Ecole Marocaine des Sciences de l'Ingénieur de Casablanca, Maroc

Modèles Statistiques

Chapitre I : Régression Linéaire simple

Variance et Écart-Type

I- Rappels et Fondements :

Variance : Mesure la dispersion des valeurs autour de leur moyenne.

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Écart-type : Racine carrée de la variance.

$$\sigma_X = \sqrt{\mathrm{Var}(X)}.$$

Interprétation

- Plus la variance est grande, plus les valeurs de X sont éloignées de leur moyenne.
- L'écart-type est plus intuitif car il est exprimé dans la même unité que la variable.

Covariance

Covariance : Mesure la *relation linéaire* entre deux variables X et Y.

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\bar{y}$$

où :

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Signification

- $\mathrm{Cov}(X,Y) > 0$: X et Y varient dans le *même* sens.
- $\mathrm{Cov}(X,Y) < 0$: X et Y varient en *sens inverse*.
- $\mathrm{Cov}(X,Y) = 0$: pas de **relation linéaire**.

Coefficient de Corrélation (r)

Définition : Le coefficient de corrélation de Pearson (r) est la *covariance* normalisée par les écarts-types :

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \, \sigma_Y} \quad \in [-1, 1].$$

Exemples d'interprétation

- $r \approx 1$: forte corrélation *positive*.
- $r \approx -1$: forte corrélation *négative*.
- $r = 0$: pas de relation linéaire.
- $|r| < 0.3$: corrélation faible.
- $|r| > 0.7$: bonne corrélation.
- $|r| > 0.9$: corrélation *excellente*.

Exemple : 5 individus (Taille/Poids)

- Exemple de base :

Considérons un échantillon de 5 individus pour lesquels nous avons mesuré la taille (en cm) et le poids (en kg). Le tableau suivant présente ces valeurs ainsi que les quantités intermédiaires nécessaires aux calculs statistiques :

Individu	Taille x_i (cm)	Poids y_i (kg)	x_i^2	y_i^2	$x_i y_i$
1	160	55	25600	3025	8800
2	170	65	28900	4225	11050
3	175	70	30625	4900	12250
4	180	80	32400	6400	14400
5	165	60	27225	3600	9900

Table – Tableau statistique des variables taille et poids

Exemple : 5 individus (Taille/Poids)

1. Moyennes

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{160 + 170 + 175 + 180 + 165}{5} = 170,$$

$$\bar{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = \frac{55 + 65 + 70 + 80 + 60}{5} = 66.$$

2. Variance de x

$$\begin{aligned}\mathrm{Var}(X) &= \overline{x^2} - \bar{x}^2 \\ &= \frac{1}{5}(25600 + 28900 + 30625 + 32400 + 27225) - (170)^2 \\ &= 50.\end{aligned}$$

3. Variance de y

$$\begin{aligned}\mathrm{Var}(Y) &= \overline{y^2} - \bar{y}^2 \\ &= \frac{1}{5}(3025 + 4225 + 4900 + 6400 + 3600) - (66)^2 \\ &= 74.\end{aligned}$$

Exemple : 5 individus (Taille/Poids)

4. Écart-type de x

$$\begin{aligned}\sigma_x &= \sqrt{\mathrm{Var}(X)} \\ &= \sqrt{50} \\ &\approx 7.07.\end{aligned}$$

5. Écart-type de y

$$\begin{aligned}\sigma_y &= \sqrt{\mathrm{Var}(Y)} \\ &= \sqrt{74} \\ &\approx 8.60.\end{aligned}$$

6. Covariance entre x et y

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \overline{xy} - \bar{x}\bar{y} \\ &= \frac{1}{5}(8800 + 11050 + 12250 + 14400 + 9900) - (170 \times 66) \\ &= 60.\end{aligned}$$

Calcul de la Corrélation

3. Coefficient de corrélation r

$$\begin{aligned} r &= \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y} \\ &= \frac{60}{(7.07 \times 8.60)} \\ &\approx 0.986. \end{aligned}$$

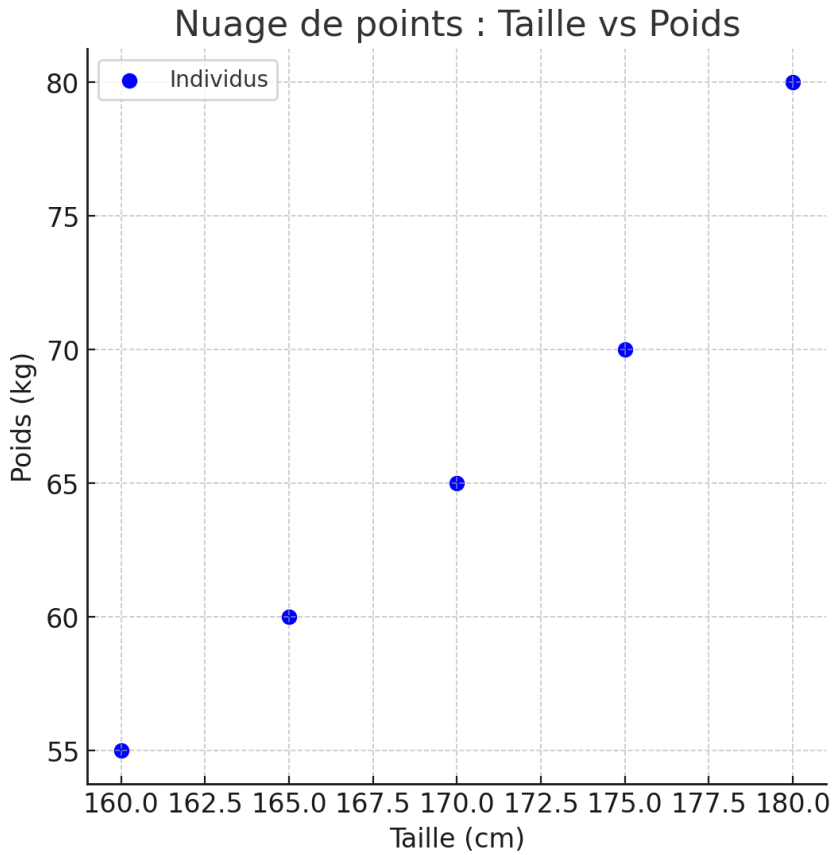
Interprétation

Le coefficient de corrélation $r \approx 0.986$ indique une forte corrélation positive entre la taille et le poids.

Exemples de Nuages de Points

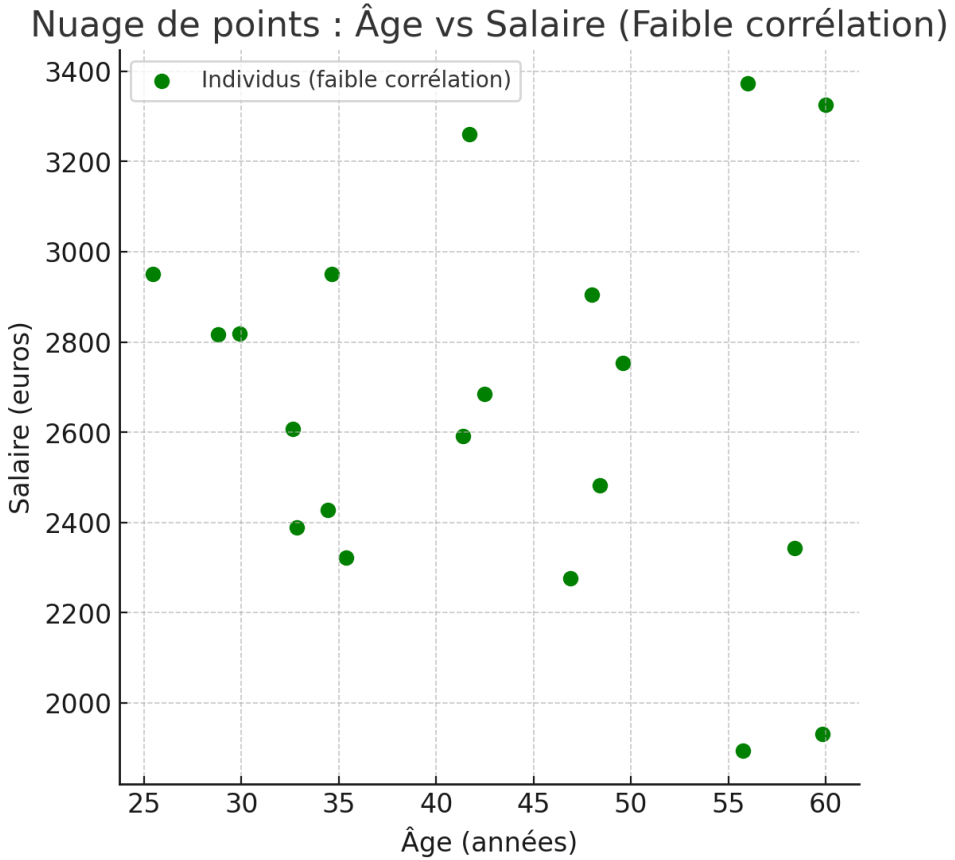
Définition : Un *nuage de points* est la représentation graphique de paires (x_i, y_i).

Exemple 1 : Corrélation positive forte



Nuages de Points (suite)

Exemple 2 : Faible corrélation / Données dispersées



Remarque : *On remarque une dispersion significative des points, ce qui indique une faible corrélation entre ces deux variables.*

Introduction

II-Régression Linéaire

1- Définition :

- La régression linéaire simple est une méthode d'analyse statistique qui permet de modéliser la relation entre deux variables :
  - Une variable indépendante X
  - Une variable dépendante Y

2-Objectif :

- L'objectif est de trouver une droite $Y = aX + b$ qui minimise l'erreur entre les valeurs observées et prédites.

Droite de Régression : Explications

3 - Droite de Régression :

Idée : Approcher la relation entre X et Y par une fonction linéaire de la forme :

$$\hat{y} = Ax + B.$$

Cette droite permet de modéliser la tendance centrale des points dans un nuage de données.

Méthode des moindres carrés : On cherche les coefficients A et B qui minimisent la somme des carrés des résidus :

$$\sum (y_i - \hat{y}_i)^2.$$

Ces coefficients sont donnés par les formules :

$$A = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)},$$

$$B = \bar{y} - A\bar{x}.$$

Exemple : Calcul des coefficients A et B

4 - Exemple de régression linéaire

À partir des valeurs déjà calculées :

$$\bar{x} = 170, \quad \bar{y} = 66, \quad \text{Var}(X) = 50, \quad \text{Cov}(X, Y) = 60.$$

Nous déterminons les coefficients de la droite de régression $\hat{y} = Ax + B$:

1. Calcul du coefficient A (pente)

$$A = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \frac{60}{50} = 1.2$$

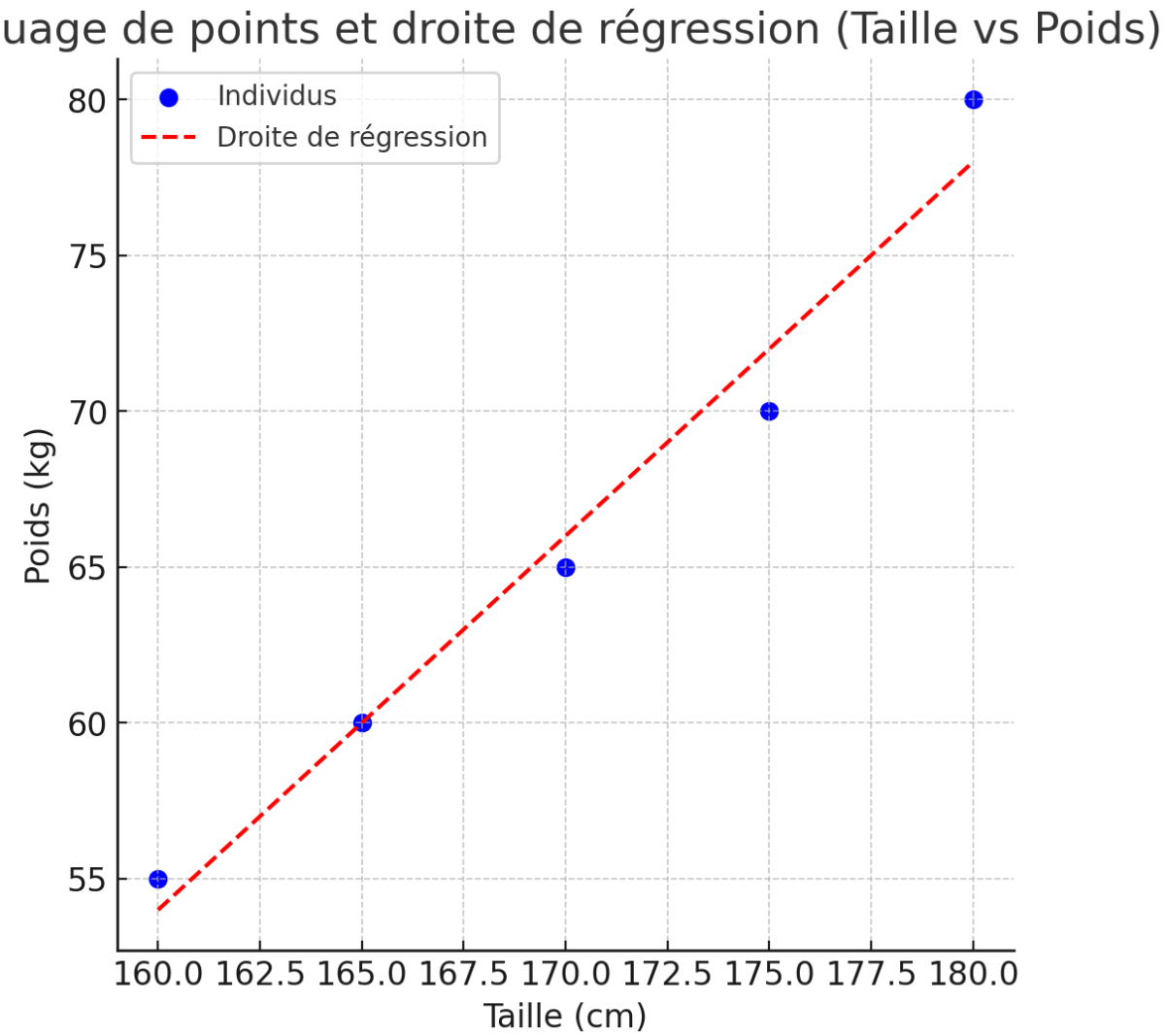
2. Calcul du coefficient B (ordonnée à l'origine)

$$B = \bar{y} - A\bar{x} = 66 - (1.2 \times 170) = -138.$$

Conclusion : L'équation de la droite de régression est donc :

$$\boxed{\hat{y} = 1.2x - 138.}$$

Représentation Graphique de la Régression



Utilité de la Droite de Régression

5 - Prédiction à l'aide de la droite de régression

La droite de régression permet non seulement de modéliser la relation entre la taille et le poids, mais aussi de prédire des valeurs inconnues. En effet, pour une valeur x (taille) qui ne figure pas dans notre échantillon, nous pouvons estimer la valeur de y (poids) à l'aide de l'équation obtenue :

$$\hat{y} = 1.2x - 138$$

Exemple de prédiction : Si une personne mesure 172 cm, son poids prédit sera :

$$\hat{y} = 1.2 \times 172 - 138 = 68.4 \text{ kg}.$$

Interprétation : Grâce à cette approche, il est possible d'estimer des valeurs même en dehors des données observées, ce qui est particulièrement utile en analyse de données et en prise de décision.

Coefficient de Détermination $\mathbf{R^2}$

6 - Le Coefficient de Détermination R^2

Le coefficient de détermination R^2 quantifie la qualité de l'ajustement du modèle de régression. Il est défini par :

$$R^2 = r^2.$$

Interprétation :

- R^2 mesure la **proportion de la variance de** Y qui est expliquée par la variable X à travers le modèle de régression linéaire.
- Il varie entre 0 et 1. Plus R^2 est proche de 1, plus le modèle est performant pour expliquer les variations de Y.

Exemple :

$$\text{Si } r = 0.9, \quad R^2 = (0.9)^2 = 0.81.$$

Cela signifie que 81% de la variabilité de Y est expliquée par X.

Conclusion :

- $R^2 \approx 1 \Rightarrow$ Le modèle explique presque toute la variance de Y.
- $R^2 \approx 0 \Rightarrow$ Le modèle n'explique quasiment rien de la variance de Y, donc la relation linéaire est faible ou inexistante.

Exercice d'Application

7 - Exercice : Régression Linéaire et Corrélation

On souhaite étudier la relation entre le nombre d'heures d'étude par semaine (X) et la note obtenue à un examen (Y) pour un groupe de 6 étudiants. Le tableau suivant présente les données recueillies :

Étudiant	Heures d'étude x_i	Note y_i	x_i^2	y_i^2	$x_i y_i$
1	5	10	..	..	..
2	8	12	..	..	..
3	10	14	..	..	..
4	12	16	..	..	..
5	15	18	..	..	..
6	18	19	..	..	..

Exercice d'Application

Questions :

1. Calculer la moyenne de X (heures d'étude) et Y (notes obtenues).
2. Déterminer la variance de X et la covariance entre X et Y.
3. Calculer le coefficient de corrélation r.
4. Déterminer l'équation de la droite de régression $\hat{y} = Ax + B$.
5. Calculer le coefficient de détermination R^2 et interpréter sa valeur.
6. Prédire la note d'un étudiant qui a étudié 11 heures par semaine.