

RÉGRESSION LINÉAIRE

Benchikh Tawfik

Faculté de Médecine, UDL, SBA

1^{ère} année Médecine

20 Octobre 2015



PLAN DU COURS

- 1 INTRODUCTION
- 2 OBJECTIF
- 3 RÉGRESSION
- 4 RÉGRESSION LINÉAIRE
- 5 RÉGRESSION NON LINÉAIRE
- 6 EXERCICE

INTRODUCTION

Exemple 1:

Afin d'étudier la relation qui pourrait exister entre l'âge et la pression sanguine, un médecin mesure sur 12 femmes d'âges (X) différents la pression sanguine systolique (Y).

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

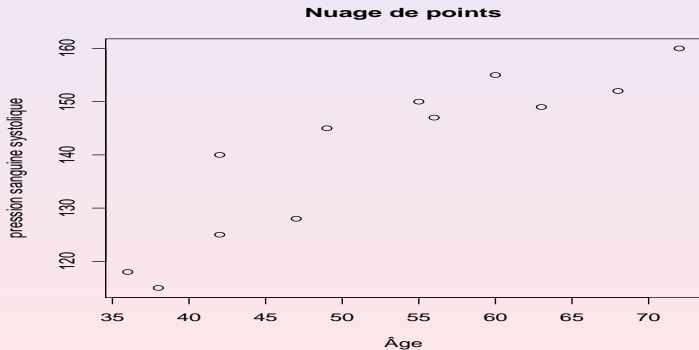
Ces observations sont représentées sur un diagramme de dispersion (nuage de points) dans lequel un point i a pour coordonnées:

$$x_i = \text{Âge}$$

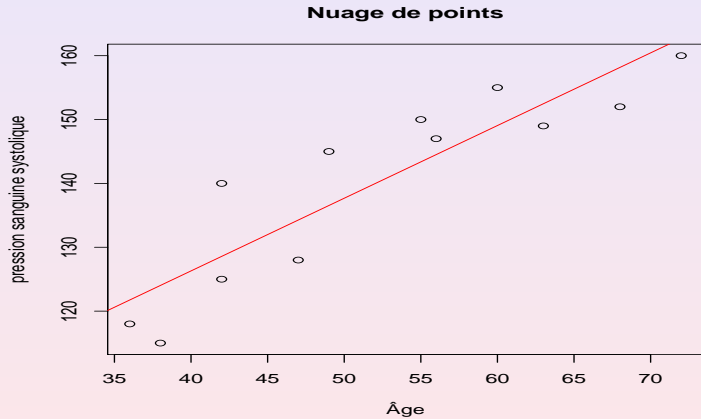
$$y_i = \text{pression sanguine systolique}$$

NUAGE DE POINTS

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



NUAGE DE POINTS: DROITE DE REGRESSION



NUAGE DE POINTS

- La figure montre qu'il semble y avoir une relation entre l'âge de l'individu et sa pression sanguine systolique et que cette relation semble être "**linéaire**".

OBJECTIF DE LA RÉGRESSION

- ① Lien entre la pression sanguine et l'âge ?
- ② Quand l'âge ↑, la pression sanguine ↑?
- ③ Connaissant l'âge, peut-on prédire la pression sanguine?
- ④ But medical: detecter la maladie !!!

OBJECTIF

- **Regression** de Y en X:

Y = pression sanguine (mm Hg)

X = âge (ans)

- Comment la pression sanguine évolue **en fonction** de l'âge ?

$$Y = \text{Pression sanguine} = f(\hat{\text{Âge}}) = f(X) + \varepsilon$$

- Comment évolue la pression sanguine?
 - = Quelle valeur de la pression sanguine?
 - \Rightarrow Pour chaque Age.
 - \Rightarrow Sachant l'âge
- Fonction $f()$ c'est une droite:

$$\mathbb{E}(\text{pression sanguine systolique} / \text{Age}) = \alpha + \beta \hat{\text{Age}}$$

- Pour chaque sujet:

$$\mathbb{E}(\text{pression sanguine systolique} / \text{Age}) = \alpha + \beta \hat{\text{Age}} + \varepsilon,$$

où ε est erreur individuelle.

RÉGRESSION: DÉFINITION

- Il s'agit ici d'étudier le **lien entre 2 variables quantitatives**.
- La variable que l'on veut modéliser est appelée variable **a expliquer** ou variable **dépendante, réponse, diagnostique (médecine)**.
- La ou les variables qui sont utilisées pour modéliser la variable a expliquer sont appelées variables **explicatives** ou variables **indépendantes** (ce terme est à éviter), **imposée** ou **symptômes** (en médecine).

RÉGRESSION

- Lorsque les valeurs prises par une variable explicative sont choisies par l'expérimentateur, on dit que la variable explicative est **contrôlée** (on parle encore de **facteur contrôlé**). Lorsque les valeurs ne sont pas choisies, mais simplement mesurées, on parle de variables **non contrôlées**.
- Les paramètres qui interviennent dans les formules de modélisation s'appellent **coefficients** du modèles.
- La partie non expliquée désignée dans les formules par ε est appelée "**reste**" ou "**résidu**" ou "**erreur**" du modèle.

DÉMARCHE POUR LA RÉGRESSION

La régression comporte 4 étapes:

- ① Choix d'un modèle $Y = f(X)$;
- ② Détermination de la valeur numérique des paramètres du modèle;
- ③ Détermination de la signification statistique des paramètres du modèle.
- ④ Validation du modèle.

RÉGRESSION: OBJECTIF

- La régression est une forme de modélisation. Elle peut avoir plusieurs objectifs:
 - **Description:** trouver le meilleur modèle fonctionnel liant la variable dépendante y à la (aux) variable(s) indépendante(s) x . Estimer la valeur la plus probable des paramètres du modèle, ainsi que leur intervalle de confiance.
 - **Inférence:** tester des hypothèses précises se rapportant aux paramètres du modèle dans la population statistique: ordonnée à l'origine, pente(s).
 - **Prédiction:** prévoir ou prédire les valeurs de la variable dépendante pour de nouvelles valeurs de la (des) variable(s) indépendante(s).

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 regression:**

- Résume le mieux le nuage de point
 - \Rightarrow La plus proche de tous les points
 - \Rightarrow Erreurs ε petits + + +

PRINCIPE DE L'ESTIMATION

- Estimer α et β tel que ε petits +++
- ε_i : écart entre la droite et le point i

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

$$\mathbb{E}(Y/X) = \alpha + \beta \times X$$

$$\Rightarrow \varepsilon_i = y_i - \mathbb{E}(Y/X)$$

PRINCIPE DE L'ESTIMATION

- Somme des Carrés des Écarts

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

- Estimer α et β tel que:

SCE minimum

ESTIMATION DE LA PENTE β

- La pente β est donnée par la formule suivante:

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

- La variance de X est estimé par (dans le cas d'un échantillon):

$$S^2(X) = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^n (x_i)^2 - n(\bar{X})^2}{n-1}.$$

- La covaïance de X et Y est estimé par:

$$\widehat{cov(X,Y)} = \frac{\sum_{i=1}^n (x_i y_i) - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n-1}.$$

ESTIMATION DE α :

- La droite passe par m_Y et m_X :

$$m_Y = a + bm_X$$

où $m_Y = \bar{Y} = \frac{\sum y_i}{n}$ et $m_X = \bar{X} = \frac{\sum x_i}{n}$

D'où:

$$a = m_Y - bm_X$$

EXEMPLE

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

$$\text{cov}(\text{pressionsanguinesystolique}, \text{Age}) = \text{Cov}(X, Y) = 160.4242$$

- Variance de l'âge: $\text{var}(\text{Age}) = S^2(X) = 140.9697$
- Estimation de β

$$b = \text{cov}(\text{pression}, \text{Age}) / \text{var}(\text{Age}) = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = 1.138005$$

- Estimation de α :

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

L'équation s'écrit donc:

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

où

$$\mathbb{E}(\text{Pression sanguine systolique} / \hat{\text{Age}}) = 80.778 + 1.138 \times \hat{\text{Age}}$$

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 $\varepsilon_i = y_i - \hat{y}_i$.

INTERPRÉTATION

- ① $\beta = 0$: pas de lien, évolutions indépendantes.
- ② $\beta < 0$: évolutions en sens contraire.
- ③ $\beta > 0$: évolutions dans le même sens
- ④ Ordonnée à l'origine: α

$$\mathbf{E}(Y/X = 0) = \alpha$$

PRÉDICTION

- Pour un âge (X) fixé, prédiction de la pression sanguine systolique (Y)

$$Y_p = a + b \times X$$

$$\text{pression sanguine systolique} = 80.778 + 1.138 * \text{Âge}$$

COEFFICIENT DE CORRÉLATION

- Le modèle est-il un bon résumé des observations ?
- Estimation du coefficient de corrélation entre X et Y :

$$r = cor(X, Y) = \frac{cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

- $|r| \leq 1$ et $|r| = 1$ ssi $y_i = a + bx_i$ pour certain $i = 1, \dots, n$.
- $r = 1$ et $r = -1$ dénotent une corrélation parfaite entre X et Y .
- $r > 0$ ssi la droite de régression est de pente positive (relation croissante: X et Y varient dans le même sens).
- $r < 0$ ssi la droite de régression est de pente négative (relation décroissante: X et Y varient dans le sens contraire).
- $r = 0$ ssi la droite de régression est horizontale: aucune tendance ne peut être déterminée .

ADÉQUATION

- Pourcentage de variance expliquée: pour interpréter les valeurs intermédiaires de r , nous avons l'égalité suivante:

$$\begin{aligned} R^2 &= \frac{\text{Part de variance expliquée par la régression}}{\text{Variance totale}} \\ &= \frac{\text{ecart}(m_{Y/X} - m_Y)}{\text{ecart}(y - m_Y)} = \frac{\sum (m_{Y/X} - m_Y)^2}{\sum (y_i - m_Y)^2} \end{aligned}$$

- Donc r^2 est la proportion de la dispersion des Y qui est expliquée par la dispersion des X .
- **Remarque:** R : estimation du coefficient de corrélation entre X et Y .

Exemple:

- Coefficient de corrélation entre X et Y

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

- 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.

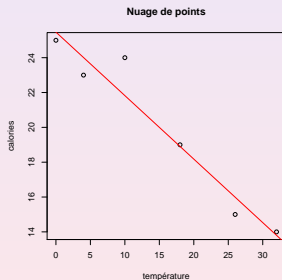
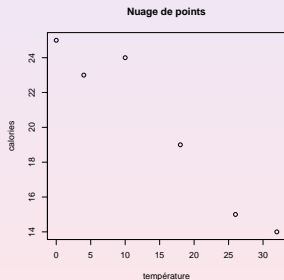
EXEMPLE 2

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

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

NUAGE DE POINTS

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



$$r = \text{cor}(\text{temperature}, \text{calories}) = -0.9682108.$$

RÉGRESSION NON LINÉAIRE

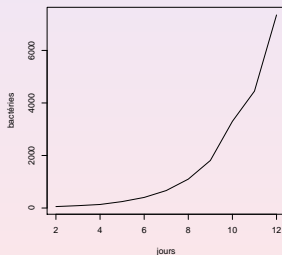
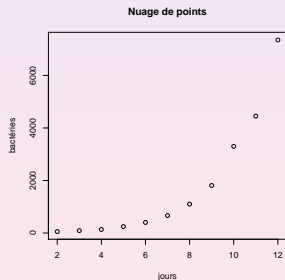
Exemple 3:

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 2^{ième} 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

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



MODÈLE NON LINÉAIRE

- D'après le graphique le nombre de bactéries croît 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

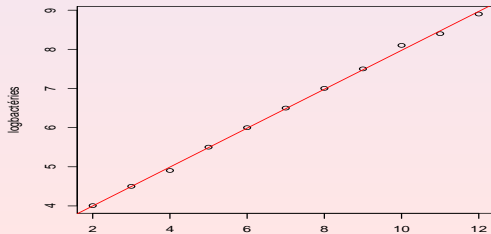
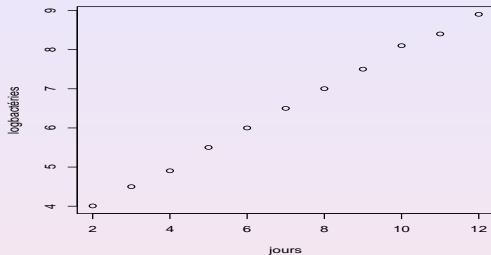
$$r = \text{cor}(\text{jours}, \text{bactries}) = 0.86474$$

- (!!!) (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 $y = \log(N)$ et $x = t$, le graphique suivant montre qu'il y a une relation linéaire en y et x .
- `logbactéries<-log(bactéries)`

NUAGE DE POINTS ET LA DROITE DE RÉGRESSION



ESTIMATION DES PARAMÈTRES DU MODÈLE

- Le coefficient de corrélation linéaire est:

$$r = \text{cor}(\text{jours}, \text{logbactries}) = 0.9996615$$

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

DROITE DE RÉGRESSION: ESTIMATION DES PARAMÈTRES DU MODÈLE

- `z4<-lm(logbactéries jours)`

- `z4`

Call: `lm(formula = logbactéries jours)`

Coefficients: (Intercept) jours 3.0142 0.4944

- Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.014162	0.032947	91.49	1.13e-14 ***
jours	0.494419	0.004289	115.27	1.41e-15 ***

- Residual standard error: 0.04499 on 9 degrees of freedom
- Multiple R-squared: 0.9993, Adjusted R-squared: 0.9992

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 :

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

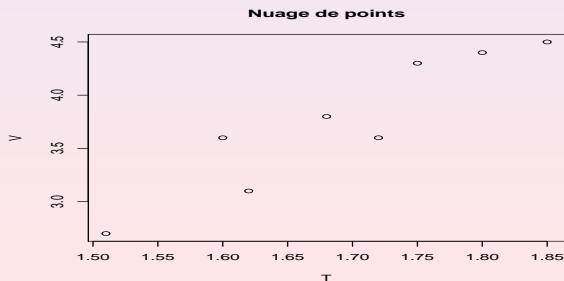
NB: $\sum t_i = 13.53$; $\sum t_i^2 = 22.9703$; $\sum v_i = 30$; $\sum v_i^2 = 115.36$;

$\sum t_i v_i = 51.205$.

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 V_{ems} :

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



SOLUTION

2. On a $n = 8$. $corr(T, V) = \frac{Cov(T, V)}{S_T \times S_V}$.

- $\bar{T} = \frac{1}{n} \sum t_i = 1.6913$, $Var(T) = \frac{1}{n-1} [\sum t_i^2 - n \times (\bar{T})^2] = 0.0125$ et $S_T = \sqrt{T} = 0.112$.
- $\bar{V} = \frac{1}{n} \sum v_i = 3.75$, $Var(V) = \frac{1}{n-1} [\sum v_i^2 - n \times (\bar{V})^2] = 0.409$ et $S_V = \sqrt{V} = 0.639$.
- $Cov(T, V) = \frac{1}{n-1} [\sum t_i v_i - n \times \bar{T} \times \bar{V}] = 0.0668$.

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

SOLUTION

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

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

où $b = \frac{\text{Cov}(T,V)}{\text{Var}(T)}$ et $a = \bar{V} - a \times \bar{T}$.

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 = $4 - 3.794 = 0.206$.