# Chap3
# Relation entre les variables : Corrélation et Coefficient de Corrélation de Pearson

Pr. A. Fadil

EMSI RABAT

30 octobre 2025

# Objectifs du cours

- Comprendre la notion de relation entre deux variables.
- Identifier les types de relations (linéaire, non linéaire).
- Définir et interpréter la corrélation.
- Calculer et interpréter le coefficient de corrélation de Pearson.
- Connaître les limites et les précautions.

# Relation entre deux variables

- Relation = lien ou dépendance entre deux variables.
- Exemples :
  - nombre des heures de préparation et la note.
  - Température et consommation d'électricité.
- Importance :
  - Prévoir, comprendre ou modéliser un phénomène.

# Types de relations

- Relation linéaire (droite).
- Relation non linéaire (courbe, parabole).
- Absence de relation (données dispersées).

# Qu'est-ce que la corrélation ?

- Mesure statistique qui exprime la force et la direction de la relation entre deux variables quantitatives.
- Corrélation $\neq$ causalité !
- Se lit souvent sur un **nuage de points** (scatter plot).

# Types de corrélation

- Positive : les deux variables augmentent ensemble.
- Négative : une variable augmente, l'autre diminue.
- Nulle : absence de relation linéaire.

# Coefficient de corrélation de Pearson

## Définition

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

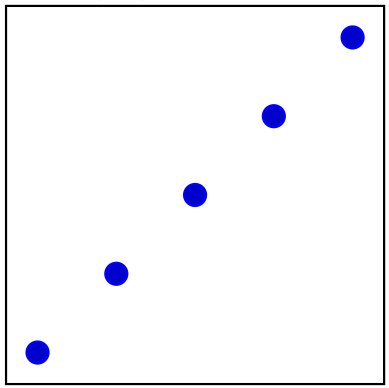
- $-1 \leq r \leq 1$.
- $r = 1$ : corrélation positive parfaite.
- $r = -1$ : corrélation négative parfaite.
- $r = 0$ : absence de corrélation linéaire.

# Interprétation de $r$

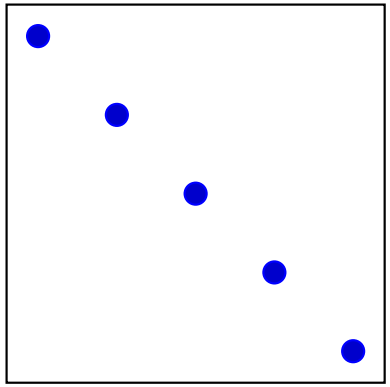
- $0.9 \leq |r| \leq 1$ : très forte corrélation.
- $0.7 \leq |r| < 0.9$ : forte corrélation.
- $0.5 \leq |r| < 0.7$ : corrélation modérée.
- $0.3 \leq |r| < 0.5$ : faible corrélation.
- $0 \leq |r| < 0.3$ : très faible ou aucune corrélation.

# Illustration graphique

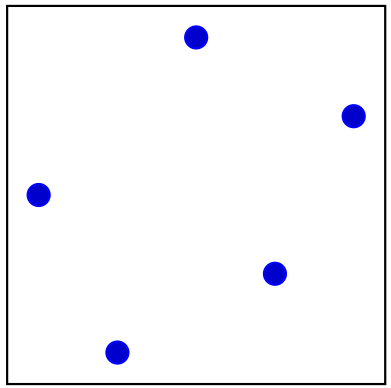
**Positive**



**Négative**



**Nulle**



# Exemple simple

Un enseignant souhaite étudier s'il existe un lien entre le *temps de révision* et la *note obtenue à un test*. Il interroge 5 étudiants et note, pour chacun, le nombre d'heures passées à réviser la veille du test ainsi que la note obtenue sur 10.

| $i$ | $X_i$ (heures) | $Y_i$ (note) |
|---|---|---|
| 1 | 1 | 2 |
| 2 | 2 | 4 |
| 3 | 3 | 5 |
| 4 | 4 | 4 |
| 5 | 5 | 5 |

- Moyennes : $\bar{x} = 3$, $\bar{y} = 4$
- Numérateur : $\sum(x_i - \bar{x})(y_i - \bar{y}) =$
  $(1-3)(2-4) + \cdots + (5-3)(5-4) = 6$
- Dénominateur :
  $\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2} = \sqrt{10 \times 6} = 7.75$
- $r = 6/7.75 \approx 0.77 \rightarrow$ forte corrélation positive.

# Précautions

- Corrélation $\neq$ causalité.
- Présence de valeurs extrêmes (*outliers*) $\rightarrow$ influence $r$.
- Relation non linéaire non détectée par Pearson.
- Toujours visualiser les données.

# Remarque

Le coefficient de Pearson ne s'applique qu'aux relations linéaires entre variables quantitatives.

- Spearman : données ordinales ou non linéaires.
- Kendall : petites séries ordinales.
- Pearson : uniquement pour relation linéaire et données quantitatives.

# Exercice d'application

**Données** :

| $i$ | Taille (cm) | Poids (kg) |
|---|---|---|
| 1 | 150 | 50 |
| 2 | 160 | 55 |
| 3 | 170 | 65 |
| 4 | 180 | 70 |
| 5 | 190 | 80 |

**Questions** :

- Calculer les moyennes $\bar{x}$ et $\bar{y}$.
- Calculer le coefficient $r$.
- Interpréter le résultat.
- Vérifier à l'aide d'un nuage de points.

ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

# Exercice d'application

**Données** :

| $i$ | Taille (cm) | Poids (kg) |
|---|---|---|
| 1 | 150 | 50 |
| 2 | 160 | 55 |
| 3 | 170 | 65 |
| 4 | 180 | 70 |
| 5 | 190 | 80 |

**Questions** :

- Calculer les moyennes $\bar{x}$ et $\bar{y}$.
- Calculer le coefficient $r$.
- Interpréter le résultat.
- Vérifier à l'aide d'un nuage de points.

ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

# Correction de l'exercice

**1. Moyennes** :

$$\bar{x} = \frac{150 + 160 + 170 + 180 + 190}{5} = 170$$

$$\bar{y} = \frac{50 + 55 + 65 + 70 + 80}{5} = 64$$

# Correction de l'exercice

**2. Calcul des écarts et produits croisés** :

| $i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|
| 1 | -20 | -14 | 280 | 400 | 196 |
| 2 | -10 | -9 | 90 | 100 | 81 |
| 3 | 0 | 1 | 0 | 0 | 1 |
| 4 | 10 | 6 | 60 | 100 | 36 |
| 5 | 20 | 16 | 320 | 400 | 256 |

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 280 + 90 + 0 + 60 + 320 = 750$$

$$\sum (x_i - \bar{x})^2 = 400 + 100 + 0 + 100 + 400 = 1000$$

$$\sum (y_i - \bar{y})^2 = 196 + 81 + 1 + 36 + 256 = 570$$

# Correction de l'exercice

**3. Calcul du coefficient $r$ :**

$$r = \frac{750}{\sqrt{1000 \times 570}} = \frac{750}{\sqrt{570000}} = \frac{750}{755} \approx 0.993$$

**4. Interprétation :**

- $r \approx 0.99$ : très forte corrélation positive entre la taille et le poids.
- En pratique, à mesure que la taille augmente, le poids augmente presque proportionnellement.

# Correction de l'exercice

**3. Calcul du coefficient $r$ :**

$$r = \frac{750}{\sqrt{1000 \times 570}} = \frac{750}{\sqrt{570000}} = \frac{750}{755} \approx 0.993$$

**4. Interprétation :**

- $r \approx 0.99$ : très forte corrélation positive entre la taille et le poids.
- En pratique, à mesure que la taille augmente, le poids augmente presque proportionnellement.

# Résumé

- Corrélation : relation entre deux variables.
- Pearson : mesure linéaire, sensible aux outliers.
- Toujours visualiser les données.
- Corrélation $\neq$ causalité.

# Questions / Discussion

- Avez-vous rencontré des variables corrélées dans votre domaine ?
- Pourquoi est-il dangereux de confondre corrélation et causalité ?

# Régression Linéaire Simple

Pr. A. Fadil

EMSI RABAT

30 octobre 2025

# Objectifs du cours

- Comprendre le concept de régression linéaire simple
- Maîtriser l'estimation des paramètres par les moindres carrés
- Évaluer la qualité de l'ajustement
- Appliquer le modèle à des exemples réels

# Définition

**Régression linéaire simple :** une méthode statistique permettant de modéliser la relation entre une variable expliquée ou réponse $Y$ et une variable explicative ou prédicteur $X$ via une relation linéaire :

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

où :

- $\beta_0$ est l'ordonnée à l'origine (intercept)
- $\beta_1$ est le coefficient directeur (pente)
- $\varepsilon$ est une erreur aléatoire

# Hypothèses du modèle

1. Relation linéaire entre $X$ et $Y$
2. Les erreurs $\varepsilon$ ont une espérance nulle : $\mathbb{E}[\varepsilon] = 0$
3. Variance constante des erreurs : homoscédasticité
4. Indépendance des erreurs
5. Normalité des erreurs (optionnelle pour estimation, nécessaire pour les tests)

# Estimation des paramètres

**Méthode des moindres carrés :** minimiser la somme des carrés des erreurs :

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

**Solutions :**

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

# Coefficient de détermination $R^2$

**Définition :**

$$R^2 = \frac{SS_{\text{exp}}}{SS_{\text{tot}}} = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

- $SS_{\text{tot}} = \sum (y_i - \bar{y})^2$ : variabilité totale
- $SS_{\text{res}} = \sum (y_i - \hat{y}_i)^2$ : variabilité non expliquée
- $SS_{\text{exp}} = \sum (\hat{y}_i - \bar{y})^2$ : variabilité expliquée

ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

# Interprétation de $R^2$

**Valeurs possibles de $R^2$ :**

| $R^2$ | Interprétation |
|---|---|
| $1$ | Le modèle explique parfaitement les données |
| $0$ | Le modèle n'explique aucune variance |
| $0 < R^2 < 1$ | Le modèle explique partiellement la variance |

Plus $R^2$ est proche de 1, plus le modèle est performant.

# Définition des résidus

**Un résidu** est la différence entre la valeur observée $y_i$ et la valeur prédite $\hat{y}_i$ :

$$e_i = y_i - \hat{y}_i$$

**Rôle des résidus :**

- Évaluer la qualité de l'ajustement
- Identifier les valeurs aberrantes
- Vérifier les hypothèses (linéarité, homoscédasticité, indépendance, normalité)

# Exemple numérique : $R^2$ et résidus

**Données :**

| $x_i$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $y_i$ | 2 | 4 | 5 | 4 |

Modèle : $\hat{y}_i = 2 + 0.7x_i$

**Résidus :**

| $x_i$ | $y_i$ | $\hat{y}_i$ | $e_i = y_i - \hat{y}_i$ |
|---|---|---|---|
| 1 | 2 | 2.7 | -0.7 |
| 2 | 4 | 3.4 | 0.6 |
| 3 | 5 | 4.1 | 0.9 |
| 4 | 4 | 4.8 | -0.8 |

$$R^2 = \frac{2.45}{4.75} = 0.51$$

$\rightarrow$ *Le modèle explique environ 51% de la variance.*

# Exemple corrigé 1 : données simples

Soit les données :

| $x_i$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $y_i$ | 2 | 4 | 5 | 4 |

**Étapes :**

- $\bar{x} = 2.5$, $\bar{y} = 3.75$
- $\hat{\beta}_1 = \frac{(13.5}{5} = 0.7$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2$
- Équation estimée : $\hat{Y} = 2 + 0.7X$

# Conclusion

- La régression linéaire est un outil puissant de modélisation
- Valider les hypothèses est essentiel
- Peut être généralisée à la régression multiple

# Régression Linéaire Multiple

Pr. A. Fadil

EMSI RABAT

30 octobre 2025

# Plan du cours



ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

# Introduction

# Pourquoi la régression multiple ?

- La régression linéaire multiple permet d'étudier l'effet de plusieurs variables explicatives (prédicteurs) sur une variable réponse.
- Elle est utile pour comprendre, modéliser et prédire une variable quantitative en fonction de plusieurs facteurs.
- **Exemple** : Prédire la pression artérielle en fonction de l'âge, du poids et du niveau de stress.

# Définitions

# Variables et concepts clés

- **Variable réponse (dépendante)** : variable que l'on cherche à modéliser ou à prédire ($Y$).
- **Variables explicatives (indépendantes)** : variables qui influencent ou expliquent la variable réponse ($X_1, X_2, \ldots, X_p$).
- **Coefficients de régression** ($\beta_j$) : mesure de l'effet de chaque variable explicative sur $Y$.
- **Erreur résiduelle** ($\varepsilon$) : différence entre la valeur observée et la valeur prédite.

# Formulation du modèle



ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

# Modèle mathématique

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

En notation matricielle :

$$\mathrm{Y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- X : matrice $n \times (p+1)$ des données (la première colonne contient des 1 pour l'intercept).
- $\boldsymbol{\beta}$ : vecteur des coefficients de régression.
- $\boldsymbol{\varepsilon}$ : vecteur des erreurs aléatoires.

# Hypothèses du modèle



ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

# Hypothèses classiques

1. **Linéarité** : la relation entre chaque $X_j$ et $Y$ est linéaire.
2. **Espérance nulle des erreurs** : $\mathbb{E}[\varepsilon_i] = 0$.
3. **Homoscédasticité** : la variance des erreurs est constante ($\text{Var}(\varepsilon_i) = \sigma^2$).
4. **Indépendance** : les erreurs sont indépendantes entre elles.
5. **Normalité** : les erreurs sont distribuées normalement (utilisé pour les tests).

# Estimation par moindres carrés

# Moindres carrés ordinaires (MCO)

- Objectif : trouver les $\beta_j$ qui minimisent la somme des carrés des résidus.

$$\min_{\boldsymbol{\beta}} \|\mathrm{Y} - \mathrm{X}\boldsymbol{\beta}\|^2$$

## Solution

$$\hat{\boldsymbol{\beta}} = (\mathrm{X}^\top \mathrm{X})^{-1} \mathrm{X}^\top \mathrm{Y}$$

ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

# Interprétation

# Interprétation des coefficients

- $\beta_j$ : variation moyenne de $Y$ lorsque $X_j$ augmente d'une unité, toutes les autres variables étant maintenues constantes.
- $\beta_0$ : valeur prédite de $Y$ lorsque toutes les variables explicatives valent 0 (intercept).

# Qualité de l'ajustement

# $R^2$ et $R^2_{\text{ajusté}}$

- **Coefficient de détermination** $R^2$ : proportion de la variabilité de $Y$ expliquée par le modèle.

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

- $R^2$ **ajusté** : corrige $R^2$ pour le nombre de variables.

$$R^2_{\text{ajusté}} = 1 - \left(\frac{1 - R^2}{n - p - 1}\right)(n - 1)$$

**Remarque :**

$$R^2_{\text{ajusté}} \leq R^2$$

# Quelle est la différence entre le R au carré ajusté et le R au carré ?

Bien que le r-carré et le r-carré ajusté évaluent tous deux la performance du modèle de régression, il existe une différence essentielle entre les deux mesures. La valeur du carré r augmente toujours ou reste la même lorsque des prédicteurs sont ajoutés au modèle, même si ces prédicteurs n'améliorent pas de manière significative le pouvoir explicatif du modèle. Ce problème peut donner une impression trompeuse de l'efficacité du modèle.

Le r-carré ajusté ajuste la valeur du r-carré pour tenir compte du nombre de variables indépendantes dans le modèle. La valeur ajustée du r-carré peut diminuer si un nouveau prédicteur n'améliore pas l'adéquation du modèle, ce qui en fait une mesure plus fiable de la précision du modèle. C'est pourquoi le r-carré ajusté peut être utilisé comme un outil par les analystes de données pour les aider à décider quels prédicteurs inclure.

# Exemple

# Exemple corrigé : Enoncé

On souhaite modéliser le score d'un étudiant ($Y$) en fonction de :

- $X_1$ : nombre d'heures d'étude par jour
- $X_2$ : nombre d'heures de sommeil

## Données

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 5 | 7 | 68 |
| 7 | 6 | 75 |
| 8 | 8 | 85 |
| 6 | 5 | 70 |
| 9 | 7 | 90 |

## Questions :

1. Estimer le modèle de régression $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
2. Interpréter les coefficients estimés
3. Calculer le score prédit pour $X_1 = 6$, $X_2 = 6$

# Exemple corrigé : Résolution

- Estimation par moindres carrés :

$$\hat{Y} = 28.45 + 5.37X_1 + 1.74X_2$$

- Interprétation :
  - Une heure d'étude supplémentaire augmente $Y$ de 5.37 points.
  - Une heure de sommeil supplémentaire augmente $Y$ de 1.74 points.
- Prédiction pour $X_1 = 6$, $X_2 = 6$ :

$$\hat{Y} = 28.45 + 5.37 \times 6 + 1.74 \times 6 = 71.11$$

# Conclusion



ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

# Résumé

- La régression multiple permet de modéliser une variable continue en fonction de plusieurs facteurs.
- La qualité du modèle se juge par $R^2$.