Linear Regression

Baptiste CRINIERE-BOIZET

Data Analysis Core

Theory

Hypotheses Checking

Interpretation

Theory

Dependent Variable and Independent Variables

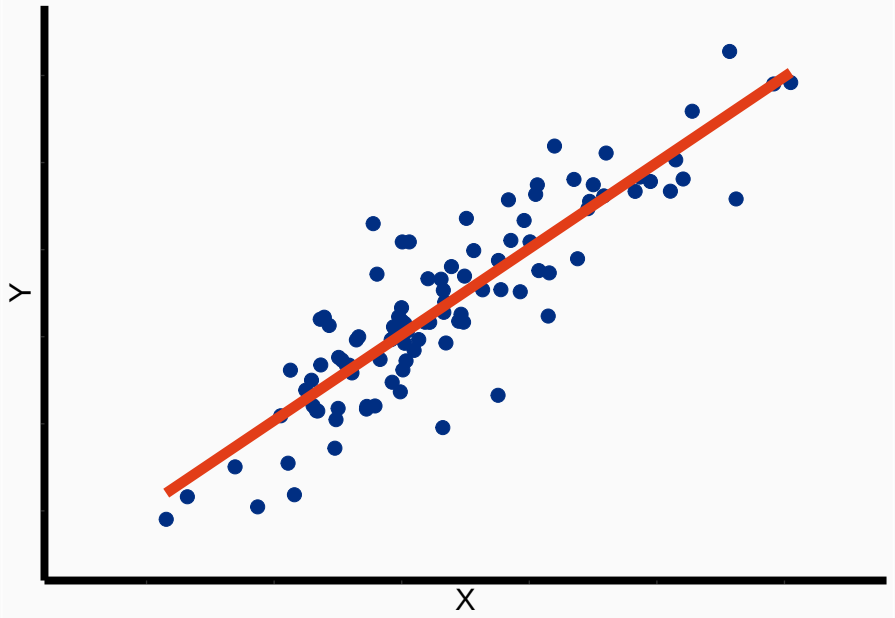
- Key Distinction: Dependent Variable (Y) vs Independent Variables (X1, X2, etc.)
- Dependent Variable = Response or Explained Variable
- Independent Variable = Predictor or Explanatory Variable
- Choice based on logic or theory, for example: child's height depending on the mother's height

Linear Equation

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

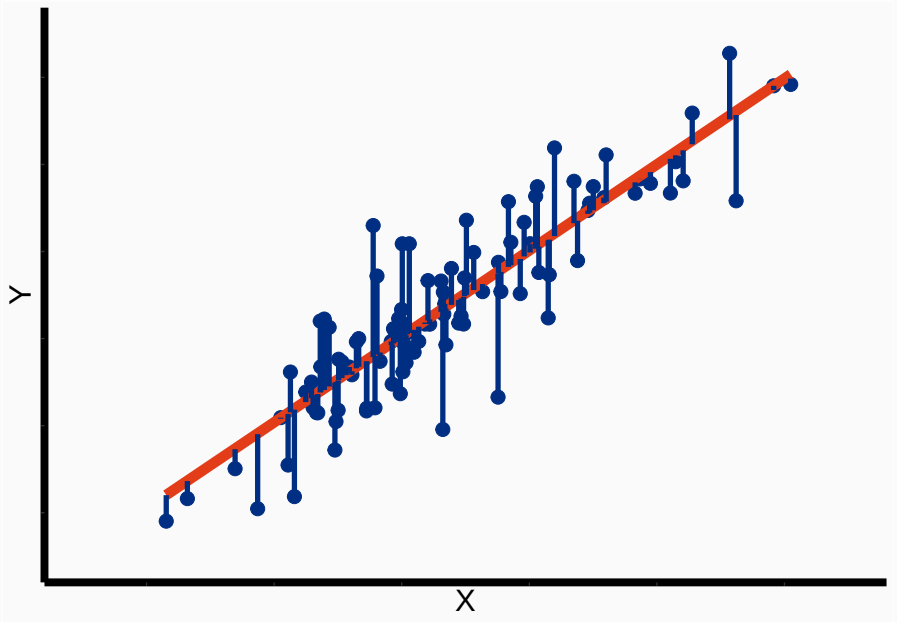
Linear Modeling

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



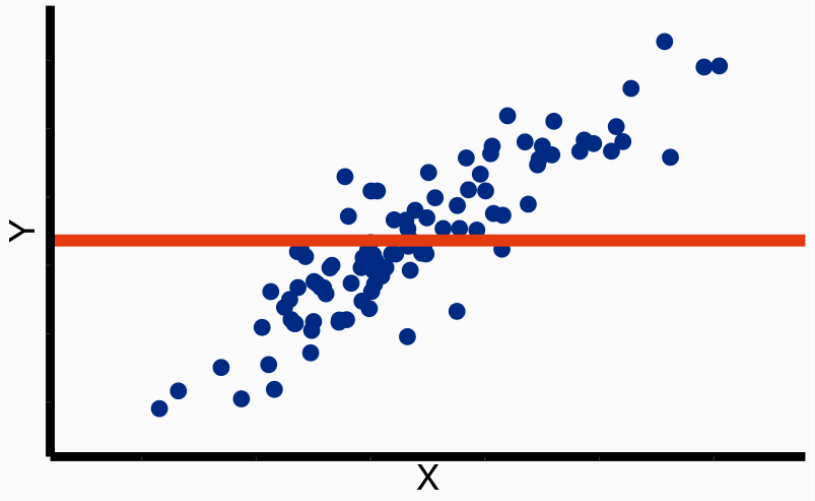
Residuals

- 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

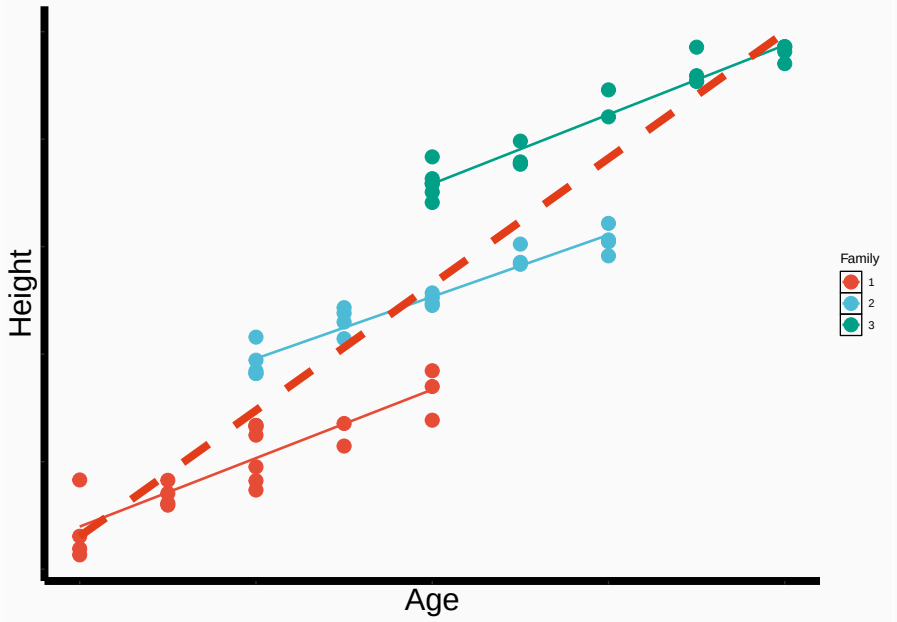
Hypotheses Checking

List of Key Hypotheses

- **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

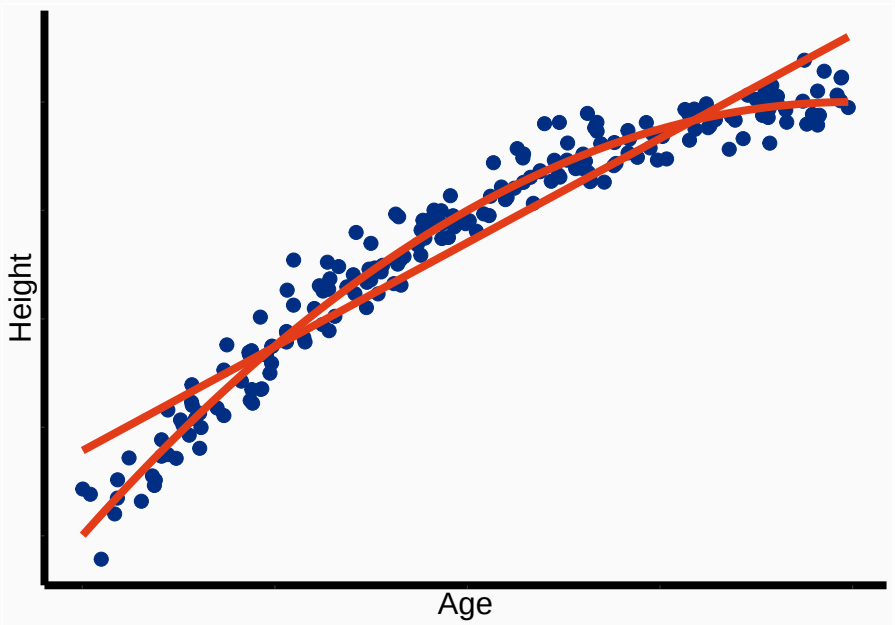
Independence

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



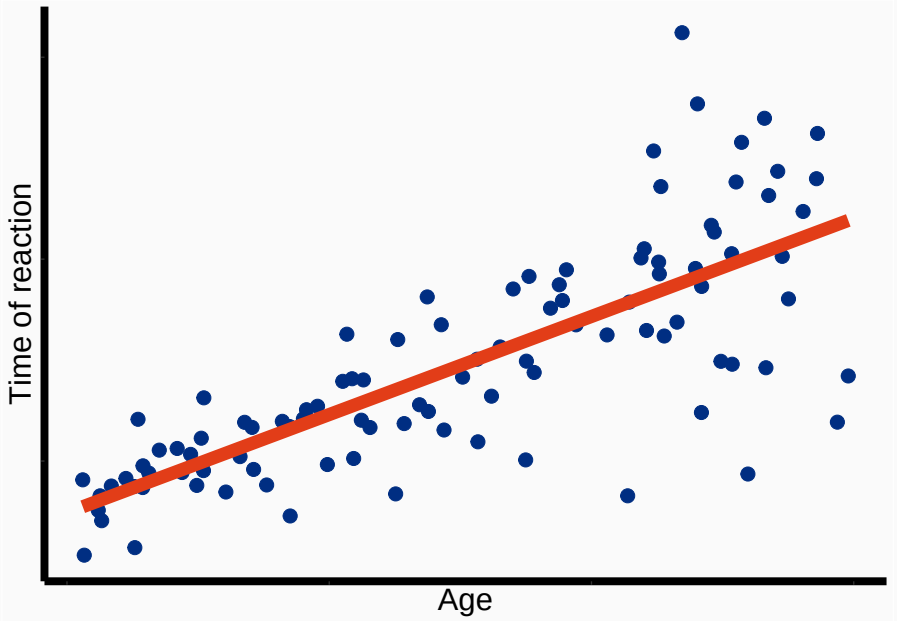
Linearity

- 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

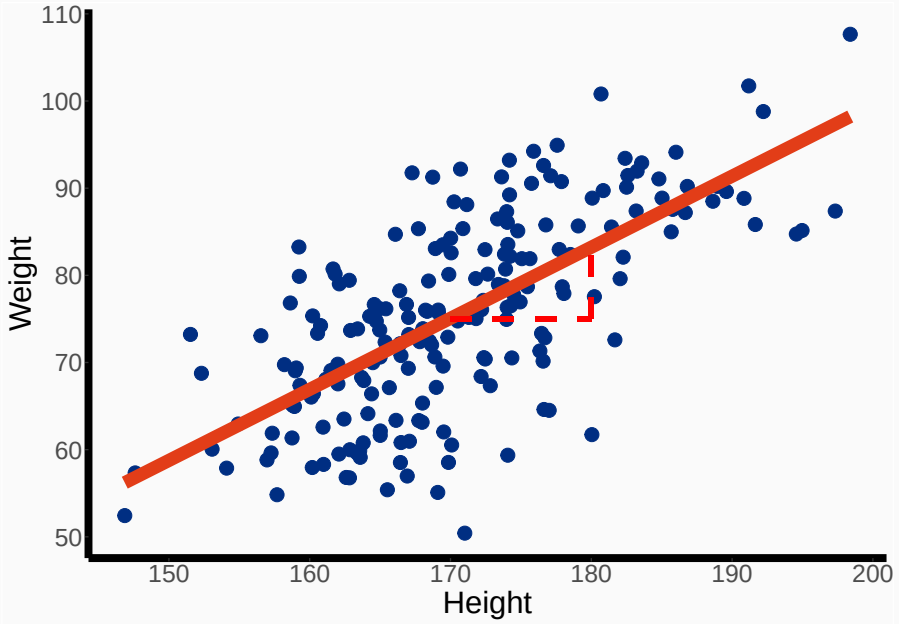
Simple regression

- **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

Simple regression

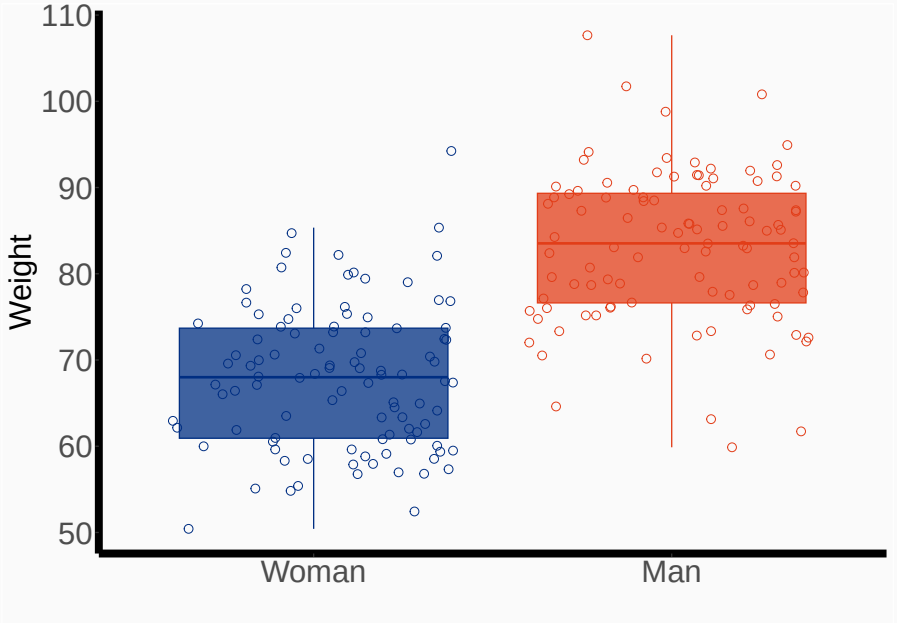
Continuous predictor

- **Model**:
  $Weight = \beta_0 + \beta_1 Height + \epsilon$
- $\beta_0 = -63$ and $\beta_1 = 0.81$,
  we also test the non-nullity
  of each coefficient
- R^2, the portion of variance
  explained by our model, is
  46%



Discrete predictor

- **Model**: $Weight = \beta_0 + \beta_1 \times 1_{\{Man\}} + \epsilon$
- $\beta_0 = 68$ and $\beta_1 = 15$, we also test the non-nullity of each coefficient
- R^2, the proportion of variance explained by our model, is 44%



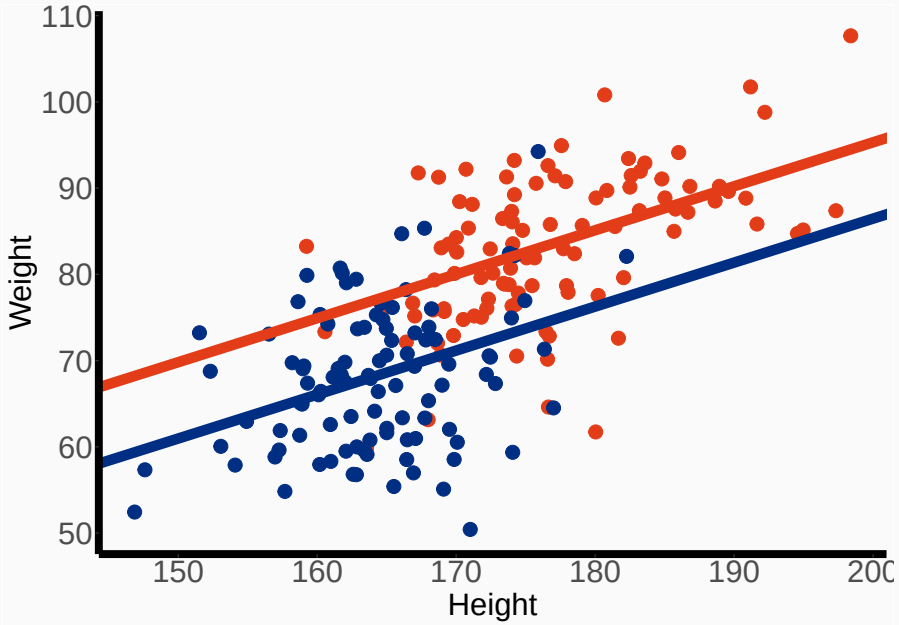
Multiple regression

- **Definition**: Regression with multiple explanatory variables
- **Adjust** our model by incorporating a new explanatory variable
- **Interaction**: when two variables simultaneously influence the variable being explained

Multiple regression : Adjust

Discrete Case

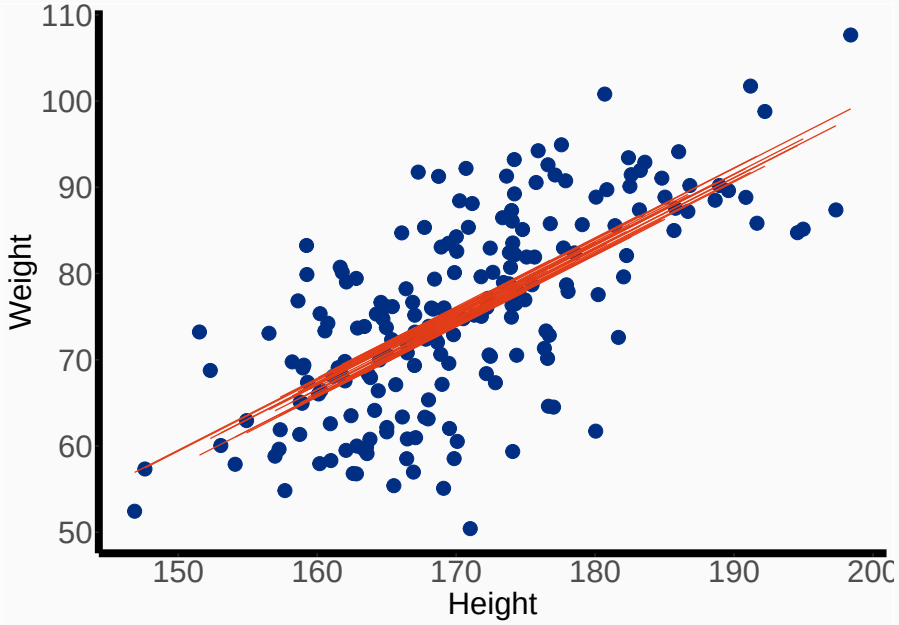
- **Model** :
  $Weight = \beta_0 + \beta_1 \times 1_{\{Man\}} + \beta_2 \times Height + \epsilon$
- Parallel lines representing each subgroup of a category
- R^2 the share of variance explained by our model is 55%



Multiple regression : Adjust

Continuous Case

- **Model**: $Weight = \beta_0 + \beta_1 \times Age + \beta_2 \times Height + \epsilon$
- Parallel lines representing each age unit



Multiple regression: interaction

Discrete Case

- **Definition** : An interaction occurs when the effect of one variable on the response variable depends on the level of another variable
- **Equation** :
  $Weight = \beta_0 + \beta_1 \times 1_{\{Man\}} + \beta_2 \times Height + \beta_3 \times 1_{\{Man\}} \times Height + \epsilon$
- Men have a steeper slope than women