UE 4 : CORRELATION – REGRESSION

Lien entre 2 variables quantitatives :

- H_0 : indépendance entre les 2 variables
- H_1 : dépendance entre les 2 variables

Corrélation = mesure la force du lien (information commune) entre 2 variables quantitatives.

- Corrélation positive : variation dans le même sens (ex : températures Montpellier et Nîmes)
- Corrélation négative : variation dans le sens contraire (ex : températures Nîmes et Melbourne)

Coefficient de corrélation $\rho_{X,Y}$:

$$\rho_{X,Y} = \frac{cov\,(X,Y)}{\sqrt{Var(X)\ \mathbf{x}\ Var(Y)}} = \frac{cov\,(X,Y)}{\sigma(X) * \sigma(Y)}$$

$-1 \leq \rho_{X,Y} \leq +1$

Indépendance entre les 2 variables → $\rho_{X,Y} = 0$
Lien entre les 2 variables (corrélation) → $\rho_{X,Y} \neq 0$

Indépendance → $\rho_{X,Y}$ = 0 pas l'inverse !!!

Estimateur r du coefficient de corrélation :

$$r = \frac{\sum(x_i - m_x) * (y_i - m_y)}{\sqrt{\sum(x_i - m_x)^2 * \sum(y_i - m_y)^2}}$$

Dans le formulaire

Statistique du test t : $t = r * \sqrt{\frac{n-2}{1-r^2}}$ et **t_α : dans la Table de Student à n-2 ddl**

n = nombre de mesures

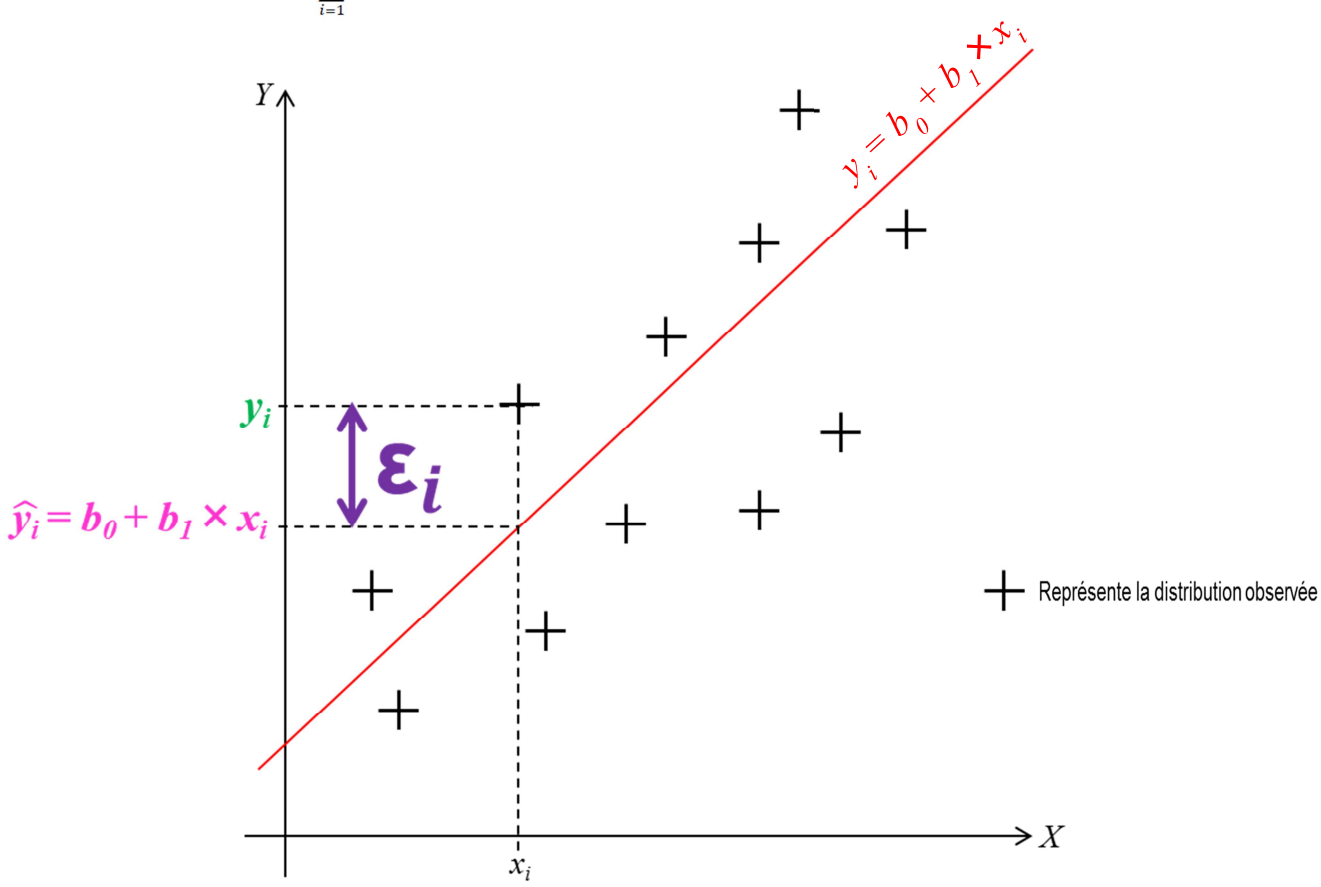
$t > t_\alpha$ → Rejet H_0 → Rejet de l'indépendance entre les 2 variables

Régression :

- X = variable explicative
- Y = variable à expliquer

- Hypothèses pour le modèle $Y = \alpha + \beta * X + \varepsilon$
  - La distribution de l'erreur ε est indépendante de X
  - α et β sont constants
  - l'erreur ε suit une loi normale centrée de variance constante σ^2 ($\varepsilon \sim \mathcal{N}(0 ; \sigma^2)$)

Méthode des moindres carrés = on cherche les valeurs b_0 et b_1 de α et β qui minimisent l'erreur $\varepsilon_i = y_i - \hat{y}_i = y_i - (\alpha + \beta * x_i)$.
On cherche donc à minimiser $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ et on obtient les estimateurs de α et β.



(Cette équation prédit la valeur de Y quand X=x)

Estimateur de **β** : $\mathbf{b_1} = \frac{S_{X,Y}}{S_X^2} = \frac{\text{cov (X,Y)}}{\text{Var (X)}} = r * \frac{S_X}{S_Y}$ avec S_X = estimateur sans biais de l'écart-réduit des x_i

Estimateur de **α** : $\mathbf{b_0} = \bar{y} - b_1 * \bar{x}$ avec $\bar{x}$ = moyenne des x_i

Test de la pente de la droite de régression : (équivalent au test de la corrélation nulle)

- $H_0 : \beta = 0$ et $H_1 : \beta \neq 0$
- $t_o = \frac{b_1}{S_b}$ suit une loi de Student à n-2 ddl

avec $\mathbf{S_b} = \sqrt{\frac{\frac{S_Y^2}{S_X^2} - b_1^2}{n-2}}$ $t_o > t_\alpha$ → Rejet H_0 → Pente significativement différente de 0 donc lien entre les 2 variables