

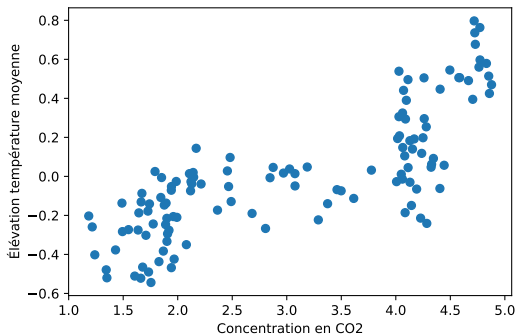
Régression Linéaire Simple

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

INSA Rouen Normandie - ITI

April 14, 2022

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

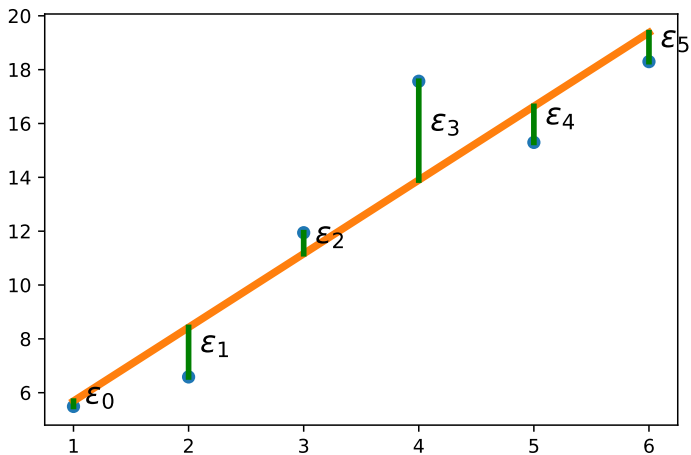
- ▶ Trouver la droite qui représente au mieux la relation linéaire

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

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

$$\hat{y}_i = ax_i + b + \varepsilon_i$$

- ▶ Modélise le **bruit**



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

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

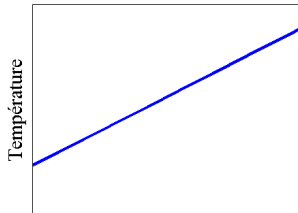
- ▶ Variables explicatives $x \in \mathbb{R}$
- ▶ Variable à expliquer $y \in \mathbb{R}$
- ▶ Erreur ou bruit ε
- ▶ Paramètres scalaires $a, b \in \mathbb{R}$,
- ▶ Paramètres (forme vectorielle) $\mathbf{a} \in \mathbb{R}^2$
- ▶ Modèle $y = f(\mathbf{x}, \mathbf{a}) + \varepsilon$
- ▶ Estimation *a star*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

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

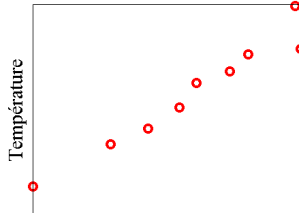
Un exemple : l'étalonnage d'un capteur

Les différentes phases de la régression

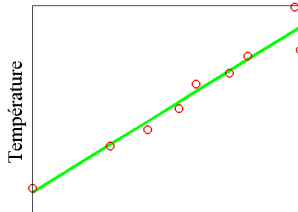
La réalité



Les données

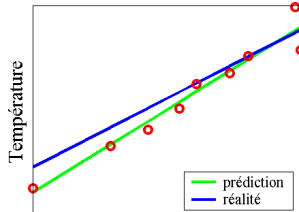


La meilleure prédiction



Taille du filament

L'erreur



Taille du filament

Les moindres carrés pour la régression simple

Moindres carrés

Fonction objectif

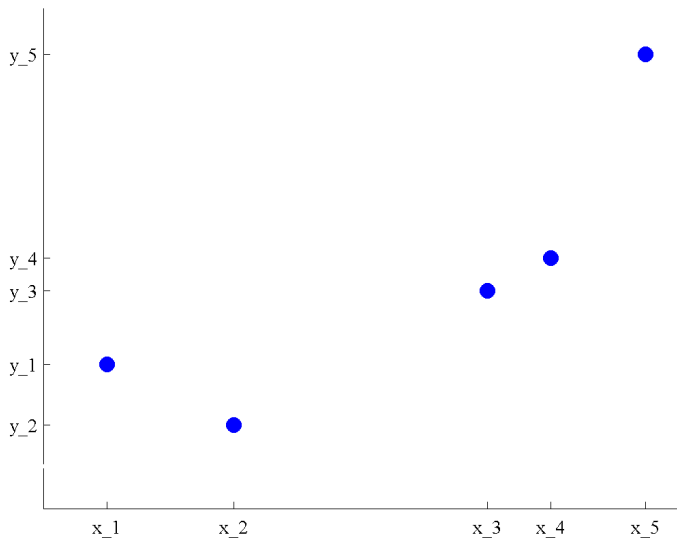
$$\min_{a,b} J(a, b)$$

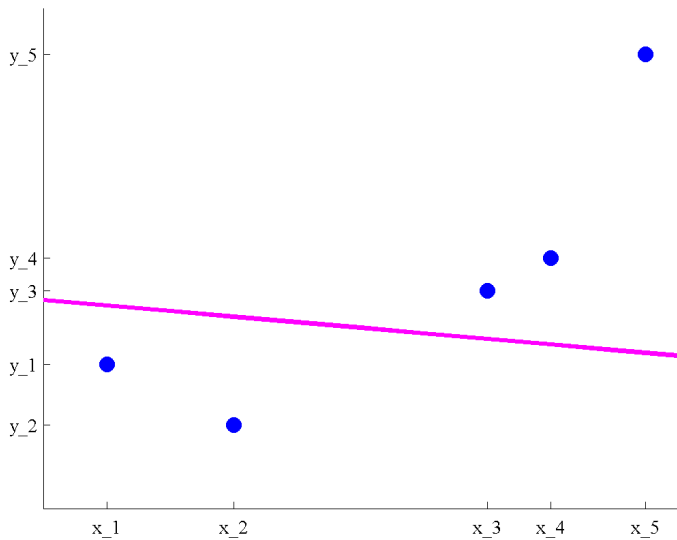
avec

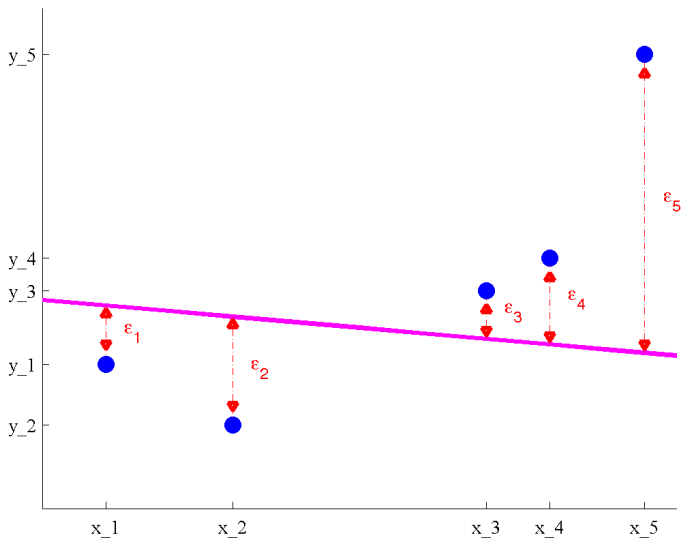
$$J(a, b) = \sum_{i=1}^n \underbrace{(y_i - ax_i - b)}_{\varepsilon_i}^2$$

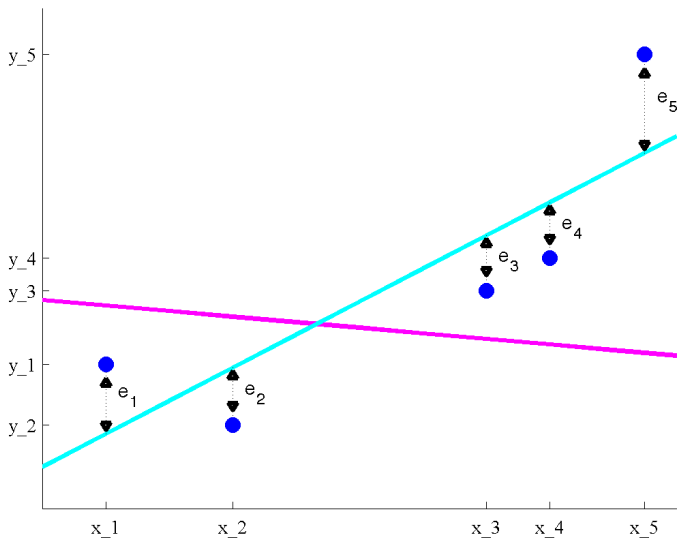
Interprétation

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 $(x_i, y_i), i = 1, \dots, n$.









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 :

$$\left\{ \begin{array}{lcl} ax_1 + b + \varepsilon_1 & = & y_1 \\ \vdots & & \vdots \\ ax_i + b + \varepsilon_i & = & y_i \\ \vdots & & \vdots \\ ax_n + b + \varepsilon_n & = & y_n \end{array} \right.$$

On recherche a et b qui minimisent simultanément tous les ε_i

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

Calcul du gradient

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

Méthode du gradient

- ▶ (a^*, b^*) est solution du problème

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

Dérivées partielles de $J(a, b)$

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

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

Calcul de a^* et b^*

Deux équations linéaires à deux inconnues

$$\begin{cases} \frac{\partial J(a^*, b^*)}{\partial a} = 0 \\ \frac{\partial J(a^*, b^*)}{\partial b} = 0 \end{cases} \Leftrightarrow \begin{aligned} a^* \sum_{i=1}^n x_i^2 + b^* \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i x_i \quad (1) \\ a^* \sum_{i=1}^n x_i + b^* n &= \sum_{i=1}^n y_i \quad (2) \end{aligned}$$

Calcul de a^* et b^*

Calcul de a^*

$$\blacktriangleright (1) * n - (2) * \sum_{i=1}^n x_i$$

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

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

Calcul de b^*

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

Réécrivons la solution

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

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

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

Le théorème des moindres carrés

Théorème : Théorème des moindres carrés

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

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

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

est donnée par a^* et b^* définis par :

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

a^* et b^* sont les estimateurs au sens des moindres carrés.

En Résumé

Apprentissage des paramètres

- ▶ Dépend de x et y
- ▶ $a^* = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{V_x}$
- ▶ $b^* = \bar{y} - a^* \bar{x}$

Modèle prédictif

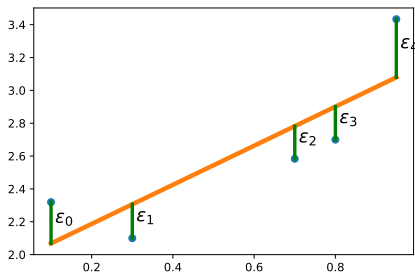
- ▶ Dépend d'une observation x , et des paramètres a^* et b^*
- ▶ Prédire une nouvelle observation x :

$$\hat{y} = f(x) = a^*x + b^*$$

- ▶ Erreur de prédiction $\varepsilon = y_i - \hat{y}_i$

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^*

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

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

Remarques

a^* et corrélation

$$a^* = \frac{\text{COV}(\mathbf{x}, \mathbf{y})}{s_x^2} = \frac{\text{COV}(\mathbf{x}, \mathbf{y})}{s_x s_x} \frac{s_y}{s_y} = \frac{\text{COV}(\mathbf{x}, \mathbf{y})}{s_x s_y} \frac{s_y}{s_x} = \text{cor}(\mathbf{x}, \mathbf{y}) \frac{s_y}{s_x}$$

Droite de régression

$$\begin{aligned} y &= a^* x + b^* \\ &= a^* x + \bar{y} - a^* \bar{x} \\ &= a^* (x - \bar{x}) + \bar{y} = \frac{\text{COV}(\mathbf{x}, \mathbf{y})}{s_x^2} (x - \bar{x}) + \bar{y} \end{aligned}$$

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

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

Coefficient de détermination R^2

Quelle quantité de la relation est expliquée ?

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

R^2

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

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

Prédiction et erreur de prédiction

Erreurs et variable aléatoire

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

y_x est également une v.a.

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

Intervalle de prédiction

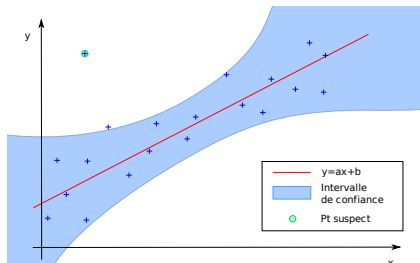
Théorème : Intervalle de prédiction

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

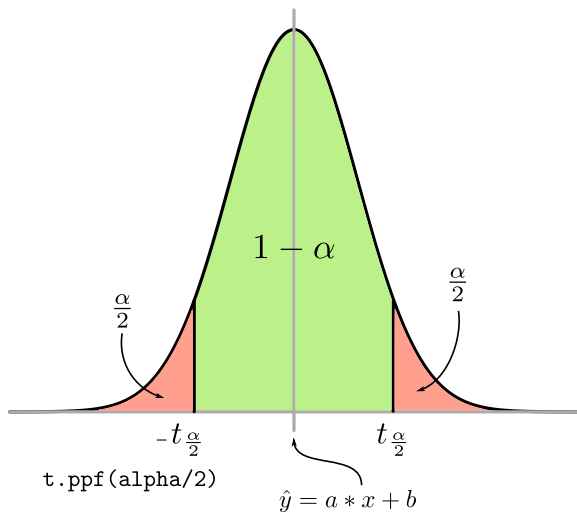
$$y_x \in \left\{ \bar{y} + a^*(x - \bar{x}) \pm t_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right\}$$

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

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



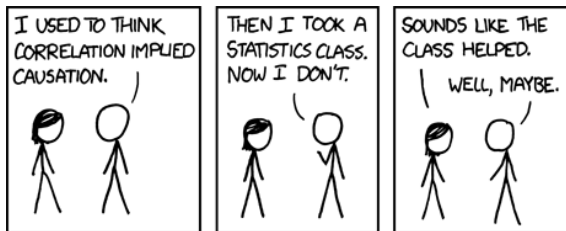
Calculer $t_{\frac{\alpha}{2}}$ avec scipy



Conclusion

La régression simple

- ▶ observation = modèle + bruit
- ▶ modélisation 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 $Var(y_x)$

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

D'où

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