REMARQUES: LES FAUSSES CORRÉLATIONS

- En outre, il y a beaucoup de corrélations statistiques qui ne résultent aucunement d'une relation de cause à effet et qui sont de ce fait trompeuses.
  - C'est notamment le cas pour les séries statistiques qui évoluent parallèlement dans le temps, avec le progrès économique et scientifique. Certes, si l'espérance de vie augmente, en même temps que diminue la fréquentation des cinémas (corrélation négative), personne n'ira soutenir que l'on vit plus vieux parce que l'on va moins souvent au cinéma.
  - Mais dans bien des cas, surtout si l'on veut prouver quelque chose, on n'hésitera pas à voir une relation de cause à effet là où il n'y a rien d'autre que l'évolution parallèle de deux séries statistiques.

Modèle non linéaire

- D'après le graphique le nombre de bactéries (exemple 5) croit de manière rapide (exponentiel).
- On peut donc déduire que le coefficient de corrélation linéaire entre le nombre de bactéries N et la variable temps t est positif; en effet on trouve

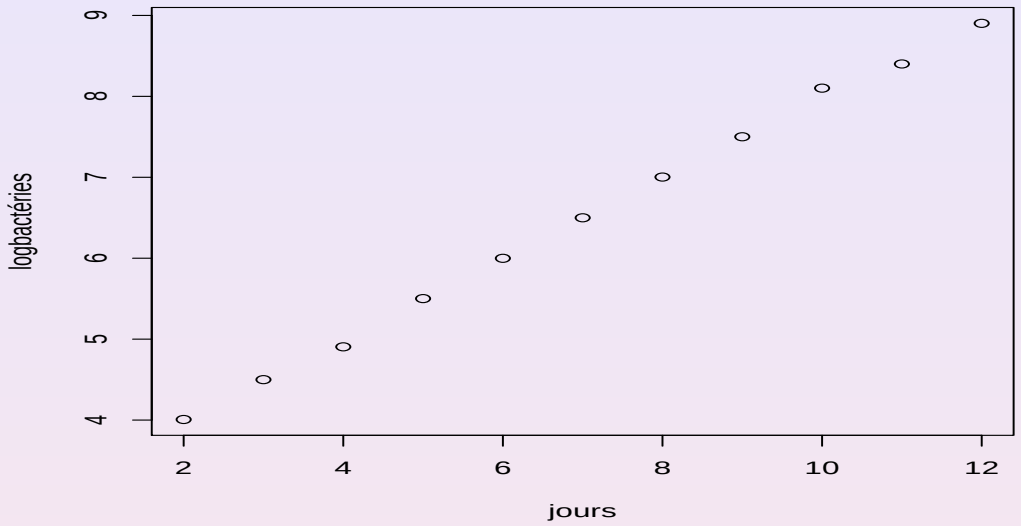
$$\hat{r} = \hat{cor}(jours, bactries) = 0.86474$$

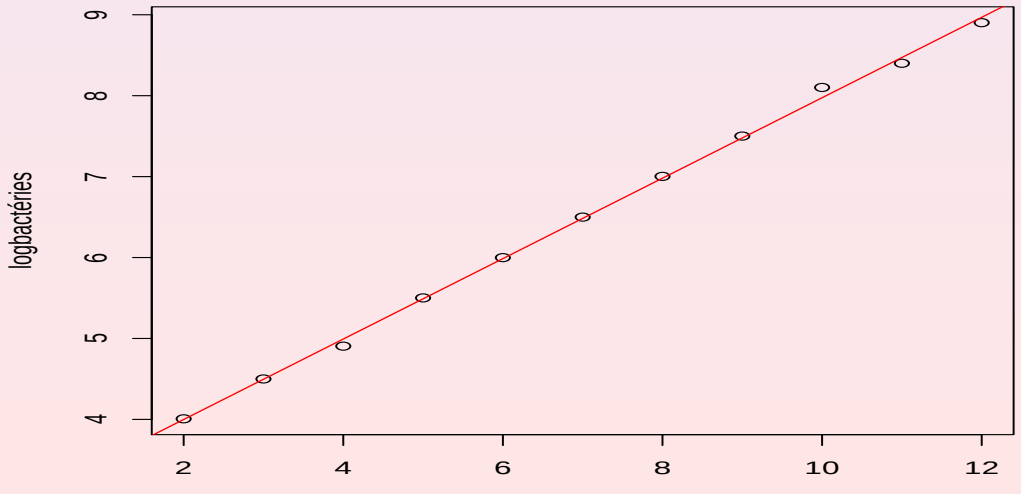
- (!!!) (attention le modèle est non linéaire).

TRANSFORMATION

- Pour expliquer N en fonction de t, nous allons faire une transformation logarithmique seulement de la variable N (car c'est la variable qui a des valeurs très grandes).
- En effet en posant $y = \ln(N)$ et $x = t$, le graphique suivant montre qu'il y a une relation linéaire en y et x.

NUAGE DE POINTS ET LA DROITE DE RÉGRESSION





ESTIMATION DES PARAMÈTRES DU MODÈLE

- Le coefficient de corrélation linéaire est:

  $r = cor(jours, logbactries) = 0.9996615$

- ajustement linéaire de $Y = \ln(N)$ en X est bien justifié.

ESTIMATION DES PARAMÈTRES DU MODÈLE

- On trouve $a = 3.014$ et $b = 0.494$.
- La droite des moindres carrés est donnée par

$$Y = 0.494X + 3.014$$

- La somme des carrés des résidus $SSR = 0.04499$ est très faible.
- Le coefficient $R^2 = 0.9993$ est très proche de 1, on peut donc affirmer que l'ajustement est de très bonne qualité.
- En résumé, on déduit que l'évolution du nombre de bactéries en fonction des jours suit l'équation:

$$N(t) = e^{0.494t+3.014} = 20.36871e^{0.494t}.$$

EXERCICE 1

L'une des mesures qui sont faites lors de l'investigation des affections respiratoires est celle du volume expiratoire moyen par seconde, appelé Vems. Sur 8 sujets tirés au sort parmi la population saine d'âge compris entre 30 et 35 ans, on a mesuré la taille T (en mètres) et le Vems V (en litres par seconde), et obtenu les résultats suivants :

Sujet	1	2	3	4	5	6	7	8
T	1,85	1,72	1,51	1,62	1,60	1,80	1,75	1,68
V	4,5	3,6	2,7	3,1	3,6	4,4	4,3	3,8

Exercice 1

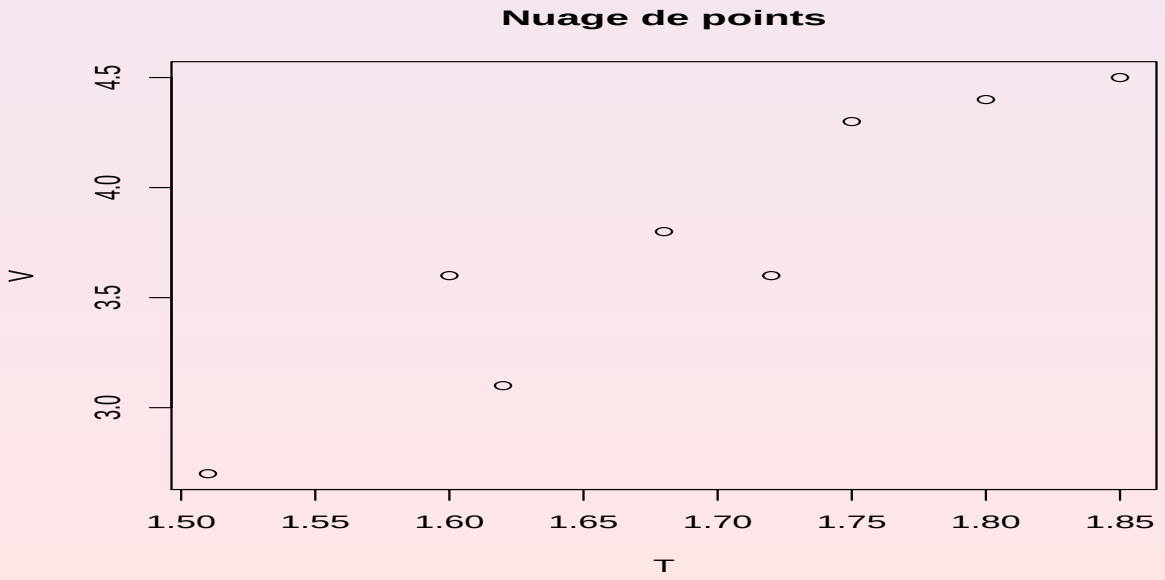
1. Dessiner et commenter le nuage des points de ces observations (T en abscisse et V en ordonnée) .
2. Calculer le coefficient de corrélation linéaire de T et Vems .
3. Sur le même repère, tracer la droite de régression observée de V par rapport à T.
4. Un neuvième sujet survient qui mesure 1,70 m. Quel Vems peut on prévoir pour lui ? En faite, son Vems est de 4 litres. Quelle erreur a-t-on commise ?

NB: $\sum t_i = 13.53$; $\sum t_i^2 = 22.9703$; $\sum v_i = 30$; $\sum v_i^2 = 115.36$; $\sum t_i v_i = 51.205$.

SOLUTION

1. Nuage des Points: on remarque que les points sont parfaitement alignés, donc on peut déduire qu'il existe une **relation de type linéaire** entre la taille T et Vems:

$$V = b \times T + a\,.$$



SOLUTION

2. On a $n = 8$. $cor(T,V) = \frac{S_{T,V}}{S_T \times S_V}$.
   - $\overline{T} = \frac{1}{n}\sum t_i = 1.6913$, $S^2(T) = \frac{1}{n}[\sum t_i^2] - \times(\overline{T})^2 = 0.0125$ et $S_T = 0.112$.
   - $\overline{V} = \frac{1}{n}\sum v_i = 3.75$, $S^2(V) = \frac{1}{n}[\sum v_i^2] - \times(\overline{V})^2 = 0.409$ et $S_V^2 = 0.639$.
   - $cov(T,V) = S_{T,V} = \frac{1}{n}[\sum t_i\, v_i] - \times\overline{T} \times \overline{V} = 0.0668$.

   Donc: $cor(T,V) = \frac{S_{T,V}}{S_T \times S_V} = 0.9335332 \simeq 0.93$.

SOLUTION

3. La droite de régression de V par rapport à T est donnée par:

$$V = a \times T - b\,,$$

où $b = \frac{S_{T,V}}{S_T}$ et $a = \overline{V} - a \times \overline{T}$.

On trouve: $b = 5.33$ et $a = -5.267$.

D'où l'équation de la droite de régression est:

$$V = 5.33 \times T - 5.267.$$

4. Si $T = 1.7$, alors, suivant la droite de régression,

$V = 5.33 \times 1.7 - 5.267 = 3.794$.

5. L'erreur = valeur observé - valeur estimé $= 4 - 3.794 = 0.206$.