

# Analyse de Corrélation et de Régression

## Application aux Sciences Médicales

### 1. Corrélation

#### Définition

La corrélation mesure l'association (linéaire ou non) entre deux variables quantitatives. Elle ne prouve pas un lien de causalité.

#### Coefficient de corrélation de Pearson ( $r$ )

- $r \in [-1, 1]$
- $r > 0$  : corrélation positive
- $r < 0$  : corrélation négative
- $r = 0$  : absence de corrélation linéaire

#### Interprétation

Force de la corrélation	Positive	Négative
Faible	$0.1 \leq r < 0.3$	$-0.3 < r \leq -0.1$
Modérée	$0.3 \leq r < 0.7$	$-0.7 < r \leq -0.3$
Forte	$r \geq 0.7$	$r \leq -0.7$

### 2. Régression Linéaire Simple

#### Objectif

Modéliser une variable  $Y$  (dépendante) à partir d'une variable explicative  $X$  (indépendante).

#### Modèle

$$Y = a + bX + \varepsilon$$

- $a$  : constante (ordonnée à l'origine)
- $b$  : pente (variation de  $Y$  pour une unité de  $X$ )
- $\varepsilon$  : erreur aléatoire

## Utilisation en médecine

- Étudier l'influence d'un facteur de risque (IMC, âge) sur une variable biologique (pression artérielle, cholestérol).
- Prédire une valeur clinique.

## 3. Exemple Médical

Étude de la relation entre l'IMC ( $\text{kg/m}^2$ ) et la pression artérielle systolique (PAS) en mmHg :

Patient	IMC	PAS
1	22	120
2	24	125
3	26	130
4	28	135
5	30	140

Corrélation de Pearson  $r \approx 1 \Rightarrow$  corrélation linéaire très forte.  
Modèle de régression :  $PAS = 100 + 1.33 \times IMC$

## 4. Exercices

### Exercice 1 : Corrélation

On étudie le lien entre la glycémie à jeun (mg/dL) et le tour de taille (cm) chez 10 patients diabétiques.

Glycémie	Tour de taille
110	85
115	90
120	93
130	95
125	92
118	89
135	96
140	99
145	102
150	104

#### Questions :

1. Tracer le nuage de points.
2. Calculer le coefficient de corrélation de Pearson.
3. Interpréter le résultat obtenu.

## Exercice 2 : Régression linéaire simple

Prédiction du cholestérol (mg/dL) en fonction de l'âge (années) :

Âge	Cholestérol
35	180
40	190
45	195
50	200
55	210
60	220

### Questions :

1. Déterminer l'équation de la droite de régression.
2. Estimer le cholestérol pour un patient de 48 ans.
3. Interpréter la pente du modèle.

## Exercice 3 : Interprétation médicale

Une étude trouve  $r = 0.75$  entre le score de dépression et le taux de cortisol chez 150 patients.

### Questions :

1. Qualifier la force de la relation.
2. Peut-on affirmer que le cortisol cause la dépression ?
3. Quelle méthode utiliser pour explorer une relation causale ?

## 5. Régression Linéaire Multiple (introduction)

Lorsque plusieurs facteurs influencent une variable :

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n + \varepsilon$$

**Exemple :** prédire la pression artérielle à partir de l'âge, de l'IMC et du niveau d'activité physique.

*Pour les corrections ou les scripts en R/Python, contactez l'enseignant ou le responsable du module.*

# Solutions des Exercices d'Analyse de Corrélation et Régression

## Exercice 1 : Corrélation entre la glycémie à jeun et le tour de taille

### 1. Nuage de points

### 2. Calcul du coefficient de corrélation de Pearson

Nous utilisons la formule :

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}}$$

Avec les données fournies, nous calculons :

- $\sum X = 85 + 90 + \dots + 104 = 945$
- $\sum Y = 110 + 115 + \dots + 150 = 1288$
- $\sum XY = 110 \times 85 + 115 \times 90 + \dots + 150 \times 104 = 122570$
- $\sum X^2 = 85^2 + 90^2 + \dots + 104^2 = 89761$
- $\sum Y^2 = 110^2 + 115^2 + \dots + 150^2 = 167424$
- $n = 10$

En substituant dans la formule :

$$r \approx \frac{10 \times 122570 - 945 \times 1288}{\sqrt{(10 \times 89761 - 945^2)(10 \times 167424 - 1288^2)}} \approx 0.98$$

### 3. Interprétation

Le coefficient de corrélation  $r \approx 0.98$  indique une **corrélation linéaire très forte et positive** entre la glycémie à jeun et le tour de taille.

## Exercice 2 : Régression linéaire simple

### 1. Équation de la droite de régression

Nous cherchons l'équation  $Y = a + bX$  où :

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2}$$
$$a = \bar{Y} - b\bar{X}$$

Calculs préliminaires :

- $\sum X = 35 + 40 + \dots + 60 = 285$
- $\sum Y = 180 + 190 + \dots + 220 = 1195$
- $\sum XY = 35 \times 180 + 40 \times 190 + \dots + 60 \times 220 = 58025$
- $\sum X^2 = 35^2 + 40^2 + \dots + 60^2 = 13875$
- $n = 6$

Calcul des coefficients :

$$b = \frac{6 \times 58025 - 285 \times 1195}{6 \times 13875 - 285^2} \approx 2.14$$

$$a = \frac{1195}{6} - 2.14 \times \frac{285}{6} \approx 100.71$$

L'équation de régression est donc :

$$\text{Cholestérol} = 100.71 + 2.14 \times \hat{\text{Age}}$$

### 2. Estimation pour un patient de 48 ans

$$Y = 100.71 + 2.14 \times 48 \approx 203.43 \text{ mg/dL}$$

### 3. Interprétation de la pente

La pente  $b = 2.14$  indique que le taux de cholestérol augmente en moyenne de **2.14 mg/dL** pour chaque année supplémentaire d'âge.

## Exercice 3 : Interprétation médicale

### 1. Force de la relation

Le coefficient  $r = 0.75$  indique une **corrélation forte et positive** entre le score de dépression et le taux de cortisol.

### 2. Lien de causalité ?

Non, une corrélation élevée ne permet pas d'affirmer une relation causale. D'autres facteurs pourraient influencer cette association.

### 3. Méthode pour explorer une relation causale

Pour étudier une relation causale, on pourrait utiliser :

- Une **étude expérimentale** randomisée contrôlée
- Une **analyse de médiation**
- Des modèles d'équations structurelles