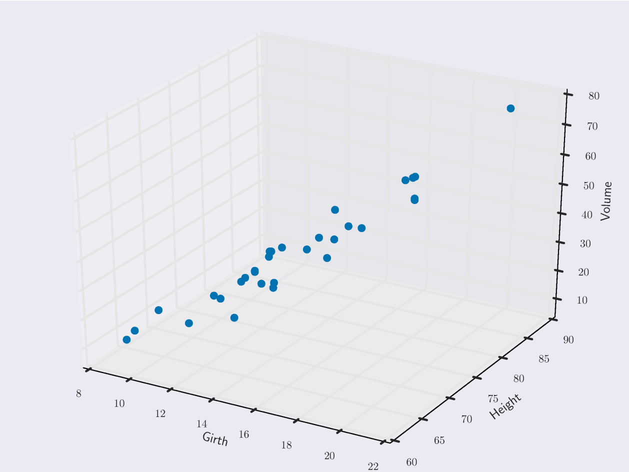
MS BGD
MDI 720 : Statistiques

Joseph Salmon
http://josephsalmon.eu
Télécom Paristech, Institut Mines-Télécom

Plan

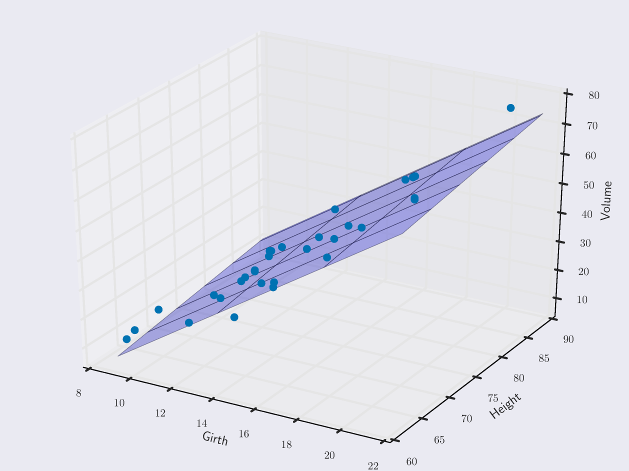
Vers des modèles multi-variés

Volume d'arbres en fonction de leur hauteur / circonférence



Vers des modèles multi-variés

Volume d'arbres en fonction de leur hauteur / circonférence



Commandes sous python

```
# Load data
url = 'http://vincentarelbundock.github.io/
Rdatasets/csv/datasets/trees.csv'
dat3 = pd.read_csv(url)
# Fit regression model
X = dat3[['Girth', 'Height']]
X = sm.add_constant(X)
y = dat3['Volume']
results = sm.OLS(y, X).fit().params
XX = np.arange(8, 22, 0.5)
YY = np.arange(64, 90, 0.5)
xx, yy = np.meshgrid(XX, YY)
zz = results[0] + results[1]*xx + results[2]*yy
fig = plt.figure()
ax = Axes3D(fig)
ax.plot(X['Girth'],X['Height'],y,'o')
ax.plot_wireframe(xx, yy, zz, rstride=10, cstride=10)
plt.show()
```

`results` renvoie `const:-57.98, Girth: 4.70, Height: 0.33`

Sommaire

Modélisation

On dispose de p variables explicatives

Modèle en dimension p

$$y_i = \theta_0^* + \sum_{j=1}^{p} \theta_j^* x_{i,j} + \varepsilon_i$$

$$\varepsilon_i \overset{i.i.d}{\sim} \varepsilon, \text{ pour } i = 1, \ldots, n$$

$$\mathbb{E}(\varepsilon) = 0$$

Rem: on fait l'hypothèse qu'il existe un vrai paramètre (point de vue fréquentiste)

Dimension p

Modèle matriciel

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \theta_0^* \\ \vdots \\ \theta_p^* \end{pmatrix}}_{\boldsymbol{\theta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

De manière équivalente : $\boxed{\mathbf{y} = X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}}$

<u>Notation colonne</u> : $X = (\mathbf{1}_n, \mathbf{x}_1, \dots, \mathbf{x}_p)$ avec $\mathbf{x}_0 = \mathbf{1}_n$

<u>Notation ligne</u> : $X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix}$

Vocabulaire

$$\boxed{\mathbf{y} = X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}}$$

- $\mathbf{y} \in \mathbb{R}^n$: vecteur des observations
- $X \in \mathbb{R}^{n\times(p+1)}$: la matrice des variables explicatives (design)
- $\boldsymbol{\theta}^* \in \mathbb{R}^{p+1}$: le **vrai** paramètre (inconnu) du modèle que l'on veut retrouver
- $\boldsymbol{\varepsilon} \in \mathbb{R}^n$: vecteur de bruit

point de vue "observations" : $y_i = \langle x_i, \boldsymbol{\theta}^* \rangle + \varepsilon_i$ pour $i = 1, \dots, n$

point de vue "variables explicatives" : $\mathbf{y} = \sum_{j=0}^p \theta_j^* \mathbf{x}_j + \boldsymbol{\varepsilon}$

Sommaire

Estimateur des moindres carrés

Un estimateur des moindres carrées est solution du problème d'optimisation :

$$\hat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \left(\frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 \right)$$

$$\hat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \frac{1}{2} \sum_{i=1}^{n} \left[y_i - \left(\theta_0 + \sum_{j=1}^{p} \theta_j x_{i,j} \right) \right]^2$$

$$\hat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \frac{1}{2} \sum_{i=1}^{n} \left[y_i - (\langle x_i, \boldsymbol{\theta} \rangle) \right]^2$$

Rem: le minimiseur n'est pas toujours unique !
Rem: le terme $\frac{1}{2}$ ne change rien au problème de minimisation, mais facilite certains calculs

Sommaire

Condition nécessaire du premier ordre pour un minimum local (CNO)

Théorème : règle de Fermat

Si f est différentiable en un minimum local $\boldsymbol{\theta}^\star$ alors le gradient de f est nul en $\boldsymbol{\theta}^\star$, *i.e.*, $\nabla f(\boldsymbol{\theta}^\star) = 0$.

<u>Rem</u>: ce n'est une condition suffisante que si f est en plus convexe

Ici $f : \boldsymbol{\theta} \mapsto \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$

$$
\begin{aligned}
f(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 &= \frac{1}{2}\|\mathbf{y}\|^2 - \langle X\boldsymbol{\theta}, \mathbf{y}\rangle + \frac{1}{2}\boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta} \\
&= \frac{1}{2}\|\mathbf{y}\|^2 - \langle \boldsymbol{\theta}, X^\top\mathbf{y}\rangle + \frac{1}{2}\boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta}
\end{aligned}
$$

Calcul du gradient de f

Le gradient de f en $\boldsymbol{\theta}$ est défini comme le vecteur $\nabla f(\boldsymbol{\theta})$ tel que :

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + O(h)$$

Le calcul pour f donne

$$f(\boldsymbol{\theta} + h) = \frac{1}{2}\|\mathbf{y}\|^2 - \langle \boldsymbol{\theta} + h, X^\top \mathbf{y} \rangle + \frac{1}{2}(\boldsymbol{\theta} + h)^\top X^\top X(\boldsymbol{\theta} + h)$$

Calcul du gradient de f

Le gradient de f en $\boldsymbol{\theta}$ est défini comme le vecteur $\nabla f(\boldsymbol{\theta})$ tel que :

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + O(h)$$

Le calcul pour f donne

$$\begin{aligned} f(\boldsymbol{\theta} + h) =& \frac{1}{2}\|\mathbf{y}\|^2 - \langle \boldsymbol{\theta} + h, X^\top \mathbf{y} \rangle + \frac{1}{2}(\boldsymbol{\theta} + h)^\top X^\top X(\boldsymbol{\theta} + h) \\ =& \frac{1}{2}\|\mathbf{y}\|^2 - \langle \boldsymbol{\theta}, X^\top \mathbf{y} \rangle - \langle h, X^\top \mathbf{y} \rangle \\ &+ \frac{1}{2}\boldsymbol{\theta}^\top X^\top X \boldsymbol{\theta} + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \end{aligned}$$

Calcul du gradient de f

Le gradient de f en $\boldsymbol{\theta}$ est défini comme le vecteur $\nabla f(\boldsymbol{\theta})$ tel que :

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + O(h)$$

Le calcul pour f donne

$$\begin{aligned}
f(\boldsymbol{\theta} + h) =& \frac{1}{2}\|\mathbf{y}\|^2 - \langle \boldsymbol{\theta} + h, X^\top \mathbf{y} \rangle + \frac{1}{2}(\boldsymbol{\theta} + h)^\top X^\top X(\boldsymbol{\theta} + h) \\
=& \frac{1}{2}\|\mathbf{y}\|^2 - \langle \boldsymbol{\theta}, X^\top \mathbf{y} \rangle - \langle h, X^\top \mathbf{y} \rangle \\
& + \frac{1}{2}\boldsymbol{\theta}^\top X^\top X \boldsymbol{\theta} + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \\
=& f(\boldsymbol{\theta}) - \langle h, X^\top \mathbf{y} \rangle + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h
\end{aligned}$$

Calcul du gradient de f

Le gradient de f en $\boldsymbol{\theta}$ est défini comme le vecteur $\nabla f(\boldsymbol{\theta})$ tel que :

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + O(h)$$

Le calcul pour f donne

$$\begin{aligned}
f(\boldsymbol{\theta} + h) =& \frac{1}{2}\|\mathbf{y}\|^2 - \langle \boldsymbol{\theta} + h, X^\top \mathbf{y} \rangle + \frac{1}{2}(\boldsymbol{\theta} + h)^\top X^\top X(\boldsymbol{\theta} + h) \\
=& \frac{1}{2}\|\mathbf{y}\|^2 - \langle \boldsymbol{\theta}, X^\top \mathbf{y} \rangle - \langle h, X^\top \mathbf{y} \rangle \\
&+ \frac{1}{2}\boldsymbol{\theta}^\top X^\top X \boldsymbol{\theta} + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \\
=& f(\boldsymbol{\theta}) - \langle h, X^\top \mathbf{y} \rangle + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \\
=& f(\boldsymbol{\theta}) + \langle h, \underbrace{X^\top X \boldsymbol{\theta} - X^\top \mathbf{y}}_{\nabla f(\boldsymbol{\theta})} \rangle + \underbrace{\frac{1}{2} h^\top X^\top X h}_{O(h)}
\end{aligned}$$

Calcul du gradient de f

Le gradient de f en $\boldsymbol{\theta}$ est défini comme le vecteur $\nabla f(\boldsymbol{\theta})$ tel que :

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + O(h)$$

Le calcul pour f donne

$$\begin{aligned}
f(\boldsymbol{\theta} + h) =& \frac{1}{2}\|\mathbf{y}\|^2 - \langle \boldsymbol{\theta} + h, X^\top \mathbf{y} \rangle + \frac{1}{2}(\boldsymbol{\theta} + h)^\top X^\top X(\boldsymbol{\theta} + h) \\
=& \frac{1}{2}\|\mathbf{y}\|^2 - \langle \boldsymbol{\theta}, X^\top \mathbf{y} \rangle - \langle h, X^\top \mathbf{y} \rangle \\
& + \frac{1}{2}\boldsymbol{\theta}^\top X^\top X \boldsymbol{\theta} + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \\
=& f(\boldsymbol{\theta}) - \langle h, X^\top \mathbf{y} \rangle + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \\
=& f(\boldsymbol{\theta}) + \langle h, \underbrace{X^\top X \boldsymbol{\theta} - X^\top \mathbf{y}}_{\nabla f(\boldsymbol{\theta})} \rangle + \underbrace{\frac{1}{2} h^\top X^\top X h}_{O(h)}
\end{aligned}$$

Ainsi, $\boxed{\nabla f(\boldsymbol{\theta}) = X^\top X \boldsymbol{\theta} - X^\top \mathbf{y} = X^\top (X\boldsymbol{\theta} - \mathbf{y})}$

Calcul du gradient de f

Le gradient de f en $\boldsymbol{\theta}$ est défini comme le vecteur $\nabla f(\boldsymbol{\theta})$ tel que :

$$f(\boldsymbol{\theta}+h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta})\rangle + O(h)$$

Le calcul pour f donne

$$\begin{aligned}
f(\boldsymbol{\theta}+h) =& \frac{1}{2}\|\mathbf{y}\|^2 - \langle \boldsymbol{\theta}+h, X^\top \mathbf{y}\rangle + \frac{1}{2}(\boldsymbol{\theta}+h)^\top X^\top X(\boldsymbol{\theta}+h) \\
=& \frac{1}{2}\|\mathbf{y}\|^2 - \langle \boldsymbol{\theta}, X^\top \mathbf{y}\rangle - \langle h, X^\top \mathbf{y}\rangle \\
&+ \frac{1}{2}\boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta} + \frac{1}{2}h^\top X^\top Xh + \boldsymbol{\theta}^\top X^\top Xh \\
=& f(\boldsymbol{\theta}) - \langle h, X^\top \mathbf{y}\rangle + \frac{1}{2}h^\top X^\top Xh + \boldsymbol{\theta}^\top X^\top Xh \\
=& f(\boldsymbol{\theta}) + \langle h, \underbrace{X^\top X\boldsymbol{\theta} - X^\top \mathbf{y}}_{\nabla f(\boldsymbol{\theta})}\rangle + \underbrace{\frac{1}{2}h^\top X^\top Xh}_{O(h)}
\end{aligned}$$

Ainsi, $$\boxed{\nabla f(\boldsymbol{\theta}) = X^\top X\boldsymbol{\theta} - X^\top \mathbf{y} = X^\top(X\boldsymbol{\theta} - \mathbf{y})}$$

Rappel sur le gradient

Le gradient de f en $\boldsymbol{\theta}$ est défini comme le vecteur $\nabla f(\boldsymbol{\theta})$ tel que :

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + O(h)$$

Propriété : le gradient peut aussi être défini comme le vecteur des dérivées partielles

$$\nabla f(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial f}{\partial \theta_0} \\ \vdots \\ \frac{\partial f}{\partial \theta_p} \end{pmatrix}$$

Moindres carrés - équation(s) normale(s)

$$\nabla f(\boldsymbol{\theta}) = 0 \Leftrightarrow X^\top X\boldsymbol{\theta} - X^\top \mathbf{y} = X^\top (X\boldsymbol{\theta} - \mathbf{y}) = 0$$

Théorème

La CNO nous assure qu'un minimiseur $\hat{\boldsymbol{\theta}}$ satisfait l'équation :

Équation(s) normale(s) : $\boxed{(X^\top X)\hat{\boldsymbol{\theta}} = X^\top \mathbf{y}}$

$\hat{\boldsymbol{\theta}}$ est donc solution d'un système linéaire "$Ax = b$" pour une matrice $A = X^\top X$ et un second membre $b = X^\top y$

<u>Rem</u>: si les variables sont redondantes il n'y pas unicité de la solution, tout comme cela arrivait en dimension un

---

Exo: coder en `python` une descente de gradient pour résoudre le problème des moindres carrées

---

Vocabulaire (et abus de langage)

Définition

On appelle **matrice de Gram** (🇬🇧 : *Gramian matrix*) la matrice

$$X^\top X$$

dont le terme général est $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$. Elle est parfois aussi appelée matrice des corrélations

Rem: si on normalise les variables pour que $\forall j \in [\![0, p]\!], \|\mathbf{x}_j\|^2 = n$, la diagonale de la matrice est $(n, \dots, n)$

Le terme $X^\top \mathbf{y} = \begin{pmatrix} \langle \mathbf{x}_0, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \mathbf{y} \rangle \end{pmatrix}$ représente le vecteur des corrélations entre variables explicatives et observations

Sommaire

Estimateur des moindres carrés et unicité

Prenons $\hat{\boldsymbol{\theta}}$ (une) solution de $\boxed{(X^\top X)\hat{\boldsymbol{\theta}} = X^\top \mathbf{y}}$

Non unicité : si $\mathrm{Ker}(X) = \{\boldsymbol{\theta} \in \mathbb{R}^{p+1} : X\boldsymbol{\theta} = 0\} \neq \{0\}$ (noyau non trivial), prenons $\boldsymbol{\theta}_K \in \mathrm{Ker}(X)$ non nul, alors

$$X(\hat{\boldsymbol{\theta}} + \boldsymbol{\theta}_K) = X\hat{\boldsymbol{\theta}}$$

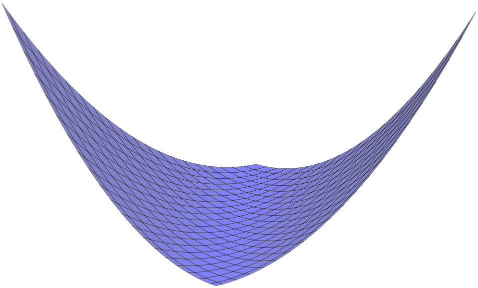
$$\text{puis} \quad (X^\top X)(\hat{\boldsymbol{\theta}} + \boldsymbol{\theta}_K) = X^\top \mathbf{y}$$

Cela montre que l'espace des solutions de l'équation normale peut s'écrire comme un sous espace (affine) :

$$\boxed{\hat{\boldsymbol{\theta}} + \mathrm{Ker}(X)}$$

Optimisation dans $\mathbb{R}^d$

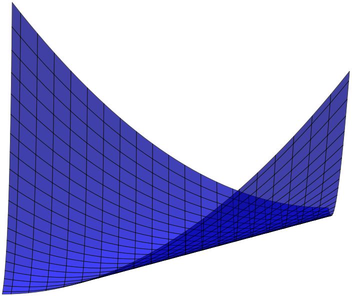
Cas d'une fonction convexe, *e.g.*, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



Rem: l'ensemble des minimiseurs est dans ce cas une droite

Optimisation dans $\mathbb{R}^d$

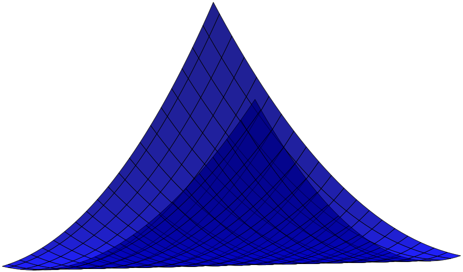
Cas d'une fonction convexe, *e.g.*, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



<u>Rem</u>: l'ensemble des minimiseurs est dans ce cas une droite

Optimisation dans $\mathbb{R}^d$

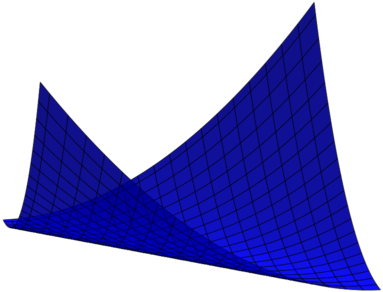
Cas d'une fonction convexe, *e.g.*, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



<u>Rem</u>: l'ensemble des minimiseurs est dans ce cas une droite

Optimisation dans $\mathbb{R}^d$

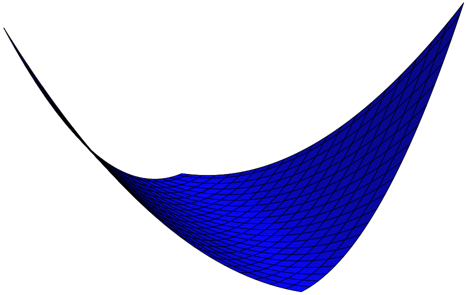
Cas d'une fonction convexe, *e.g.*, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



Rem: l'ensemble des minimiseurs est dans ce cas une droite

Optimisation dans $\mathbb{R}^d$

Cas d'une fonction convexe, *e.g.*, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



Rem: l'ensemble des minimiseurs est dans ce cas une droite

Non unicité : interprétation pour une variable

Rappel : $$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Si $\mathrm{Ker}(X) = \{\boldsymbol{\theta} \in \mathbb{R}^2 : X\boldsymbol{\theta} = 0\} \neq \{0\}$ il existe $(\theta_0, \theta_1) \neq (0, 0)$:

$$\begin{cases} \theta_0 + \theta_1 x_1 & = 0 \\ \vdots \qquad \vdots & = \vdots \\ \theta_0 + \theta_1 x_n & = 0 \end{cases}$$

1. si $\theta_1 = 0$ **absurde**, car alors $\theta_0 = 0$, et donc $(\theta_0, \theta_1) \neq (0, 0)$
2. si $\theta_1 \neq 0$
   2.1 si $\forall i, x_i = 0$ alors $X = (\mathbf{1}_n, 0)$
   2.2 sinon il existe $x_{i_0} \neq 0$ puis $\forall i, x_i = -\theta_0/\theta_1 = x_{i_0}$, *i.e.*, $X = (\mathbf{1}_n \quad x_{i_0} \cdot \mathbf{1}_n)$

Interprétation en dimension quelconque

Rappel : on note $X = (\mathbf{1}_n, \mathbf{x}_1, \dots, \mathbf{x}_p)$, les colonnes étant les variables explicatives (de taille n)

La propriété $\mathrm{Ker}(X) = \{\boldsymbol{\theta} \in \mathbb{R}^{p+1} : X\boldsymbol{\theta} = 0\} \neq \{0\}$ signifie qu'il existe une relation linéaire entre les variables explicatives $\mathbf{1}_n, \mathbf{x}_1, \dots, \mathbf{x}_p$ (on dit aussi que les variables sont liées),*i.e.*, il existe un vecteur non nul $\boldsymbol{\theta} = (\theta_0, \dots, \theta_p)^\top \in \mathbb{R}^p$ tel que

$$\theta_0 \mathbf{1}_n + \sum_{j=1}^{p} \theta_j \mathbf{x}_j = 0$$

Quelques rappels d'algèbre

Définition

Rang d'une matrice : $\quad \mathrm{rang}(X) = \dim(\mathrm{vect}(\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p))$

Propriété : $\mathrm{rang}(X) = \mathrm{rang}(X^\top)$

Théorème du rang

$$\mathrm{rang}(X) + \dim(\mathrm{Ker}(X)) = p + 1$$

$$\mathrm{rang}(X^\top) + \dim(\mathrm{Ker}(X^\top)) = n$$

Exo: $\mathrm{Ker}(X) = \mathrm{Ker}(X^\top X)$

Rem: $\boxed{\mathrm{rang}(X) \leqslant \min(n, p + 1)}$

Détails sur ce thème : *cf.* Golub et Van Loan (1996)

Quelques rappels d'algèbre (suite)

Caractérisation de l'inversion

Une matrice carrée $A \in \mathbb{R}^{m \times m}$ est inversible

- si et seulement si son noyau est nul : $\mathrm{Ker}(A) = \{0\}$
- si et seulement elle est de plein rang $\mathrm{rang}(A) = m$

Exo: Montrer que $\mathrm{Ker}(A) = \{0\}$ est équivalent au fait que la matrice $A^\top A$ est inversible.

Sommaire

Formule des moindres carrés

Formule pour le cas d'un noyau non trivial

Si la matrice X est de plein rang (*i.e.*, si $X^\top X$ inversible) alors

$$\hat{\boldsymbol{\theta}} = (X^\top X)^{-1} X^\top \mathbf{y}$$

Rem: on retrouve pour la moyenne pour le cas simple $X = \mathbf{1}_n$:
$\hat{\boldsymbol{\theta}} = (\langle \mathbf{1}_n, \mathbf{1}_n \rangle)^{-1} \langle \mathbf{1}_n, \mathbf{y} \rangle = \bar{y}_n$

Rem: dans le cas simple $X = \mathbf{x} = (x_1, \ldots, x_n)^\top$: $\hat{\boldsymbol{\theta}} = \langle \frac{\mathbf{x}}{\|\mathbf{x}\|^2}, \mathbf{y} \rangle$

Exo: retrouver le cas unidimensionnel avec constante

ATTENTION : en pratique éviter de calculer l'inverse de $X^\top X$:

- cela est coûteux en temps de calcul
- une matrice $(p+1) \times (p+1)$ peut être volumineuse, si "$p \gg n$" (*e.g.*, en biologie n patients, p gènes...)

Prédiction

Définition

Vecteurs des prédictions : $\hat{\mathbf{y}} = X\hat{\boldsymbol{\theta}}$

Rem: $\hat{\mathbf{y}}$ est une fonction linéaire des observations $\mathbf{y}$

Rappel : un **projecteur orthogonal** est une matrice H telle que

1. H est symétrique : $H^\top = H$
2. H est idempotente : $H^2 = H$

Proposition

En notant H_X le projecteur orthogonal sur l'espace engendré par les colonnes de X, on obtient que $\hat{\mathbf{y}} = H_X \mathbf{y}$

Rem: si X est de plein rang, alors $\hat{\mathbf{y}} = X(X^\top X)^{-1}X^\top \mathbf{y}$. Dans ce cas $\boxed{H_X = X(X^\top X)^{-1}X^\top}$ est souvent appelée matrice "chapeau" (: *hat matrix*)

Prédiction (suite)

Si une nouvelle observation $x_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})$ arrive, la prédiction associée est :

$$\begin{aligned}
\hat{y}_{n+1} =& \langle \hat{\boldsymbol{\theta}}, (1, x_{n+1,1}, \ldots, x_{n+1,p})^\top \rangle \\
\hat{y}_{n+1} =& \hat{\theta}_0 + \sum_{j=1}^{p} \hat{\theta}_j x_{n+1,j}
\end{aligned}$$

<u>Rem</u>: l'équation normale assure l'**équi-corrélation** entre des observations et des prédictions avec les variables explicatives :

$$(X^\top X)\hat{\boldsymbol{\theta}} = X^\top \mathbf{y} \Leftrightarrow X^\top \hat{\mathbf{y}} = X^\top \mathbf{y} \Leftrightarrow \begin{pmatrix} \langle \mathbf{x}_0, \hat{\mathbf{y}} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \hat{\mathbf{y}} \rangle \end{pmatrix} = \begin{pmatrix} \langle \mathbf{x}_0, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \mathbf{y} \rangle \end{pmatrix}$$

Exo: Soit $P = \begin{pmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} \in \mathbb{R}^{n\times n}$.

1. Vérifier que P est une matrice de projection orthogonale.
2. Déterminer $\mathrm{Im(P)}$, l'espace image de P.
3. On note $\mathbf{x} = (x_1, \dots, x_n)^\top$ et $\overline{x}_n$ la moyenne et σ_x l'écart-type (empirique) :

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad \sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}.$$

   Montrer que $\sigma_x = \|(\mathrm{Id}_n - P)x\|/\sqrt{n}$.

Résidus et équations normales

Définition

Résidu(s) : $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X\hat{\boldsymbol{\theta}} = (\mathrm{Id}_n - H_X)\mathbf{y}$

Rappel :

Équations normales : $\boxed{(X^\top X)\hat{\boldsymbol{\theta}} = X^\top \mathbf{y}}$

Grâce aux résidus on peut écrire cette équation sous la forme :

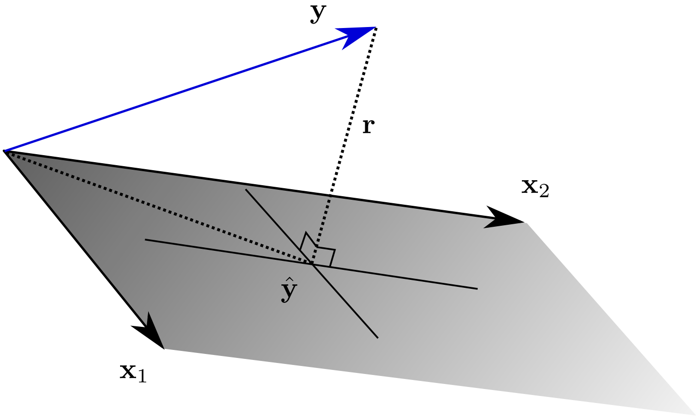
$$X^\top(X\hat{\boldsymbol{\theta}} - \mathbf{y}) = 0 \Leftrightarrow X^\top \mathbf{r} = 0 \Leftrightarrow \mathbf{r}^\top X = 0$$

Cela se réécrit avec $X = (\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p)$ de la manière suivante :

$$\forall j = 1, \ldots, p : \langle \mathbf{r}, \mathbf{x}_j \rangle = 0 \text{ et } \overline{r}_n = 0$$

Interprétation : le résidu est orthogonal aux variables explicatives

Visualisation : prédicateurs et résidus $(p = 2)$



Sommaire

Biais

Rappel : $\mathbf{y} = X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$

Proposition

Sous l'hypothèse que $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ et que la matrice X est de plein rang, alors l'estimateur des moindres carrés est sans biais :

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^*$$

Rem: l'hypothèse $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ signifie que $\forall i \in [\![1, n]\!], \mathbb{E}(\varepsilon_i) = 0$

Démonstration :

$$B = \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^* = \mathbb{E}((X^\top X)^{-1} X^\top \mathbf{y}) - \boldsymbol{\theta}^*$$

Biais

Rappel : $\mathbf{y} = X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$

Proposition

Sous l'hypothèse que $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ et que la matrice X est de plein rang, alors l'estimateur des moindres carrés est sans biais :

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^*$$

Rem: l'hypothèse $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ signifie que $\forall i \in [\![1, n]\!], \mathbb{E}(\varepsilon_i) = 0$

Démonstration :

$$B = \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^* = \mathbb{E}((X^\top X)^{-1} X^\top \mathbf{y}) - \boldsymbol{\theta}^*$$

$$B = \mathbb{E}((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \varepsilon)) - \boldsymbol{\theta}^*$$

Biais

Rappel : $\mathbf{y} = X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$

Proposition

Sous l'hypothèse que $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ et que la matrice X est de plein rang, alors l'estimateur des moindres carrés est sans biais :

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^*$$

Rem: l'hypothèse $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ signifie que $\forall i \in [\![1, n]\!], \mathbb{E}(\varepsilon_i) = 0$

Démonstration :

$$\begin{aligned}
B =& \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^* = \mathbb{E}((X^\top X)^{-1} X^\top \mathbf{y}) - \boldsymbol{\theta}^* \\
B =& \mathbb{E}((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon})) - \boldsymbol{\theta}^* \\
B =& (X^\top X)^{-1} X^\top X \boldsymbol{\theta}^* + (X^\top X)^{-1} X^\top \mathbb{E}(\boldsymbol{\varepsilon}) - \boldsymbol{\theta}^* = 0
\end{aligned}$$

Biais

Rappel : $\mathbf{y} = X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$

Proposition

Sous l'hypothèse que $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ et que la matrice X est de plein rang, alors l'estimateur des moindres carrés est sans biais :

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^*$$

Rem: l'hypothèse $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ signifie que $\forall i \in [\![1, n]\!], \mathbb{E}(\varepsilon_i) = 0$

Démonstration :

$$\begin{aligned}
B =& \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^* = \mathbb{E}((X^\top X)^{-1} X^\top \mathbf{y}) - \boldsymbol{\theta}^* \\
B =& \mathbb{E}((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon})) - \boldsymbol{\theta}^* \\
B =& (X^\top X)^{-1} X^\top X \boldsymbol{\theta}^* + (X^\top X)^{-1} X^\top \mathbb{E}(\boldsymbol{\varepsilon}) - \boldsymbol{\theta}^* = 0
\end{aligned}$$

Risque quadratique

Définition

Le **risque quadratique** est la quantité suivante :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2$$

Décomposition biais/variance

$$\boxed{\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2 = \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2}$$

<u>Démonstration</u> :

$$\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2 = \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}}) + \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2$$

Risque quadratique

Définition

Le **risque quadratique** est la quantité suivante :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2$$

Décomposition biais/variance

$$\boxed{\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2 = \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2}$$

<u>Démonstration</u> :

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2 =& \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}}) + \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\ =& \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\ &+ 2\mathbb{E}\langle \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\rangle\end{aligned}$$

Risque quadratique

Définition

Le **risque quadratique** est la quantité suivante :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2$$

Décomposition biais/variance

$$\boxed{\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2 = \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2}$$

<u>Démonstration</u> :

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2 =& \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}}) + \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\
=& \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\
& + 2\mathbb{E}\langle \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\rangle \\
=& \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2
\end{aligned}$$

Risque quadratique

Définition

Le **risque quadratique** est la quantité suivante :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2$$

Décomposition biais/variance

$$\boxed{\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2 = \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2}$$

<u>Démonstration</u> :

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2 =& \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}}) + \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\
=& \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\
&+ 2\mathbb{E}\langle \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\rangle \\
=& \mathbb{E}\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2
\end{aligned}$$

Décomposition biais/variance

Rappel : pour les moindres carrés le biais est nul sous l'hypothèse que X est de plein rang, ainsi $\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^* = 0$ et

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2 =& \|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\
\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2 =& \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2
\end{aligned}$$

Sommaire

Intermédiaire sur la trace

Définition

Soit $A \in \mathbb{R}^{n \times n}$ une matrice carrée. La **trace** de A, notée $\operatorname{tr}(A)$ vaut la somme des éléments diagonaux de A :

$$\operatorname{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Quelques propriétés :

- $\operatorname{tr}(A) = \operatorname{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n \times n}$, et tout $\alpha \in \mathbb{R}$, $\operatorname{tr}(\alpha A + B) = \alpha \operatorname{tr}(A) + \operatorname{tr}(B)$ (linéarité)

Intermédiaire sur la trace

Définition

Soit $A \in \mathbb{R}^{n\times n}$ une matrice carrée. La **trace** de A, notée $\mathrm{tr}(A)$ vaut la somme des éléments diagonaux de A :

$$\mathrm{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Quelques propriétés :

- $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n\times n}$, et tout $\alpha \in \mathbb{R}$, $\mathrm{tr}(\alpha A + B) = \alpha\,\mathrm{tr}(A) + \mathrm{tr}(B)$ (linéarité)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} A_{i,j}^2$

Intermédiaire sur la trace

Définition

Soit $A \in \mathbb{R}^{n\times n}$ une matrice carrée. La **trace** de A, notée $\mathrm{tr}(A)$ vaut la somme des éléments diagonaux de A :

$$\mathrm{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Quelques propriétés :

- $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n\times n}$, et tout $\alpha \in \mathbb{R}$, $\mathrm{tr}(\alpha A + B) = \alpha\,\mathrm{tr}(A) + \mathrm{tr}(B)$ (linéarité)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} A_{i,j}^2$
- Pour toutes matrices $A, B \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$

Intermédiaire sur la trace

Définition

Soit $A \in \mathbb{R}^{n\times n}$ une matrice carrée. La **trace** de A, notée $\mathrm{tr}(A)$ vaut la somme des éléments diagonaux de A :

$$\mathrm{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Quelques propriétés :

- $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n\times n}$, et tout $\alpha \in \mathbb{R}$, $\mathrm{tr}(\alpha A + B) = \alpha\,\mathrm{tr}(A) + \mathrm{tr}(B)$ (linéarité)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} A_{i,j}^2$
- Pour toutes matrices $A, B \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
- $\mathrm{tr}(PAP^{-1}) = \mathrm{tr}(A)$, donc si A est diagonalisable, sa trace est la somme de ses valeurs propres

Intermédiaire sur la trace

Définition

Soit $A \in \mathbb{R}^{n \times n}$ une matrice carrée. La **trace** de A, notée $\operatorname{tr}(A)$ vaut la somme des éléments diagonaux de A :

$$\operatorname{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Quelques propriétés :

- $\operatorname{tr}(A) = \operatorname{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n \times n}$, et tout $\alpha \in \mathbb{R}$, $\operatorname{tr}(\alpha A + B) = \alpha \operatorname{tr}(A) + \operatorname{tr}(B)$ (linéarité)
- $\operatorname{tr}(A^\top A) = \sum_{i=1}^{n} A_{i,j}^2$
- Pour toutes matrices $A, B \in \mathbb{R}^{n \times n}$, $\operatorname{tr}(AB) = \operatorname{tr}(BA)$
- $\operatorname{tr}(PAP^{-1}) = \operatorname{tr}(A)$, donc si A est diagonalisable, sa trace est la somme de ses valeurs propres
- Si H est un projecteur orthogonal $\operatorname{tr}(H) = \operatorname{rang}(H)$

Intermédiaire sur la trace

Définition

Soit $A \in \mathbb{R}^{n\times n}$ une matrice carrée. La **trace** de A, notée $\mathrm{tr}(A)$ vaut la somme des éléments diagonaux de A :

$$\mathrm{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Quelques propriétés :

- $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n\times n}$, et tout $\alpha \in \mathbb{R}$, $\mathrm{tr}(\alpha A + B) = \alpha\,\mathrm{tr}(A) + \mathrm{tr}(B)$ (linéarité)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} A_{i,j}^2$
- Pour toutes matrices $A, B \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
- $\mathrm{tr}(PAP^{-1}) = \mathrm{tr}(A)$, donc si A est diagonalisable, sa trace est la somme de ses valeurs propres
- Si H est un projecteur orthogonal $\mathrm{tr}(H) = \mathrm{rang}(H)$

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \operatorname{tr}((X^\top X)^{-1})$$

Démonstration :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top (\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \varepsilon) - \boldsymbol{\theta}^*)^\top ((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \varepsilon) - \boldsymbol{\theta}^*)\right]$$

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \operatorname{tr}((X^\top X)^{-1})$$

Démonstration :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1}X^\top(X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)^\top((X^\top X)^{-1}X^\top(X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1}X^\top\varepsilon)^\top((X^\top X)^{-1}X^\top\varepsilon)\right] = \mathbb{E}(\varepsilon^\top X(X^\top X)^{-2}X^\top\varepsilon)$$

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \operatorname{tr}((X^\top X)^{-1})$$

<u>Démonstration</u> :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top (\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)^\top ((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top ((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})\right] = \mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-2} X^\top \boldsymbol{\varepsilon})$$

$$= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1} (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})]$$

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \operatorname{tr}((X^\top X)^{-1})$$

Démonstration :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top (\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)^\top ((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top ((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})\right] = \mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-2} X^\top \boldsymbol{\varepsilon})$$

$$= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1} (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})]$$

$$= \mathbb{E}\left(\operatorname{tr}\left[(X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1}\right]\right)$$

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \operatorname{tr}((X^\top X)^{-1})$$

<u>Démonstration</u> :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1}X^\top(X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)^\top((X^\top X)^{-1}X^\top(X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})^\top((X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})\right] = \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-2}X^\top\boldsymbol{\varepsilon})$$

$$= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})]$$

$$= \mathbb{E}\left(\operatorname{tr}\left[(X^\top X)^{-1}X^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}\right]\right)$$

$$= \operatorname{tr}[(X^\top X)^{-1}X^\top\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)X(X^\top X)^{-1}]$$

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \operatorname{tr}((X^\top X)^{-1})$$

<u>Démonstration</u> :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top (\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)^\top ((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top ((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})\right] = \mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-2} X^\top \boldsymbol{\varepsilon})$$

$$= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1} (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})]$$

$$= \mathbb{E}\left(\operatorname{tr}\left[(X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1}\right]\right)$$

$$= \operatorname{tr}[(X^\top X)^{-1} X^\top \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) X (X^\top X)^{-1}]$$

$$= \sigma^2 \operatorname{tr}((X^\top X)^{-1})$$

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \, \mathrm{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \, \mathrm{tr}((X^\top X)^{-1})$$

Démonstration :

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top (\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right]$$
$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)^\top ((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)\right]$$
$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top ((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})\right] = \mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-2} X^\top \boldsymbol{\varepsilon})$$
$$= \mathrm{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1} (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})]$$
$$= \mathbb{E}\left(\mathrm{tr}\left[(X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1}\right]\right)$$
$$= \mathrm{tr}[(X^\top X)^{-1} X^\top \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) X (X^\top X)^{-1}]$$
$$= \sigma^2 \, \mathrm{tr}((X^\top X)^{-1})$$

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^* - \hat{\mathbf{y}}\|^2/n$

Sous l'hypothèse de modèle homoscédastique :

$$R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \frac{\operatorname{rang}(X)}{n}$$

Démonstration : début identique

$$n \cdot R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right]$$
$$= \mathbb{E}(\varepsilon^\top X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top \varepsilon)$$

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^* - \hat{\mathbf{y}}\|^2/n$

Sous l'hypothèse de modèle homoscédastique :

$$R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \frac{\operatorname{rang}(X)}{n}$$

<u>Démonstration</u> : début identique

$$\begin{aligned} n \cdot R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] \\ &= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}) \\ &= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}) \end{aligned}$$

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^* - \hat{\mathbf{y}}\|^2/n$

Sous l'hypothèse de modèle homoscédastique :

$$R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \frac{\operatorname{rang}(X)}{n}$$

<u>Démonstration</u> : début identique

$$\begin{aligned}
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\
&= \operatorname{tr}[\mathbb{E}(\varepsilon^\top H_X \varepsilon)] = \operatorname{tr}[\mathbb{E}(\varepsilon^\top H_X^\top H_X \varepsilon)]
\end{aligned}$$

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^* - \hat{\mathbf{y}}\|^2/n$

Sous l'hypothèse de modèle homoscédastique :

$$R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \frac{\operatorname{rang}(X)}{n}$$

<u>Démonstration</u> : début identique

$$\begin{aligned}
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1} (X^\top X)(X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}) \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}) \\
&= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X \boldsymbol{\varepsilon})] = \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X^\top H_X \boldsymbol{\varepsilon})] \\
&= \operatorname{tr}[\mathbb{E}(H_X \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top H_X^\top)] = \operatorname{tr}\left(H_X \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) H_X^\top\right)
\end{aligned}$$

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^* - \hat{\mathbf{y}}\|^2/n$

Sous l'hypothèse de modèle homoscédastique :

$$R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \frac{\operatorname{rang}(X)}{n}$$

Démonstration : début identique

$$\begin{aligned}
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\
&= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X \boldsymbol{\varepsilon})] = \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X^\top H_X \boldsymbol{\varepsilon})] \\
&= \operatorname{tr}[\mathbb{E}(H_X \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top H_X^\top)] = \operatorname{tr}\left(H_X \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) H_X^\top\right) \\
&= \sigma^2 \operatorname{tr}(H_X) = \sigma^2 \operatorname{rang}(H_X) = \sigma^2 \operatorname{rang}(X)
\end{aligned}$$

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^* - \hat{\mathbf{y}}\|^2/n$

Sous l'hypothèse de modèle homoscédastique :

$$R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] = \sigma^2 \frac{\operatorname{rang}(X)}{n}$$

Démonstration : début identique

$$\begin{aligned}
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\
&= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X \boldsymbol{\varepsilon})] = \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X^\top H_X \boldsymbol{\varepsilon})] \\
&= \operatorname{tr}[\mathbb{E}(H_X \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top H_X^\top)] = \operatorname{tr}\left(H_X \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) H_X^\top\right) \\
&= \sigma^2 \operatorname{tr}(H_X) = \sigma^2 \operatorname{rang}(H_X) = \sigma^2 \operatorname{rang}(X)
\end{aligned}$$

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

Sous l'hypothèse de modèle homoscédastique et que X est de plein rang :

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

<u>Démonstration</u> : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

$$V = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \varepsilon) - \boldsymbol{\theta}^*)((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \varepsilon) - \boldsymbol{\theta}^*)^\top\right]$$

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

Sous l'hypothèse de modèle homoscédastique et que X est de plein rang :

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

Démonstration : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

$$\begin{aligned}
V =& \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\right] \\
=& \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)^\top\right] \\
=& \mathbb{E}\left[((X^\top X)^{-1} X^\top \varepsilon)((X^\top X)^{-1} X^\top \varepsilon)^\top\right]
\end{aligned}$$

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

Sous l'hypothèse de modèle homoscédastique et que X est de plein rang :

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

<u>Démonstration</u> : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

$$\begin{aligned}
V =& \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\right] \\
=& \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)^\top\right] \\
=& \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top\right] \\
=& (X^\top X)^{-1} X^\top \mathbb{E}\left[\varepsilon\varepsilon^\top\right] X (X^\top X)^{-1}
\end{aligned}$$

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

Sous l'hypothèse de modèle homoscédastique et que X est de plein rang :

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

<u>Démonstration</u> : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

$$\begin{aligned}
V =& \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\right] \\
=& \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)^\top\right] \\
=& \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top\right] \\
=& (X^\top X)^{-1} X^\top \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right] X (X^\top X)^{-1} \\
=& (X^\top X)^{-1} X^\top (\sigma^2 \operatorname{Id}_n) X (X^\top X)^{-1}
\end{aligned}$$

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

Sous l'hypothèse de modèle homoscédastique et que X est de plein rang :

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

<u>Démonstration</u> : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

$$\begin{aligned}
V =& \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\right] \\
=& \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)^\top\right] \\
=& \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top\right] \\
=& (X^\top X)^{-1} X^\top \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right] X (X^\top X)^{-1} \\
=& (X^\top X)^{-1} X^\top (\sigma^2 \,\mathrm{Id}_n) X (X^\top X)^{-1} \\
=& \sigma^2 (X^\top X)^{-1}
\end{aligned}$$

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

Sous l'hypothèse de modèle homoscédastique et que X est de plein rang :

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

<u>Démonstration</u> : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

$$\begin{aligned}
V =& \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\right] \\
=& \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*)^\top\right] \\
=& \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top\right] \\
=& (X^\top X)^{-1} X^\top \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right] X (X^\top X)^{-1} \\
=& (X^\top X)^{-1} X^\top (\sigma^2 \,\mathrm{Id}_n) X (X^\top X)^{-1} \\
=& \sigma^2 (X^\top X)^{-1}
\end{aligned}$$

Sommaire

Estimateur du niveau de bruit

- On peut construire un estimateur de le variance σ^2 du bruit :

$$\boxed{\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2}$$

ou si l'on souhaite un estimateur sans biais :

$$\boxed{\hat{\sigma}^2 = \frac{1}{n - \mathrm{rg}(X)}\|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2}$$

- Motivation "débiaisage" : théorie des tests

$$\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \mathbf{y}^\top(\mathrm{Id}_n - H_X)\mathbf{y} = \boldsymbol{\varepsilon}^\top(\mathrm{Id}_n - H_X)\boldsymbol{\varepsilon} = \sum_{i=1}^{n-\mathrm{rg}(X)} \tilde{\varepsilon}_i^2$$

Cas gaussien : si $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, alors $\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$ suit une loi du χ^2 à $n - \mathrm{rg}(X)$ degrés de liberté

Rem: implicitement on fait donc encore l'hypothèse $n > p$

Sommaire

Cas hétéroscédastique

L'estimateur MCO $\hat{\boldsymbol{\theta}}$ postule implicitement que les variables $y_1, \ldots, y_n$ ont même niveau de bruit

Rem: pour cela reprendre le calcul du maximum de vraisemblance d'un modèle gaussien avec variance σ^2 fixée / connue

Modèle hétéroscédastique : on suppose que le niveau de bruit diffère pour chaque y_i et on note σ_i^2 la variance associée

$$\hat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left(\frac{y_i - \langle \boldsymbol{\theta}, x_i \rangle}{\sigma_i} \right)^2 = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} (y - X\boldsymbol{\theta})^\top \Omega (y - X\boldsymbol{\theta})$$

avec $\Omega = \operatorname{diag}(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_n^2})$

Exo: donner une formule explicite si $X^\top \Omega X$ est de plein rang

---

Sommaire

Variables qualitatives

On parle de variable qualitative, quand une variable ne prend que des modalités discrètes et/ou non-numériques.

Exemple : couleurs, genre, ville, etc.

Encodage classique : variables fictives/indicatrices
(🇬🇧 : *dummy variables*)="encodage à chaud"
(🇬🇧 : *one-hot encoder*).

Si la variable $\mathbf{x}$ peut prendre K modalités $a_1, \ldots, a_K$ on créer les K variables explicatives suivantes : $\forall k \in [\![1, K]\!], \mathbb{1}_{a_k} \in \mathbb{R}^n$ définies par

$$\forall i \in [\![1, n]\!], \quad (\mathbb{1}_{a_k})_i = \begin{cases} 1, & \text{if } x_i = a_k \\ 0, & \text{sinon} \end{cases}$$

Exemple d'encodage

Cas binaire : M/F, oui/non, j'aime/j'aime pas.

Client	Genre
1	H
2	F
3	H
4	F
5	F

$$\longrightarrow \begin{pmatrix} F & H \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}$$

Cas général : couleur, villes, etc.

Client	Couleurs
1	Bleu
2	Blanc
3	Rouge
4	Rouge
5	Bleu

$$\longrightarrow \begin{pmatrix} \text{Bleu} & \text{Blanc} & \text{Rouge} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Quelques difficultés

Corrélations : $\sum_{k=1}^{K} \mathbb{1}_{a_k} = \mathbf{1}_n$! On peut enlever une des modalités (*e.g.*, `drop_first=True` dans `get_dummies` de `pandas`)

Interprétation sans constante et avec toutes les modalités :
$X = [\mathbb{1}_{a_1}, \dots, \mathbb{1}_{a_K}]$. Si $x_{n+1} = a_k$ alors $\hat{y}_{n+1} = \hat{\boldsymbol{\theta}}_k$

Interprétation sans constante et avec une modalité en moins :
$X = [\mathbf{1}_n, \mathbb{1}_{a_2}, \dots, \mathbb{1}_{a_K}]$, en choisissant d'enlever la première modalité

Si $x_{n+1} = a_k$ alors $\hat{y}_{n+1} = \begin{cases} \hat{\boldsymbol{\theta}}_0, & \text{si } k = 1 \\ \hat{\boldsymbol{\theta}}_0 + \hat{\boldsymbol{\theta}}_k, & \text{sinon} \end{cases}$

Rem: peut créer une colonne nulle en CV

Rem: difficultés limitées par régularisation (*e.g.*, Lasso, Ridge)

Exo: Calculer l'estimateur des moindres carrés avec $X = [\mathbb{1}_{a_1}, \dots, \mathbb{1}_{a_K}]$ obtenu par des *dummy variables* avec une seule variable explicative ayant K modalités

Sommaire

Et si $n < p$?

Beaucoup des choses vues avant ont besoin d'être révisées :

Par exemple : si $\mathrm{rg}(X) = n$, alors $H_X = \mathrm{Id}_n$ et $\hat{\mathbf{y}} = X\hat{\boldsymbol{\theta}} = \mathbf{y}$!
En effet, l'espace engendré par les colonnes $[\mathbf{x}_0, \ldots, \mathbf{x}_p]$ est $\mathbb{R}^n$, et donc le signal observé et le signal prédit sont **identiques**

Rem: c'est un problème inhérent à la grande dimension (grand nombre de variables explicatives p)

Solutions possibles : sélection de variables, *cf.* cours sur le Lasso et méthodes gloutonnes (à venir)

Sites webs et livres pour aller plus loin

- Packages Python pour les moindres carrés :
  `statsmodels`
  `sklearn.linear_model.LinearRegression`
- McKinney (2012) concernant `python` pour les statistiques
- Lejeune (2010) concernant le modèle linéaire (notamment)
- Delyon (2015) cours plus avancé sur la régression :
  https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf

Références I

- B. Delyon.
  Régression, 2015.
  https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf.
- G. H. Golub and C. F. van Loan.
  Matrix computations.
  Johns Hopkins University Press, Baltimore, MD, third edition, 1996.
- M. Lejeune.
  Statistiques, la théorie et ses applications.
  Springer, 2010.
- W. McKinney.
  Python for Data Analysis : Data Wrangling with Pandas, NumPy, and IPython.
  O'Reilly Media, 2012.