Linear Regression

Baptiste CRINIERE-BOIZET

Data Analysis Core

Theory

Hypothesis Checking

Interpretation

Theory

Dependent Variable and Independent Variables

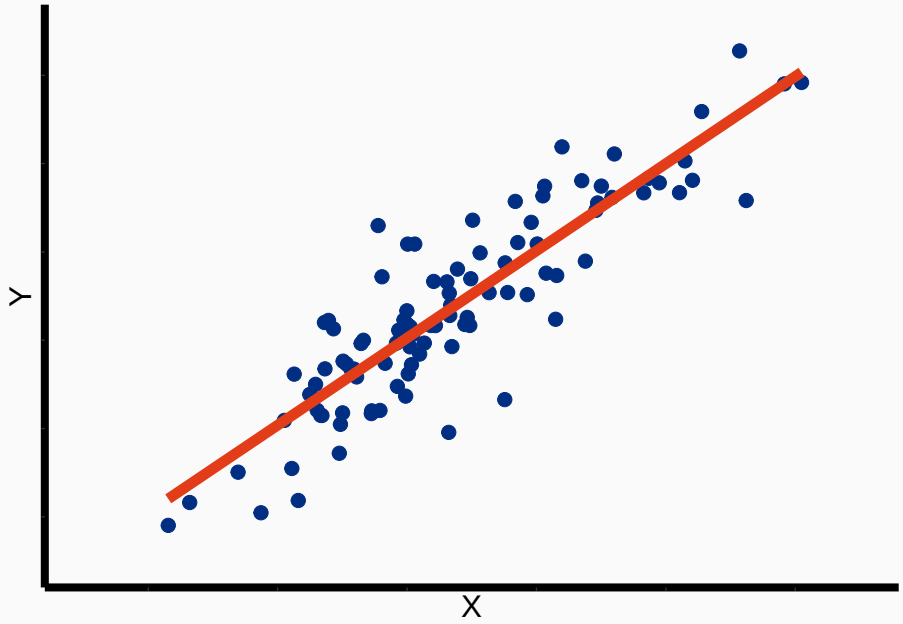
- Key Distinction: Dependent Variable (Y) vs Independent Variables (X1, X2, etc.)
- Dependent Variable = Response or Explained Variable
- Independent Variable = Predictor or Explanatory
- Choice based on logic or theory, for example: child's height depending on the mother's height

Linear Equation

- $Y = b + a \times X$
- b : Intercept,
  a : Slope
- **Utility:** Allows determining Y when the value of X is known

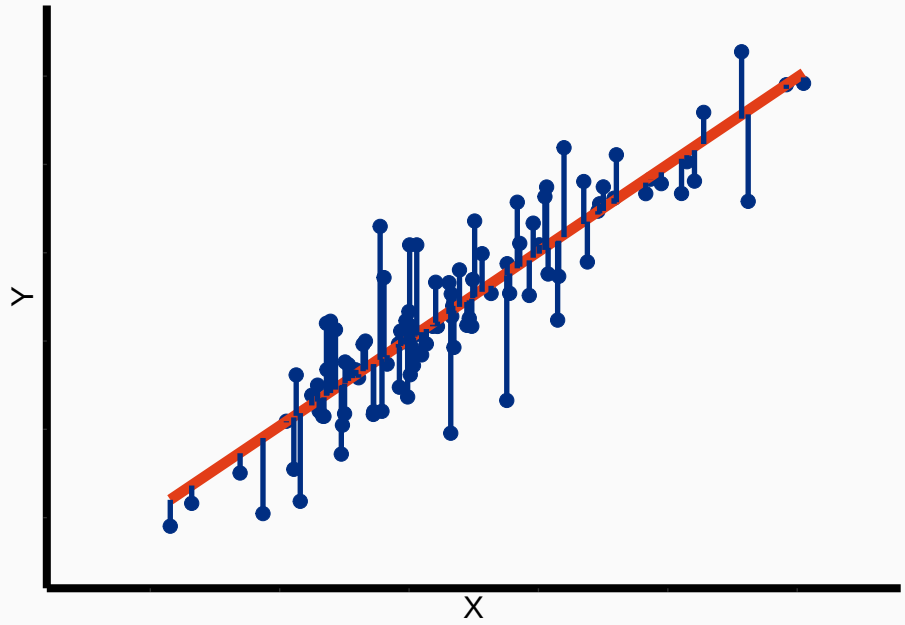
Linear Modeling

- Perfect Linear Relationships Rare in Real Data (NOISE)
- **Objective**: Plot a line or predictive model capturing the observed trend



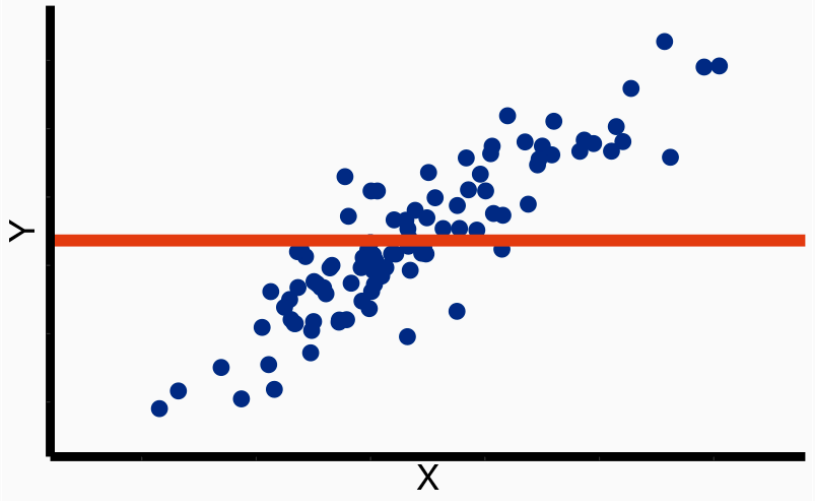
Residuals

- Linear Regression: Fitting a line to data with two variables
- Limited Accuracy: Inaccuracies between observed and predicted values
- **Residuals :** Differences between actual and predicted values
- $e_i = y_i - \hat{y}_i$



The best-fit line

- Ensure balanced prediction errors (negative and positive)
- Residual distribution centered around 0
- Position the line as close as possible to the data points
- In practice, minimize the sum of squared residuals



Criteria Lowest sum of squared residuals, representing the least squares estimator

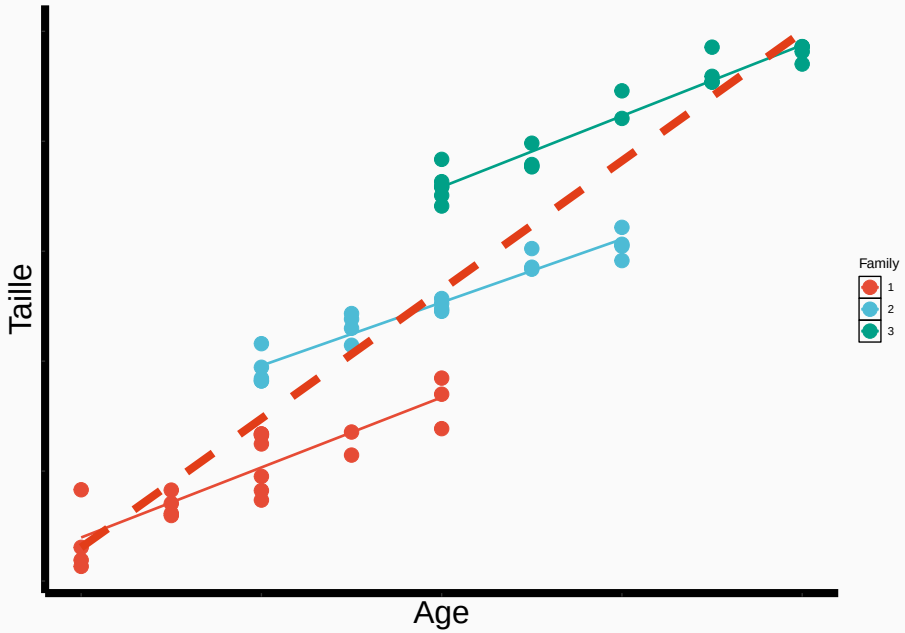
Hypothesis Checking

List of Key Hypothesis

- **Independence**: The observations must be independent of each other
- **Linearity**: The relationship between variables must be linear
- **Homoscedasticity**: The residuals should exhibit constant variance across all predicted values
- **Normality**: The distribution of residuals should follow a normal distribution

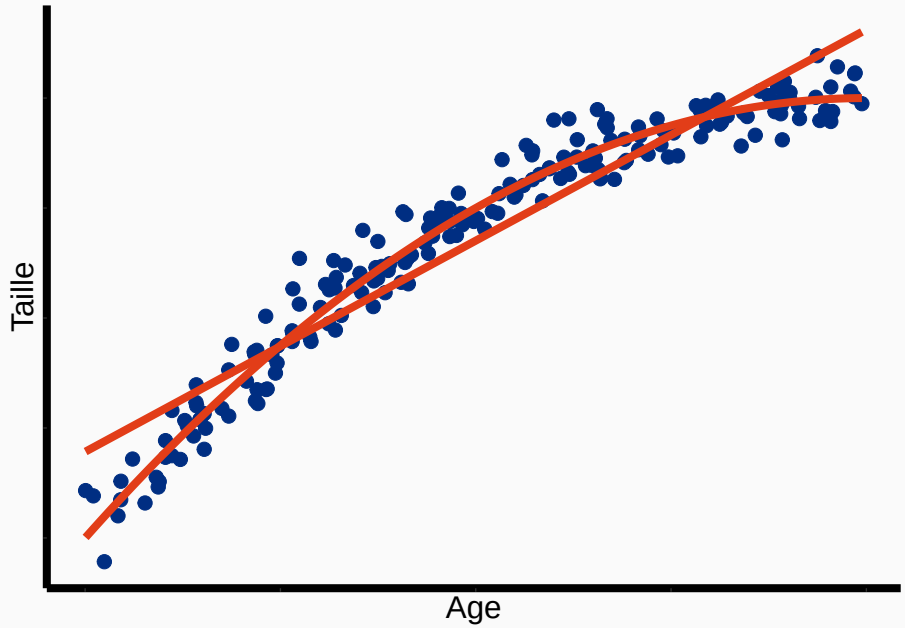
Independence

- Ensure that observations are not related to each other
- **Consequences**: Estimation bias, reliability of tests
- **Check** : no regrouping in the distribution of residues



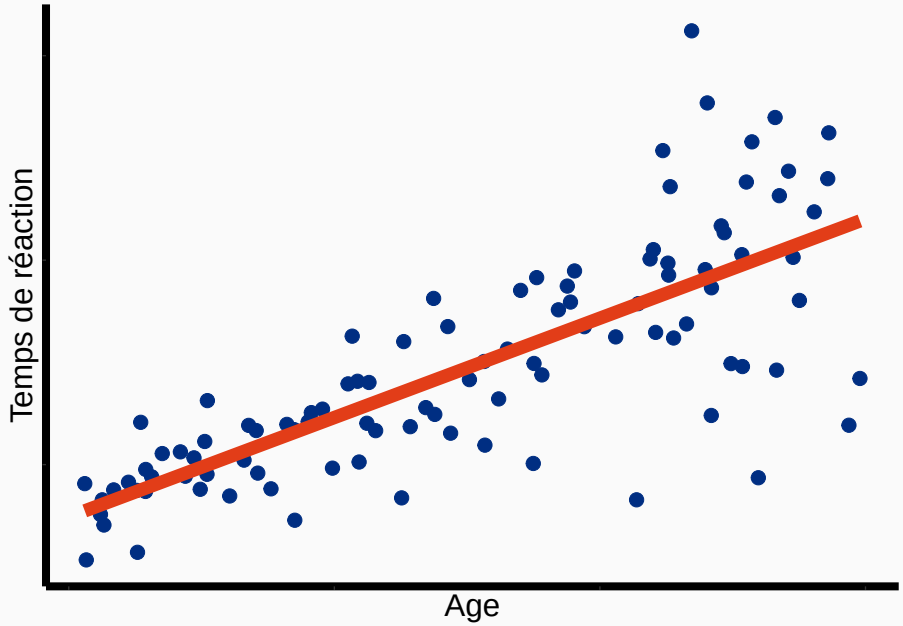
Linearity

- Concerns the relationship between variables and coefficients
- The linear model can be adjusted with transformations (quadratic, cubic)"
- **Consequences** : Risk of biased estimates, inaccurate predictions



Homoscedasticity

- Variability of residuals remains constant regardless of the value of the independent variable
- **Check** : Funnel-shaped pattern or increasing/decreasing trend in the dispersion of residuals
- **Consequences** : Risk of bias in estimations and statistical tests
- **Solution** : Data transformation (log)



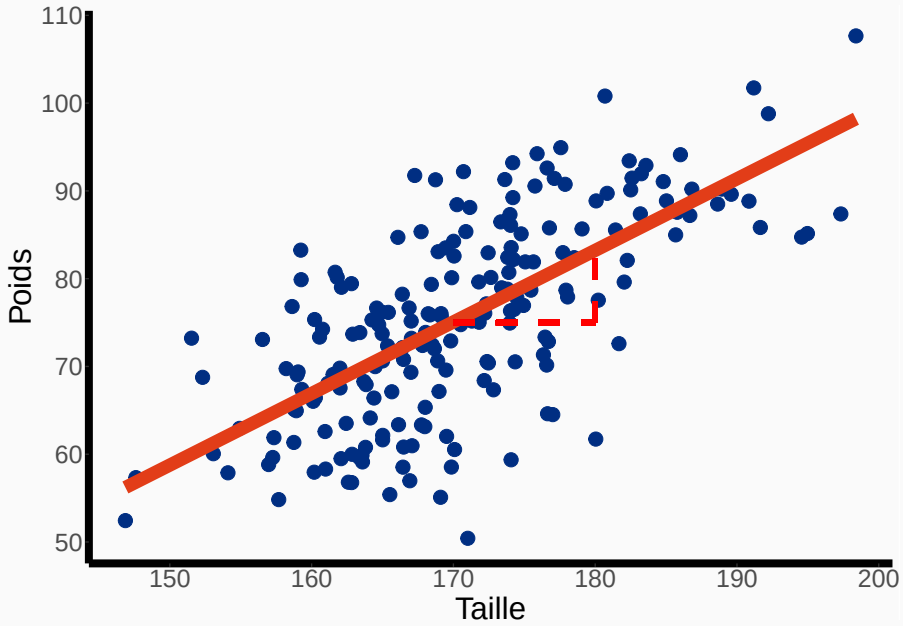
Interpretation

Imple regression

- **Definition** : A simple linear regression uses a single explanatory variable
- Types of explanatory variables: Continuous or Categorical
- Example: Analysis of the relationship between weight and height & weight and sex

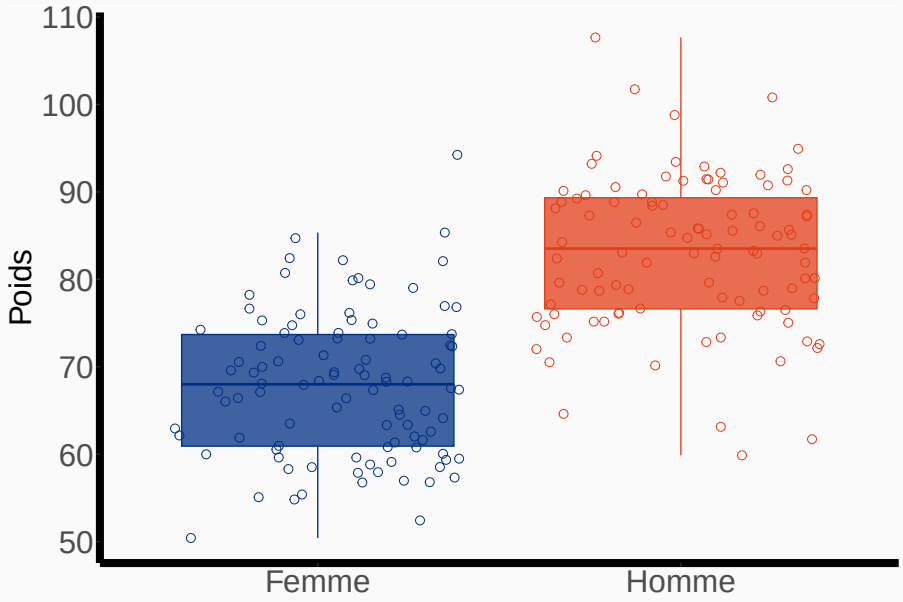
Prédicteur continu

- **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%



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%



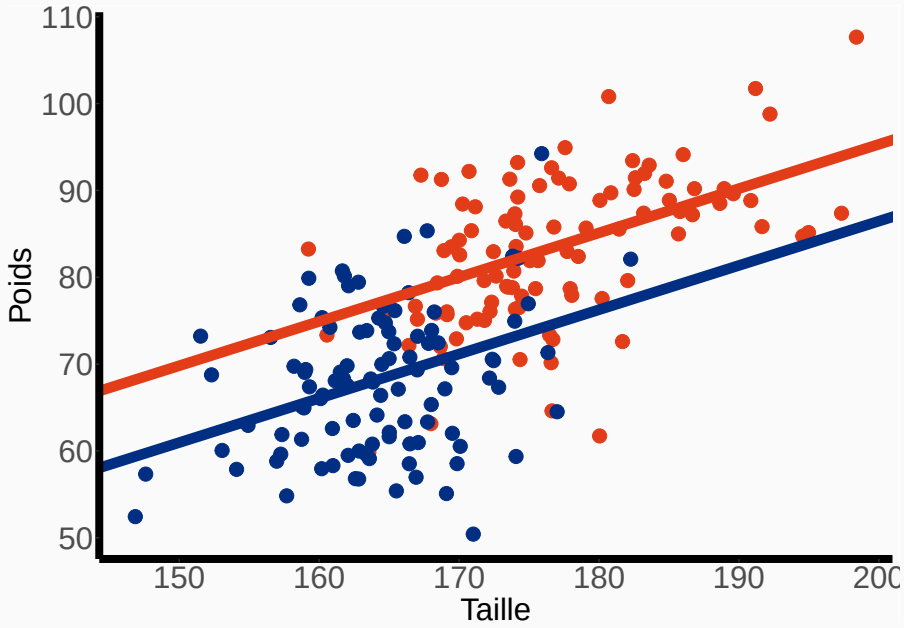
Régression multiple

- **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

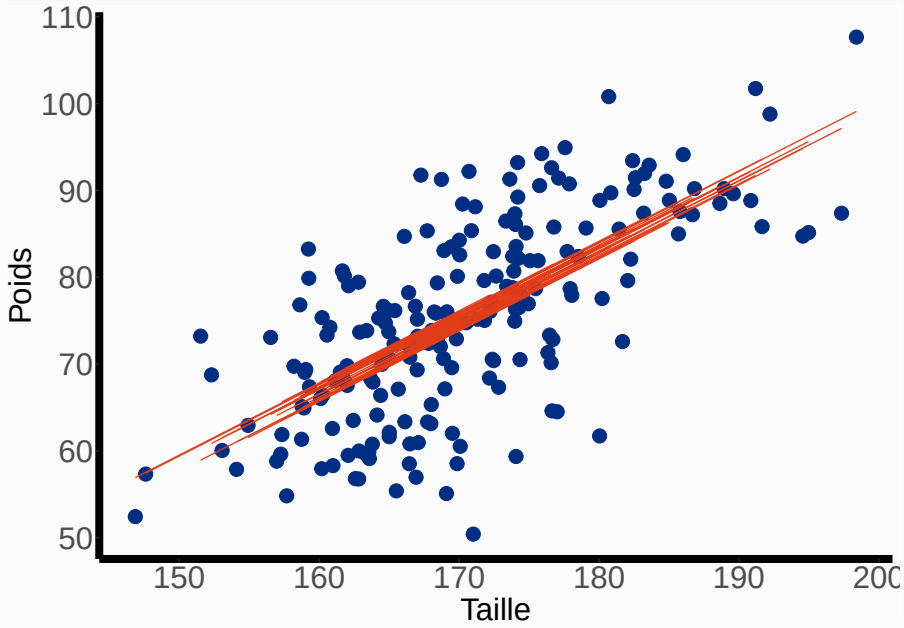
- **Modèle** :
  $Poids = \beta_0 + \beta_1 \times 1_{\{Homme\}} + \beta_2 \times Taille + \epsilon$
- Droites parallèles représentant chaque sous-groupe d'une catégorie
- R^2 la part de variance expliquée par notre modèle est de 55%



Régression multiple : ajustement

Cas continu

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



Régression multiple : 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** :
  $Poids = \beta_0 + \beta_1 \times 1_{\{Homme\}} + \beta_2 \times Taille + \beta_3 \times 1_{\{Homme\}} \times Taille + \epsilon$
- Les hommes ont une pente plus prononcée que les femmes