

MS BGD: MDI720
Modèle linéaire multidimensionnel

Anne Sabourin
Telecom ParisTech

Septembre 2018

1. Moindres carrés pour deux variables explicatives

2. Moindres carrés multi-dimensionnels

- Modélisation matricielle

- Définition des moindres carrés

- Optimisation

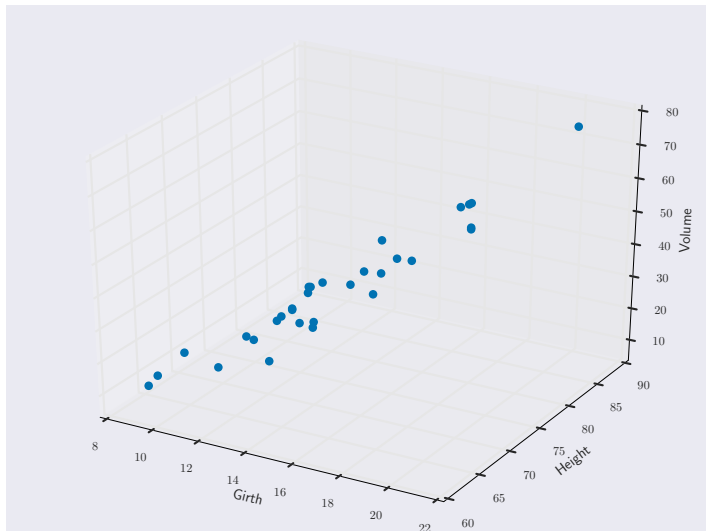
- Questions d'unicité

- Formule explicite, prédiction et résidus

- Coefficient de détermination

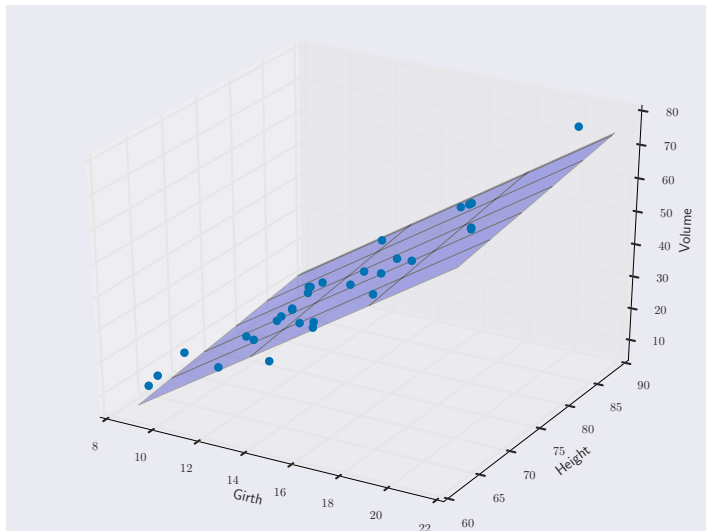
Vers des modèles multi-variés

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

```
from matplotlib.mplot3d import Axes3D
# 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

1. Moindres carrés pour deux variables explicatives

2. Moindres carrés multi-dimensionnels

- Modélisation matricielle

- Définition des moindres carrés

- Optimisation

- Questions d'unicité

- Formule explicite, prédiction et résidus

- Coefficient de détermination

1. Moindres carrés pour deux variables explicatives

2. Moindres carrés multi-dimensionnels

Modélisation matricielle

Définition des moindres carrés

Optimisation

Questions d'unicité

Formule explicite, prédiction et résidus

Coefficient de détermination

Modélisation

On dispose de p variables explicatives $(\mathbf{x}_1, \dots, \mathbf{x}_p)$

Modèle en dimension p

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

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

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

Rem: on fait l'hypothèse qu'il existe un vrai paramètre

$$\boldsymbol{\theta}^* = (\theta_0^*, \dots, \theta_p^*)^\top$$

Dimension p

Écriture matricielle

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

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

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

Notation ligne : $X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} = (x_1, \dots, x_n)^\top$

Rem: parfois \mathbf{x}_0 sera omis par simplicité

Vocabulaire

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

- $\mathbf{y} \in \mathbb{R}^n$: 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
- $\boldsymbol{\epsilon} \in \mathbb{R}^n$: vecteur de bruit

Point de vue “observations” : $y_i = \langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle + \varepsilon_i$ pour $i = 1, \dots, n$

Point de vue “vectoriel” : $\mathbf{y} = \sum_{j=0}^p \theta_j^* \mathbf{x}_j + \boldsymbol{\epsilon}$

1. Moindres carrés pour deux variables explicatives

2. Moindres carrés multi-dimensionnels

Modélisation matricielle

Définition des moindres carrés

Optimisation

Questions d'unicité

Formule explicite, prédiction et résidus

Coefficient de détermination

Estimateur des moindres carrés (ordinaires)

Un estimateur des moindres carrés est solution du problème :

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

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

$$\hat{\boldsymbol{\theta}} \in \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \frac{1}{2} \sum_{i=1}^n [y_i - \langle \mathbf{x}_i, \boldsymbol{\theta} \rangle]^2$$

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

1. Moindres carrés pour deux variables explicatives

2. Moindres carrés multi-dimensionnels

Modélisation matricielle

Définition des moindres carrés

Optimisation

Questions d'unicité

Formule explicite, prédiction et résidus

Coefficient de détermination

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

Théorème : règle de Fermat

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

Rem: ce n'est une condition suffisante que si f est en plus convexe

Pour notre problème $f : \boldsymbol{\theta} \mapsto \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$ ou encore :

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

Calcul du gradient de f

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

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

Calcul du gradient de f

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

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

Pour notre fonction f , cela donne

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

Calcul du gradient de f

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

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

Pour notre fonction f , cela donne

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

Calcul du gradient de f

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

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

Pour notre fonction f , cela donne

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

Calcul du gradient de f

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

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

Pour notre fonction f , cela donne

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

Calcul du gradient de f

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

$$f(\theta + h) = f(\theta) + \langle h, \nabla f(\theta) \rangle + o(h)$$

Pour notre fonction f , cela donne

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

Ainsi,

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

Rappel sur le gradient

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

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + o(h) \quad \text{pour tout } h$$

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

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

Moindres carrés - équation(s) normale(s)

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

Théorème

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

Équation(s) normale(s) :

$$\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\theta}} = \mathbf{X}^\top \mathbf{y}$$

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

Rem: 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 carrés

Vocabulaire (et abus de langage)

Définition

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

$$X^T X$$

dont le terme général est $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$.

Rem: $X^T X$ est parfois aussi appelée matrice des corrélations

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

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

1. Moindres carrés pour deux variables explicatives

2. Moindres carrés multi-dimensionnels

Modélisation matricielle

Définition des moindres carrés

Optimisation

Questions d'unicité

Formule explicite, prédiction et résidus

Coefficient de détermination

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 : cela se produit quand

$\text{Ker}(X) = \{\boldsymbol{\theta} \in \mathbb{R}^{p+1} : X\boldsymbol{\theta} = 0\} \neq \{0\}$ (noyau non trivial).

Prenons $\boldsymbol{\theta}_K \in \text{Ker}(X)$ non nul, alors

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

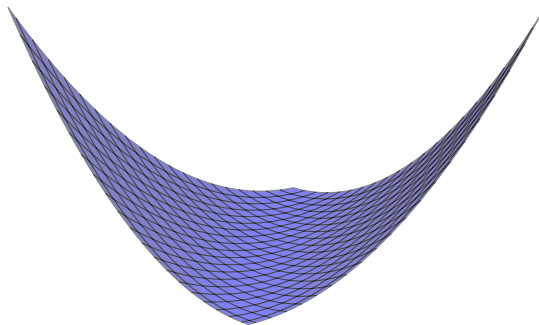
$$\text{puis } (X^\top X)(\hat{\boldsymbol{\theta}} + \boldsymbol{\theta}_K) = X^\top \mathbf{y}$$

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

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

Optimisation dans \mathbb{R}^d

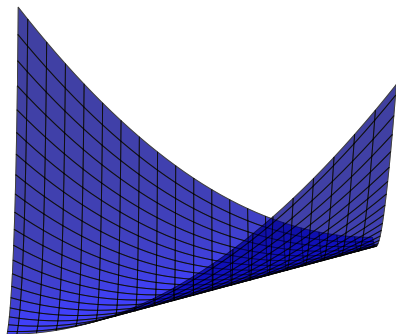
Cas d'une fonction convexe, *e.g.* $f(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



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

Optimisation dans \mathbb{R}^d

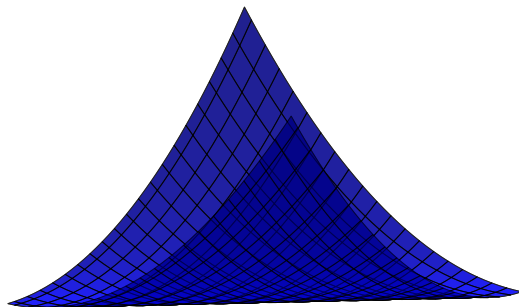
Cas d'une fonction convexe, *e.g.* $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



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

Optimisation dans \mathbb{R}^d

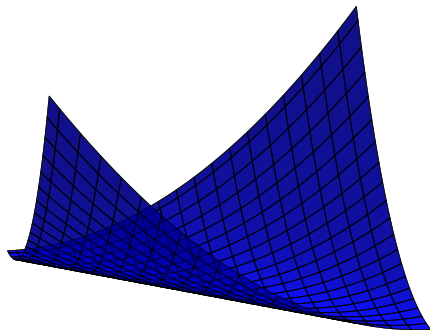
Cas d'une fonction convexe, *e.g.* $f(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



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

Optimisation dans \mathbb{R}^d

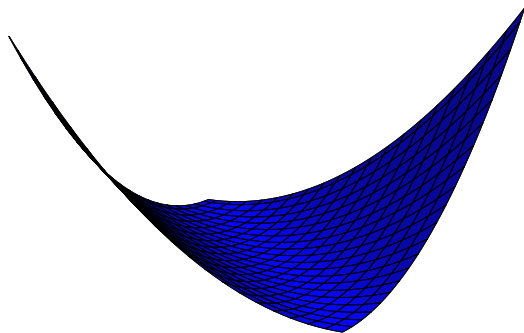
Cas d'une fonction convexe, *e.g.* $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



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

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

$$\begin{cases} \theta_0 + \theta_1 x_1 & = 0 \\ \vdots & \vdots & = \vdots \\ \theta_0 + \theta_1 x_n & = 0 \end{cases} \quad (\star)$$

1. cas $\theta_1 = 0$: $(\star) \Rightarrow \theta_0 = 0$, donc $(\theta_0, \theta_1) = (0, 0)$: **absurde** !

2. cas $\theta_1 \neq 0$:

2.1 si $\forall i, x_i = 0$ alors $X = (\mathbb{1}_n, 0)$ et $\theta_0 = 0$

2.2 sinon il existe $x_{i_0} \neq 0$ puis $\forall i, x_i = -\theta_0/\theta_1 = x_{i_0}$, i.e. $X = [\mathbb{1}_n \quad x_{i_0} \cdot \mathbb{1}_n]$

Interprétation : $\mathbf{x}_1 \propto \mathbb{1}_n$, i.e. \mathbf{x}_1 est constante

Interprétation en dimension quelconque

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

La propriété $\text{Ker}(X) = \{\boldsymbol{\theta} \in \mathbb{R}^{p+1} : X\boldsymbol{\theta} = 0\} \neq \{0\}$ signifie qu'il existe une relation linéaire entre les variables explicatives $\mathbb{1}_n, \mathbf{x}_1, \dots, \mathbf{x}_p$ (on dit aussi que les variables sont liées),

Reformulation : $\exists \boldsymbol{\theta} = (\theta_0, \dots, \theta_p)^\top \in \mathbb{R}^{p+1} \setminus \{0\}$ t.q.

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

Quelques rappels d'algèbre

Définition

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

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

Théorème du rang

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

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

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

Rem:

$\text{rang}(X) \leq \min(n, p + 1)$

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

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

1. Moindres carrés pour deux variables explicatives

2. Moindres carrés multi-dimensionnels

Modélisation matricielle

Définition des moindres carrés

Optimisation

Questions d'unicité

Formule explicite, prédiction et résidus

Coefficient de détermination

Formule des moindres carrés

Formule pour le cas d'un noyau trivial

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

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

Rem: on retrouve la moyenne quand $X = \mathbb{1}_n : \hat{\theta} = \frac{\langle \mathbb{1}_n, y \rangle}{\langle \mathbb{1}_n, \mathbb{1}_n \rangle} = \bar{y}_n$

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

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

- cela est coûteux en temps de calcul
- la matrice $X^\top X$ peut être volumineuse si “ $p \gg n$ ”, e.g. en biologie n patients (≈ 100), p gènes (≈ 10000)

Exo: retrouver le cas unidimensionnel avec constante

Prédiction

Définition

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


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 $H_X = X(X^\top X)^{-1}X^\top$ est appelée la matrice “chapeau” ( : *hat matrix*)

Prédiction (suite)

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

$$\hat{y}_{n+1} = \langle \hat{\boldsymbol{\theta}}, (1, x_{n+1,1}, \dots, x_{n+1,p})^\top \rangle$$

$$\hat{y}_{n+1} = \hat{\theta}_0 + \sum_{j=1}^p \hat{\theta}_j x_{n+1,j}$$

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

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

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 $\text{Im}(P)$, l'espace image de P .
3. On note $\mathbf{x} = (x_1, \dots, x_n)^\top$ et \bar{x}_n la moyenne et $\sigma_{\mathbf{x}}$ l'écart-type (empirique) :

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

Montrer que $\sigma_{\mathbf{x}} = \|(\text{Id}_n - P)\mathbf{x}\|/\sqrt{n}$.

Résidus et équations normales

Définition

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

Rappel :

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

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

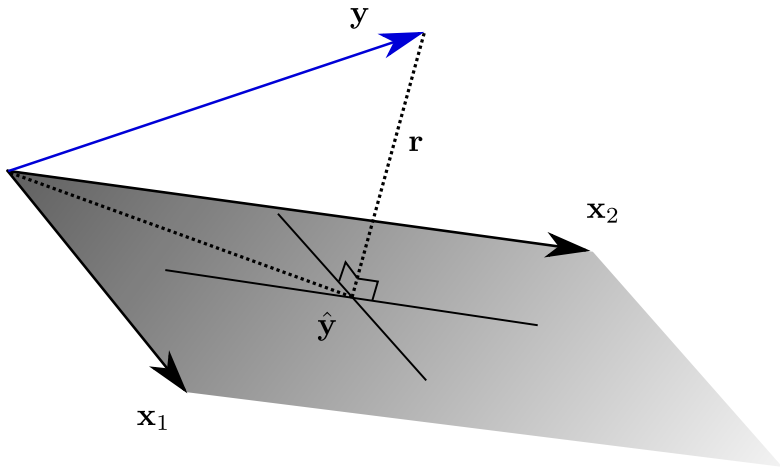
$$X^\top (X\hat{\boldsymbol{\theta}} - \mathbf{y}) = 0 \Leftrightarrow X^\top \mathbf{r} = 0 \Leftrightarrow \mathbf{r}^\top X = 0$$

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

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

Interprétation : le résidu est orthogonal aux variables explicatives

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



1. Moindres carrés pour deux variables explicatives

2. Moindres carrés multi-dimensionnels

Modélisation matricielle

Définition des moindres carrés

Optimisation

Questions d'unicité

Formule explicite, prédiction et résidus

Coefficient de détermination

Résumé de la pertinence du modèle : le " R^2 "

- On suppose les (y_i) non constants, *i.e.* $\sum_1^n (y_i - \bar{y}_n)^2 > 0$.
- Le Coefficient de détermination R^2 est le ratio entre variance 'expliquée' et variance 'totale',

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

- Rem: \bar{y}_n est la moyenne empirique des y_i **et** celle des \hat{y}_i d'après la propriété de centrage des résidus.
- $R^2 \in [0, 1]$ d'après Pythagore et l'orthogonalité des résidus et des prédicteurs :

$$\|Y - \bar{y}_n \mathbb{1}_n\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \bar{y}_n \mathbb{1}_n\|^2$$

(car $\mathbb{1}_n \in \text{vect}(\mathbb{1}_n, x_1, \dots, x_p)$)

le R^2 : comparaison avec le prédicteur constant

$$\|Y - \bar{y}_n \mathbb{1}_n\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \bar{y}_n \mathbb{1}_n\|^2$$

On en déduit

$$R^2 = 1 - \frac{\|Y - \hat{Y}\|^2}{\|Y - \bar{y}_n \mathbb{1}_n\|^2}$$

Avec $\bar{y}_n \mathbb{1}_n$: meilleur prédicteur constant de Y au sens des moindres carrés.

- $R^2 = 1$ si prédiction parfaite ($Y = \hat{Y}$)
- $R^2 = 0$ si le prédicteur constant est une solution des moindres carrés.

Références I

- [Gv96] G. H. Golub and C. F. van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.