Modélisation linéaire simple

Définition

Soient $x = (x_1, \dots, x_n)$ et $y = (y_1, \dots, y_n)$ des données statistiques.
La variable x est appelée **variable exogène** ou *expliquée*.
La variable y est appelée **variables endogène** ou *à expliquer*.
Une **modélisation linéaire simple** consiste à considérer les variables aléatoires

$$Y_i = ax_i + b + \varepsilon_i$$

où es ε_i sont des variables aléatoires i.i.d. appelés **termes d'erreurs** et suivent une loi normale $\mathcal{N}(0, \sigma)$.

Définition Covariance

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \quad \Big| \quad \sigma_{x,y} = \overline{(x - \overline{x})(y - \overline{y})}$$

Lemme :

Soient X, Y et Z des variables aléatoires réelles et α et β des nombres réels.

1. $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
2. $X \perp\!\!\!\perp Y \Rightarrow \mathrm{Cov}(X, Y) = 0$
3. $\mathrm{Cov}(X, X) = \mathbb{V}(X)$
4. $\mathrm{Cov}(X + \alpha, Y) = \mathrm{Cov}(X, Y)$
5. $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$
6. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
7. $\mathbb{V}(\alpha X + \beta Y) = \alpha^2 \mathbb{V}(X) + 2\alpha\beta \mathrm{Cov}(X, Y) + \beta^2 \mathbb{V}(Y)$

Soient $x = (x_1, \dots, x_n)$ et $y = (y_1, \dots, y_n)$ des données statistiques et α et β des nombres réels. Notons $xy = (x_1 y_1, \dots x_n y_n)$.

1. $\sigma_{x,y} = \overline{xy} - \overline{x}.\overline{y}$
2. $\sigma_{x,x} = \sigma_x^2$
3. $\sigma_{x+\alpha,y} = \sigma_{x,y}$
4. $\sigma_{x+y,z} = \sigma_{x,z} + \sigma_{y,z}$
5. $\sigma_{x,y} = \sigma_{y,x}$
6. $\sigma_{\alpha x+\beta y} = \alpha^2 \sigma_x^2 + 2\alpha\beta\sigma_{x,y} + \beta^2 \sigma_y^2$

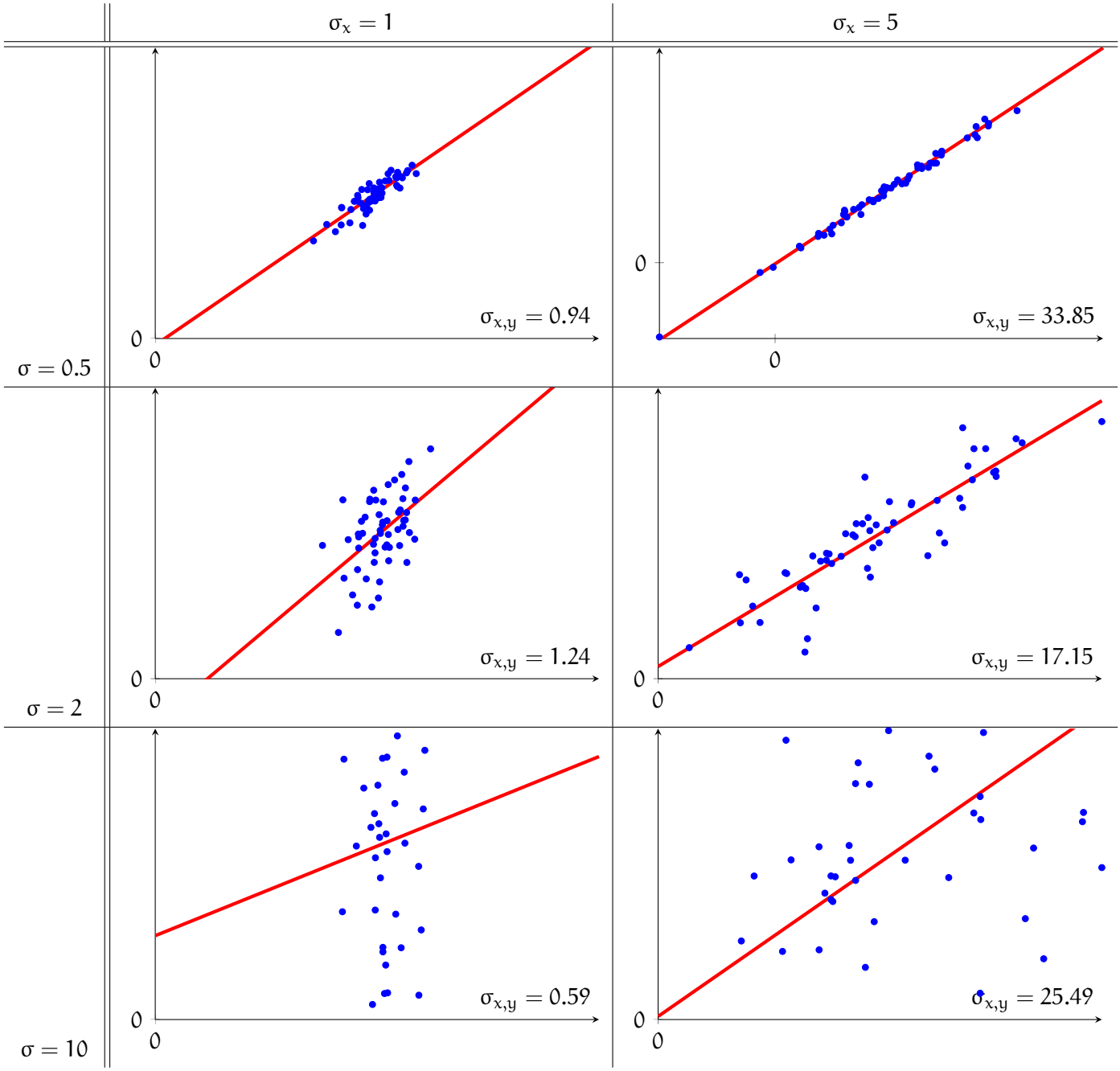
Théorème Inégalité de Cauchy-Schwartz

$$\mathrm{Cov}(X, Y)^2 \leqslant \mathbb{V}(X)\mathbb{V}(Y) \qquad \sigma_{x,y}^2 \leqslant \sigma_x^2 \sigma_y^2$$

Théorème

$$a = \hat{a} = \frac{\sigma_{x,y}}{\sigma_x^2} \qquad b = \hat{b} = \overline{y} - \hat{a}\overline{x}$$

$d(x) = \hat{a}x + \hat{b}$ est appelé la **droite de régression linéaire**.



Définition

Notons $\hat{y}_i = \hat{a}x_i + \hat{b}$.

On appel **résidus du modèle** les valeurs $\hat{\varepsilon}_i = y_i - \hat{y}_i$

Proposition

Les résidus d'une modélisations linéaire simple ont une moyenne nulle.

Définition

coefficient de détermination du modèle

$$R^2_{x,y} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Proposition

$$R^2_{x,y} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Si $R^2_{x,y}$ est proche de 1 alors $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ es proche de 0 ce qui signifie que le modèle est très proche des valeurs : c'est un bon modèle.

Si $R^2_{x,y}$ est proche de 0 alors $\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$ est proche de 0 et les $\hat{y}_i$ approche la moyenne : ce modèle n'est pas bon.

Proposition

Les estimateurs suivants

$$A_n = \frac{\sigma_{x,Y}}{\sigma_x^2} = a + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \frac{x_i - \overline{x}}{\sigma_x^2}$$

et

$$B_n = \overline{Y}_n - A_n \overline{x} = b + \overline{\varepsilon}_n + (A_n - a)\overline{x}$$

sont des estimateurs convergents et dans biais de a et b. De plus

$$\mathbb{V}(A_n) = \frac{\sigma^2}{n} \left(\frac{1}{\sigma_x^2} \right) \qquad \text{et} \qquad \mathbb{V}(B_n) = \frac{\sigma^2}{n} \left(1 + \frac{\overline{x}}{\sigma_x^2} \right)$$

Théorème

Soit $\hat{\mathfrak{E}}_i = Y_i - \hat{y}_i = ax_i + b + \varepsilon_i - \hat{a}x_i - \hat{b}$.

$$\mathbb{E}\left(\sum_{i=1}^{n} \hat{\mathfrak{E}}_i^2\right) = (n-2)\sigma^2$$

Corollaire

La variable aléatoire

$$S_n = \frac{\sum_{i=1}^{n} \hat{\mathfrak{E}}_i^2}{n-2}$$

est un estimateur convergent et sans biais de σ^2. En particulier :

1. $S_n^a = S_n \dfrac{1}{n\sigma_x^2}$ est un estimateur convergent et sans biais de $\mathbb{V}(A_n)$.
2. $S_n^b = S_n \dfrac{\sigma_x^2 + \overline{x}}{n\sigma_x^2}$ est un estimateur convergent sans biais de $\mathbb{V}(B_n)$.

De plus $(n-2)\dfrac{S_n}{\sigma^2} \sim \chi^2(n-2)$.

Corollaire

$$\frac{A_n - a}{\sqrt{S_n^a}} \sim \mathcal{T}(n-2) \qquad \text{et} \qquad \frac{B_n - b}{\sqrt{S_n^b}} \sim \mathcal{T}(n-2)$$

Corollaire

Soient $0 < \beta < \alpha < 1$, $t_1 = Q_{\mathcal{T}(n-2)}(\beta)$ et $t_2 = Q_{\mathcal{T}(n-2)}(1-\alpha+\beta)$ alors

$$\left[A_n - t_2\sqrt{S_n^a}; A_n - t_1\sqrt{S_n^a}\right]$$

est un intervalle de confiance $1-\alpha$ de a et

$$\left[B_n - t_2\sqrt{S_n^b}; B_n - t_1\sqrt{S_n^b}\right]$$

est un intervalle de confiance de niveau $1-\alpha$ de b.