Relation entre variables

Analyse de régression

variables avec lien de causalité

ex. variable indépendante (x) et dépendante (y)

Analyse de corrélation

variable interdépendantes, sans hiérarchie

« Correlation is not causation »

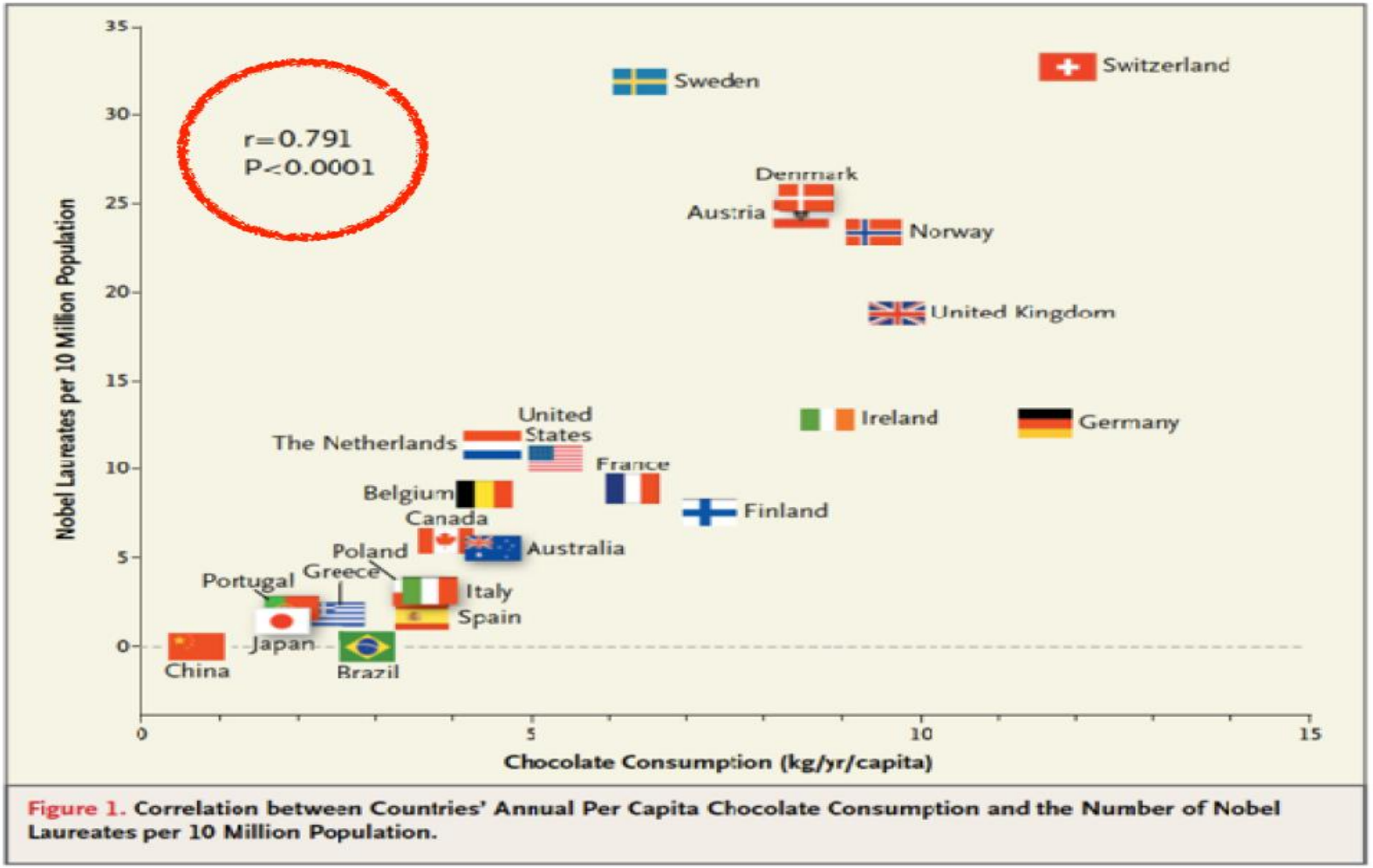
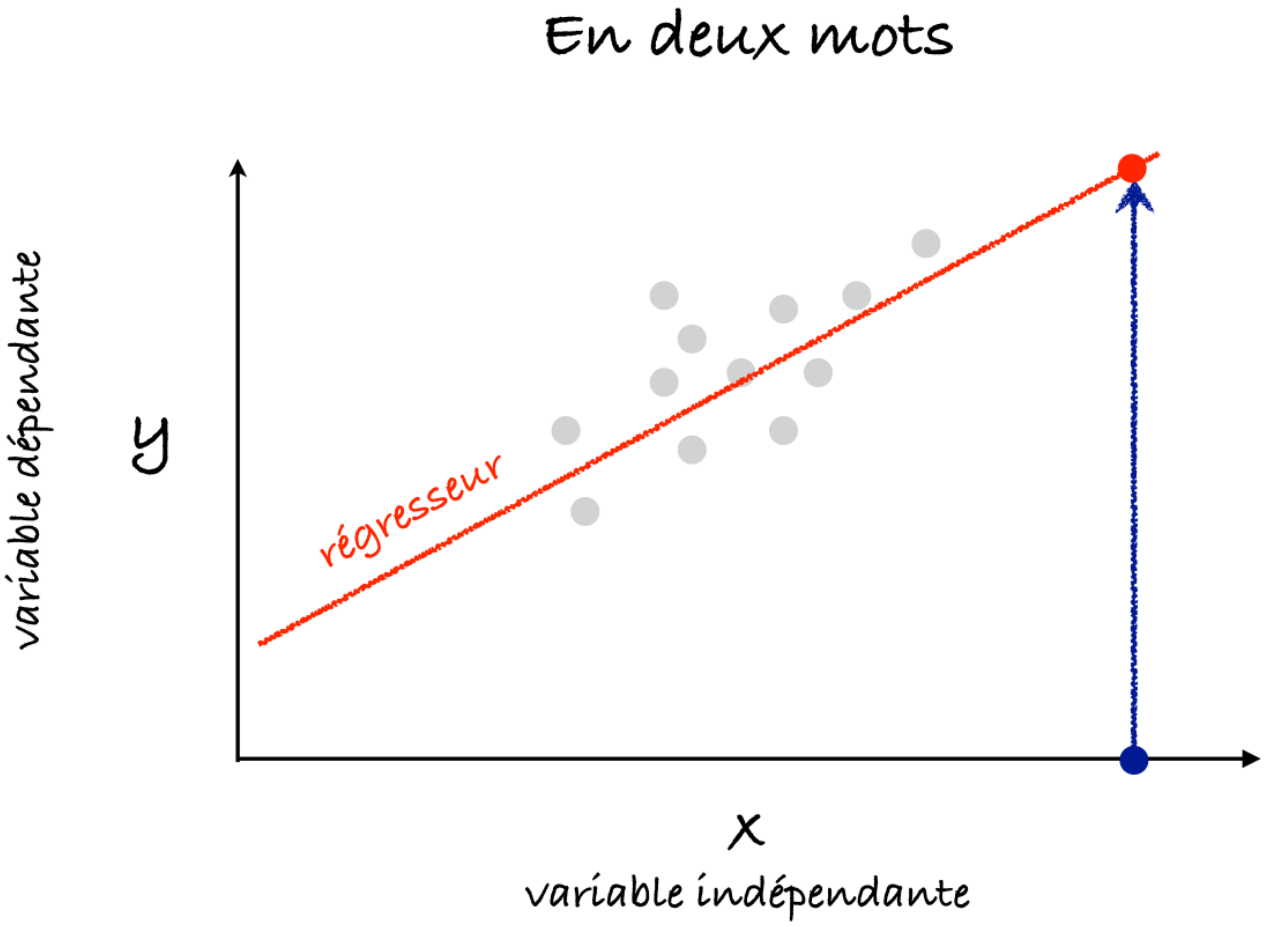


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Image credit: New England Journal of Medicine

Régression


En deux mots
variable dépendante
y
régresseur
x
variable indépendante

Modèles linéaires

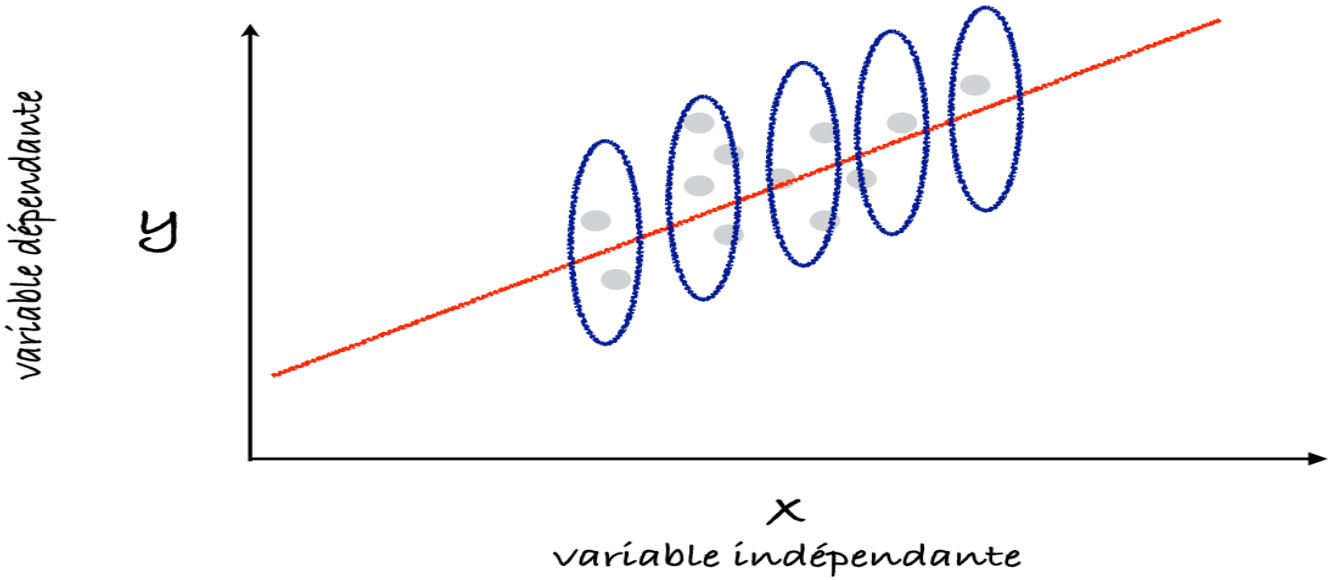
Régresseur

ligne théorique $y = a + b_1 x_1 + \cdots + b_n x_n$

Modèle

observations $y_i = a + b_1 x_{1,i} + \cdots + b_n x_{1,n} + \epsilon_i$

Méthode des moindres carrés

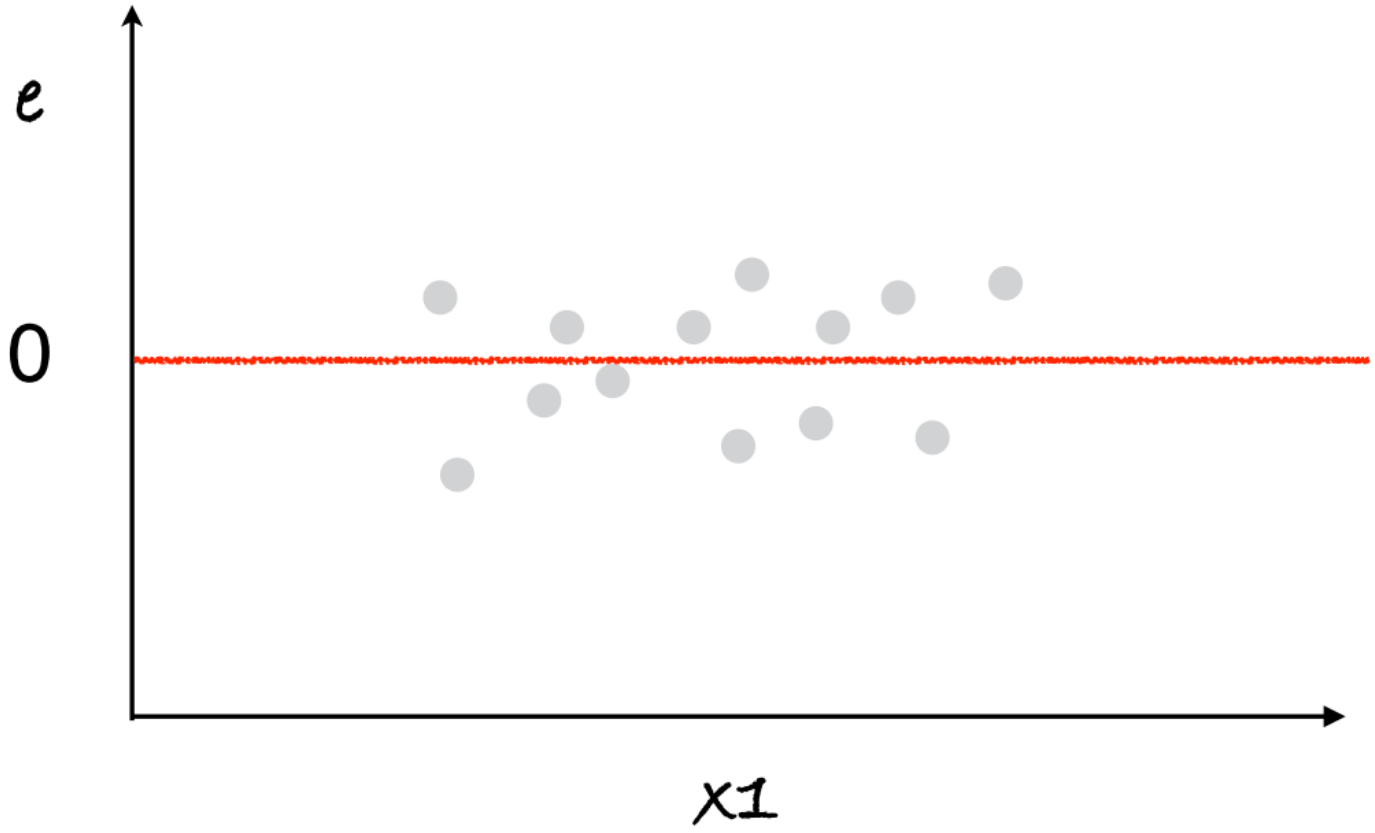


minimiser la somme quadratique des déviations des observations aux prédictions

$$\sum_i (y_i - y)^2$$

Résidus

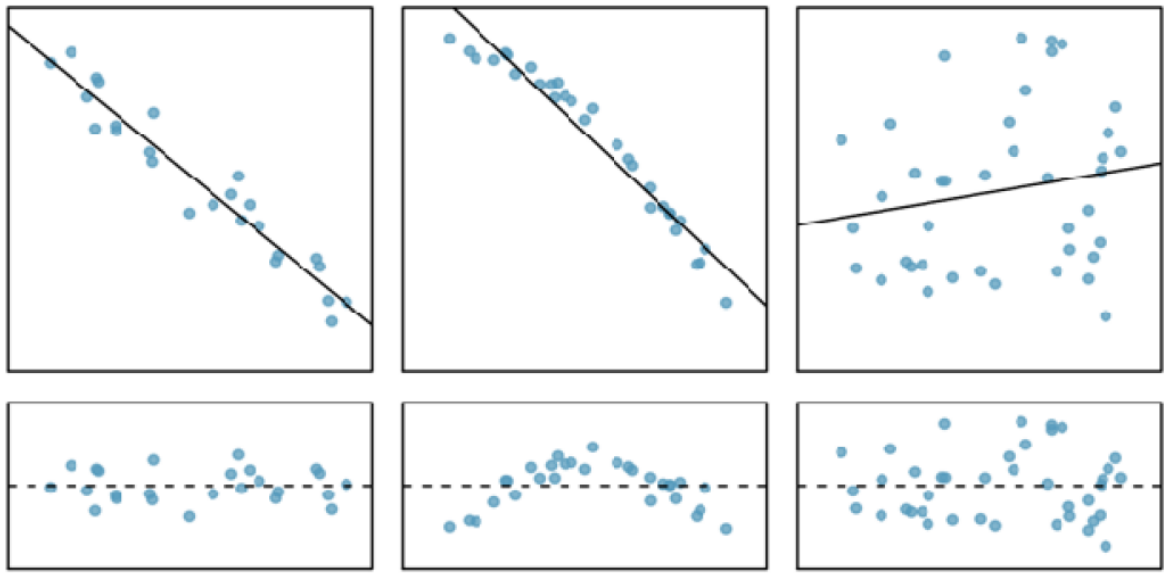
$$e_i = y_i - y$$



Résidus

$$e_i = y_i - y$$

y, yi



e

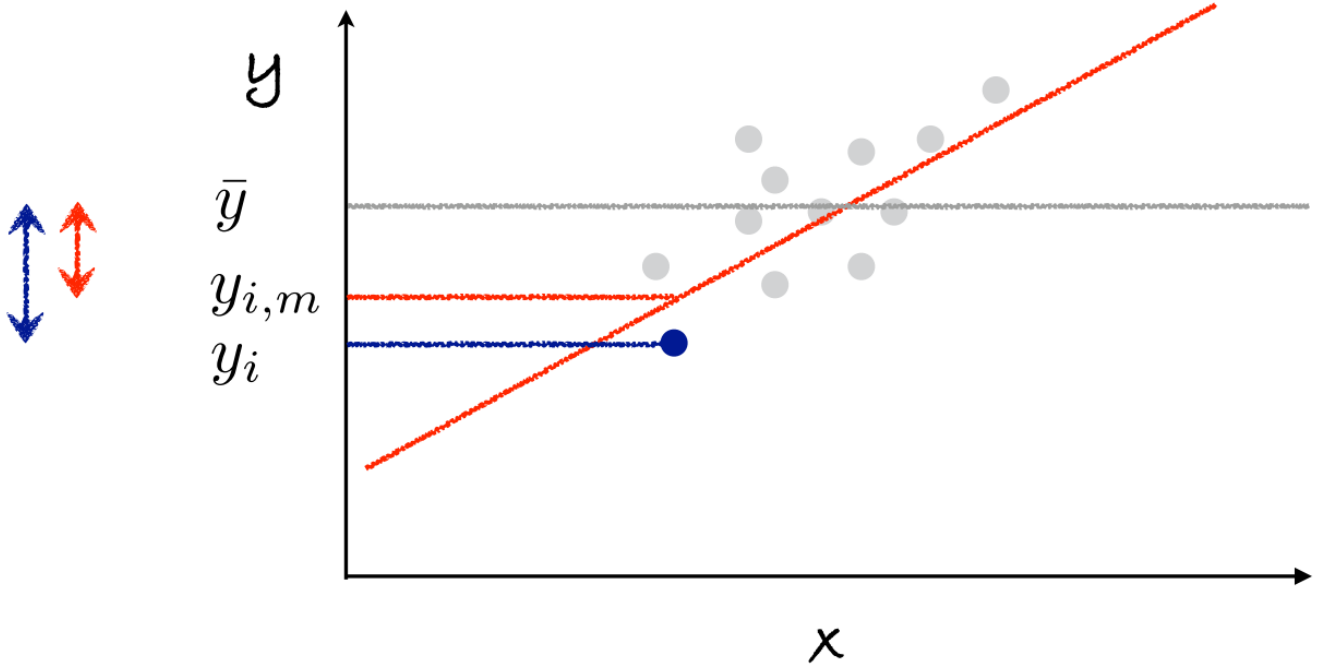
Figure 7.9: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

x1

Coefficient de corrélation (R)

$$R^2 = \frac{\sum_i (y_{i,m} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

fraction de la variation de y expliquée par la régression

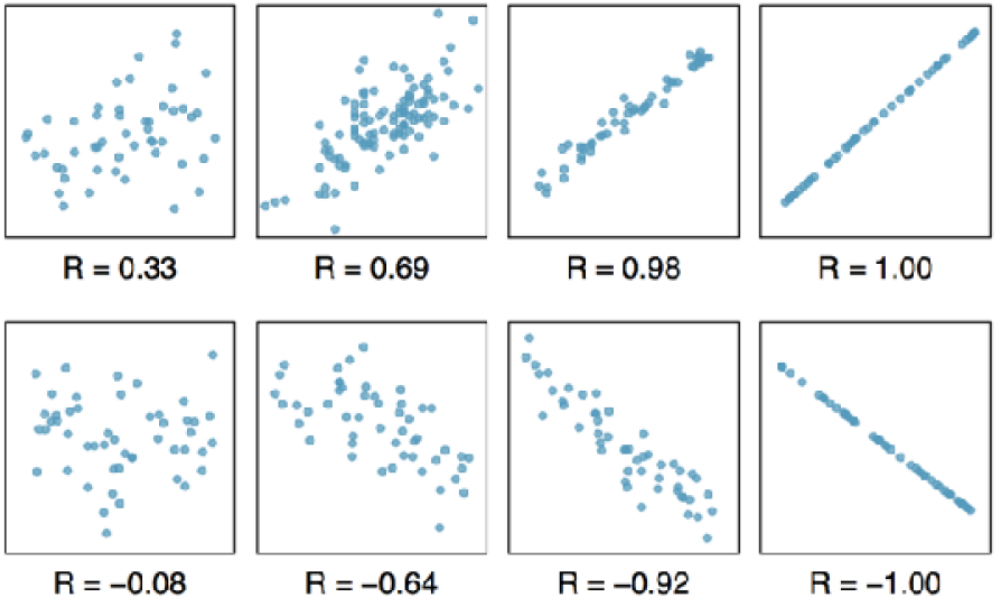


Coefficient de corrélation (R)

$$R^2 = \frac{\sum_i (y_{i,m} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

fraction de la variation de y expliquée par la régression

$$0 \leq R^2 \leq 1$$

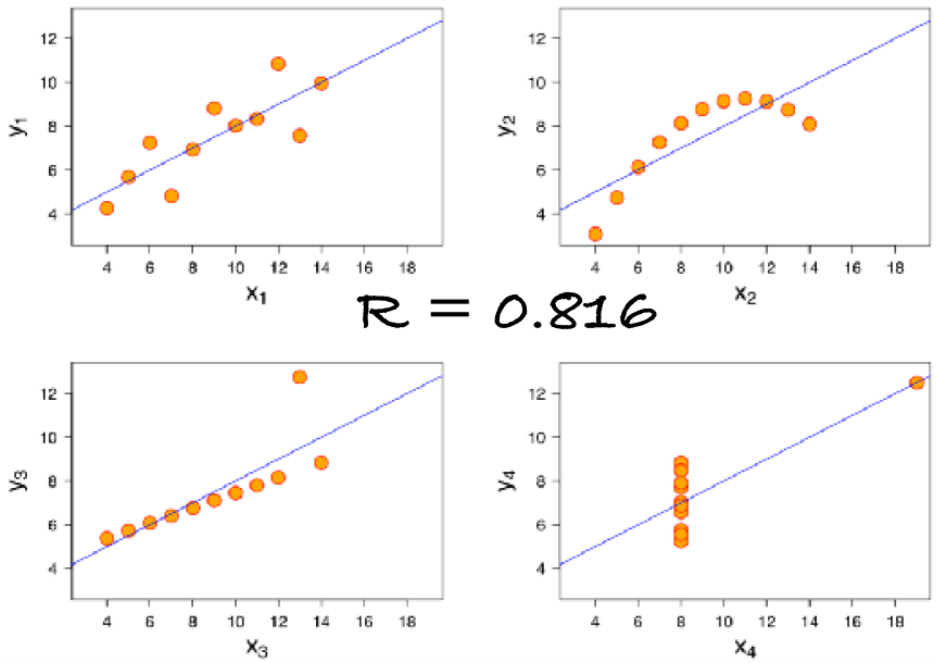


ATTENTION

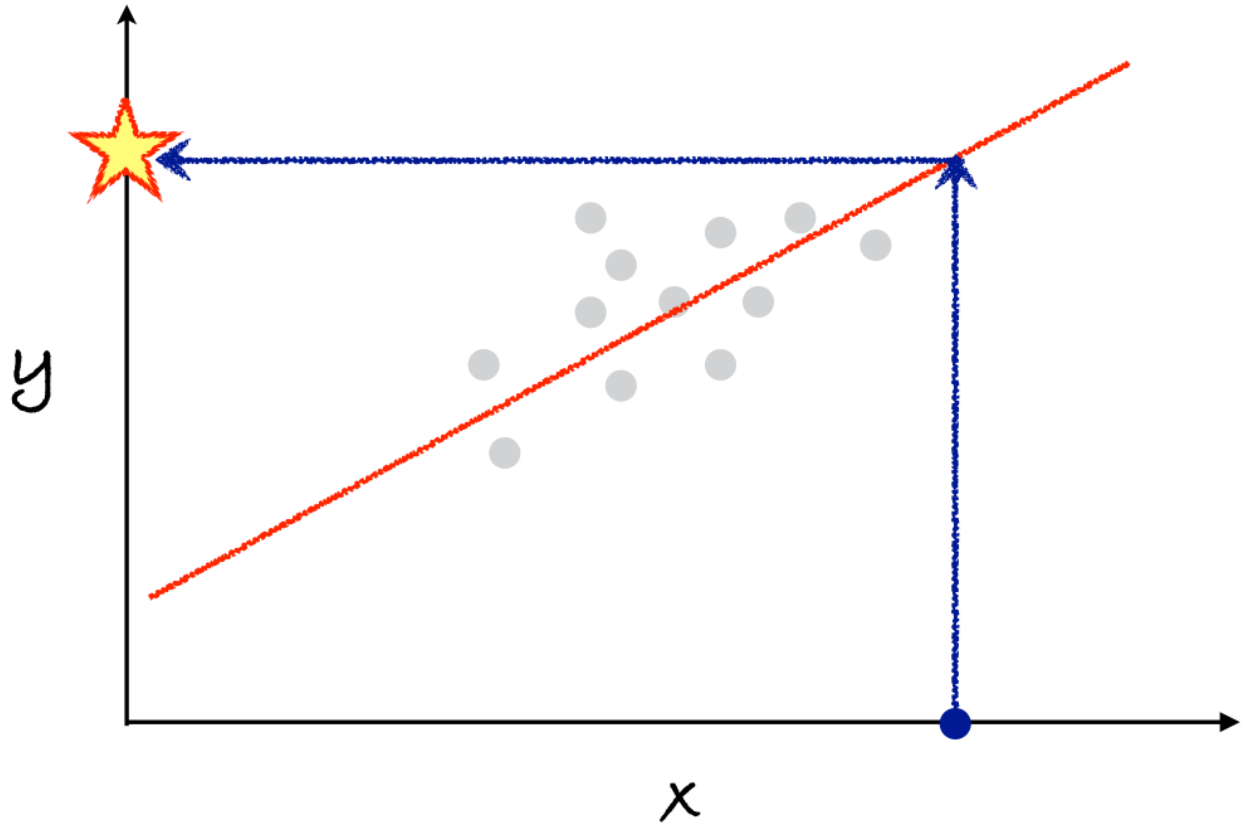
Regardez les données avant d'interpreter R

Quartet d'Anscombe

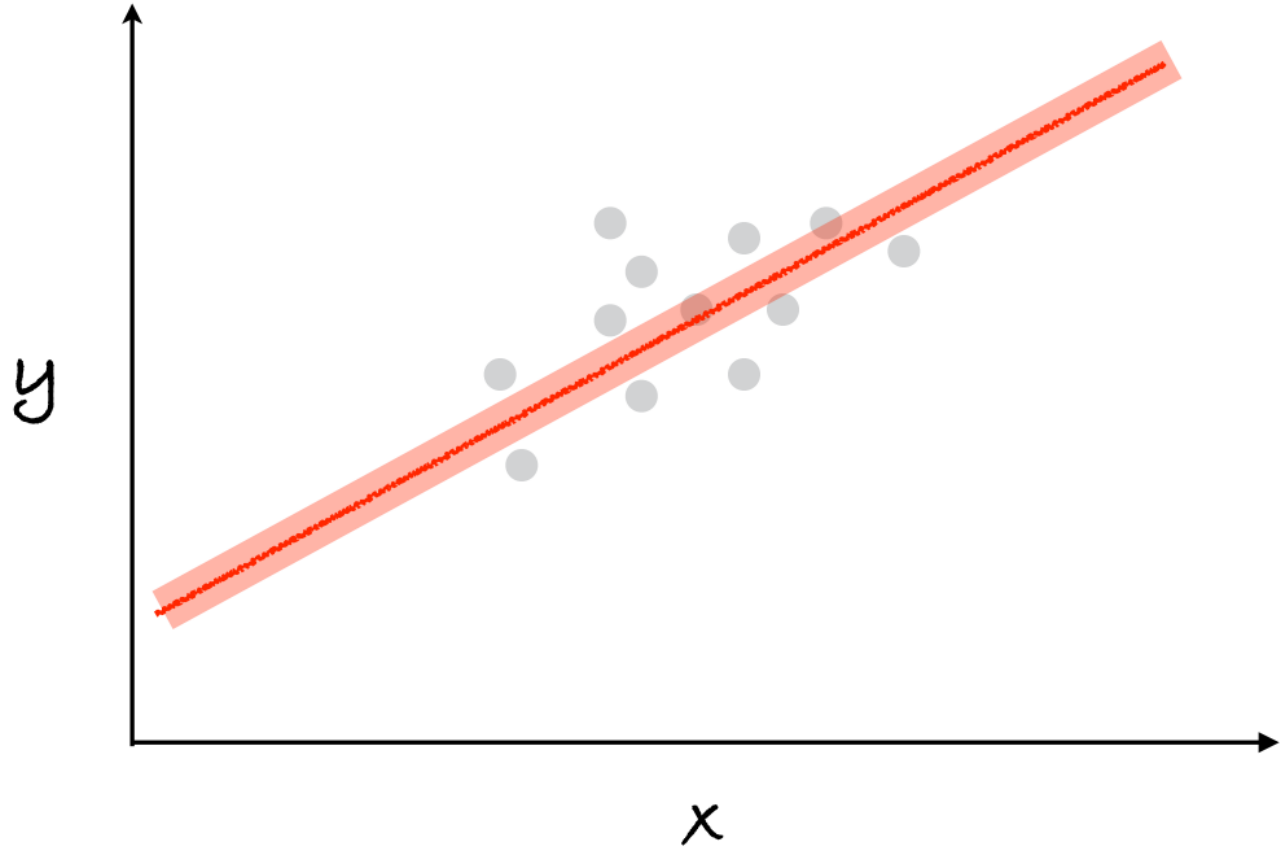
4 datasets très différents, mais avec même propriétés statistiques de base



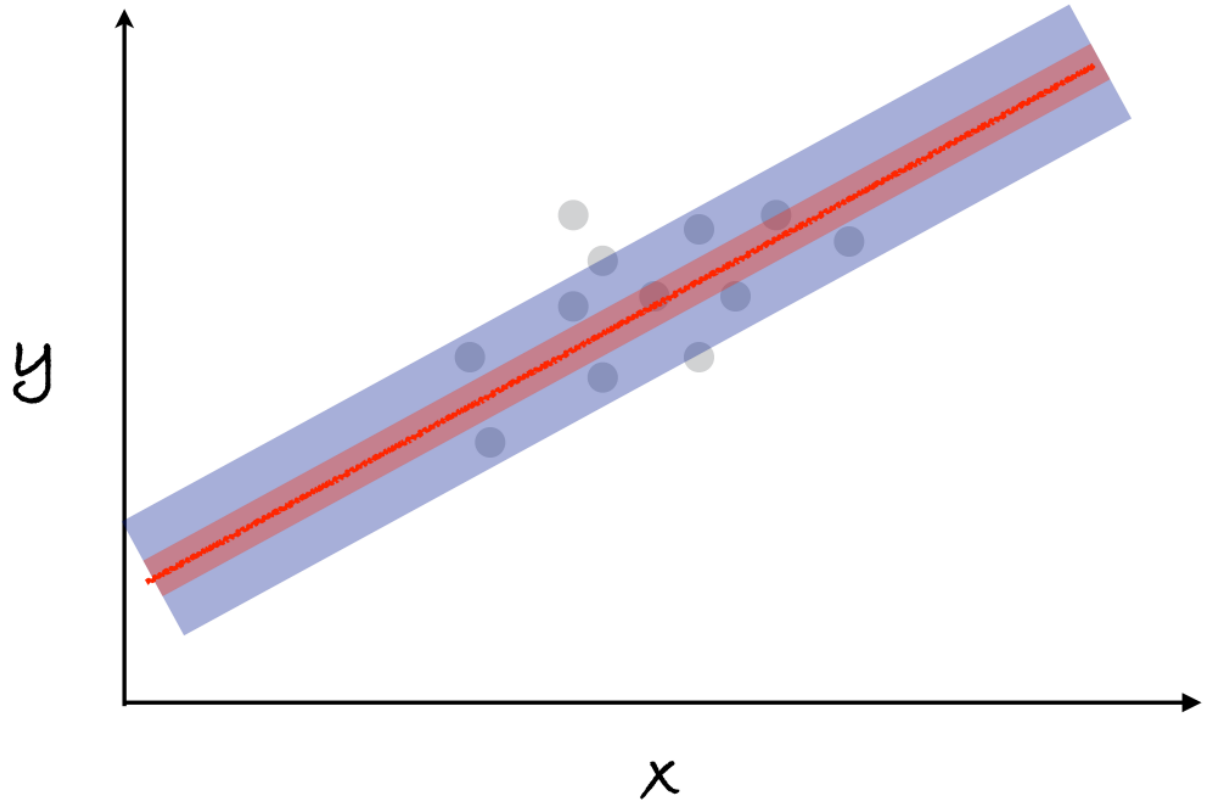
Prédictions



Erreur d’échantillonnage



Erreur de prediction



PREDICTION