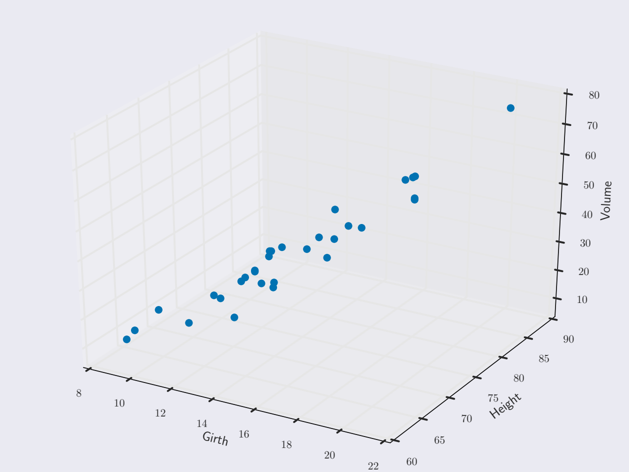
MS BGD
MDI 720 : Statistiques

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

Plan

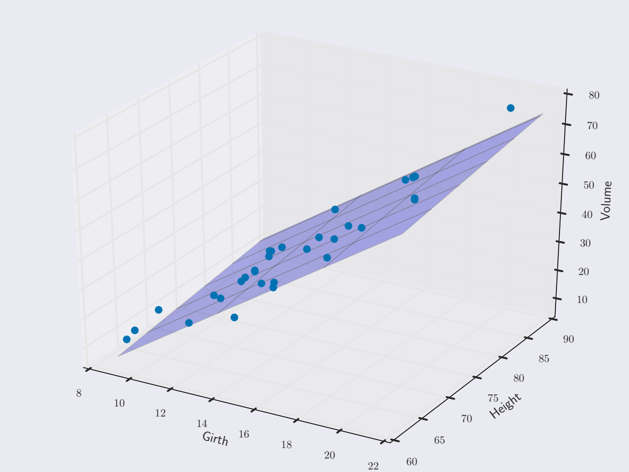
Vers des modèles multi-variés

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 $(\mathbf{x}_1, \ldots, \mathbf{x}_p)$

Modèle en dimension p

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

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

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

<u>Rem</u>: on fait l'hypothèse (point de vue fréquentiste) qu'il existe un vrai paramètre $\boldsymbol{\theta}^\star = (\theta_0^\star, \ldots, \theta_p^\star)^\top$

Dimension p

Modèle matriciel

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

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

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

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

Rem: parfois x_0 sera omis par simplicité

Vocabulaire

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

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

<u>Point de vue</u> "observations" : $y_i = \langle x_i, \boldsymbol{\theta}^\star \rangle + \varepsilon_i$ pour $i = 1, \dots, n$

<u>Point de vue</u> "variables explicatives" : $\mathbf{y} = \sum_{j=0}^{p} \theta_j^\star \mathbf{x}_j + \boldsymbol{\varepsilon}$

Sommaire

Estimateur des moindres carrés (ordinaires)

Un estimateur des moindres carrés est solution du problème :

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

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

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

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

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

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

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

Calcul du gradient de f

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

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

Pour notre fonction f, cela 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 :

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

Pour notre fonction f, cela 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 :

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

Pour notre fonction f, cela 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 :

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

Pour notre fonction f, cela 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 :

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

Pour notre fonction f, cela donne

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

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 :

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

Pour notre fonction f, cela donne

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

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

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 $b = X^\top \mathbf{y}$

<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 carrés

---

Vocabulaire (et abus de langage)

Définition

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

$$X^\top X$$

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

Rem: $X^\top X$ 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 $(n, \ldots, n)$

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 : cela se produit quand
$\mathrm{Ker}(X) = \{\boldsymbol{\theta} \in \mathbb{R}^{p+1} : X\boldsymbol{\theta} = 0\} \neq \{0\}$ (noyau non trivial).
Prenons $\boldsymbol{\theta}_K \in \mathrm{Ker}(X)$ non nul, alors

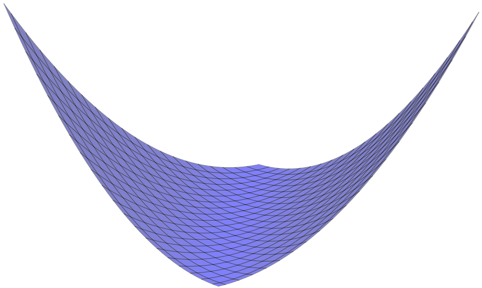
$$X(\hat{\boldsymbol{\theta}} + \boldsymbol{\theta}_K) = X\hat{\boldsymbol{\theta}}$$
$$\text{puis} \quad (X^\top X)(\hat{\boldsymbol{\theta}} + \boldsymbol{\theta}_K) = X^\top \mathbf{y}$$

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 $\mathbb{R}^d$

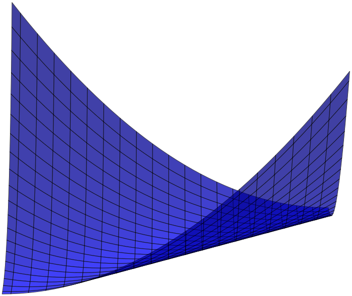
Cas d'une fonction convexe, *e.g.*, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



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

Optimisation dans $\mathbb{R}^d$

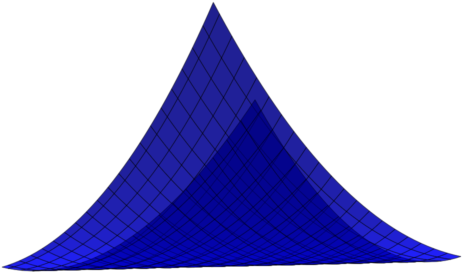
Cas d'une fonction convexe, *e.g.*, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



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

Optimisation dans $\mathbb{R}^d$

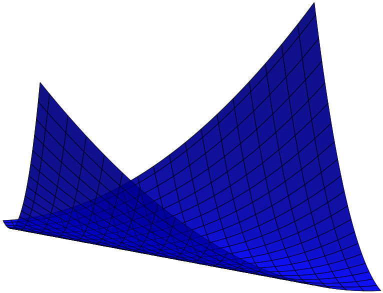
Cas d'une fonction convexe, *e.g.*, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



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

Optimisation dans $\mathbb{R}^d$

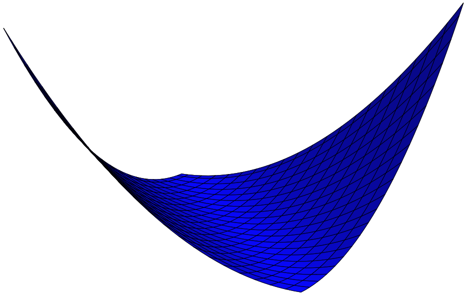
Cas d'une fonction convexe, *e.g.*, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, dont l'ensemble des minimiseurs n'est pas unique :



<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 :



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 $X = (\mathbf{1}_n, 0)$ 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.*, $X = [\mathbf{1}_n \quad x_{i_0} \cdot \mathbf{1}_n]$

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

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),

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

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

Quelques rappels d'algèbre

Définition

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

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

Théorème du rang

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

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

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

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

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 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 la moyenne quand $X = \mathbf{1}_n$: $\hat{\boldsymbol{\theta}} = \frac{\langle \mathbf{1}_n, \mathbf{y} \rangle}{\langle \mathbf{1}_n, \mathbf{1}_n \rangle} = \bar{y}_n$

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

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

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

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

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

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

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

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

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 : prédicteurs et résidus $(p = 2)$