Introduction à la régression
cours n°1

ENSM.SE – 1A

Olivier Roustant - Laurent Carraro

Problématique de la régression

- Expliquer une **réponse** y
  grâce à des **prédicteurs** $x_1, \ldots, x_p$

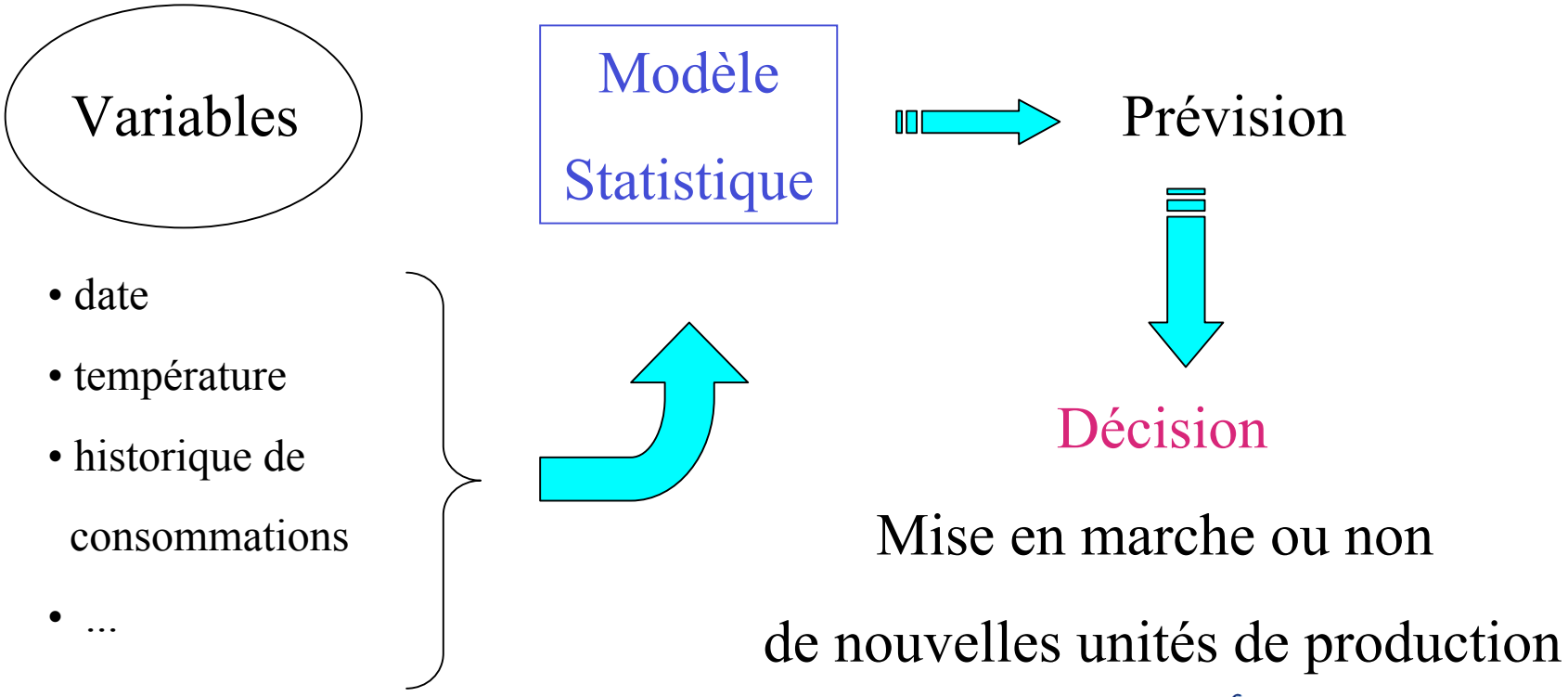
… dans un contexte incertain

… à partir d'expériences

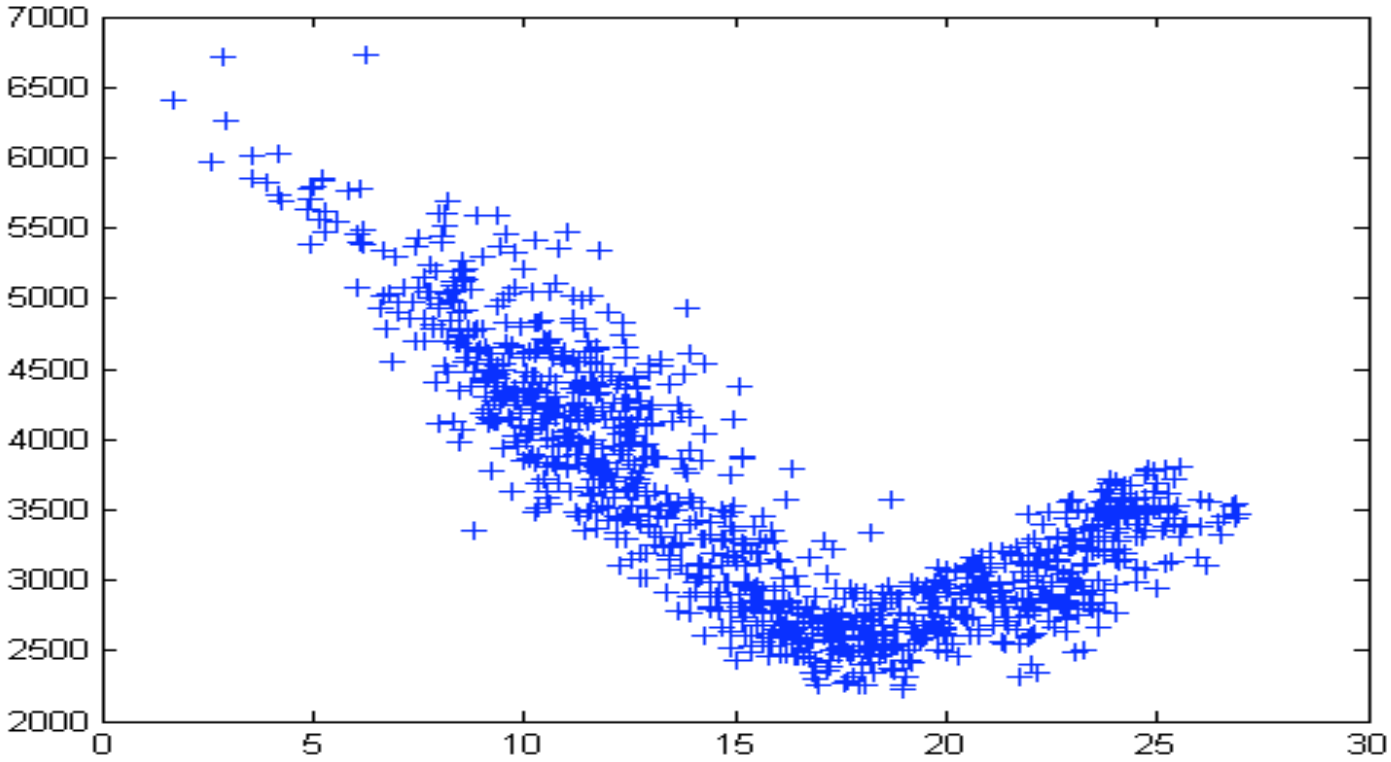
… dans un but **prédictif**

Exemple 1 : entreprise d'énergie

- But : prévoir la consommation électrique



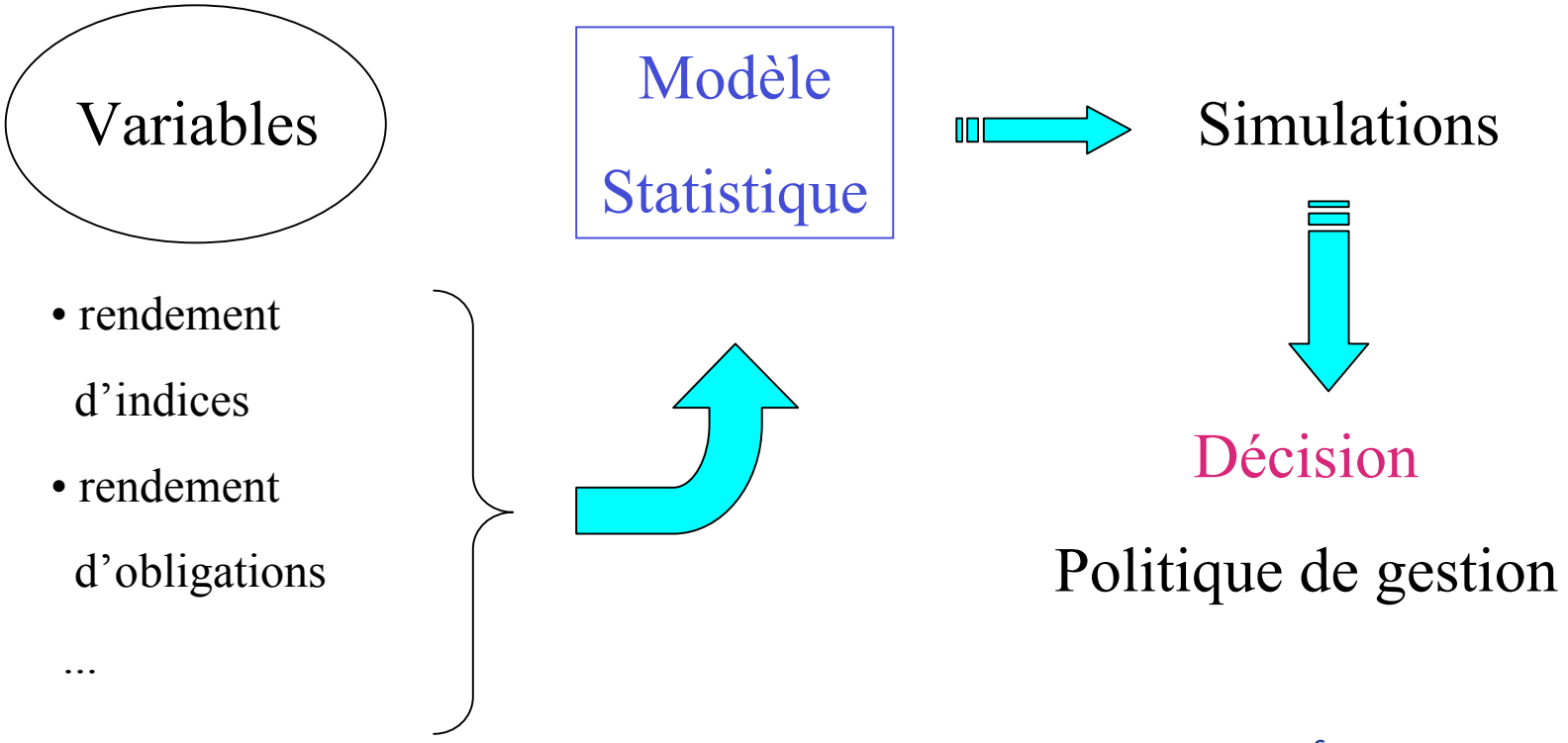
Consommation électrique (MgW)



Température (°C)

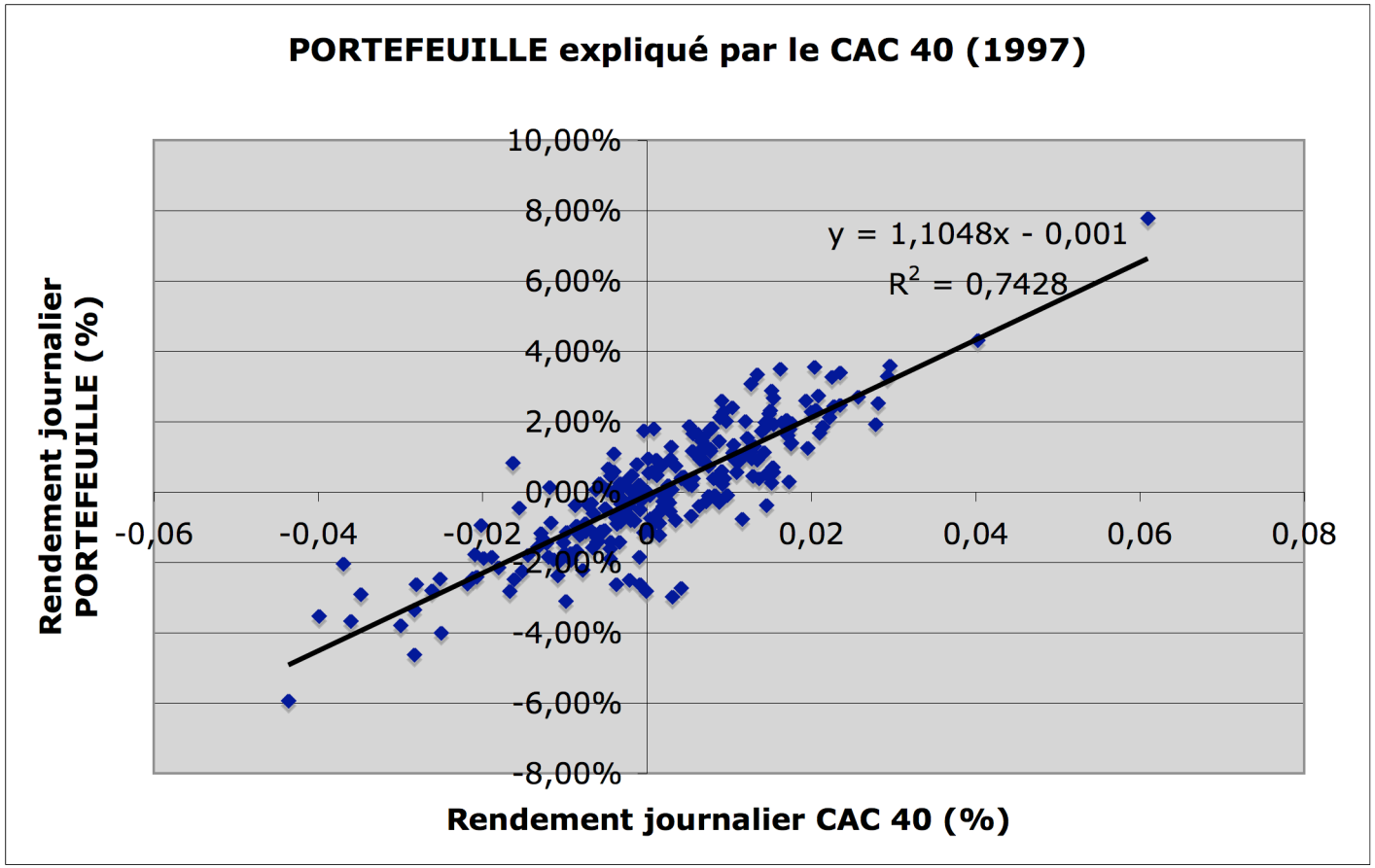
Exemple 2 : société financière

- But : estimer le risque d'un portefeuille



- Exemple. Soit un portefeuille constitué de 40% d'actions Lafarge et 60% Carrefour
  - Quel(s) prédicteur(s) choisir ?
  - Quel modèle proposer ?

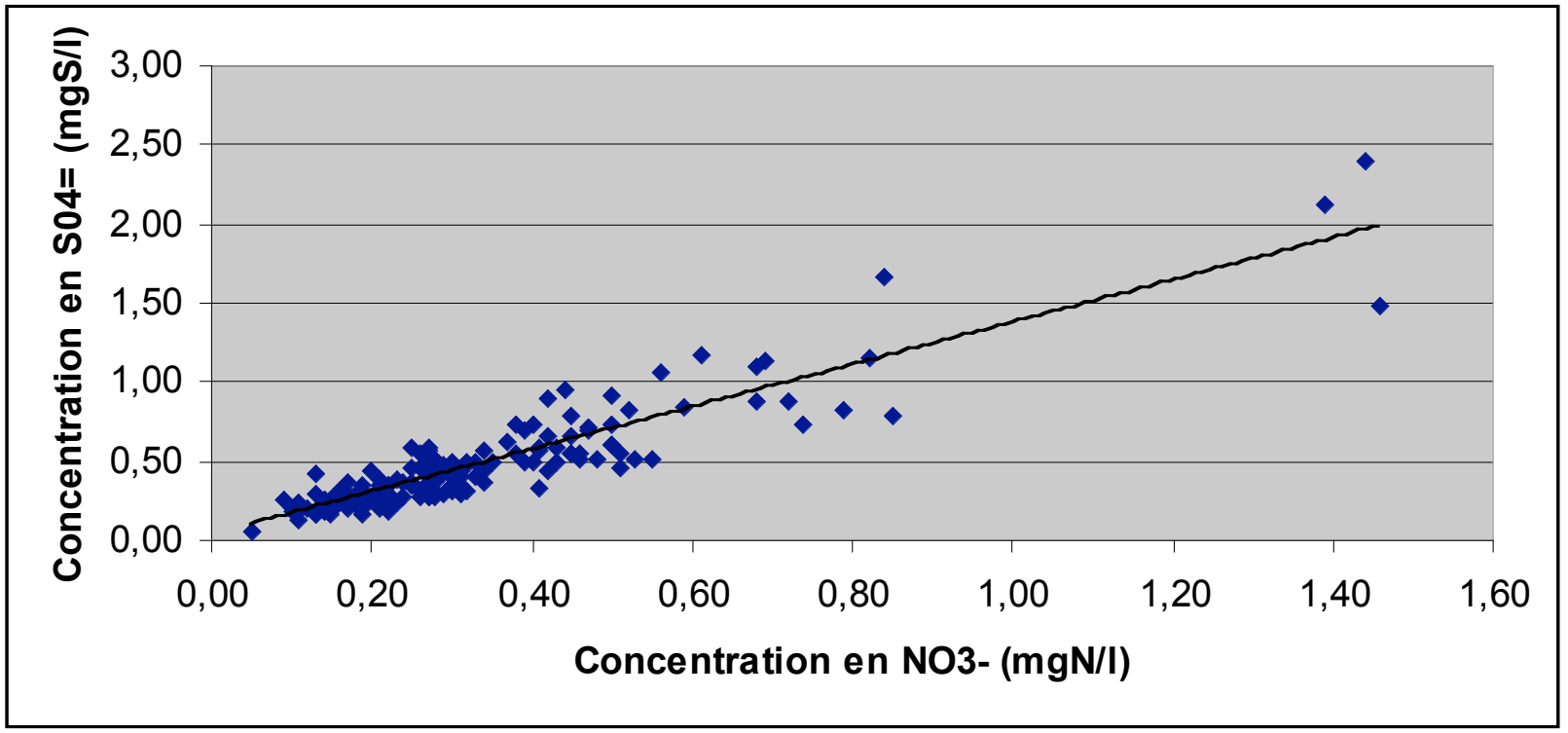
Portefeuille : 40% Lafarge ; 60% Carrefour



Exemple 3

- Réponse : concentration en ions SO4 --
- Prédicteur : concentration en ions NO3-

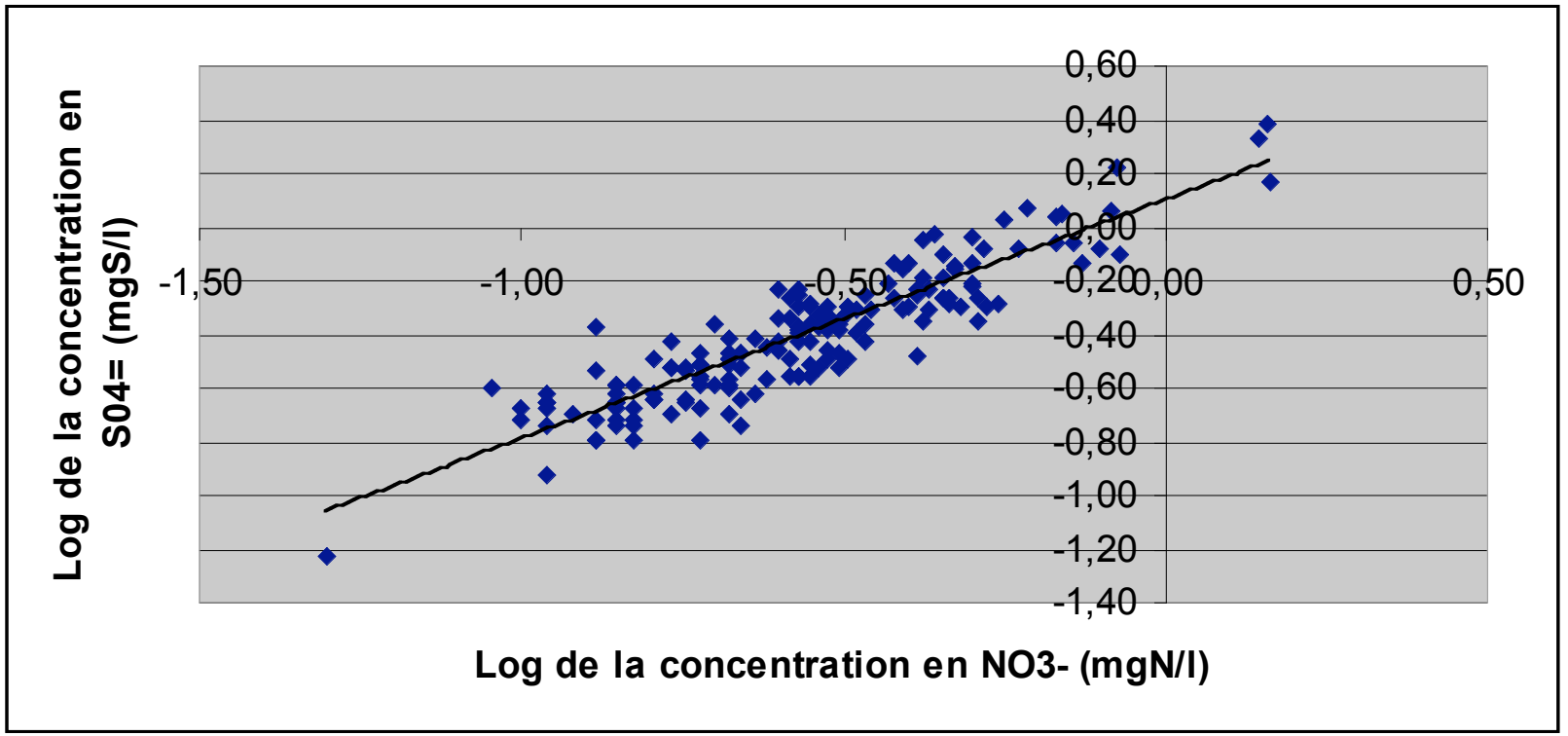
- Objectif : prévoir la concentration en ions S04-- connaissant celle en ions NO3-



Source : école des mines de Douai

Hétéroscédasticité = variance non constante

Résolution du problème d'hétéroscédasticité



Modélisation statistique : démarche

- Travail préliminaire
  - Etude des données
  - Choix des prédicteurs

  ⇒ proposition d'un modèle de régression
  - Plusieurs modèles possibles

- Etude statistique
  - Estimation, analyse des résultats
  - Validation du modèle : vérification des hypothèses
  - Si non validé, retour à la 1ère étape !

Dans ce cours…

- On se limite à 1 ou 2 prédicteurs
- On n'étudiera pas de données temporelles sauf si elles sont indépendantes dans le temps
  - Nécessite des connaissances en séries temporelles

Définition du modèle linéaire

- Y vecteur des réponses
- X matrice du plan d'expériences
- β vecteurs des paramètres
- ε vecteurs des écarts au modèle :

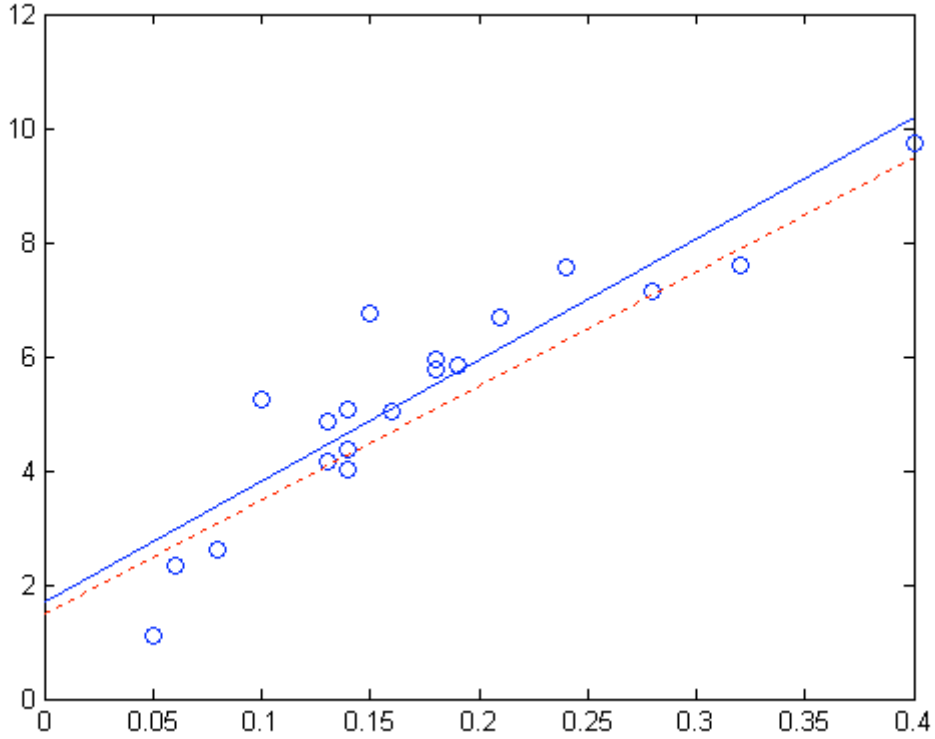
$$Y = X\beta + \varepsilon$$

$\varepsilon_1, \ldots, \varepsilon_n$ indépendants et de même loi $N(0, \sigma^2)$

Interprétation

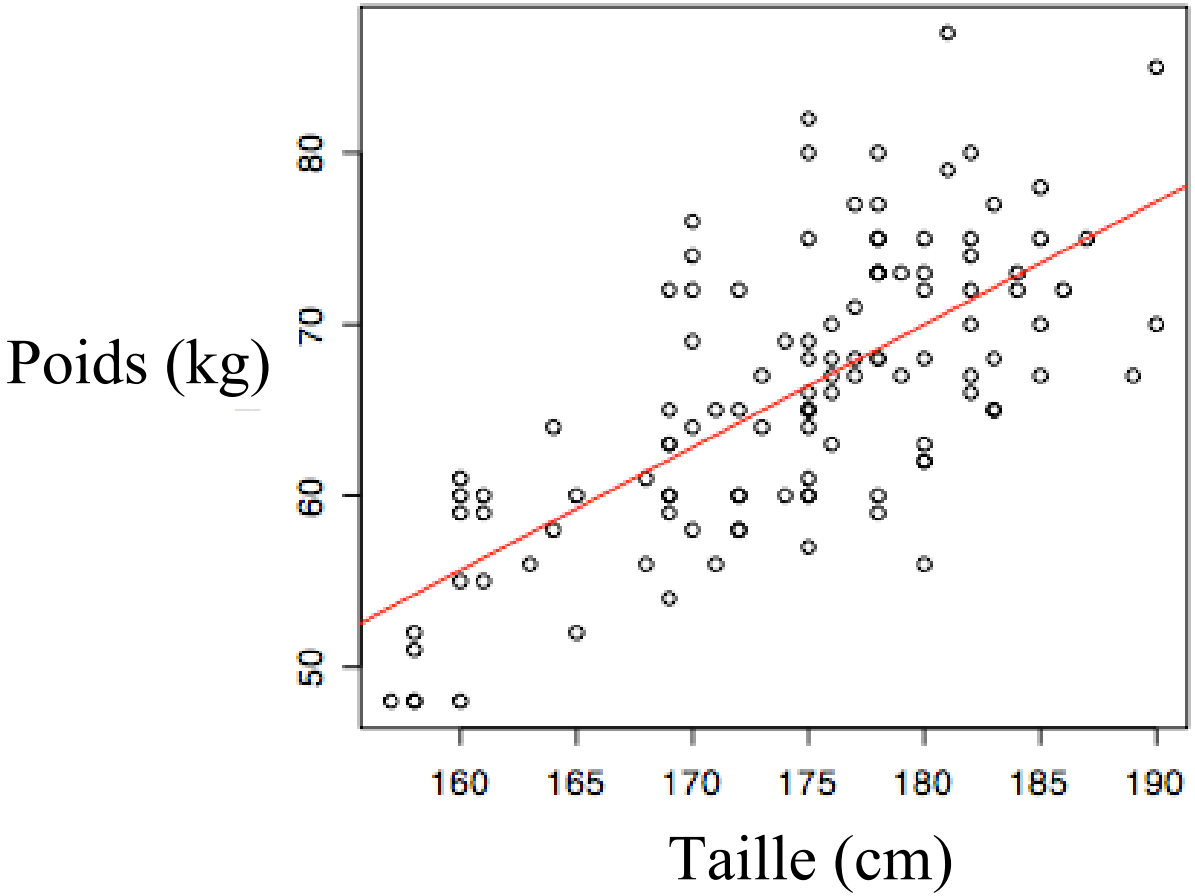
i.i.d. = indépendantes et identiquement distribuées

Générateur de jeux de *n* réponses i.i.d et de loi normale



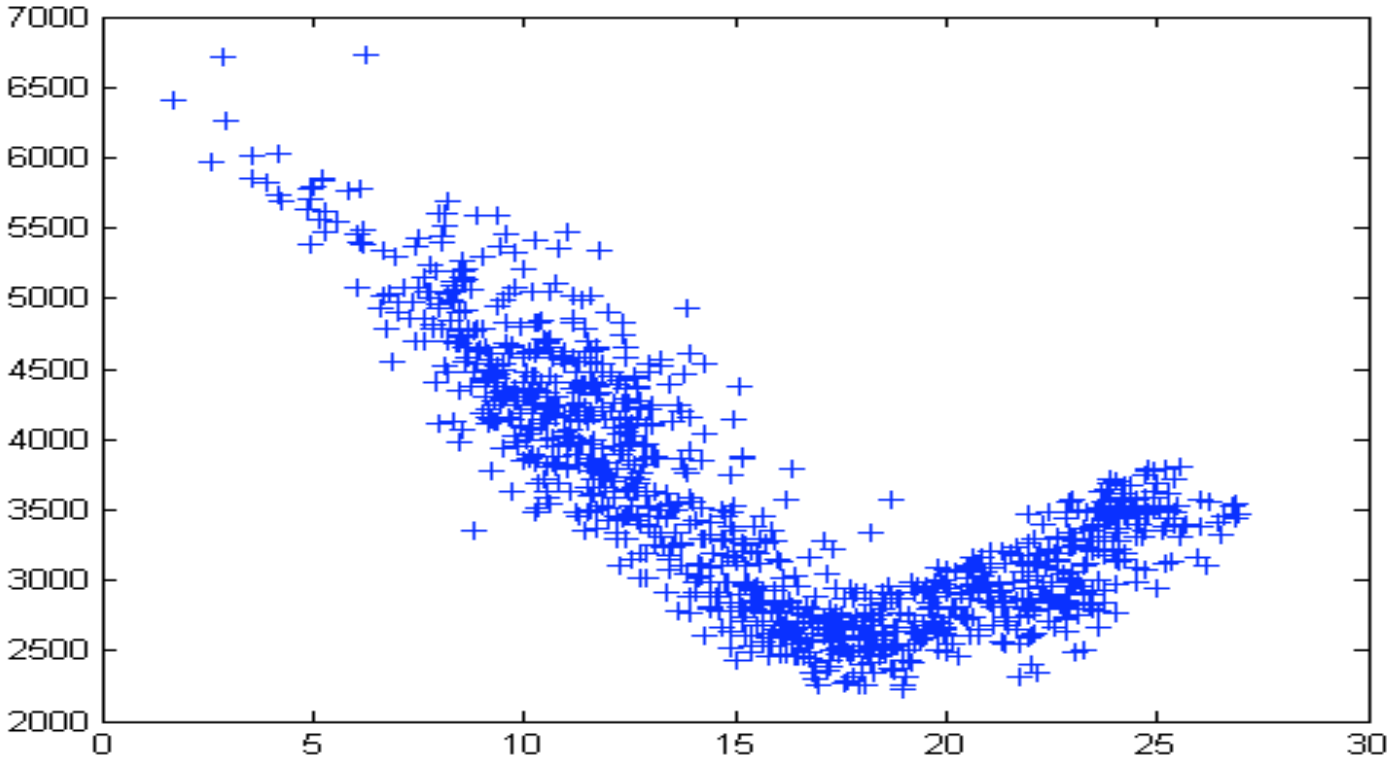
Exemples (1/2)

Que penser du modèle sur ces données ?



Exemples (2/2)

Et là ?



Exercice

On suppose que $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$ sont des v.a. i.i.d de loi $N(0, \sigma^2)$. Dans quels cas peut-on se ramener à un modèle de régression linéaire ?

1. $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$
2. $\log(y_i) = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$
3. $y_i = \beta_0 \exp(\beta_1 x_i) \times |\varepsilon_i|$
4. $y_i = \beta_0 / (1 + \beta_1 x_i) + \varepsilon_i$