

Corrélation et régression linéaire – Fiche de cours

1. Présentation

Pour évaluer l'association entre 2 variables quantitatives, on peut établir 3 niveaux :

- dépendance
- dépendance monotone (corrélation et régression linéaire)
- concordance

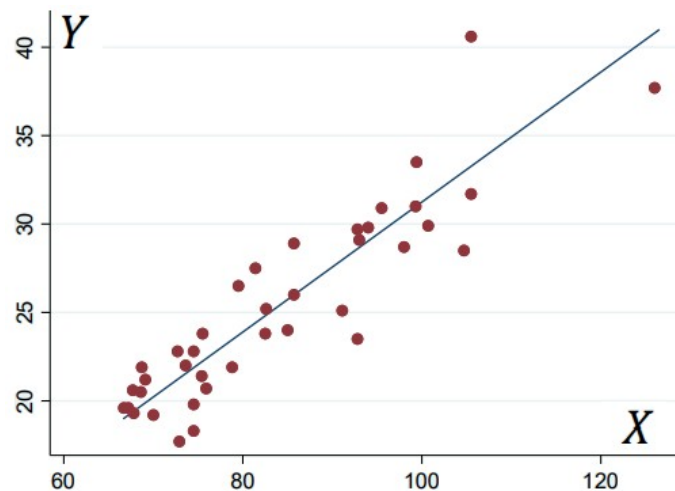
2. Coefficient de corrélation (Pearson)

a. Série statistique à 2 variables

On définit une série statistique double en observant deux critères sur une même population de dimension n :

$$x = (x_1, x_2, \dots, x_n) \quad y = (y_1, y_2, \dots, y_n)$$

Valeur x_i	x_1	x_2	...	x_n
Valeur y_i	y_1	y_2	...	y_n



b. Sur une population

- covariance

La covariance sur une série de dimension n liée à la population est définie par :

$$COV(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x) \cdot (y_i - \mu_y) = \frac{1}{n} \sum_{i=1}^n x_i \cdot y_i - \mu_x \cdot \mu_y$$

- variance et écart type

La variance et l'écart-type sur une série de dimension n liée à la population sur X est définie par :

$$V(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \mu_x^2 \quad \sigma_x = \sqrt{V(X)}$$

De la même manière on définit $V(Y)$ et $\sigma_y = \sqrt{V(Y)}$

- coefficient de corrélation

Le coefficient de corrélation linéaire

$$\rho = \frac{COV(X, Y)}{\sigma_x \sigma_y} \quad \rho \in [-1; 1]$$

Si ρ proche de -1 ou de 1 alors il y a association linéaire

c. Sur un échantillon

- covariance

La covariance sur une série de dimension n liée à un échantillon est définie par :

$$cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - m_x) \cdot (y_i - m_y) = \frac{1}{n-1} \sum_{i=1}^n x_i \cdot y_i - m_x \cdot m_y$$

- variance et écart type

La variance et l'écart-type sur une série de dimension n liée à un échantillon sur X est définie par :

$$V(X) = \frac{1}{n-1} \sum_{i=1}^n (x_i - m_x)^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - m_x^2 \quad s_x = \sqrt{V(X)}$$

De la même manière on définit $V(Y)$ et $s_y = \sqrt{V(Y)}$

- coefficient de corrélation

Le coefficient de corrélation linéaire

$$r = \frac{\text{cov}(X, Y)}{s_x \cdot s_y} \quad r \in [-1; 1]$$

Si r proche de -1 ou de 1 alors il y a association linéaire

d. Test de significativité du coefficient de corrélation linéaire

On souhaite estimer au risque α si le coefficient de corrélation obtenu lors d'un échantillonnage est un bon estimateur dans la population générale

Hypothèse nulle : $H_0: \rho = 0$ Hypothèse alternative : $H_1: \rho \neq 0$

$$\text{Statistique de test : } t_0 = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \quad \text{Seuil : } s = t_{(n-2); \frac{\alpha}{2}}$$

3. Droite de régression linéaire

a. Définition

La droite de régression linéaire par la méthode des moindres carrés de l'échantillon a pour expression : $y = a \cdot x + b$

L'équation de la droite pour la population a pour expression :

$$y = p \cdot x + q$$

a est un estimateur de p et b un estimateur de q

$$\text{avec } a = \frac{\text{cov}(X, Y)}{s_x^2} \quad \text{et } b = m_y - a m_x$$

x est la variable explicative (connue sans erreur)

y est la variable réponse (connue avec un écart ou un résidu)

b. Test de significativité de la pente de la droite de régression

On souhaite estimer au risque α s'il y a association entre les variables X et Y dans la population

Hypothèse nulle : $H_0: p = 0$ Hypothèse alternative : $H_1: p \neq 0$

$$\text{Statistique de test : } t_0 = \frac{a}{s_a} \quad \text{Seuil : } s = t_{(n-2); \frac{\alpha}{2}}$$

$$\text{avec } s_a^2 = \frac{\frac{s_y^2}{n-2} - a^2}{s_x^2}$$