Corrélation - Régression

Dr Mathieu Fauvernier, d'après un cours du Pr Delphine Maucort-Boulch

Service de Biostatistique, HCL
UMR CNRS 5558

PASS UFR Lyon Est

Plan

Plan du cours

Rappels de probabilités

Espérance

Soit X une variable aléatoire (VA), on note $E(X)$ (ou μ_X) son espérance mathématique. C'est à dire la moyenne des valeurs prises par la variable, pondérée par les probabilités correspondantes.

$$\begin{aligned} \mu_X &= E[X] \\ &= \sum_{i=1}^{\infty} x_i P(X = x_i) \qquad \text{(cas discret)} \\ &= \int_{-\infty}^{+\infty} x f(x) \mathrm{d}x \qquad \text{(cas continu)} \end{aligned}$$

Rappels de probabilités

Variance

$var(X)$ (ou σ_X^2) mesure la dispersion d'une VA X autour de son espérance mathématique (moyenne) $E(X)$ (ou μ_X)

$$\begin{aligned} \sigma_X^2 &= E[(X - \mu_X)^2] \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

Rappels de statistiques

Échantillonnage et estimation

- On recrute n individus **représentatifs** de la population d'étude
- On mesure une variable X chez chaque sujet. On obtient n **réalisations** de cette variable : $(x_1, \ldots, x_n)$
- On cherche à estimer des paramètres de la loi de X : espérance, variance, ...

Rappels de statistiques

Estimation de la vraie moyenne μ_X

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

Estimation de la vraie variance σ_X^2

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_X)^2$$

Plan du cours

Cadre

Différentes notions

- Corrélation entre cancer du poumon et tabagisme
  $\rightarrow$ 2 VA qualitatives
- Corrélation entre poids à la naissance et sexe
  $\rightarrow$ 1 VA quantitative et 1 VA qualitative
- **Corrélation entre poids et taille à la naissance**
  $\rightarrow$ **2 VA quantitatives (objet de ce cours)**

Remarque

Corrélation entre des mesures faites avec 2 appareils différents $\rightarrow$ concordance entre 2 méthodes

Définition

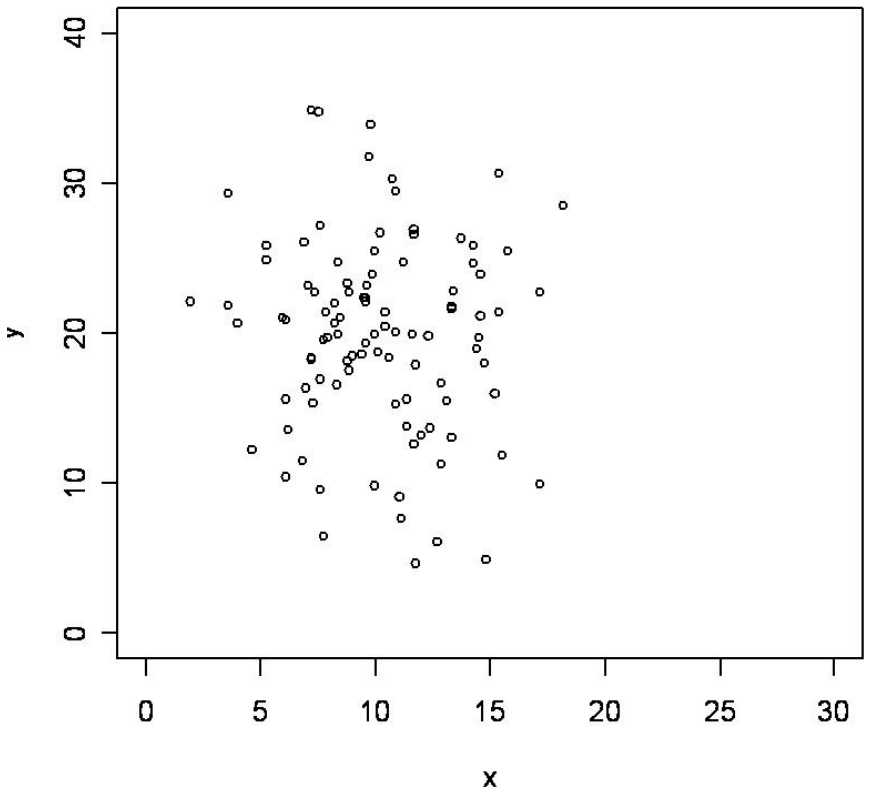
Corrélation

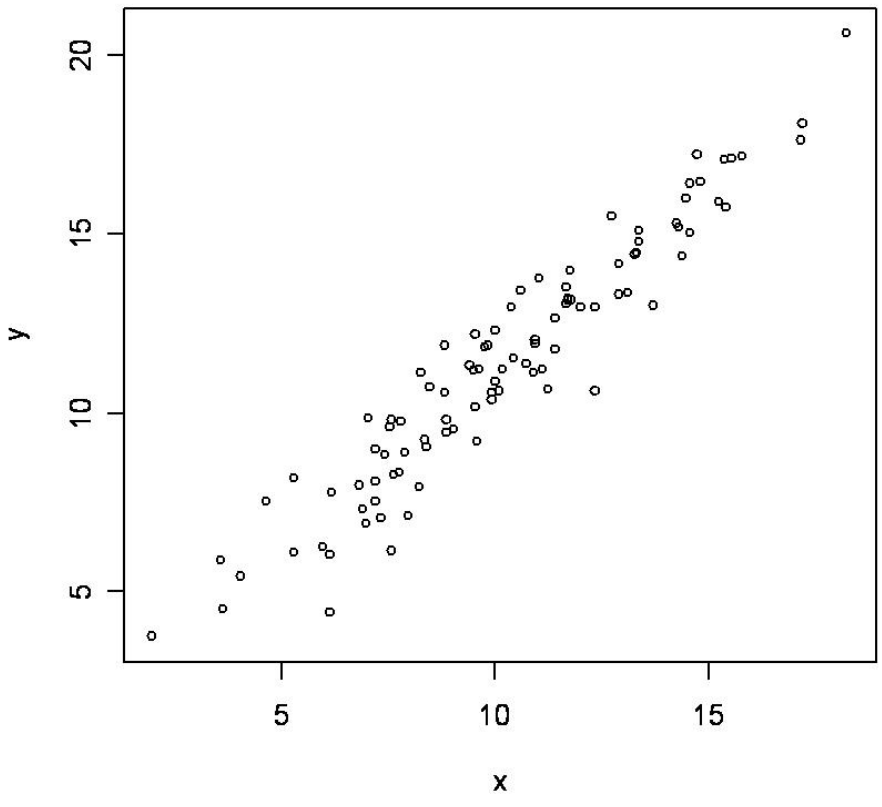
- Co-relation = dépendance réciproque de deux phénomènes qui varient conjointement
- Degré de liaison de deux variables aléatoires quantitatives X et Y
- Outils pour l'analyse :
  - analyse graphique : Nuage de points
  - quantification mathématique : covariance et coefficient de corrélation

Très important

Dans le cadre de la corrélation, les variables X et Y jouent des **rôles symétriques**. On n'étudie pas l'influence de l'une sur l'autre (contrairement à la régression).

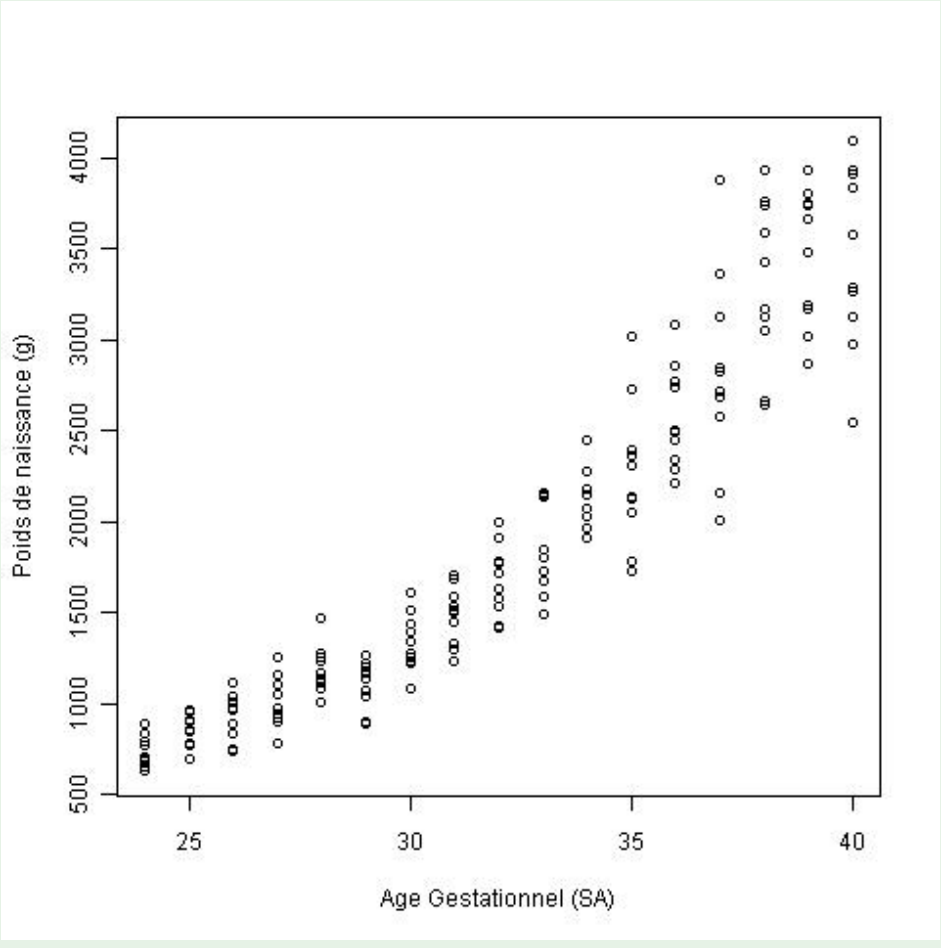
Analyse graphique : nuages de points





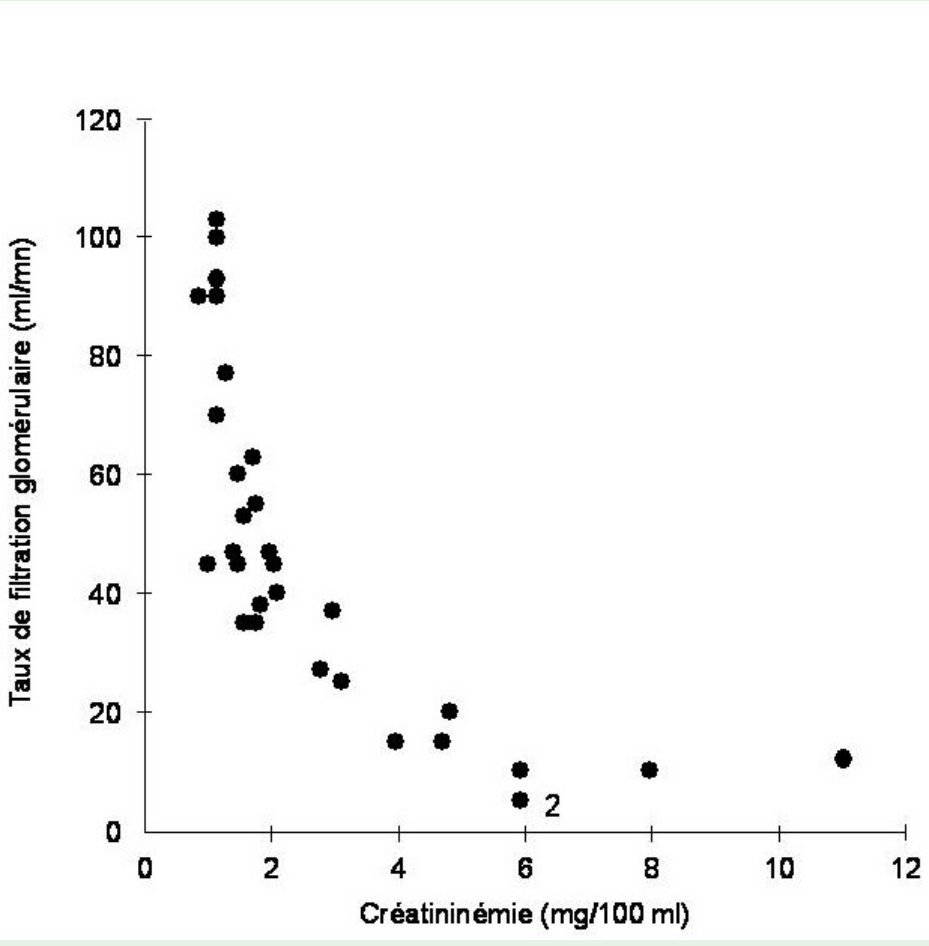
Exemple 1

Age gestationnel et poids à la naissance



Exemple 2

Taux de filtration glomérulaire et créatininémie



Quantification de la relation

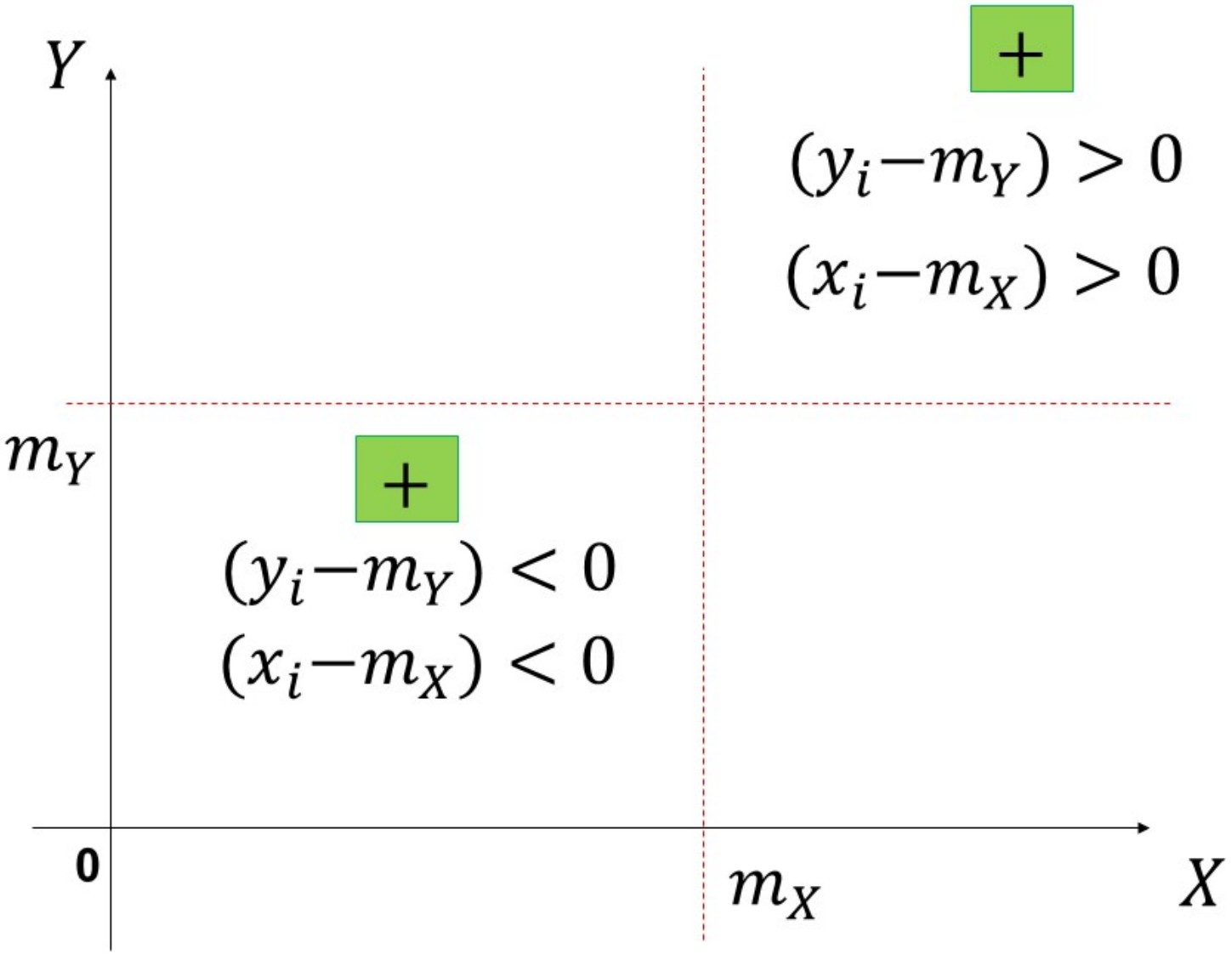
Covariance de deux variables X et Y

$$\begin{aligned} cov(X, Y) = \sigma_{X,Y} &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

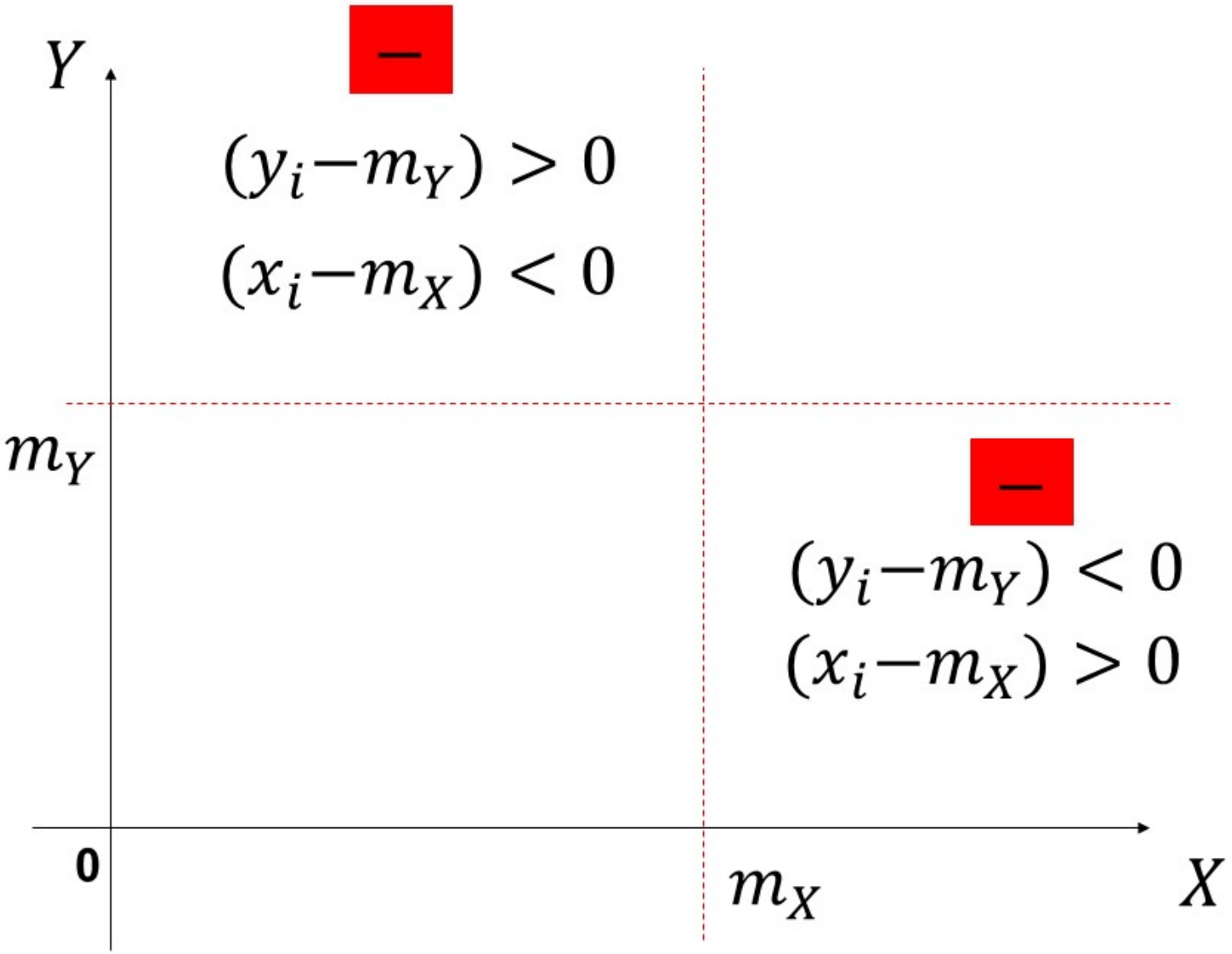
Estimation de la vraie covariance à partir d'un échantillon

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m_X)(y_i - m_Y)$$

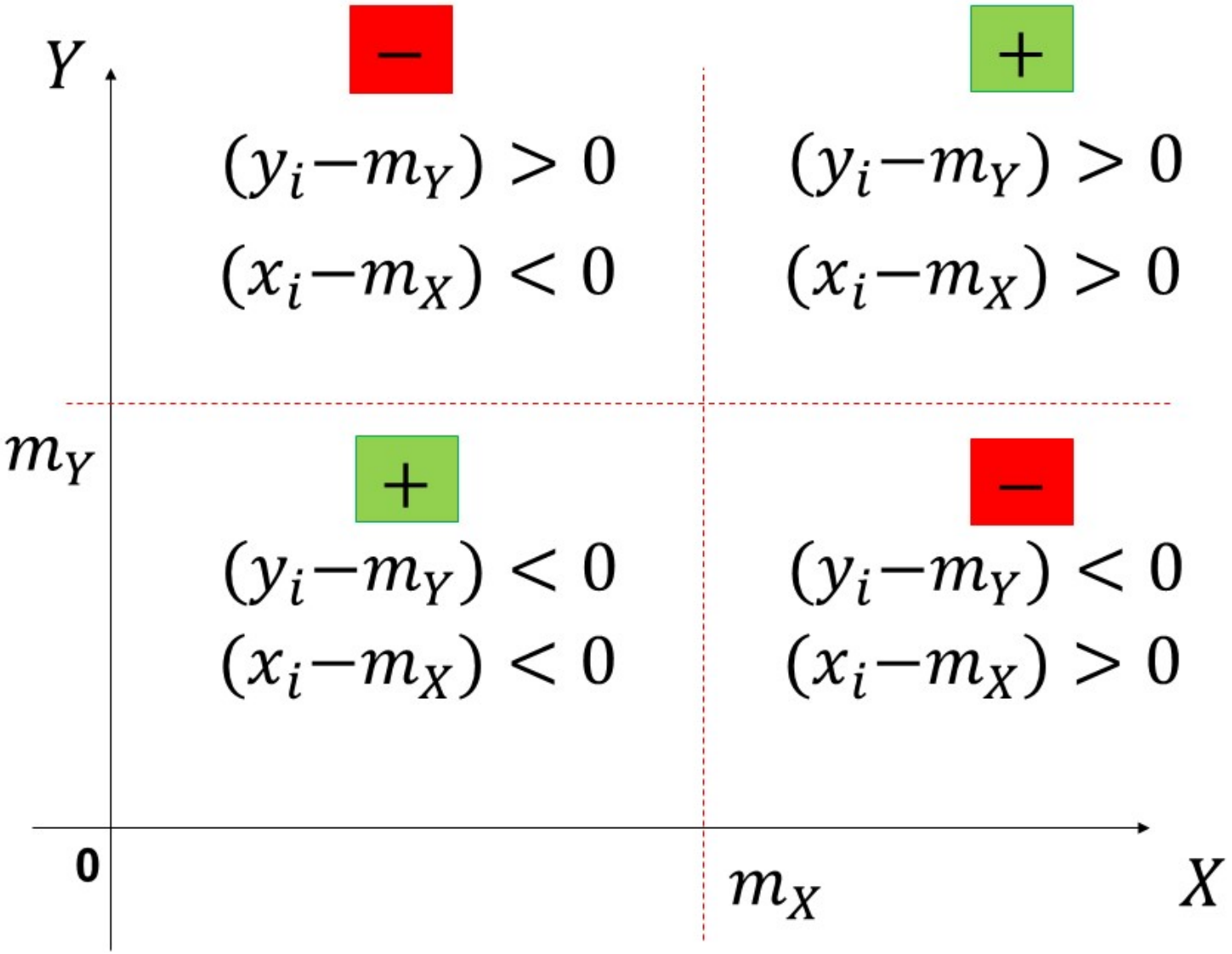
Interprétation



Interprétation



Interprétation



Covariance

Propriétés

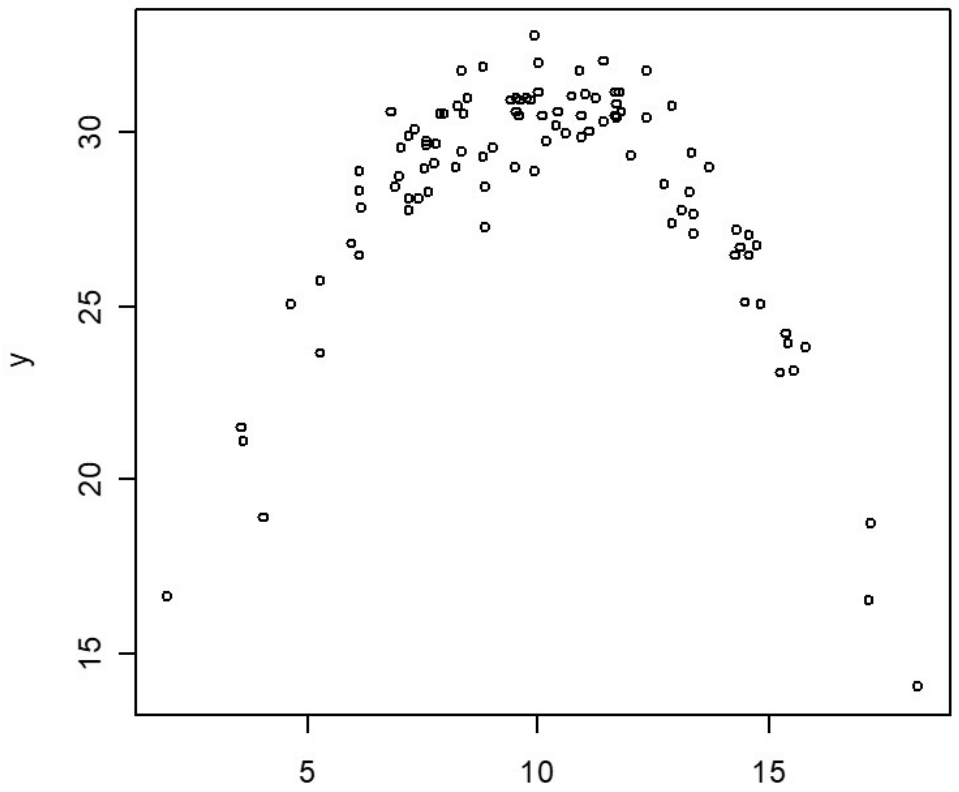
- cov(X,Y) = cov(Y,X)
- cov(aX,Y)=a cov(X,Y), a constante
- cov(X,X)=$\sigma_{X,X}$=σ_X^2 = var(X)
- var(X+Y) = var(X)+ var(Y) + 2 cov(X,Y)

Covariance

Propriétés

- X et Y indépendantes $\Rightarrow$ cov(X,Y) = 0

l'inverse n'est pas nécessairement vrai



Exemple

Age gestationnel et poids à la naissance (g)

AG (SA)	36	37	38	39	40
PN (g)	2589	2868	3133	3360	3480

Covariance

$$
\begin{aligned}
m_{AG} &= \frac{36+37+38+39+40}{5} = 38 \\
m_{PN} &= \frac{2589+2868+3133+3360+3480}{5} = 3086 \\
s_{AG,PN} &= \frac{\begin{array}{r}(36-38)(2589-3086) + (37-38)(2868-3086) + (38-38)(3133-3086) \\ + (39-38)(3360-3086) + (40-38)(3480-3086)\end{array}}{4} \\
&= 568.5(\text{g.SA})
\end{aligned}
$$

En moyenne: AG $\nearrow \Rightarrow$ PN $\nearrow$

Exemple

Age gestationnel et poids à la naissance (**kg**)

AG (SA)	36	37	38	39	40
PN (kg)	2.589	2.868	3.133	3.360	3.480

Covariance

$$s_{AG,PN} = 0.5685(\text{kg.SA})$$

En moyenne: AG $\nearrow \Rightarrow$ PN $\nearrow$

Limite de la covariance

La covariance dépend des unités de mesure des variables ! Elle n'a donc **pas de valeurs limites** et ne permet pas de réaliser des comparaisons.

Coefficient de corrélation de Pearson (1)

Definition

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

Estimation de $\rho_{X,Y}$

$$r_{X,Y} = \frac{s_{XY}}{s_X s_Y}$$

Avec $s_X = \sqrt{s_X^2}$ et $s_Y = \sqrt{s_Y^2}$

Coefficient de corrélation de Pearson (2)

Propriétés

- $-1 \leq \rho_{X,Y} \leq 1$
- $\rho_{X,Y} = \rho_{Y,X}$
- Signe de $\rho_{X,Y}$ = signe de $\sigma_{X,Y}$
- Pas d'unité
- Si X et Y indépendantes alors $\rho_{X,Y} = 0$

Cas particulier

Si X et Y $\sim \mathcal{N}$ alors $\rho_{X,Y} = 0 \Rightarrow$ X, Y indépendantes

Test du coefficient de corrélation linéaire

Hypothèses de validité

- (X,Y) $\sim \mathcal{N}_2$
- Ou
  1. Y $\sim \mathcal{N}$ avec σ^2_Y cste $\forall x$ et vice versa
  2. Relation linéaire entre X et Y

Statistique de test (paramétrique)

$\mathcal{H}_0 : \rho = 0$ $\mathcal{H}_1 : \rho \neq 0$

Sous $\mathcal{H}_0$

$$t = \frac{|r - \rho|}{s_r} = \frac{|r|}{s_r} = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} \sim Student_{n-2\text{ddl}}$$

Avec $s_r = \sqrt{\frac{1-r^2}{n-2}}$ (écart-type estimé de l'estimateur de ρ)

- $|t| \geq t_{\alpha,\ n-2\text{ddl}}$ $\rightarrow$ on rejette $\mathcal{H}_0$ au seuil α
- $|t| < t_{\alpha,\ n-2\text{ddl}}$ $\rightarrow$ on ne peut pas rejeter $\mathcal{H}_0$

Application (1)

Age gestationnel et poids à la naissance

- s_X^2=2.5
  s_Y^2=131763.5
- n=5
- $r = \frac{568.5}{\sqrt{2.5 * 131763.5}}$=0.99

Test

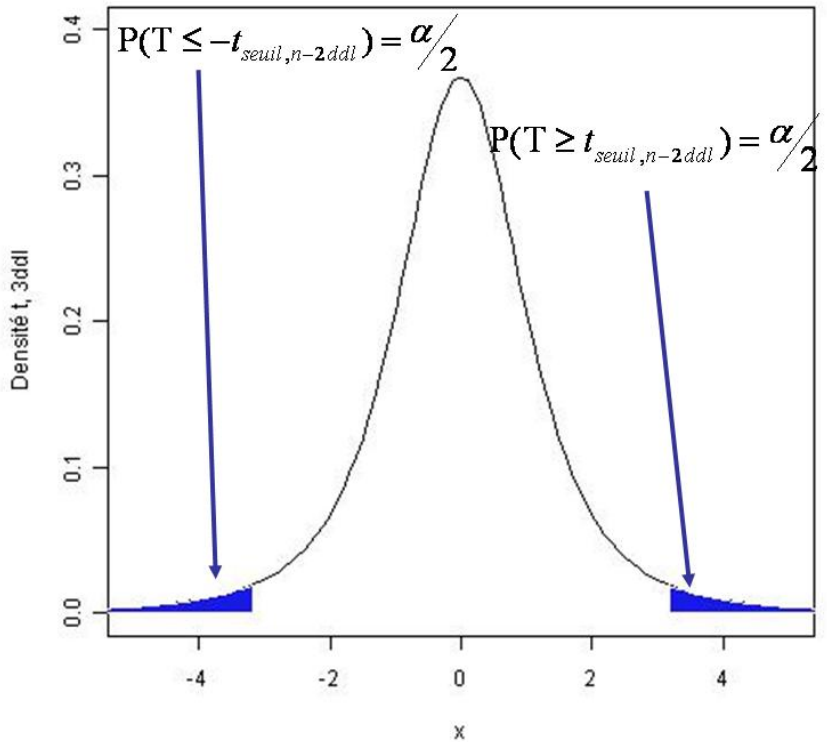
- $\mathcal{H}_0$: $\rho = 0$, $\mathcal{H}_1$: $\rho \neq 0$
- $t = \frac{0.99 * \sqrt{5-2}}{\sqrt{1-0.99^2}} = 12.16$
- Test à 3 ddl, risque α= 0.05

Application (2)

Lecture dans la table de la loi de Student

$$P(|T| \geq t_{\alpha,n-2ddl}) = \alpha \Leftrightarrow$$

$$P(T \geq t_{\alpha,n-2ddl}) = \alpha/2 \text{ ou } P(T \leq -t_{\alpha,n-2ddl}) = \alpha/2$$



Application (3)

Lecture dans la table de la loi de Student (cf. cours compa. moyennes

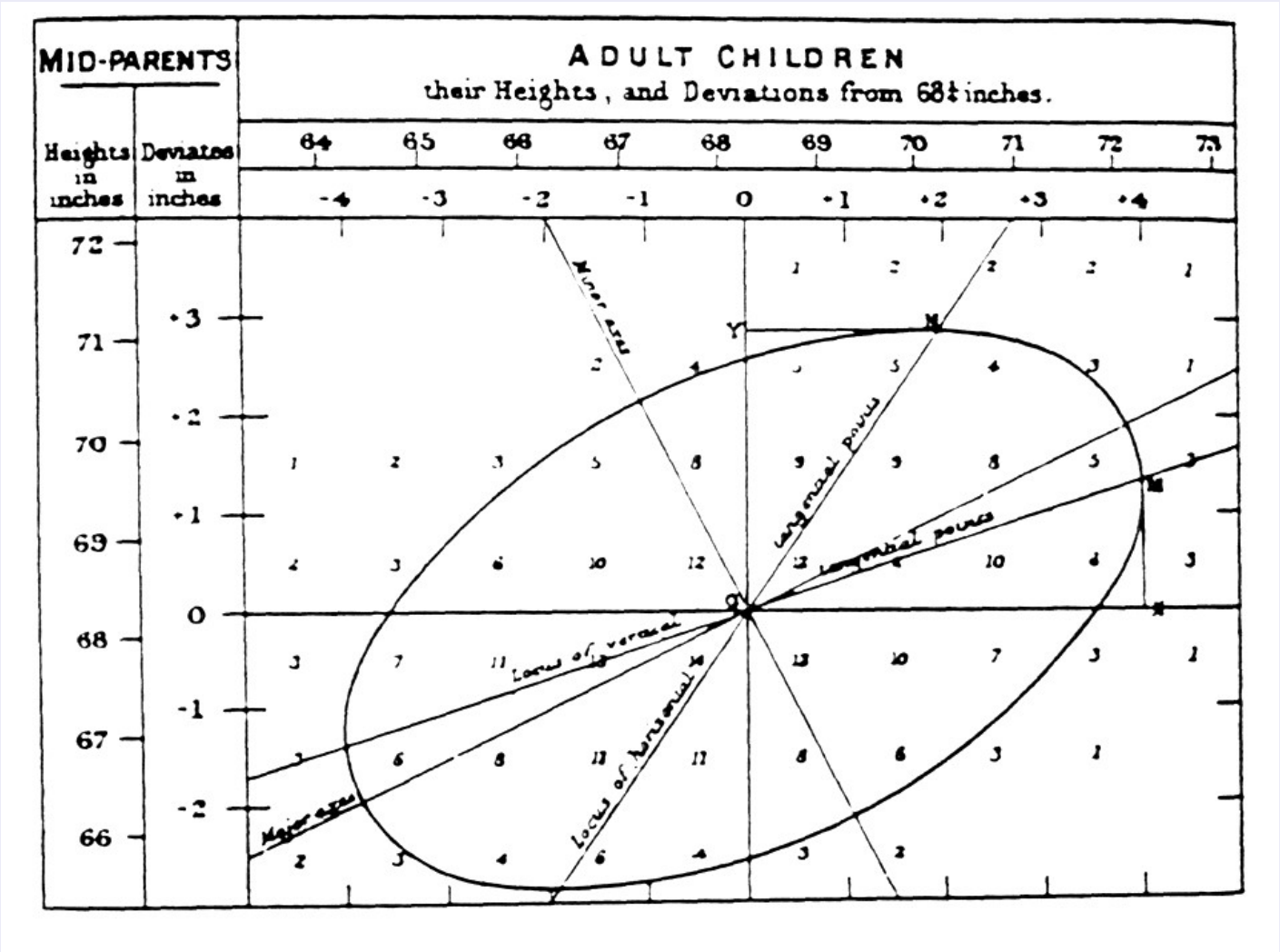
ddl \ p	0,9	...	0,1	0,05	0,02	0,01	0,005	0,001
1	0,1584		6,3138	12,7062	31,8205	63,6567	127,3213	636,6192
2	0,1421		2,9200	4,3027	6,9646	9,9248	14,0890	31,5991
3	0,1366		2,3534	3,1824	4,5407	5,8409	7,4533	12,9240

Avec 3 ddl, le seuil de rejet de $\mathcal{H}_0$ est égal à 3.18 pour un risque consenti α=5%.
Or, t = 12.16 $>$ 3.18, on rejette donc l'hypothèse nulle et on conclut à une dépendance entre l'âge gestationnel et le poids à la naissance.
! Conditions de validité, robustesse du test

Plan du cours

Historique

Galton (1822-1911)



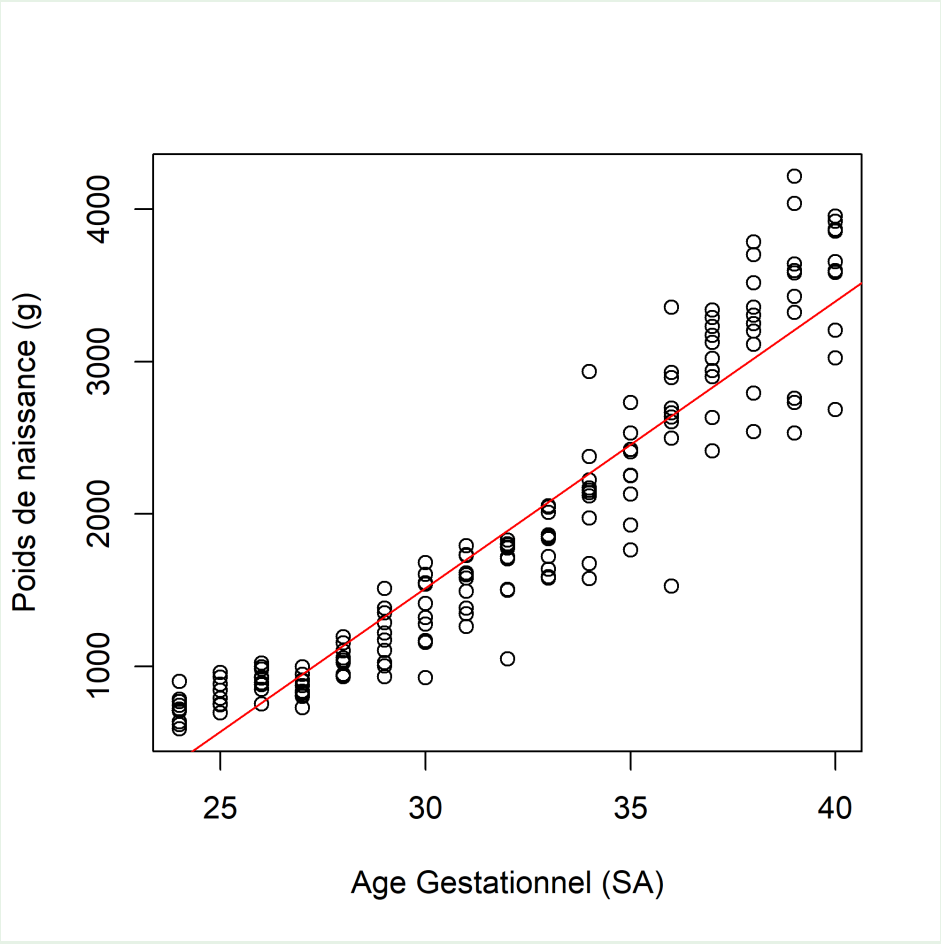
Régression linéaire simple

Définition

- 2 variables aléatoires X et Y
- L'une à expliquer=variable dépendante Y
- L'autre explicative=variable indépendante X
- Droite décrivant les variations de Y en fonction de X = droite de régression de Y en X

Exemple

Age gestationnel et poids à la naissance



Principe

Modèle théorique

Pour tous les individus i, on cherche à prédire la VA Y_i sachant que $X_i = x_i$. Le modèle **aléatoire** s'écrit :

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Avec

$$\epsilon_i \sim \mathcal{N}(0, \sigma) \quad \text{et} \quad \epsilon_i \quad \text{indépendants}$$

Autrement dit,

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma)$$

Principe

Valeurs prédites et résidus

Dans le cadre d'un échantillon, on dispose d'observations (x_i, y_i). La relation entre y_i et x_i s'écrit :

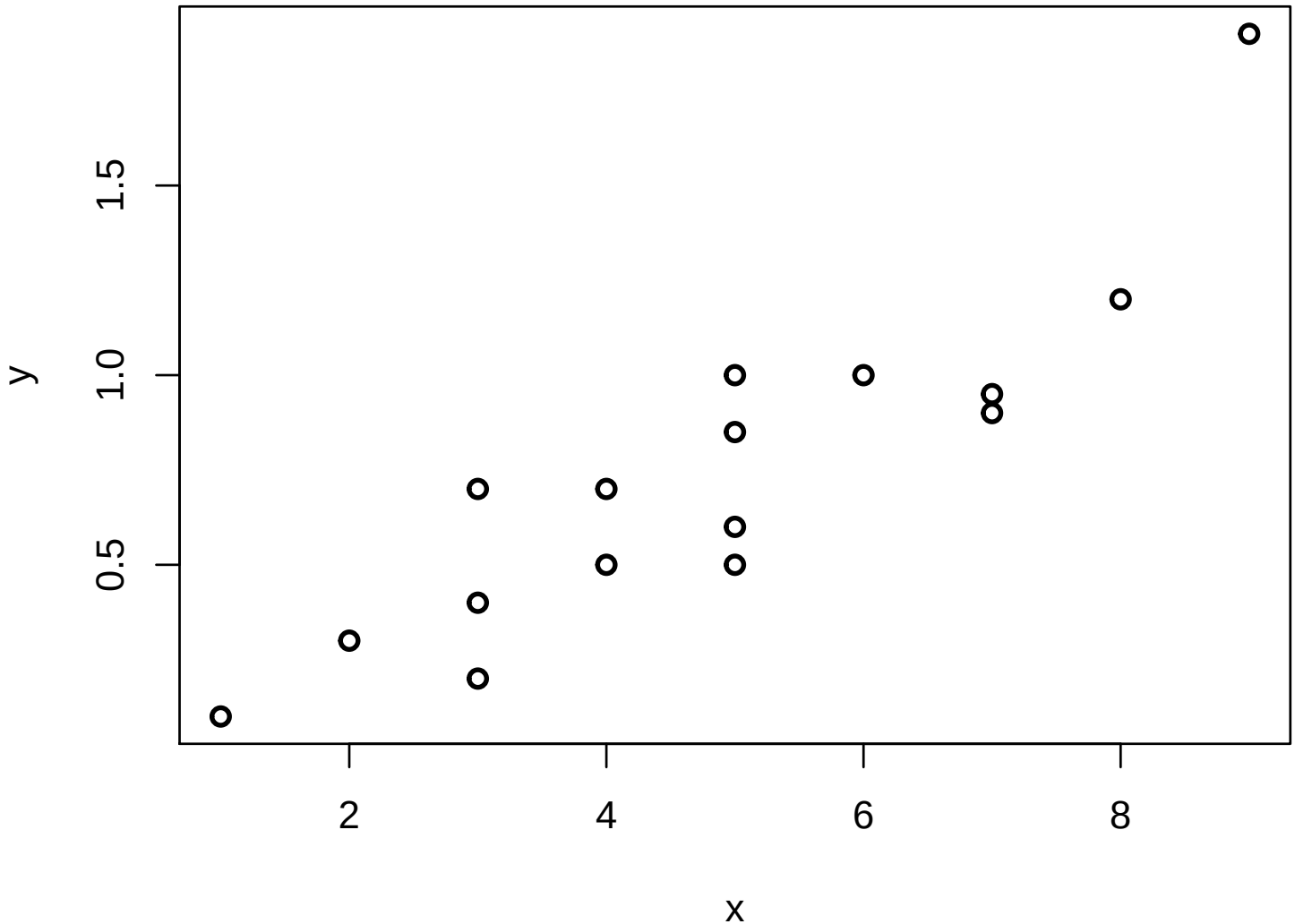
$$y_i = b_0 + b_1 x_i + e_i$$

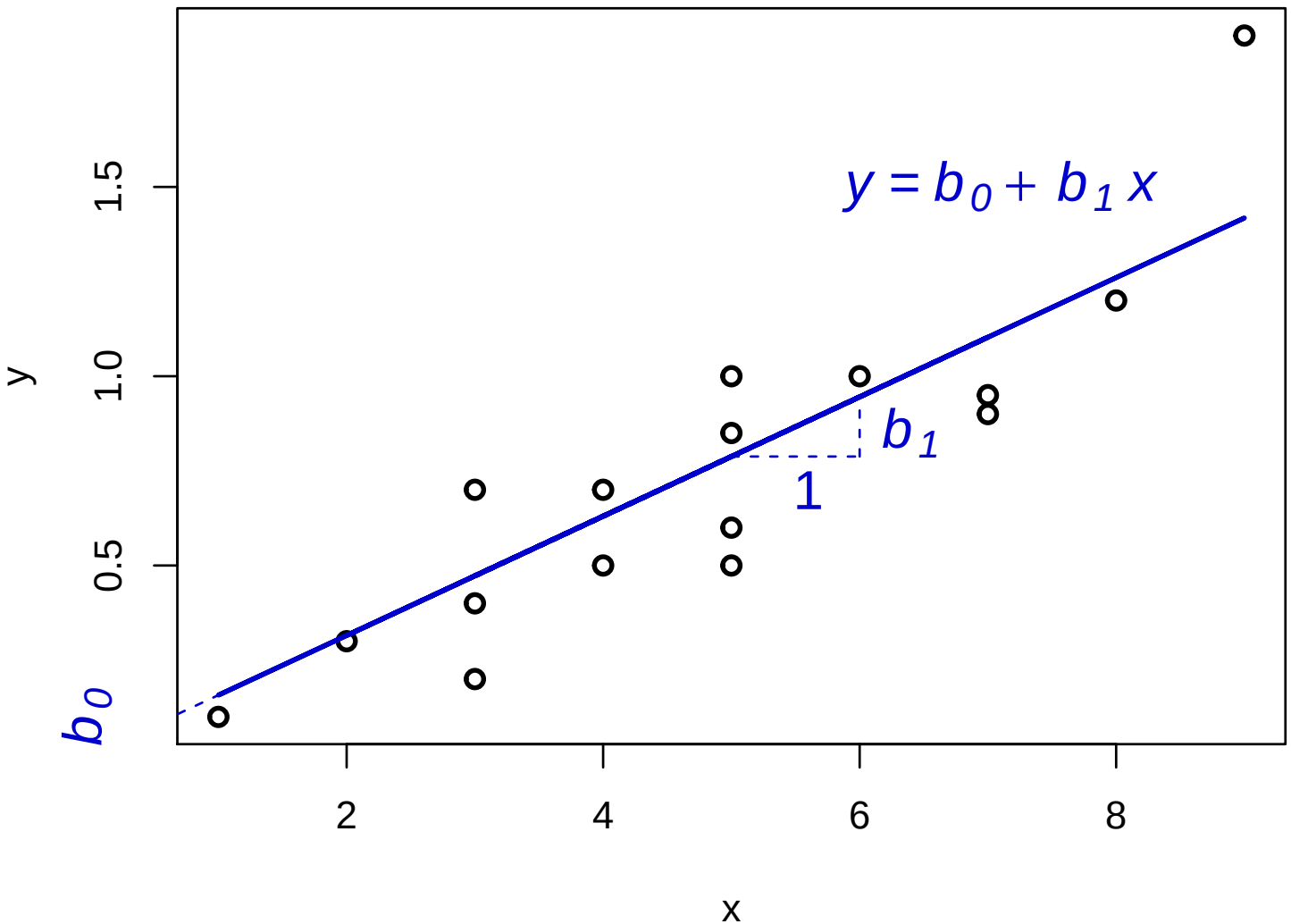
Avec

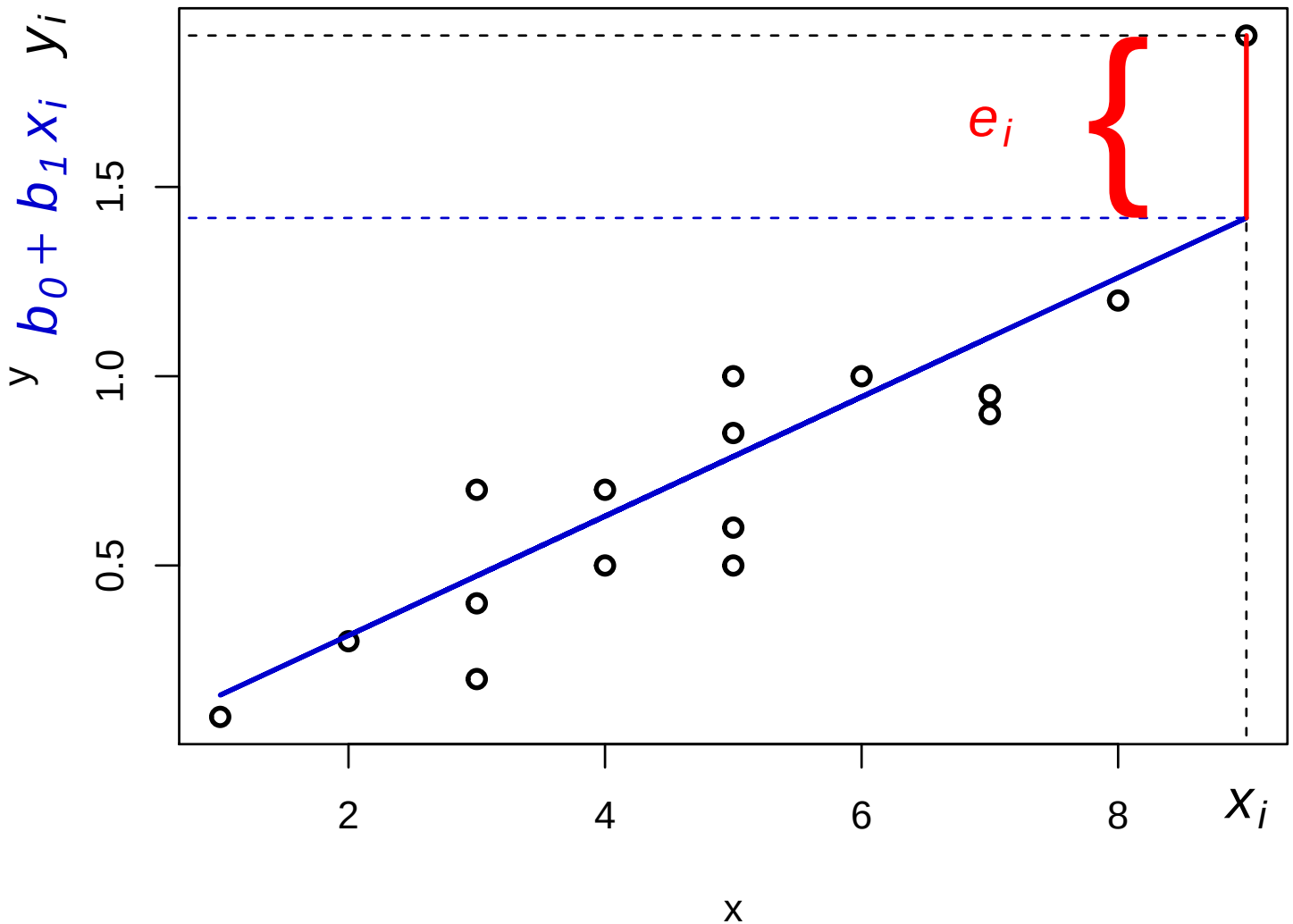
- b_0 : ordonnée à l'origine ou intercept (estimation de β_0)
- b_1 : pente (estimation de β_1)
- $b_0 + b_1 x_i$: valeur prédite par le modèle de y_i
- e_i : résidu (écart entre la valeur observée et la valeur prédite)

Objectif

Calculer b_0 et b_1 en minimisant les résidus e_i.


y
0.5
1.0
1.5
2
4
6
8
x


$y = b_0 + b_1 x$
b_1
1
b_0
y
x
0.5
1.0
1.5
2
4
6
8


y_i
$b_0 + b_1 x_i$
e_i
y
0.5
1.0
1.5
2
4
6
8
x_i
x

Méthode des Moindres Carrés Ordinaires

Principe

- Minimiser les écarts e_i
- Eviter que les écarts positifs et négatifs ne se compensent
- $E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1.x_i)^2$

Résolution

$$\begin{cases} \frac{\partial E}{\partial b_1} = 0 \\ \frac{\partial E}{\partial b_0} = 0 \end{cases}$$

Une solution unique $\begin{cases} b_1 = \frac{s_{XY}}{s_X^2} \\ b_0 = m_Y - b_1.m_X \end{cases}$

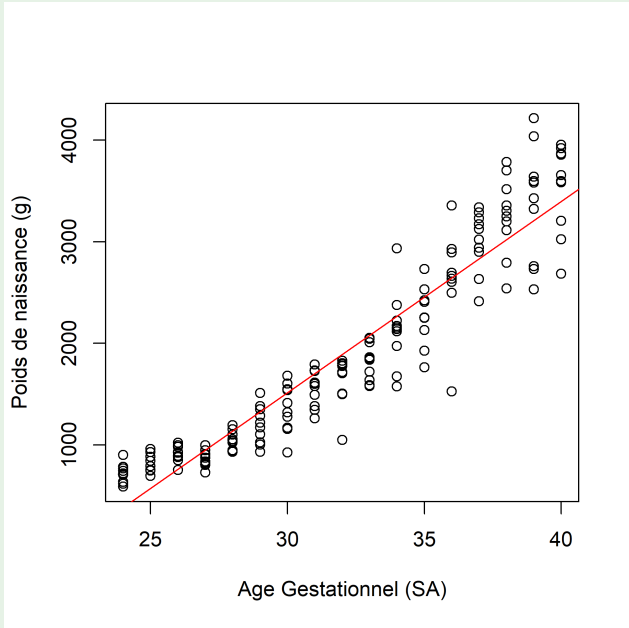
Remarques

- La droite passe par le point moyen (m_X, m_Y)
- b_0 représente la valeur prédite quand X vaut zéro
- b_1 correspond à l'augmentation de la valeur prédite quand X augmente d'une unité. En l'absence de relation entre X et Y, $b_1 = 0$
- e_i correspond à l'erreur commise en résumant y_i par la valeur prédite par le modèle $b_0 + b_1.x_i$

Age gestationnel et poids à la naissance

Droite de régression

- Droite de régression
  $PN = b_0 + b_1.AG$
- $b_1 = \frac{s_{XY}}{s_X^2} = \frac{4364.86}{24.14}$
  $= 180.80$
- $b_0 = m_Y - b_1.m_X$
  $= 1932.64 - 180.80 \times 32$
  $= -3852.94$



Un foetus prend en moyenne 180.8g par semaine d'aménorrhée supplémentaire
Le poids moyen estimé à 0 SA est de -3852,94g. Il n'a pas de sens ici!

Coefficients de régression et de corrélation

On sait que

$$b_1 = \frac{s_{XY}}{s_X^2}$$

Or,

$$r_{X,Y} = \frac{s_{XY}}{s_X s_Y}$$

Donc

$$r_{X,Y} = b_1 . \frac{s_X}{s_Y}$$

Tests (1)

La relation entre Y et X est-elle significative?

- $\mathcal{H}_0 \ \ \beta_1 = 0, \mathcal{H}_1 \ \ \beta_1 \neq 0$

$$t = \frac{b_1 - 0}{s_{b_1}} \ \sim \ t_{n-2\text{ddl}}$$

Avec $s_{b_1} = \sqrt{\frac{\frac{s_Y^2}{s_X^2} - b_1^2}{n-2}}$ (écart-type estimé de l'estimateur de β_1)

- $|t| \geq t_{\alpha,\, n-2\text{ddl}} \ \rightarrow$ on rejette $\mathcal{H}_0$ au seuil α
  $|t| < t_{\alpha,\, n-2\text{ddl}} \ \rightarrow$ on ne peut pas rejeter $\mathcal{H}_0$
- Tester la pente équivaut à tester la corrélation

Tests (2)

Lien entre test et intervalle de confiance

- Intervalle de confiance de la pente

$$IC_{(1-\alpha)}(\beta_1) = b_1 \pm t_{\alpha,\ n-2\text{ddl}} \cdot s_{b_1}$$

- Pente significativement différente de 0 si et seulement si $IC_{(1-\alpha)}(\beta_1)$ ne contient pas 0.

Tests (3)

La relation entre Y et X est-elle réellement linéaire? (1)

- Validité de la droite de régression
- Si la relation est linéaire, les résidus e_i ne contiennent plus d'information structurée $\Rightarrow$ exploration des résidus

Normalité des résidus

- Tests de normalité : Kolmogorov-Smirnov, Shapiro-Wilks ...
- Tests peu puissants $\Rightarrow$ procédure empirique: graphiques
- Droite de Henry: quantiles théoriques de la loi normale vs. quantiles de la distribution des résidus estimée sur les données

Tests (4)

La relation entre Y et X est-elle réellement linéaire? (2)

Homoscédasticité des résidus (variance constante)

- Répartition homogène des résidus, indépendante des valeurs prédites
- Tests formels
- Approche empirique graphique: résidus standardisés ($\frac{e_i}{s_{e_i}}$) en fonction des valeurs prédites

Indépendance des résidus

- Absence de corrélation entre les résidus
- Hors programme

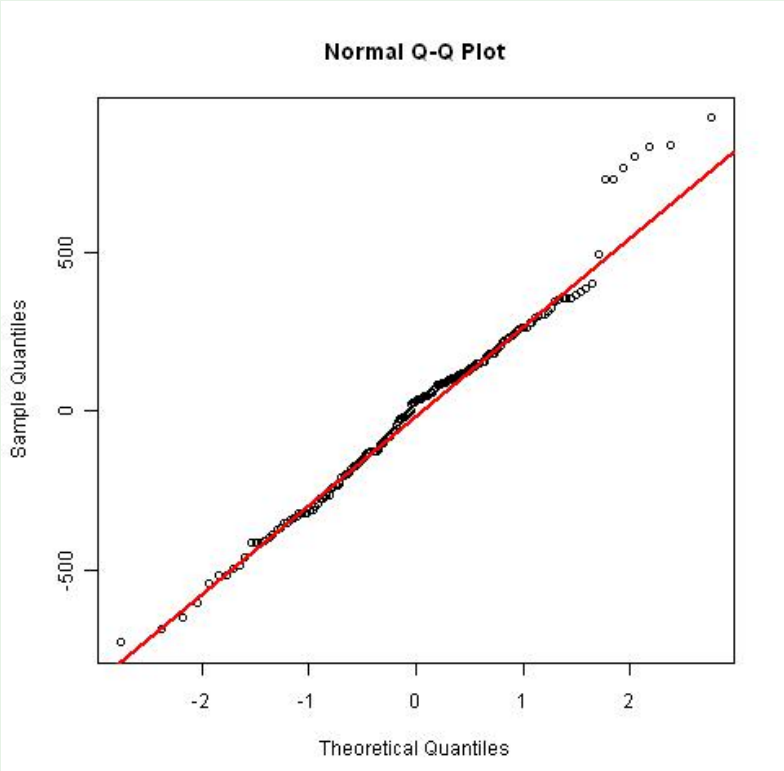
Age gestationnel et poids à la naissance

Significativité de la relation

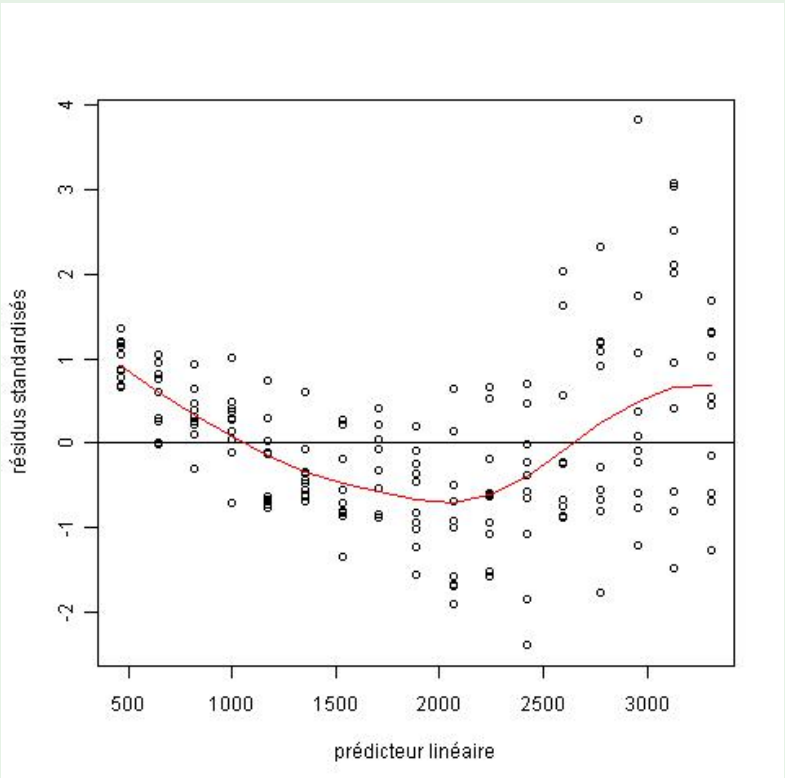
- $b_1 = 180.80$ $s_{b_1} = 4.72$
- Sous $\mathcal{H}_0$, $t = \frac{180.8}{4.72} = 38.34$
- ddl = $n - 2 = 168$
- $|t| > 1.96 = t_{5\%,\infty \text{ddl}}$
  Au risque 5%, on rejette donc $\mathcal{H}_0$. Le poids à la naissance dépend de l'âge gestationnel.
- Rappel : si $ddl > 100$ on applique les seuils de la loi normale (dernière ligne de la table de Student).

Age gestationnel et poids à la naissance

Droite de Henry



Résidus standardisés



Plan du cours

Comprendre et retenir

Comprendre

- Corrélation
  - Quantifie la force de la relation entre X et Y
  - Mesure de la relation symétrique
- Régression
  - Estime l'effet d'une variable explicative X sur l'évolution d'une variable à expliquer Y
  - Mesure de la relation asymétrique

Retenir

- Coefficient de corrélation
  - Compris entre -1 et 1
  - Test $\sim Student_{n-2}$ddl
- Régression linéaire
  - Pente liée au coefficient de corrélation
  - Test de la pente $\sim Student_{n-2}$ddl
  - Examen des résidus