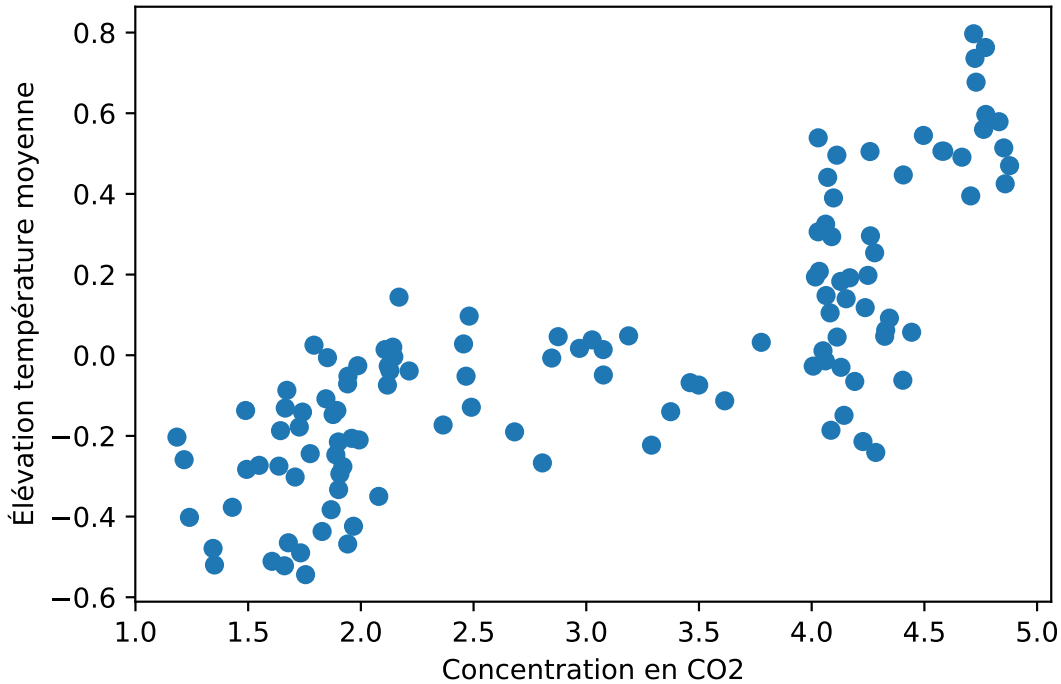
Régression Linéaire Simple

Benoit Gaüzère, Stéphane Canu
benoit.gauzere@insa-rouen.fr

INSA Rouen Normandie - ITI

March 20, 2025

Observation de deux variables



Statistiques descriptives

- Analyse univariée
- Analyse bivariée : dépendance linéaire

Comment expliciter cette relation ?

Le problème de la régression linéaire

Les données

n couple d'observations:

- $x_i \in \mathbb{R}$: la variable explicative
- $y_i \in \mathbb{R}$: la variable à expliquer, prédire

Le problème

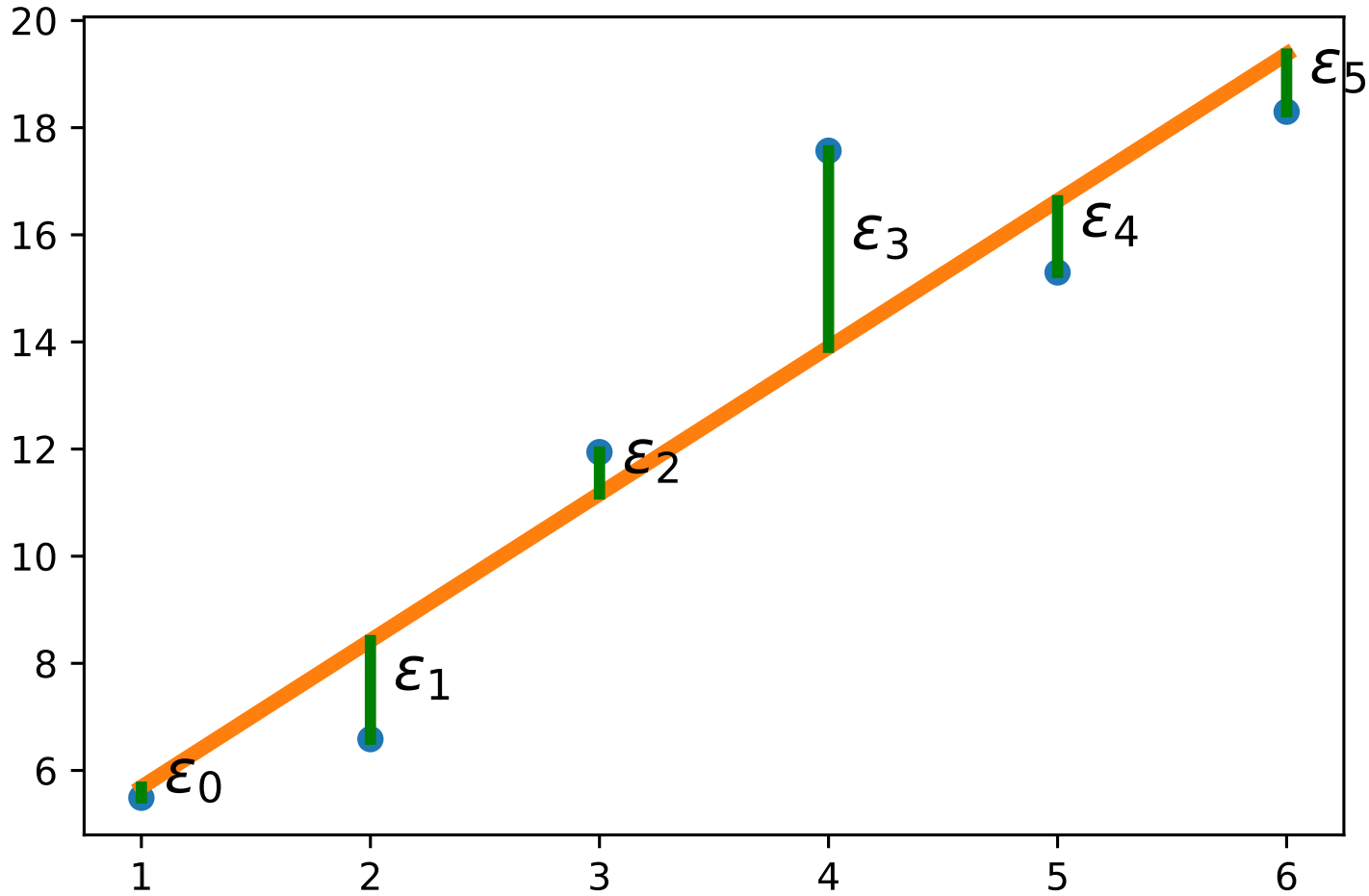
- Trouver la droite qui représente au mieux la relation linéaire

$$\widehat{y} = f(x) \simeq ax + b$$

- **Mais** les points ne sont pas forcément sur la droite

$$\widehat{y_i} = ax_i + b + \varepsilon_i$$

- Modélise le **bruit**



- $a \simeq 2,73$
- $b \simeq 2.95$
- $\varepsilon_i \simeq [-0.197, -1.840, 0.781, 3.670, -1.339, -1.076]$

Le Modèle Linéaire

Définition : Modèle Linéaire

Le modèle linéaire pose la relation suivante entre la variable explicative x et la variable à expliquer y avec les paramètres inconnus (a, b, ε)

$$y = ax + b + \varepsilon \text{ avec } \mathbf{a} = (a, b) \in \mathbb{R}^2$$

- Hypothèse : Observations = modèle + bruit
- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

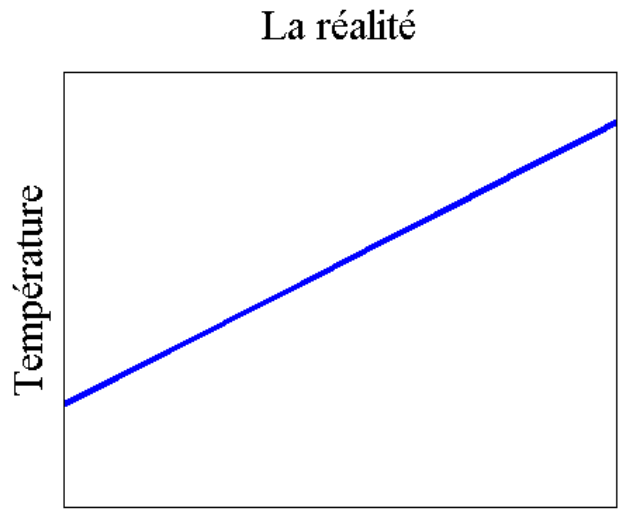
Résumé du vocabulaire

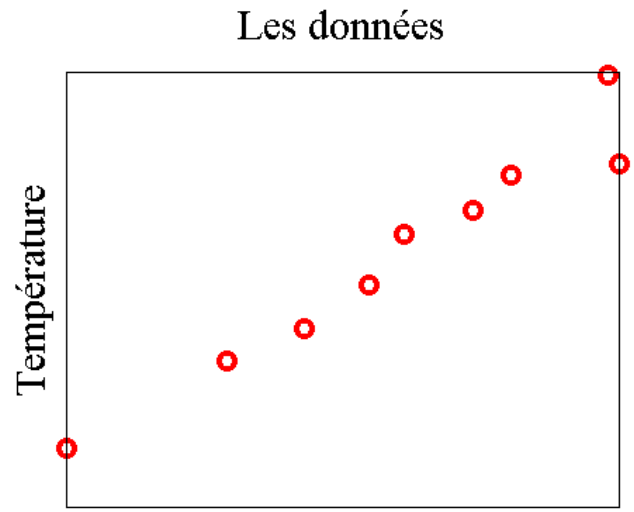
- Variables explicatives ..... $x \in \mathbb{R}$
- Variable à expliquer ..... $y \in \mathbb{R}$
- Erreur ou bruit ..... ε
- Paramètres scalaires ..... $a, b \in \mathbb{R},$
- Paramètres (forme vectorielle) ..... $\mathbf{a} \in \mathbb{R}^2$
- Modèle ..... $y = f(\mathbf{x}, \mathbf{a}) + \varepsilon$
- Estimation ..... $a^\star$
- Prédiction ..... $\widehat{y} = f(x, a^\star)$
- Variables aléatoires ..... ε et donc y et donc $a^\star$

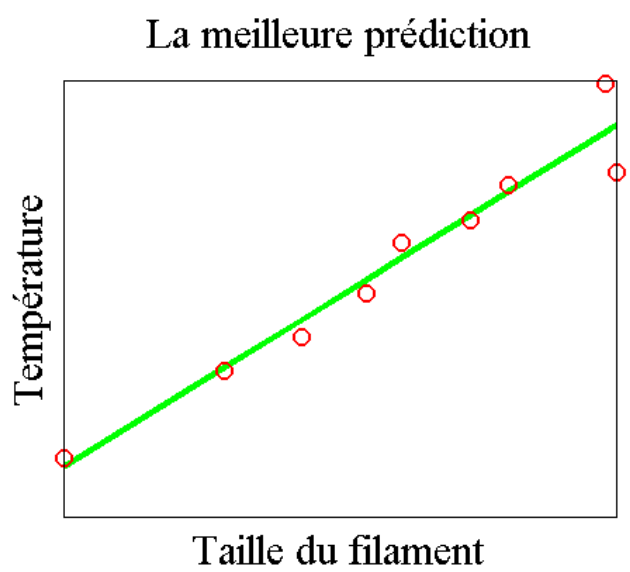
$$y = \underbrace{ax + b}_{f(x,\mathbf{a})} + \varepsilon$$

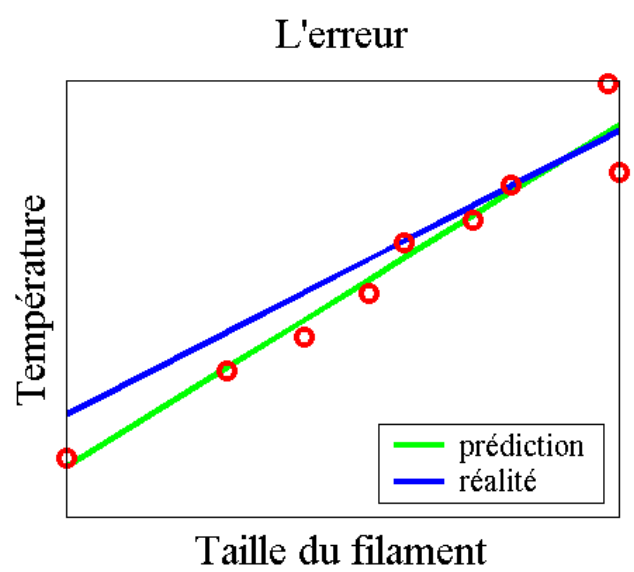
Un exemple : l'étalonnage d'un capteur

Les différentes phases de la régression









Les moindres carrés pour la régression simple

Moindres carrés

Fonction objectif

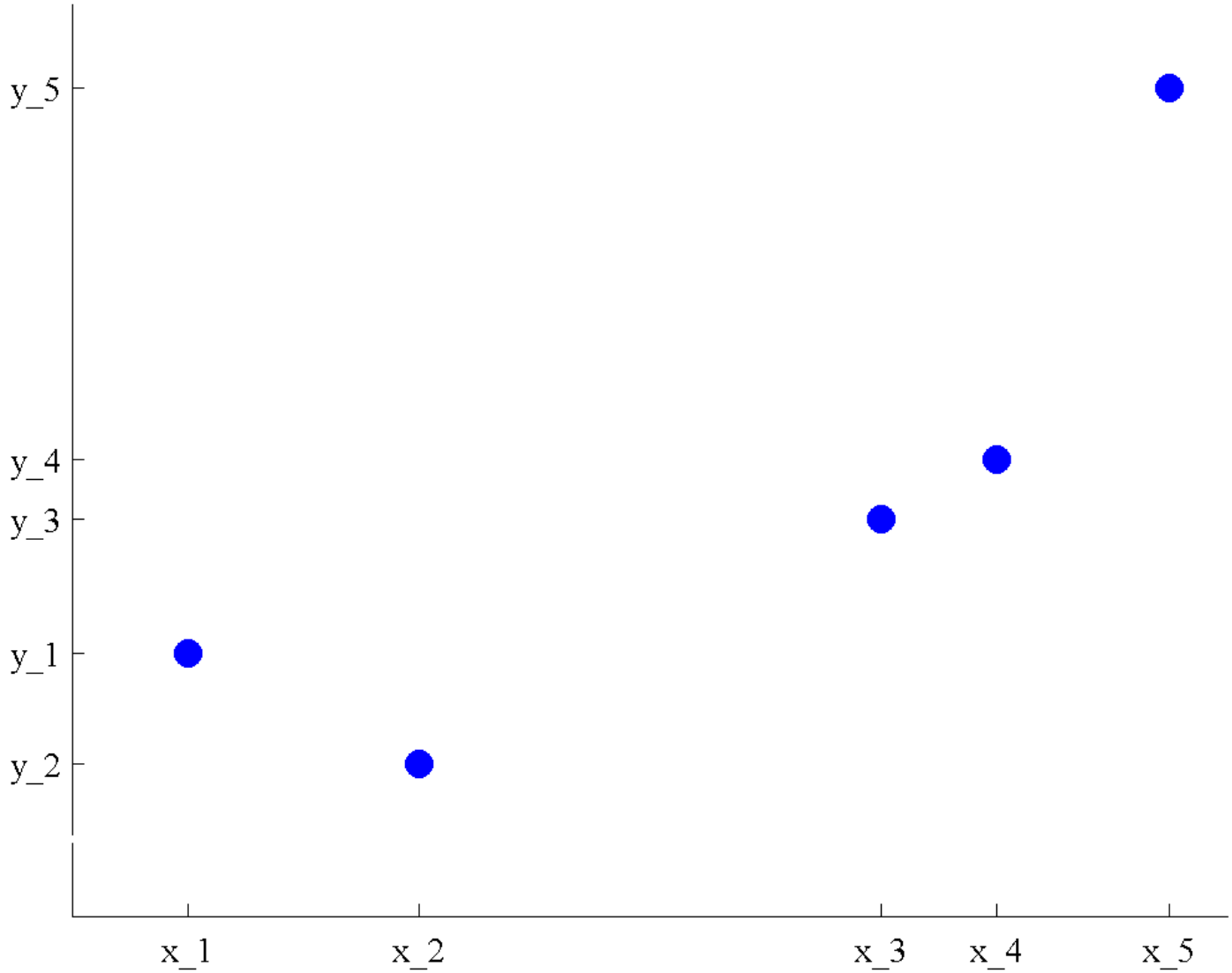
$$\min_{a,b} J(a,b)$$

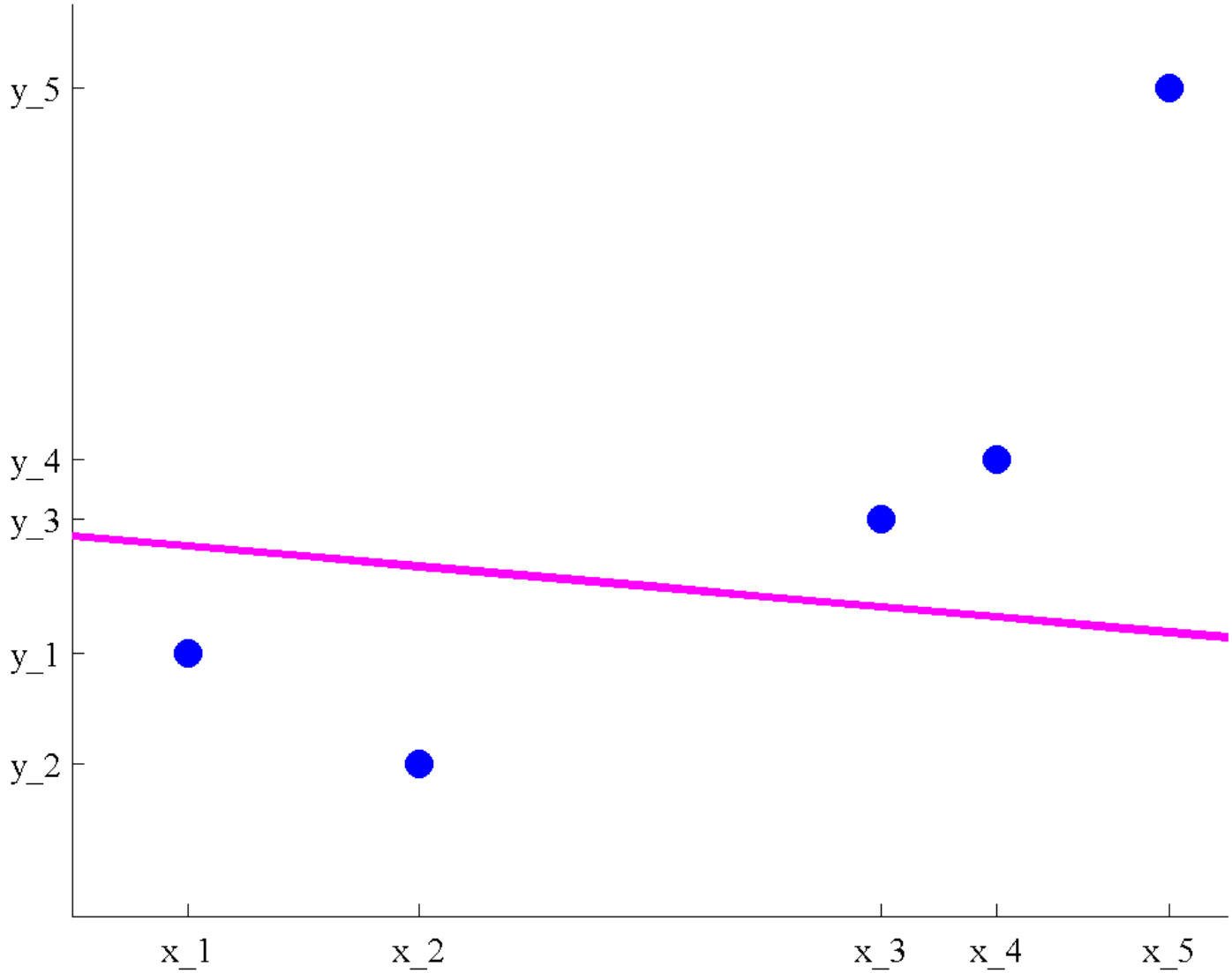
avec

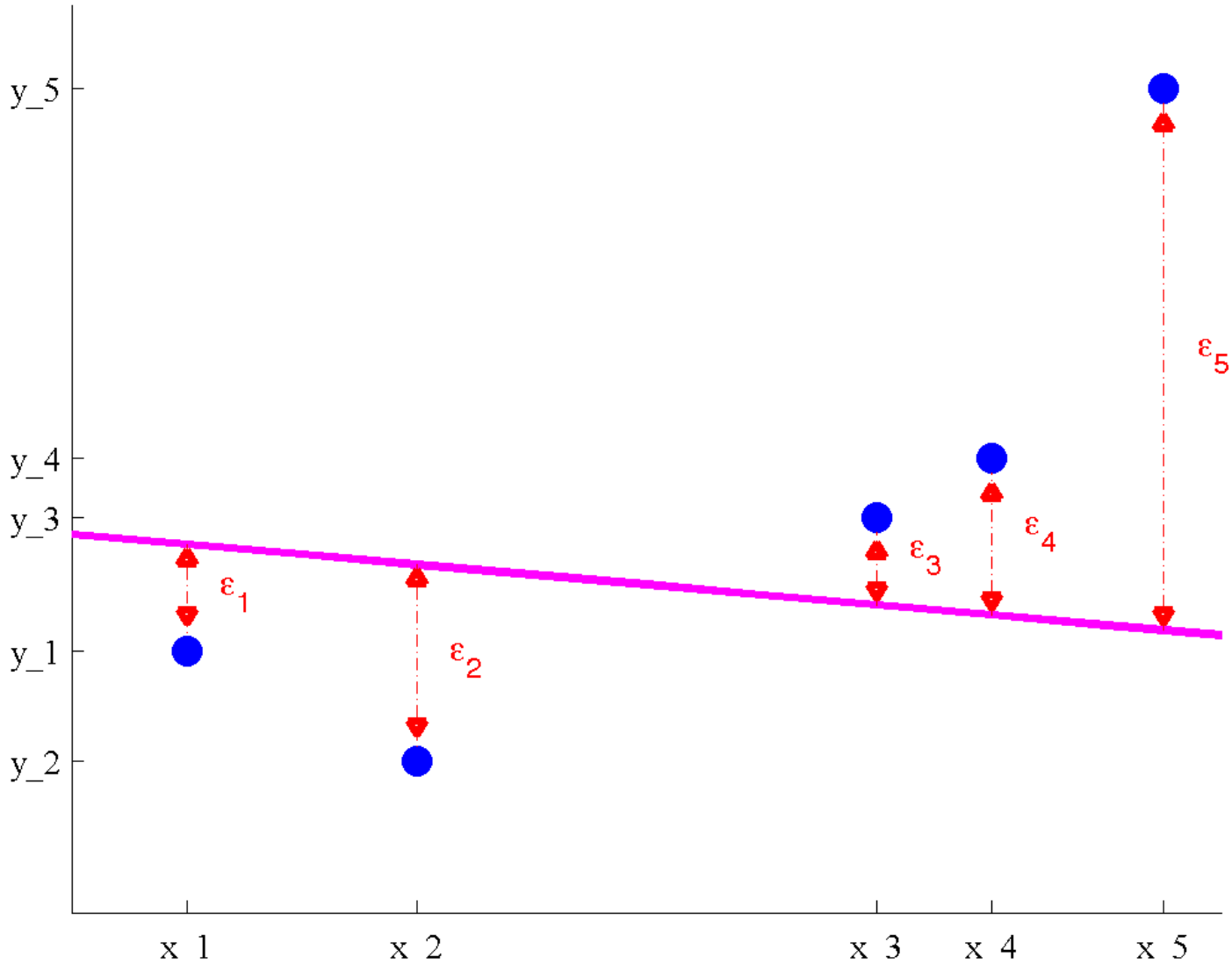
$$J(a,b) = \sum_{i=1}^{n} \underbrace{\left(y_i - ax_i - b\right)}_{\varepsilon_i}^2$$

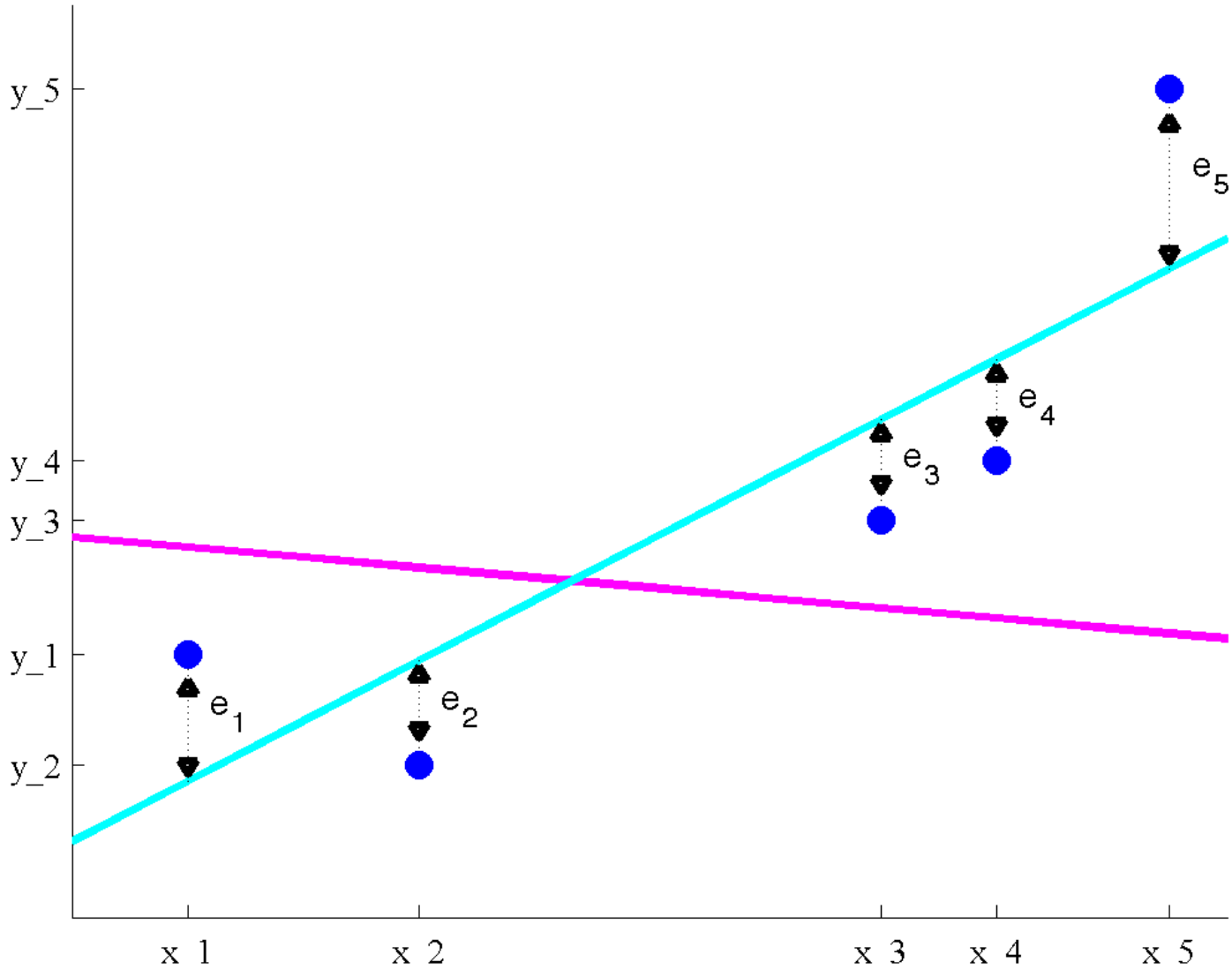
Interprétation

Ce problème peut s'interpréter comme la recherche de la droite d'équation $ax + b$ passant "au mieux" (au sens des moindres carrés) parmi le nuage des observations $(x_i, y_i)\,, i = 1, \ldots, n$.


y_5
y_4
y_3
y_1
y_2
x_1
x_2
x_3
x_4
x_5


y_5
y_4
y_3
y_1
y_2
x_1
x_2
x_3
x_4
x_5


y_5
y_4
y_3
y_1
y_2
x_1
x_2
x_3
x_4
x_5
ε1
ε2
ε3
ε4
ε5


y_5
y_4
y_3
y_1
y_2
e_1
e_2
e_3
e_4
e_5
x_1
x_2
x_3
x_4
x_5

Le problème des moindres carrés

Les données dont nous disposons peuvent aussi être vue comme un système de n équations à $2+n$ inconnues (a, b et les ε_i). Ce système s'écrit de la manière suivante :

$$\begin{cases} ax_1 + b + \varepsilon_1 & = y_1 \\ \vdots & \vdots \\ ax_i + b + \varepsilon_i & = y_i \\ \vdots & \vdots \\ ax_n + b + \varepsilon_n & = y_n \end{cases}$$

On recherche a et b qui minimisent simultanément tous les ε_i

$$J(a,b) = \sum_{i=1}^{n} \big(\underbrace{y_i - ax_i - b}_{\varepsilon_i} \big)^2$$

Calcul du gradient

$$\arg\min_{a,b} J(a,b) \qquad \text{avec} \qquad J(a,b) = \frac{1}{2}\sum_{i=1}^{n} (ax_i + b - y_i)^2$$

Méthode du gradient

- $(a^\star, b^\star)$ est solution du problème

$$\arg\min_{a,b} J(a,b) \quad \Leftrightarrow \begin{cases} \dfrac{\partial J(a^\star, b^\star)}{\partial a} = 0 \\ \dfrac{\partial J(a^\star, b^\star)}{\partial b} = 0 \end{cases}$$

Dérivées partielles de $J(a, b)$

$$\begin{aligned}
\frac{\partial J(a,b)}{\partial a} &= \sum_{i=1}^{n} (ax_i + b - y_i)\, x_i \\
&= \sum_{i=1}^{n} \left(ax_i^2 + bx_i - y_i x_i\right) \\
&= \sum_{i=1}^{n} ax_i^2 + \sum_{i=1}^{n} bx_i - \sum_{i=1}^{n} y_i x_i \\
&= a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J(a,b)}{\partial b} &= \sum_{i=1}^{n} (ax_i + b - y_i) \\
&= \sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} b - \sum_{i=1}^{n} y_i \\
&= a \sum_{i=1}^{n} x_i + bn - \sum_{i=1}^{n} y_i
\end{aligned}$$

Calcul de $a^\star$ et $b^\star$

Deux équations linéaires à deux inconnues

$$\begin{cases} \dfrac{\partial J(a^\star, b^\star)}{\partial a} = 0 \\ \dfrac{\partial J(a^\star, b^\star)}{\partial b} = 0 \end{cases} \Leftrightarrow \begin{array}{lll} a^\star \displaystyle\sum_{i=1}^{n} x_i^2 & + b^\star \displaystyle\sum_{i=1}^{n} x_i & = \displaystyle\sum_{i=1}^{n} y_i x_i \quad (1) \\ a^\star \displaystyle\sum_{i=1}^{n} x_i & + b^\star \, n & = \displaystyle\sum_{i=1}^{n} y_i \quad (2) \end{array}$$

Calcul de $a^\star$ et $b^\star$

Calcul de $a^\star$

- $(1) * n - (2) * \sum_{i=1}^{n} x_i$
- $a^\star \left(n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right) = n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i$
- $$a^\star = \frac{n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

Calcul de $b^\star$

$$b^\star = \frac{\sum_{i=1}^{n} y_i - a^\star \sum_{i=1}^{n} x_i}{n}$$

Réécrivons la solution

Soit $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ et $V_x = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$.

$$a^\star = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\sum_{i=1}^{n} y_i \ \frac{1}{n}\sum_{i=1}^{n} x_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \overline{y}\ \overline{x}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$a^\star = \frac{\mathsf{cov}(\mathbf{x}, \mathbf{y})}{V_x}$$

$$
\begin{aligned}
b^\star &= \frac{\sum_{i=1}^{n} y_i - a^\star \sum_{i=1}^{n} x_i}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i - a^\star \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \overline{y} - a^\star \, \overline{x}
\end{aligned}
$$

Le théorème des moindres carrés

Théorème : Théorème des moindres carrées

Soit $(x_i, y_i), i = 1, n$ un ensemble de couples d'observations.

La solution du problème de minimisation de la somme des carrées des erreurs

$$\min_{a,b} \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

est donnée par $\mathbf{a}^\star$ et $\mathbf{b}^\star$ définis par :

$$a^\star = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\mathsf{cov}(\mathbf{x}, \mathbf{y})}{V_x} \text{ et } b^\star = \overline{y} - a^\star \, \overline{x}$$

$a^\star$ et $b^\star$ sont les estimateurs au sens des moindres carrés.

En Résumé

Apprentissage des paramètres

- Dépend de x et y
- $a^\star = \dfrac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{V_x}$
- $b^\star = \overline{y} - a^\star \, \overline{x}$

Modèle prédictif

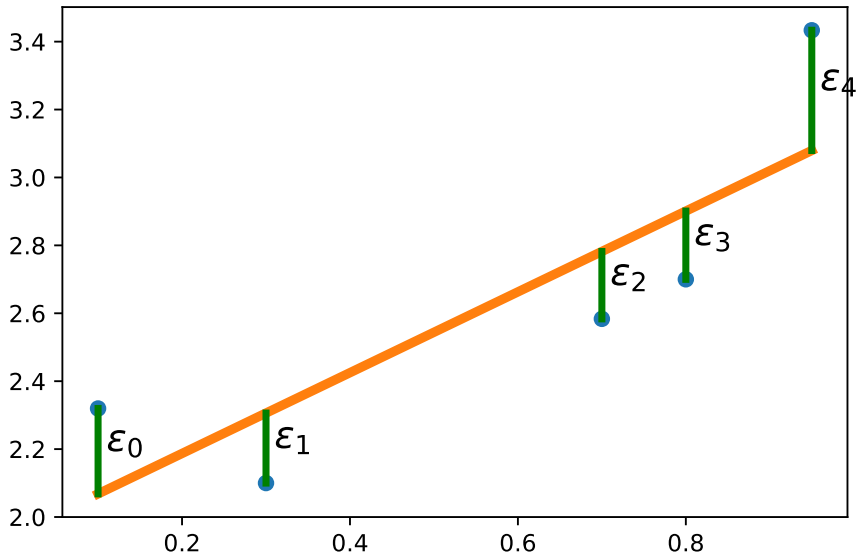
- Dépend d'une observation x, et des paramètres $a^\star$ et $b^\star$
- Prédire une nouvelle observation x :

$$\widehat{y} = f(x) = a^\star x + b^\star$$

- Erreur de prédiction $\varepsilon = y_i - \widehat{y_i}$

Mise en oeuvre

```
mx = np.mean(x)
my = np.mean(y)
sxx = np.sum((x-mx)**2)
sxy = (x-mx).T @ (y-my)
a = sxy/sxx
b = my - a*mx
yp = a*x+b
e = y - yp
```



x	y	y_p	ε_i
0.1	2.3196	2.0681	0.2514
0.3	2.1000	2.3061	−0.2061
0.7	2.5836	2.7820	−0.1984
0.8	2.7000	2.9010	−0.2010
0.95	3.4335	3.0795	0.3540

Le poids des observations

Influence des observations dans le calcul de $a^\star$

$$
\begin{aligned}
a^\star &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} &&= \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\, y_i \\
&= \sum_{i=1}^{n} g(x_i)\, y_i &&= \sum_{i=1}^{n} w_i\, y_i
\end{aligned}
$$

$w_i = \dfrac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ mesure le poids de l'observation x_i dans le calcul de $a^\star$.

Remarques

$a^\star$ et corrélation

$$a^\star = \frac{\mathsf{cov}(\mathbf{x},\mathbf{y})}{s_x^2} = \frac{\mathsf{cov}(\mathbf{x},\mathbf{y})}{s_x s_x}\frac{s_y}{s_y} = \frac{\mathsf{cov}(\mathbf{x},\mathbf{y})}{s_x s_y}\frac{s_y}{s_x} = \mathsf{cor}(\mathbf{x},\mathbf{y})\frac{s_y}{s_x}$$

Droite de régression

$$\begin{aligned} y &= a^\star x + b^\star \\ &= a^\star x + \bar{y} - a^\star\, \bar{x} \\ &= a^\star(x-\bar{x}) + \bar{y} \quad = \frac{\mathsf{cov}(\mathbf{x},\mathbf{y})}{s_x^2}(x-\bar{x}) + \bar{y} \end{aligned}$$

- La droite de régression passe par le point $(\bar{x}, \bar{y})$:

$$f(\bar{x}) = a^\star \bar{x} + b^\star = \frac{\mathsf{cov}(\mathbf{x},\mathbf{y})}{s_x^2}\underbrace{(x-\bar{x})}_{=0} + \bar{y}$$

Coefficient de détermination R^2

Quelle quantité de la relation est expliquée ?

- Écart expliqué par le modèle : $\sum_{i=1}^{n}(\widehat{y_i} - \bar{y})^2$
- Écart total : $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Écart résiduel : $\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$

R^2

Le coefficient de détermination $R^2 \in \{0, 1\}$ est le rapport de l'écart expliqué versus l'écart total

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y_i} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Prédiction et erreur de prédiction

Erreurs et variable aléatoire

- Les ε_i suivent une loi normale $\mathcal{N}(0, \sigma^2)$
- $\widehat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - a^\star x_i - b^\star)^2$

y_x est également une v.a.

- $y_x = a^\star x + b^\star + \varepsilon = \bar{y} + a^\star(\mathbf{x} - \bar{x}) + \varepsilon$
- $\widehat{\mathbb{E}(y_x)} = a^\star x + b^\star = \bar{y} + a^\star(x - \bar{x})$
- $V(y_x) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$

Intervalle de prédiction

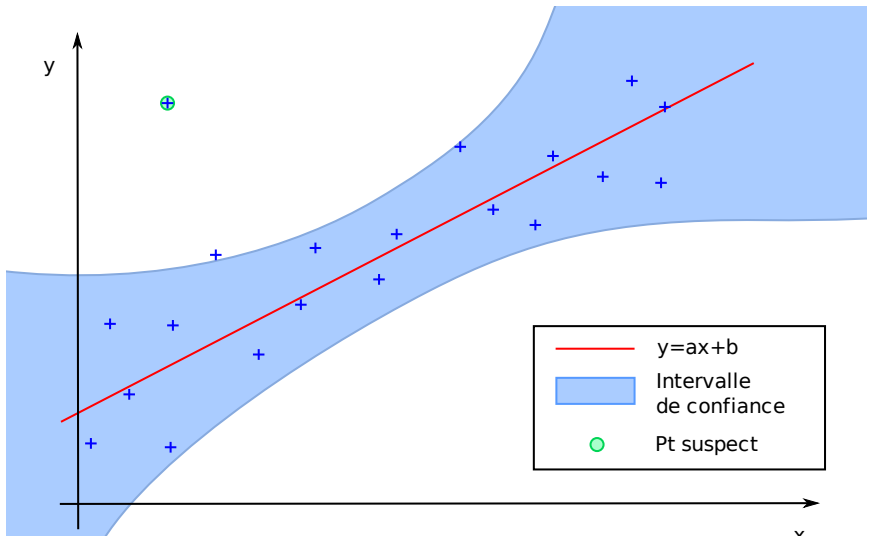
Théorème : Intervalle de prédicition

Pour un x donné, il est probable, avec une probabilité $1-\alpha$, que

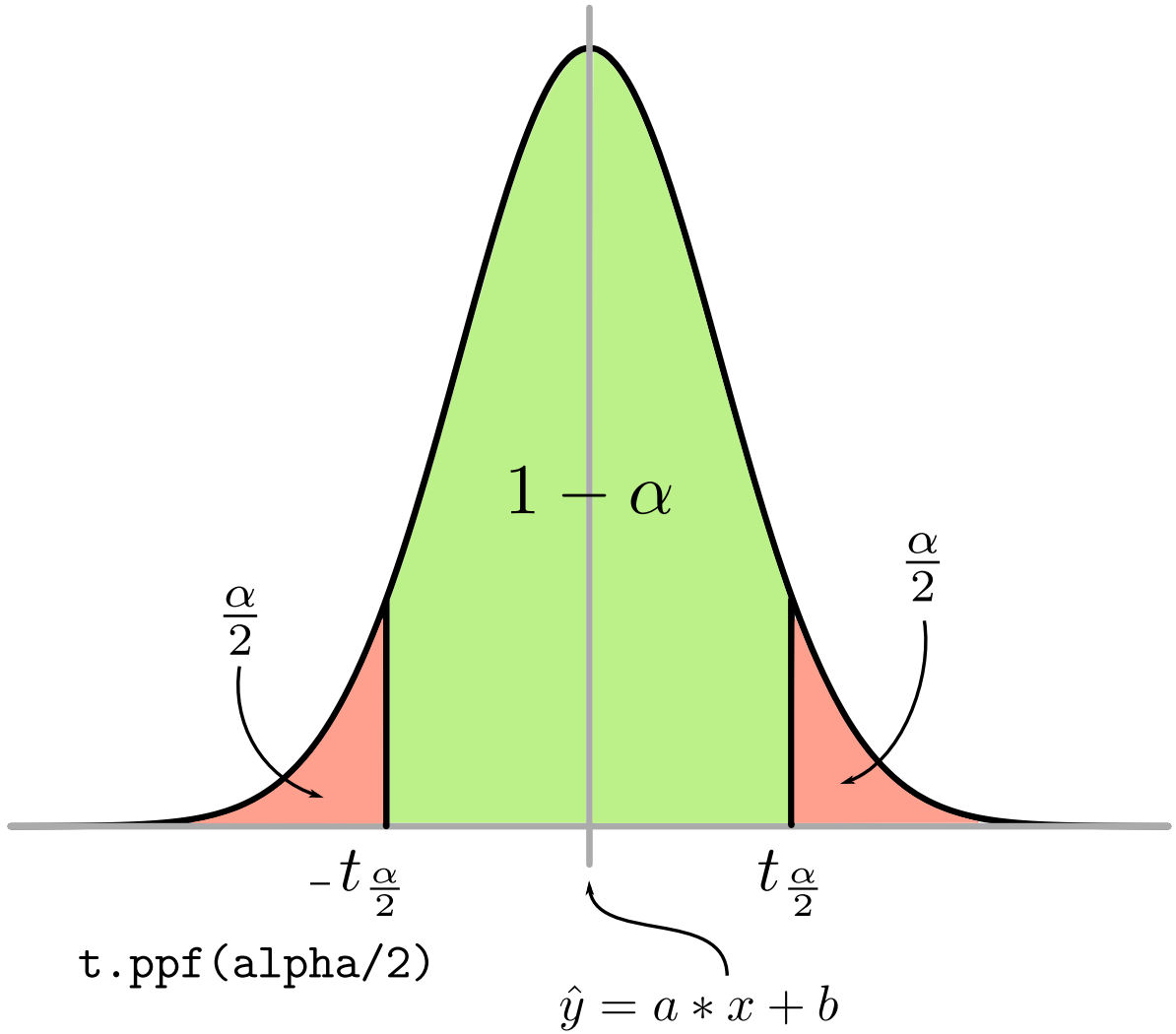
$$y_x \in \left\{ \bar{y} + a^\star(x-\bar{x}) \pm t_{\frac{\alpha}{2}}\, \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \right\}$$

où t_α vérifie, relativement à la loi de Student à $n-2$ degrès de libertés, $P(T \leq t_\alpha) = \alpha$.

Plus on s'éloigne de la moyenne des x, plus l'incertitude grandit



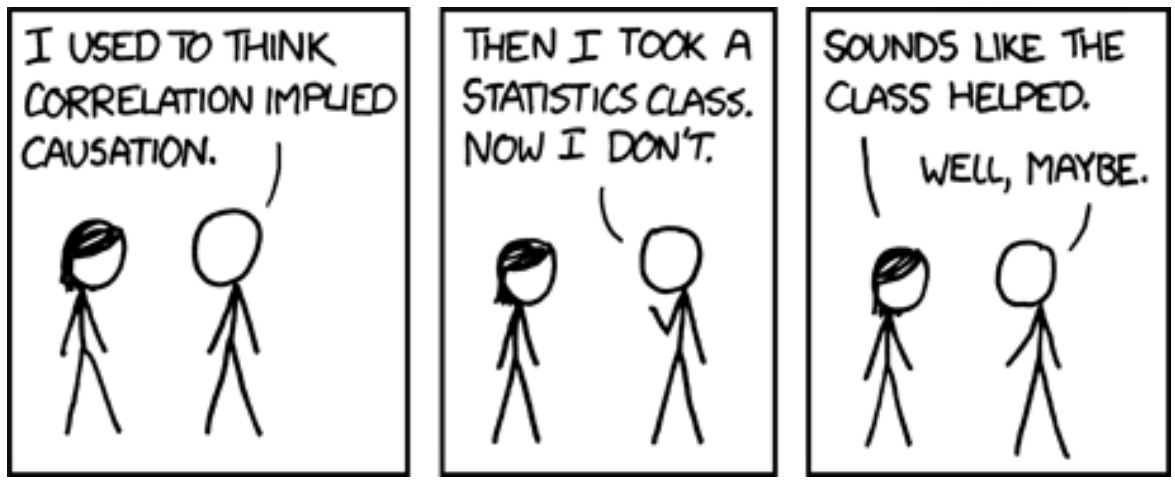
Calculer $t_{\frac{\alpha}{2}}$ avec scipy



Conclusion

La régression simple

- observation = modèle + bruit
- modèlisation linéaire d'une relation
- certains modèles non linéaires peuvent être linéarisés (projection)
- Moindres carrés : facile à calculer



La suite

- Passer au multi dimensionnel (plusieurs variables explicatives)
- Valider le modèle : diagnostic de la régression

Éléments de démonstration de $Var(y_x)$

$$
\begin{aligned}
V(a^\star) &= V\left(\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(\mathbf{x}_i-\bar{x})^2}\right)\\
&= \left(\frac{1}{\sum_{i=1}^{n}(\mathbf{x}_i-\bar{x})^2}\right)^2 V\left(\sum_{i=1}^{n}(a(x_i-\bar{x})+\varepsilon)(x_i-\bar{x})\right)\\
&= \left(\frac{1}{\sum_{i=1}^{n}(\mathbf{x}_i-\bar{x})^2}\right)^2 V\left(\sum_{i=1}^{n}\varepsilon(x_i-\bar{x})\right)\\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(\mathbf{x}_i-\bar{x})^2}
\end{aligned}
$$

D'où

$$
\begin{aligned}
V(y_x) &= V(\bar{y}) + V(a^\star)(\mathbf{x}-\bar{x})^2 + \sigma^2\\
&= \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^{n}(\mathbf{x}_i-\bar{x})^2}(\mathbf{x}-\bar{x})^2 + \sigma^2\\
&= \sigma^2\left(1+\frac{1}{n}+\frac{(\mathbf{x}-\bar{x})^2}{\sum_{i=1}^{n}(\mathbf{x}_i-\bar{x})^2}\right)
\end{aligned}
$$

Dérivée partielle $J(a,b)$ de par rapport à a

$$
\begin{aligned}
J(a,b) &= \tfrac{1}{2}\sum_{i=1}^{n}\left(ax_i+b-y_i\right)^2 \\
&= \tfrac{1}{2}\sum_{i=1}^{n} a^2x_i^2 + bax_i - ax_iy_i + bax_i + b^2 - by_i - y_iax_i - b \\
\frac{\partial J(a,b)}{\partial a} &= \sum_{n} ax_i^2 + bx_i - y_iax_i \\
&= \sum_{i=1} \left(ax_i+b-y_i\right)x_i
\end{aligned}
$$

Calcul de $a^\star$

- $(1) * n - (2) * \sum_{i=1}^{n} x_i$ (pour supprimer $b^\star$)
  $n(a^\star \sum_{i=1}^{n} x_i^2 + b^\star \sum_{i=1}^{n} x_i) - (\sum_{i=1}^{n} x_i)(a^\star \sum_{i=1}^{n} x_i + b^\star n) =$
  $n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$
  $\Rightarrow a^\star n \sum x_i^2 + b^\star n \sum x_i - a^\star (\sum x_i)^2 - b^\star n \sum x_i$
  $\Rightarrow a^\star (n(\sum x_i^2) - (\sum x_i)^2 = n \sum y_i x_i - \sum x_i \sum y_i$
  $\Rightarrow a^\star = \frac{n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$

Calcul de $b^\star$

Connaissant $a^\star$, on reprend (2)

Réécriture de $a^\star$

$$
\begin{aligned}
a^\star &= \frac{n\sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \Rightarrow *\frac{1}{n^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n y_i x_i - \frac{1}{n^2}\sum_{i=1}^n y_i \sum_{i=1}^n x_i}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^n x_i\right)^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n y_i x_i - \frac{1}{n}\sum_{i=1}^n y_i \frac{1}{n}\sum_{i=1}^n x_i}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n y_i x_i - \bar{y}\bar{x}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2} \\
&= \frac{\mathsf{cov}(\mathbf{x}, \mathbf{y})}{V_x}
\end{aligned}
$$

Théorème de Konig-Huygens

$$(\frac{1}{n}\sum x_i^2) - \bar{x}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

https://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_de_K%C3%B6nig-Huygens

$$\frac{1}{n}\sum y_i x_i - \bar{y}\bar{x} = \frac{1}{n}\sum (y_i - \bar{y})(x_i - \bar{x})$$

$$
\begin{aligned}
\frac{1}{n}\sum(y_i-\bar{y})(x_i-\bar{x}) &= \frac{1}{n}\sum y_i x_i - y_i\bar{x} - \bar{y}x_i + \bar{y}\bar{x} \\
&= \frac{1}{n}\sum y_i x_i - \frac{1}{n}\sum y_i\bar{x} - \frac{1}{n}\sum \bar{y}x_i + \bar{y}\bar{x} \\
&= \frac{1}{n}\sum y_i x_i - \bar{x}\frac{1}{n}\sum y_i - \bar{y}\frac{1}{n}\sum x_i + \bar{y}\bar{x} \\
&= \frac{1}{n}\sum y_i x_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{y}\bar{x} \\
&= \frac{1}{n}\sum y_i x_i - \bar{x}\bar{y}
\end{aligned}
$$

$$\sum(x_i-\bar{x})(y_i-\bar{y}) = \sum(x_i-\bar{x})y_i$$

$$
\begin{aligned}
\sum (x_i - \bar{x})(y_i - \bar{y}) &= \sum x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y} \\
&= \sum x_i y_i - \sum x_i \bar{y} - \sum \bar{x} y_i + \sum \bar{x}\bar{y} \\
&= \sum x_i y_i - \sum x_i \frac{1}{n} \sum y_j - \sum \frac{1}{n} \sum x_j y_i + \sum \bar{x} \\
&= \sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i - \frac{1}{n} \sum x_i y_i + \sum \bar{x}\bar{y} \\
&= \sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i - \frac{1}{n} \sum x_i y_i + \sum \frac{1}{n} \sum \\
&= \sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i - \frac{1}{n} \sum x_i y_i + \frac{1}{n} \sum x_i \\
&= \sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i \\
&= \sum x_i y_i - \bar{x} y_i \\
&= \sum (x_i - \bar{x}) y_i
\end{aligned}
$$