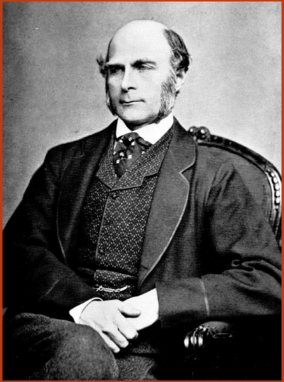
MOOC Statistique pour ingénieur
Thème 4 : Régression linéaire
Vidéo 1 : Mettre en œuvre la régression linéaire simple

Anca Badea   Lomig Hamon   François Seyte   Audrey Villot

Institut Mines-Télécom
Mines Saint-Étienne, Mines Nantes, Mines Alès, Mines Nantes

Sommaire

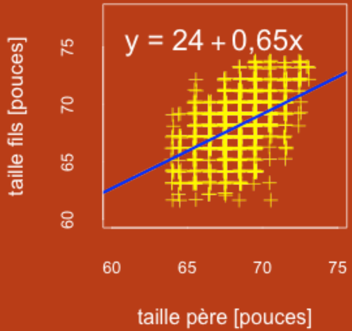
Origine



Francis Galton
(1822-1911)

Regression towards mediocrity in hereditary stature
Journal of the Anthropological Institute 15 : 246-63 (1886)

Origine



la taille des enfants nés des parents très grands (ou petits) se rapproche de la taille moyenne de la population
→ elle **régresse**

Objet

analyser la relation entre

- une variable expliquée
  - variable dépendante
  - réponse
  - variable endogène

et

- une ou plusieurs variables explicatives
  - variables indépendantes
  - prédicteurs
  - variables exogènes

dans un but

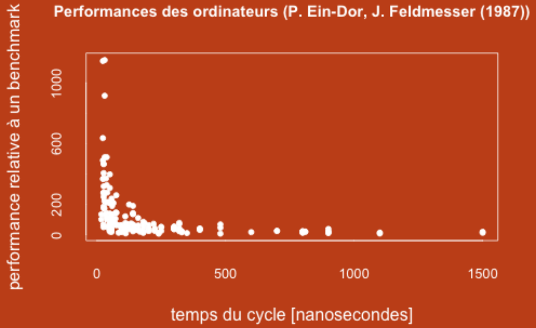
- explicatif
- de prévision
- . . .

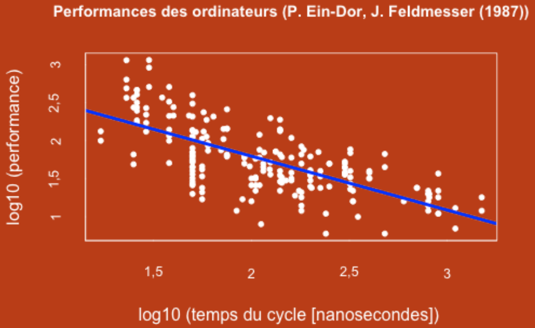
sous l'hypothèse que la relation est
linéaire en ses paramètres

et
à partir de données

Exemples

- industrie



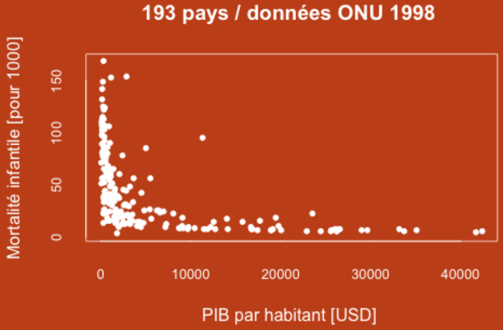


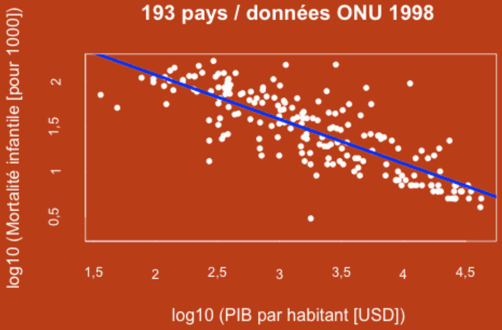
Exemples

- économie

Exemples

- économie



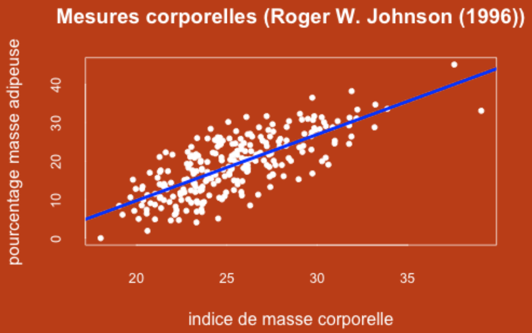


Exemples

- médecine

Exemples

- médecine



- contrôle qualité
- sociologie
- métrologie
- marketing
- ...

Sommaire

Vocabulaire

- **régression linéaire simple** :

  une seule variable expliquée
  une seule variable explicative

- régression linéaire multiple :

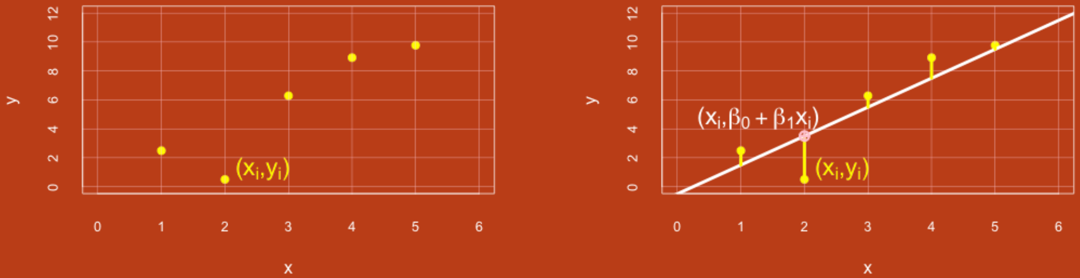
  une seule variable expliquée
  plusieurs variables explicatives

Notations

- y : la variable expliquée
- x : la variable explicative
- $y \approx \beta_0 + \beta_1 x$: la relation approximative entre y et x
- $\beta_0, \ \beta_1$: les paramètres (coefficients) du modèle
- $\hat{\beta}_0, \ \hat{\beta}_1$: les estimations des paramètres
- $y = \beta_0 + \beta_1 x + \varepsilon$ avec ε : l'erreur
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$: la prédiction de la variable expliquée

Sommaire

Les paramètres du modèle sont à estimer à partir des données $(x_i, y_i)_{i=1,\dots,n}$



interprétation : β_0 : l'ordonnée à l'origine, β_1 : la pente

Méthode des moindres carrés ordinaires

$$(\hat{\beta}_0, \hat{\beta}_1) = \operatorname{argmin} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\left\{ \begin{array}{rcl} \dfrac{\partial}{\partial \beta_0} \displaystyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 & = & 0 \\ \dfrac{\partial}{\partial \beta_1} \displaystyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 & = & 0 \end{array} \right.$$

$$\left\{ \begin{array}{rcl} \hat{\beta}_0 & = & \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 & = & \dfrac{\mathbb{C}\mathrm{ov}(x, y)}{s_x^2} \end{array} \right.$$

les estimations des paramètres

Méthode des moindres carrés ordinaires

notations usuelles

- moyennes : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- variances : $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$
- covariance : $\mathbb{C}\mathrm{ov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Exemple

$(x_i, y_i)_{i=1,\ldots,n}$

$$(1;\ 3,36), (2;\ 0,71), (3;\ 5,27), (4;\ 7,55), (5;\ 9,01)$$

$$\hat{\beta}_0 = -0,27\,;\, \hat{\beta}_1 = 1,82$$

Sommaire

Somme des carrés

$y_i = \hat{y}_i + \varepsilon_i$
avec

$\hat{y}_i$ la part expliquée par le modèle et

ε_i la part inexpliquée : l'erreur

Et les sommes des carrés correspondantes ?

Somme des carrés

- somme des carrés totaux : $SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- somme des carrés expliqués : $SCE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
- somme des carrés résiduels : $SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Equation d'analyse de la variance

$$SCT = SCE + SCR$$

Indicateur de qualité d'une régression

coefficient de détermination : $R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$

R^2 doit être proche de 1

rôle du statisticien dans l'interprétation

INSTITUT
Mines-Télécom