

Régression linéaire

Baptiste CRINIERE-BOIZET

Data Analysis Core



Théorie

Vérification des hypothèses

Interprétation

Exercice

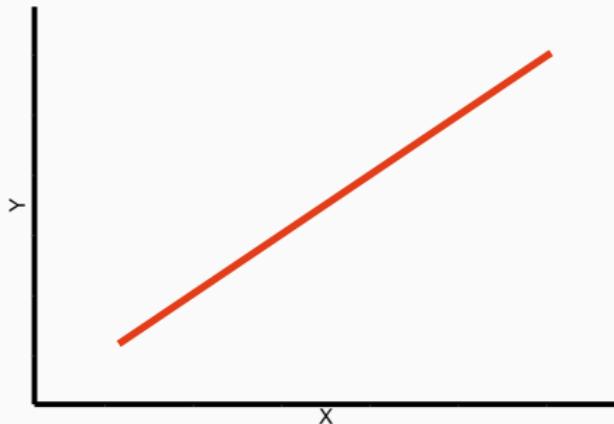
Théorie

Variable dépendante et variables indépendantes

- Distinction clé : Variable Dépendante (Y) vs Variables Indépendantes (X₁, X₂, etc.)
- Variable Dépendante = Variable Réponse ou à Expliquer
- Variable Indépendante = Prédicteur ou Explicative
- Choix basé sur logique ou théorie, exemple : taille de l'enfant dépendant de la taille de la mère

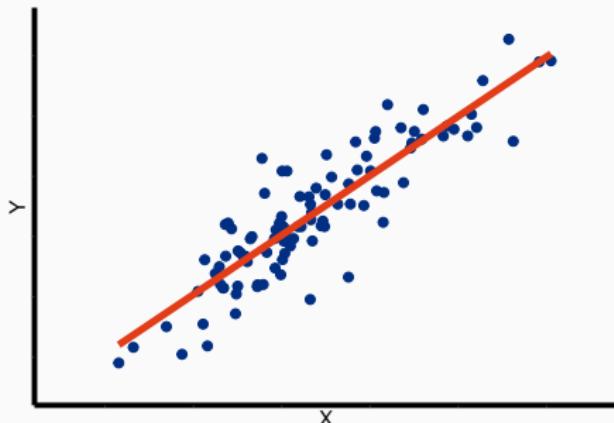
Equation linéaire

- $Y = b + a \times X$
- b : Ordonnée à l'origine,
 a : Coefficient directeur de
la droite
- **Utilité** : Permet de
déterminer Y en
connaissant la valeur de X



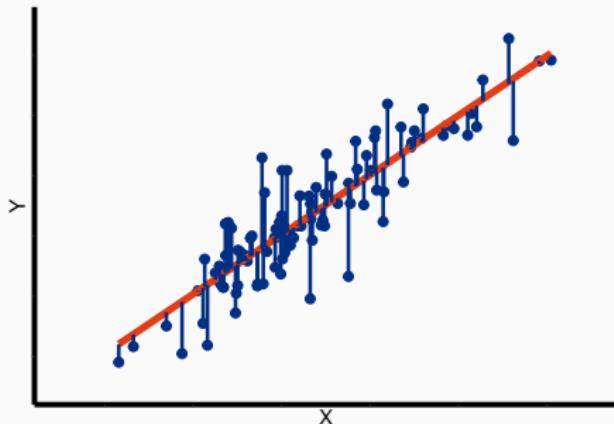
Modélisation linéaire

- Relations linéaires parfaites rares dans les données réelles (BRUIT)
- **Objectif :** Tracer une ligne ou modèle prédictif capturant la tendance observée



Résidus

- Régression linéaire : ajuster une droite à des données avec deux variables
- Précision limitée : Imprécisions entre valeurs observées et prédites
- **Résidus** : Différences entre valeurs réelles et prédites
- $e_i = y_i - \hat{y}_i$



La meilleure droite

- Assurer des erreurs de prédiction équilibrées (négatives et positives).
- Distribution des résidus centrée autour de 0
- Positionner la droite au plus près des points de données
- En pratique, minimiser la somme des résidus au carré

Critère Somme des carrés des résidus la plus faible, représentant l'estimateur des moindres carrés

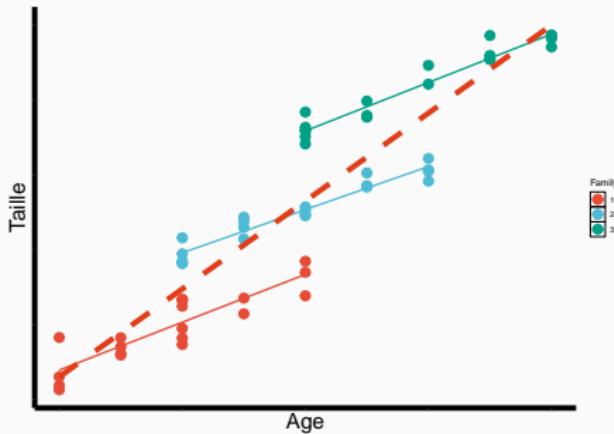
Vérification des hypothèses

Liste des principales hypothèses

- **Indépendance** : Les observations doivent être indépendantes les unes des autres
- **Linéarité** : La relation entre les variables doit être linéaire.
- **Homoscédasticité** : Les résidus doivent présenter une variance constante à travers toutes les valeurs prédictives.
- **Normalité** : La distribution des résidus doit suivre une distribution normale

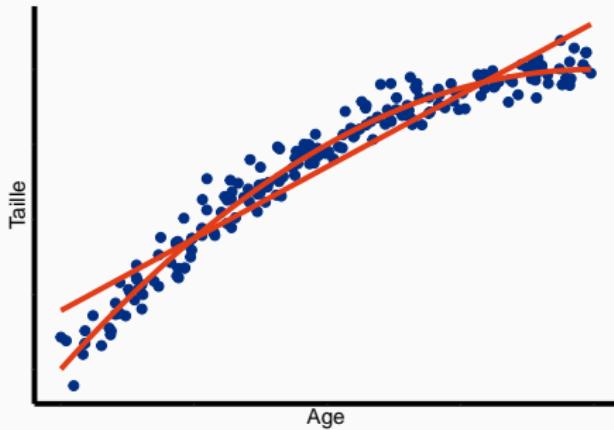
Indépendance

- S'assurer que les observations ne sont pas liées entre elles
- **Conséquences** : biais estimations, fiabilité des tests
- **Vérification** : regroupement dans la distribution des résidus



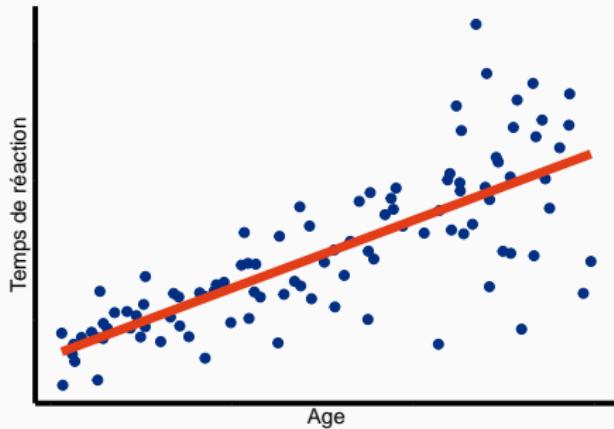
Linéarité

- Porte sur la relation entre variables et coefficients
- Le modèle linéaire peut s'ajuster avec des transformations (quadratique, cubique)
- **Conséquences** : risque d'estimations biaisées, prédictions inexactes
- **Vérification** : dispersion aléatoire des résidus



Homoscédasticité

- Variabilité des résidus reste constante quelle que soit la valeur de la variable indépendante
- **Détection** : Motif en entonnoir ou tendance croissante/décroissante de la dispersion des résidus
- **Conséquences** : Risque de biais dans les estimations et les tests statistiques
- **Solution** : Transformation des données



Interprétation

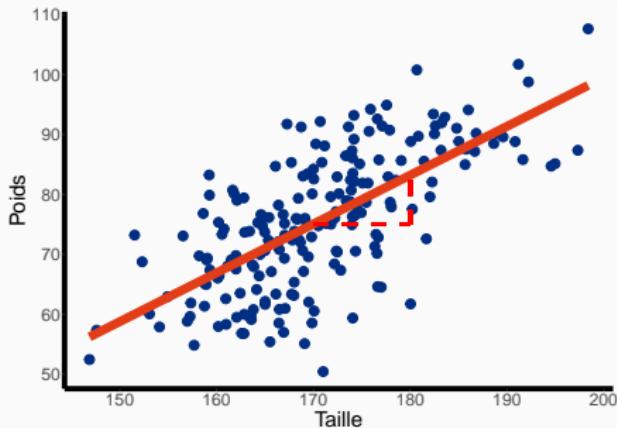
Régression simple

- **Définition** : Une régression linéaire simple utilise une seule variable explicative.
- Types de variables explicatives : Continue ou Catégorielle
- Exemple : Analyse du lien entre le poids et la taille & le poids et le sexe

Régression simple

Prédicteur continu

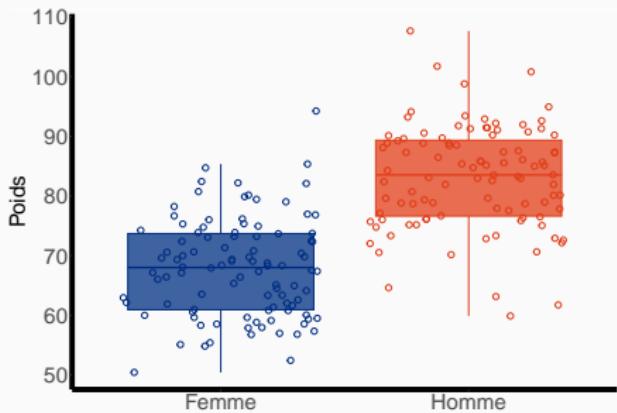
- **Modèle :**
 $Poids = \beta_0 + \beta_1 Taille + \epsilon$
- $\beta_0 = -63$ et $\beta_1 = 0.81$, on teste aussi la non-nullité de chaque coefficient
- R^2 la part de variance expliquée par notre modèle est de 46%



Régression simple

Prédicteur discret

- **Modèle :**
 $Poids = \beta_0 + \beta_1 Taille + \epsilon$
- $\beta_0 = -63$ et $\beta_1 = 0.81$, on teste aussi la non-nullité de chaque coefficient
- R^2 la part de variance expliquée par notre modèle est de 46%



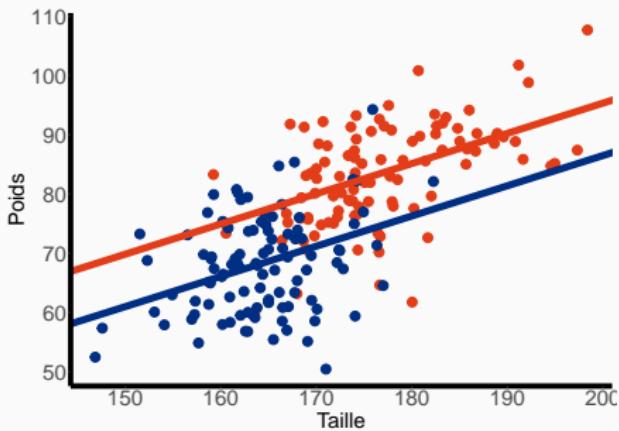
Régression multiple

- **Définition** : régression avec plusieurs variables explicatives
- **Ajuster** notre modèle par une nouvelle variable explicative
- **Interaction** lorsque deux variables influencent simultanément la variable à expliquer

Régression multiple : ajustement

Cas discret

- Modèle :
 $Poids = \beta_0 + \beta_1 \times 1_{\{Homme\}} + \beta_2 \times Taille + \epsilon$
- Droites parallèles représentant chaque sous-groupe d'une catégorie
- R^2 la part de variance expliquée par notre modèle est de 55%



Régression multiple : ajustement

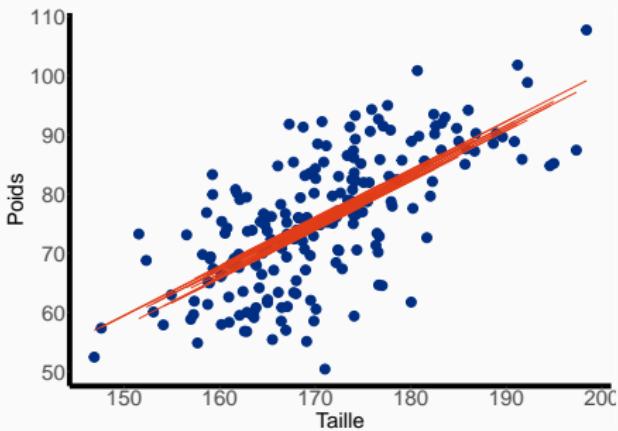
Cas continu

- **Modèle :**

$$\text{Poids} = \beta_0 + \beta_1 \times 1_{\{\text{Homme}\}} + \beta_2 \times \text{Taille} + \epsilon$$

- Droites parallèles représentant chaque sous-groupe d'une catégorie

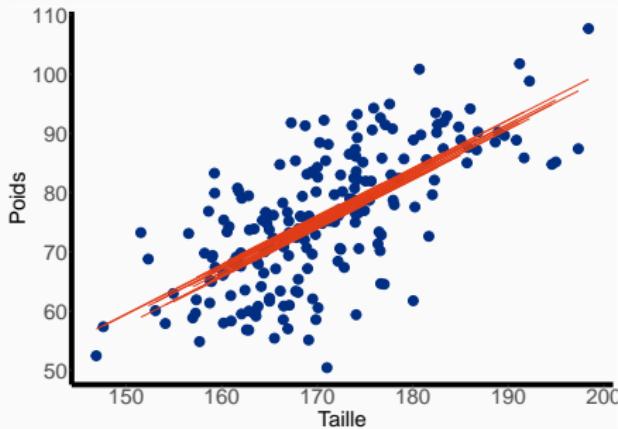
- R^2 la part de variance expliquée par notre modèle est de 55%



Régression multiple : interaction

Cas discret

- **Définition :** Une interaction survient lorsque l'effet d'une variable sur la variable réponse dépend du niveau d'une autre variable
- **Equation :**
$$\text{Poids} = \beta_0 + \beta_1 \times 1_{\{\text{Homme}\}} + \beta_2 \times \text{Taille} + \beta_3 \times 1_{\{\text{Homme}\}} \times \text{Taille} + \epsilon$$
- Les hommes ont une pente plus prononcée que les femmes



Exercice

Contexte