

# MAP 433 : Introduction aux méthodes statistiques. Cours 5

7 mars 2014

MAP 433 : Introduction aux méthodes statistiques. Cours 5

Méthode d'estimation dans le modèle de régression

Sélection de variables

Régression non-linéaire

Bilan provisoire : modèles paramétriques dominés

# Aujourd'hui

- 1 Méthode d'estimation dans le modèle de régression
  - Modèle de régression, notion de « design »
  - Régression à design déterministe
  - La droite des moindres carrés
  - Régression linéaire multiple
  - Le cas gaussien
  - Modèle linéaire gaussien
- 2 Sélection de variables
  - Backward Stepwise Regression
  - LASSO
- 3 Régression non-linéaire
- 4 Bilan provisoire : modèles paramétriques dominés

MAP 433 : Introduction aux méthodes statistiques. Cours 5

Méthode d'estimation dans le modèle de régression

Sélection de variables

Régression non-linéaire

Bilan provisoire : modèles paramétriques dominés

# Influence d'une variable sur une autre

- Principe : on part de l'observation d'un  $n$ -échantillon

$$Y_1, \dots, Y_n \quad (Y_i \in \mathbb{R})$$

- A chaque observation  $Y_i$  est associée une observation **auxiliaire**  $\mathbf{X}_i \in \mathbb{R}^k$ .
- On **suspecte** l'échantillon

$$\mathbf{X}_1, \dots, \mathbf{X}_n \quad (\mathbf{X}_i \in \mathbb{R}^k)$$

de contenir la « majeure partie de la variabilité des  $Y_i$  ».

MAP 433 : Introduction aux méthodes statistiques. Cours 5

Méthode d'estimation dans le modèle de régression

Modèle de régression, notion de « design »

Régression à design déterministe

La droite des moindres carrés

Régression linéaire multiple

Le cas gaussien

Modèle linéaire gaussien

Sélection de variables

Régression non-linéaire

Bilan provisoire :

# Modélisation de l'influence

- Si  $\mathbf{X}_i$  contient **toute la variabilité** de  $Y_i$ , alors  $Y_i$  est mesurable par rapport à  $\mathbf{X}_i$  : il existe  $r : \mathbb{R}^k \rightarrow \mathbb{R}$  telle que

$$Y_i = r(\mathbf{X}_i),$$

mais peu réaliste (ou alors **problème d'interpolation numérique**).

- Alternative : représentation précédente avec **erreur additive** : on **postule**

$$Y_i = r(\mathbf{X}_i) + \xi_i,$$

$\xi_i$  erreur aléatoire centrée (pour des raisons d'identifiabilité).

MAP 433 : Introduction aux méthodes statistiques. Cours 5

Méthode d'estimation dans le modèle de régression

Modèle de régression, notion de « design »

Régression à design déterministe

La droite des moindres carrés

Régression linéaire multiple

Le cas gaussien

Modèle linéaire gaussien

Sélection de variables

Régression non-linéaire

Bilan provisoire :

## Motivation : meilleure approximation $L^2$

- Meilleure approximation  $L^2$ . Si  $\mathbb{E}[Y^2] < +\infty$ , la meilleure approximation de  $Y$  par une variable aléatoire  $\mathbf{X}$ -mesurable est donnée par l'espérance conditionnelle  $\mathbb{E}[Y|\mathbf{X}]$  :

$$\mathbb{E}[(Y - r(\mathbf{X}))^2] = \min_h \mathbb{E}[(Y - h(\mathbf{X}))^2]$$

- où

$$r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}], \quad \mathbf{x} \in \mathbb{R}^k.$$

- On appelle  $r(\cdot)$  **fonction de régression de  $Y$  sur  $\mathbf{X}$** .

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
<< design >>

Régression à  
design  
déterministe.

La droite des  
moindres carrés  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :

## Régression

- On définit :

$$\xi = Y - \mathbb{E}[Y|\mathbf{X}] \implies \mathbb{E}[\xi] = 0.$$

- On a alors naturellement la représentation désirée

$$Y = r(\mathbf{X}) + \xi, \quad \mathbb{E}[\xi] = 0$$

si l'on pose

$$r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}], \quad \mathbf{x} \in \mathbb{R}^k$$

- On observe alors un  $n$ -échantillon

$$(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$$

où

$$Y_i = r(\mathbf{X}_i) + \xi_i, \quad \mathbb{E}[\xi_i] = 0$$

avec comme **paramètre la fonction  $r(\cdot)$**  + un **jeu d'hypothèses** sur la loi des  $\xi_i$ .

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
<< design >>

Régression à  
design  
déterministe.

La droite des  
moindres carrés  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :

## Modèle de régression à design aléatoire

### Définition

Modèle de régression à design aléatoire = donnée de l'observation

$$(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$$

avec  $(Y_i, \mathbf{X}_i) \in \mathbb{R} \times \mathbb{R}^k$  *i.i.d.*, et

$$Y_i = r(\vartheta, \mathbf{X}_i) + \xi_i, \quad \mathbb{E}[\xi_i|\mathbf{X}_i] = 0, \quad \vartheta \in \Theta \subset \mathbb{R}^d.$$

- $\mathbf{x} \rightsquigarrow r(\vartheta, \mathbf{x})$  fonction de **régression**, connue au paramètre  $\vartheta$  près.
- $\mathbf{X}_i$  = variables explicatives, co-variables, prédicteurs ;  $(\mathbf{X}_1, \dots, \mathbf{X}_n)$  = design.

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
<< design >>

Régression à  
design  
déterministe.

La droite des  
moindres carrés  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :

## Modèle alternatif : signal+bruit

- Principe : **sur un exemple**. On observe

$$Y_i = r(\vartheta, i/n) + \xi_i, \quad i = 1, \dots, n$$

où  $r(\vartheta, \cdot) : [0, 1] \rightarrow \mathbb{R}$  est une fonction connue au paramètre  $\vartheta \in \Theta \subset \mathbb{R}^d$  près, et les  $\xi_i$  sont *i.i.d.*,  $\mathbb{E}[\xi_i] = 0$ .

- **But** : reconstruire  $r(\vartheta, \cdot)$  c'est-à-dire **estimer  $\vartheta$** .
- Plus généralement, on observe

$$Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i, \quad i = 1, \dots, n$$

où  $\mathbf{x}_1, \dots, \mathbf{x}_n$  sont des points de  $\mathbb{R}^k$  **déterministes**.

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
<< design >>

Régression à  
design  
déterministe.

La droite des  
moindres carrés  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :

## Modèle de régression à design déterministe

### Définition

Modèle de régression à design déterministe = donnée de l'observation

$$(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)$$

avec  $Y_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^k$ , et

$$Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i, \quad \mathbb{E}[\xi_i] = 0, \quad \vartheta \in \Theta \subset \mathbb{R}^d.$$

- $\mathbf{x}_i$  déterministes, donnés (ou choisis) : plan d'expérience, points du « design ».
- Hypothèses sur les  $\xi_i$  : à débattre. *Pour simplifier*, les  $\xi_i$  sont i.i.d. (*hypothèse restrictive*).
- $\implies$  les  $Y_i$  ne sont pas i.i.d.

Question : Comment estimer  $\theta$  dans ce modèle ?

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
« design »

Régression à  
design  
déterministe

La droite des  
moindres carrés  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire

## Régression gaussienne

- Modèle de régression à design déterministe :

$$Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i, \quad \vartheta \in \Theta \subset \mathbb{R}^d.$$

- Supposons :  $\xi_i \sim \mathcal{N}(0, \sigma^2)$ , i.i.d.
- On a alors le modèle de **régression gaussienne**. Comment estimer  $\vartheta$ ? **On sait expliciter la loi de l'observation**  $Z = (Y_1, \dots, Y_n) \implies$  appliquer le principe du maximum de vraisemblance.
- La loi de  $Y_i$  :

$$\mathbb{P}^{Y_i}(dy) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - r(\vartheta, \mathbf{x}_i))^2\right) dy \ll dy.$$

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
« design »

Régression à  
design  
déterministe

La droite des  
moindres carrés  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire

## EMV pour régression gaussienne

- Le modèle  $\{\mathbb{P}_\vartheta^n = \text{loi de } (Y_1, \dots, Y_n), \vartheta \in \mathbb{R}^k\}$  est dominé par  $\mu^n(dy_1 \dots dy_n) = dy_1 \dots dy_n$ .
- D'où

$$\begin{aligned} \frac{d\mathbb{P}_\vartheta^n}{d\mu^n}(y_1, \dots, y_n) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - r(\vartheta, \mathbf{x}_i))^2\right) \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - r(\vartheta, \mathbf{x}_i))^2\right). \end{aligned}$$

- La fonction de vraisemblance

$$\mathcal{L}_n(\vartheta, Y_1, \dots, Y_n) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - r(\vartheta, \mathbf{x}_i))^2\right)$$

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
« design »

Régression à  
design  
déterministe

La droite des  
moindres carrés  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire

## Estimateur des moindres carrés

Maximiser la **vraisemblance** en régression gaussienne = minimiser la somme des carrés :

$$\sum_{i=1}^n (Y_i - r(\vartheta, \mathbf{x}_i))^2 \rightarrow \min_{\vartheta \in \Theta}.$$

### Définition

Estimateur des **moindres carrés** : tout estimateur  $\hat{\vartheta}_n^{\text{mc}}$  t.q.  
 $\hat{\vartheta}_n^{\text{mc}} \in \arg \min_{\vartheta \in \Theta} \sum_{i=1}^n (Y_i - r(\vartheta, \mathbf{x}_i))^2$ .

- L'EMC est un M-estimateur. Pour le modèle de régression gaussienne :  $\boxed{\text{EMV} = \text{EMC}}$ .
- **Existence, unicité.**
- Propriétés remarquables si la régression est linéaire :  $r(\vartheta, \mathbf{x}_i) = \vartheta^T \mathbf{x}_i$ .

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
« design »

Régression à  
design  
déterministe

La droite des  
moindres carrés  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire

## Droite de régression

- Modèle le plus simple  $r(\vartheta, x) = a + bx$

$$Y_i = a + bx_i + \xi_i, \quad i = 1, \dots, n$$

avec  $\vartheta = (a, b)^T \in \Theta = \mathbb{R}^2$  et les  $(x_1, \dots, x_n)$  donnés.

- L'estimateur des moindres carrés :

$$\hat{\vartheta}_n^{mc} = (\hat{a}, \hat{b}) = \arg \min_{(a,b) \in \mathbb{R}^2} \sum_{i=1}^n (Y_i - a - bx_i)^2.$$

- **Solution explicite** existe toujours, sauf cas pathologique quand tous les  $x_i$  sont les mêmes (Poly, page 112).

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

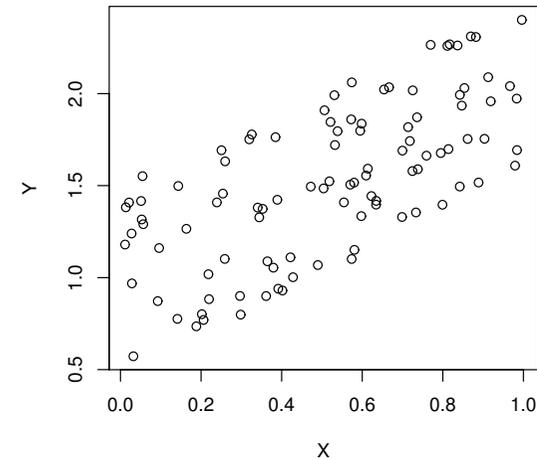
Modèle de  
régression,  
notion de  
<< design >>  
Régression à  
design  
déterministe.  
**La droite des  
moindres carrés**  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :

## Régression linéaire simple



MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

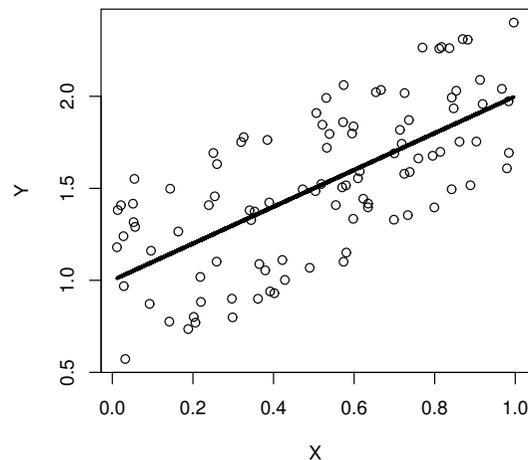
Modèle de  
régression,  
notion de  
<< design >>  
Régression à  
design  
déterministe.  
**La droite des  
moindres carrés**  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :

## Régression linéaire simple



MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
<< design >>  
Régression à  
design  
déterministe.  
**La droite des  
moindres carrés**  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :

## Régression linéaire multiple (=Modèle linéaire)

- La fonction de régression est  $r(\vartheta, \mathbf{x}_i) = \vartheta^T \mathbf{x}_i$ . On observe

$$(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)$$

avec

$$Y_i = \vartheta^T \mathbf{x}_i + \xi_i, \quad i = 1, \dots, n$$

où  $\vartheta \in \Theta = \mathbb{R}^k$ ,  $\mathbf{x}_i \in \mathbb{R}^k$ .

- **Matriciellement**

$$\mathbf{Y} = \mathbb{M}\vartheta + \boldsymbol{\xi}$$

avec  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ ,  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$  et  $\mathbb{M}$  la matrice  $(n \times k)$  dont les **lignes** sont les  $\mathbf{x}_i$ .

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
<< design >>  
Régression à  
design  
déterministe.  
**La droite des  
moindres carrés**  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :

## EMC en régression linéaire multiple

- Estimateur des **moindres carrés** en régression linéaire multiple : tout estimateur  $\hat{\vartheta}_n^{\text{mc}}$  satisfaisant

$$\sum_{i=1}^n (Y_i - (\hat{\vartheta}_n^{\text{mc}})^T \mathbf{x}_i)^2 = \min_{\vartheta \in \mathbb{R}^k} \sum_{i=1}^n (Y_i - \vartheta^T \mathbf{x}_i)^2.$$

- En notations matricielles :

$$\begin{aligned} \|\mathbf{Y} - \mathbb{M} \hat{\vartheta}_n^{\text{mc}}\|^2 &= \min_{\vartheta \in \mathbb{R}^k} \|\mathbf{Y} - \mathbb{M} \vartheta\|^2 \\ &= \min_{v \in V} \|\mathbf{Y} - v\|^2 \end{aligned}$$

où  $V = \text{Im}(\mathbb{M}) = \{v \in \mathbb{R}^n : v = \mathbb{M} \vartheta, \vartheta \in \mathbb{R}^k\}$ .  
Projection orthogonale sur  $V$ .

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
<< design >>  
Régression à  
design  
déterministe.  
La droite des  
moindres carrés  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :

## Géométrie de l'EMC

- L'EMC vérifie

$$\mathbb{M} \hat{\vartheta}_n^{\text{mc}} = P_V \mathbf{Y}$$

où  $P_V$  est le projecteur orthogonal sur  $V$ .

- Mais  $\mathbb{M}^T P_V = \mathbb{M}^T P_V^T = (P_V \mathbb{M})^T = \mathbb{M}^T$ . On en déduit **les équations normales des moindres carrés** :

$$\mathbb{M}^T \mathbb{M} \hat{\vartheta}_n^{\text{mc}} = \mathbb{M}^T \mathbf{Y}.$$

- Remarques.

- L'EMC est un Z-estimateur.
- Pas d'**unicité** de  $\hat{\vartheta}_n^{\text{mc}}$  si la matrice  $\mathbb{M}^T \mathbb{M}$  n'est pas inversible.

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
<< design >>  
Régression à  
design  
déterministe.  
La droite des  
moindres carrés  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :

## Géométrie de l'EMC

### Proposition

Si  $\mathbb{M}^T \mathbb{M}$  (matrice  $k \times k$ ) inversible, alors  $\hat{\vartheta}_n^{\text{mc}}$  est **unique** et

$$\hat{\vartheta}_n^{\text{mc}} = (\mathbb{M}^T \mathbb{M})^{-1} \mathbb{M}^T \mathbf{Y}$$

- Contient le cas précédent de la droite de régression simple.
- Résultat géométrique, **non stochastique**.
- $\mathbb{M}^T \mathbb{M} \geq 0$ ;  $\mathbb{M}^T \mathbb{M}$  inversible  $\iff \mathbb{M}^T \mathbb{M} > 0$ ;

$$\mathbb{M}^T \mathbb{M} > 0 \iff \text{rang}(\mathbb{M}) = k \iff \dim(V) = k.$$

$$\mathbb{M}^T \mathbb{M} > 0 \implies n \geq k.$$

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
<< design >>  
Régression à  
design  
déterministe.  
La droite des  
moindres carrés  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :

## Géométrie de l'EMC

Soit  $\mathbb{M}^T \mathbb{M} > 0$ . Alors, la matrice  $n \times n$

$$A = \mathbb{M} (\mathbb{M}^T \mathbb{M})^{-1} \mathbb{M}^T$$

est dite **matrice chapeau** (hat matrix).

### Proposition

Si  $\mathbb{M}^T \mathbb{M} > 0$ , alors  $A$  est le projecteur sur  $V$  :

$$A = P_V$$

et  $\text{rang}(A) = k$ .

### Preuve :

- $A = A^T$ ,  $A = A^2$ , donc  $A$  est un projecteur.
- $\text{Im}(A) = V$ , donc  $A = P_V$ ;  $\text{rang}(P_V) = \dim(V) = k$ .

<< **Chapeau** >>, car  $A$  génère la prévision de  $\mathbb{M} \vartheta$  notée  $\hat{\mathbf{Y}}$  :

$$\hat{\mathbf{Y}} = \mathbb{M} \hat{\vartheta}_n^{\text{mc}} = A \mathbf{Y}.$$

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Modèle de  
régression,  
notion de  
<< design >>  
Régression à  
design  
déterministe.  
La droite des  
moindres carrés  
Régression  
linéaire multiple  
Le cas gaussien  
Modèle linéaire  
gaussien

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :

## Régression gaussienne

**Régression gaussienne** : on suppose  $\xi \sim \mathcal{N}(0, \sigma^2 \text{Id}_n)$ . Alors on a plusieurs propriétés remarquables :

- Estimateur des moindres carrés  $\hat{\vartheta}_n^{\text{mc}}$  et estimateur du maximum de vraisemblance **coïncident**.  
*Preuve* : écriture de la fonction de vraisemblance.
- On sait expliciter la loi **exacte** (non-asymptotique!) de  $\hat{\vartheta}_n^{\text{mc}}$ .  
*Ingrédient* : **loi des vecteurs gaussiens sont caractérisés par leur moyenne et matrice de variance-covariance.**

MAP 433 :  
Introduction aux méthodes statistiques.  
Cours 5

Méthode d'estimation dans le modèle de régression

Modèle de régression, notion de « design »  
Régression à design déterministe  
La droite des moindres carrés  
Régression linéaire multiple  
Le cas gaussien  
Modèle linéaire gaussien

Sélection de variables

Régression non-linéaire

Bilan provisoire :

## Cadre gaussien : loi des estimateurs

- Hyp. 1 :  $\xi \sim \mathcal{N}(0, \sigma^2 \text{Id}_n)$ .
- Hyp. 2 :  $\mathbb{M}^T \mathbb{M} > 0$ .

### Proposition

- (i)  $\hat{\vartheta}_n^{\text{mc}} \sim \mathcal{N}(\vartheta, \sigma^2 (\mathbb{M}^T \mathbb{M})^{-1})$
- (ii)  $\|\mathbf{Y} - \mathbb{M} \hat{\vartheta}_n^{\text{mc}}\|^2 \sim \sigma^2 \chi^2(n - k)$  **loi du Chi 2 à  $n - k$  degrés de liberté**
- (iii)  $\hat{\vartheta}_n^{\text{mc}}$  et  $\mathbf{Y} - \mathbb{M} \hat{\vartheta}_n^{\text{mc}}$  sont indépendants.

- *Preuve* : **Thm. de Cochran** (Poly, page 18). Si  $\xi \sim \mathcal{N}(0, \text{Id}_n)$  et  $A_j$  matrices  $n \times n$  projecteurs t.q.  $A_j A_i = 0$  pour  $i \neq j$ , alors :  $A_j \xi \sim \mathcal{N}(0, A_j)$ , **indépendants**,  $\|A_j \xi\|^2 \sim \chi^2(\text{Rang}(A_j))$ .

MAP 433 :  
Introduction aux méthodes statistiques.  
Cours 5

Méthode d'estimation dans le modèle de régression

Modèle de régression, notion de « design »  
Régression à design déterministe  
La droite des moindres carrés  
Régression linéaire multiple  
Le cas gaussien  
Modèle linéaire gaussien

Sélection de variables

Régression non-linéaire

Bilan provisoire :

## Preuve de la proposition

- (i)  $\hat{\vartheta}_n^{\text{mc}} = \vartheta + (\mathbb{M}^T \mathbb{M})^{-1} \mathbb{M}^T \xi$ .  
On vérifie :  $\mathbb{E}[\hat{\vartheta}_n^{\text{mc}}] = \vartheta$ ,

$$\begin{aligned} & \mathbb{E} [(\mathbb{M}^T \mathbb{M})^{-1} \mathbb{M}^T \xi ((\mathbb{M}^T \mathbb{M})^{-1} \mathbb{M}^T \xi)^T] \\ &= \sigma^2 (\mathbb{M}^T \mathbb{M})^{-1}. \end{aligned}$$

- (ii)

$$\begin{aligned} \mathbf{Y} - \mathbb{M} \hat{\vartheta}_n^{\text{mc}} &= \mathbb{M} (\vartheta - \hat{\vartheta}_n^{\text{mc}}) + \xi \\ &= -\mathbb{M} (\mathbb{M}^T \mathbb{M})^{-1} \mathbb{M}^T \xi + \xi \\ &= \sigma (\text{Id}_n - A) \xi', \quad \xi' \sim \mathcal{N}(0, \text{Id}_n). \end{aligned}$$

- (iii) le vecteur  $(\hat{\vartheta}_n^{\text{mc}}, \mathbf{Y} - \mathbb{M} \hat{\vartheta}_n^{\text{mc}})$  est gaussien. On calcule explicitement sa matrice de variance-covariance.

MAP 433 :  
Introduction aux méthodes statistiques.  
Cours 5

Méthode d'estimation dans le modèle de régression

Modèle de régression, notion de « design »  
Régression à design déterministe  
La droite des moindres carrés  
Régression linéaire multiple  
Le cas gaussien  
Modèle linéaire gaussien

Sélection de variables

Régression non-linéaire

Bilan provisoire :

## Propriétés de l'EMC : cadre gaussien

Estimateur de la variance  $\sigma^2$  :

$$\hat{\sigma}_n^2 = \frac{\|\mathbf{Y} - \mathbb{M} \hat{\vartheta}_n^{\text{mc}}\|^2}{n - k} = \frac{1}{n - k} \sum_{i=1}^n (Y_i - (\hat{\vartheta}_n^{\text{mc}})^T \mathbf{x}_i)^2$$

D'après la dernière Proposition :

- $\hat{\sigma}_n^2 / \sigma^2 \sim \chi^2(n - k)$  **loi du Chi 2 à  $n - k$  degrés de liberté**
- C'est un estimateur **sans biais** :

$$\mathbb{E}_{\vartheta} [\hat{\sigma}_n^2] = \sigma^2.$$

- $\hat{\sigma}_n^2$  est **indépendant** de  $\hat{\vartheta}_n^{\text{mc}}$ .

MAP 433 :  
Introduction aux méthodes statistiques.  
Cours 5

Méthode d'estimation dans le modèle de régression

Modèle de régression, notion de « design »  
Régression à design déterministe  
La droite des moindres carrés  
Régression linéaire multiple  
Le cas gaussien  
Modèle linéaire gaussien

Sélection de variables

Régression non-linéaire

Bilan provisoire :

## Propriétés de l'EMC : cadre gaussien

- Lois des coordonnées de  $\hat{\vartheta}_n^{mc}$  :

$$(\hat{\vartheta}_n^{mc})_j - \vartheta_j \sim \mathcal{N}(0, \sigma^2 b_j)$$

où  $b_j$  est le  $j$ ème élément diagonal de  $(M^T M)^{-1}$ .

$$\frac{(\hat{\vartheta}_n^{mc})_j - \vartheta_j}{\hat{\sigma}_n \sqrt{b_j}} \sim t_{n-k}$$

loi de Student à  $n - k$  degrés de liberté.

$$t_q = \frac{\xi}{\sqrt{\eta/q}}$$

où  $q \geq 1$  un entier,  $\xi \sim \mathcal{N}(0, 1)$ ,  $\eta \sim \chi^2(q)$  et  $\xi$  **indépendant** de  $\eta$ .

MAP 433 :  
Introduction aux méthodes statistiques.  
Cours 5

Méthode d'estimation dans le modèle de régression

Modèle de régression, notion de « design »  
Régression à design déterministe

La droite des moindres carrés  
Régression linéaire multiple  
Le cas gaussien  
Modèle linéaire gaussien

Sélection de variables

Régression non-linéaire

Bilan provisoire

## Exemple de données de régression

### Données de diabète

Patient	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	Response
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	82	221
442	36	1	19.6	71	250	132.2	97	3	4.6	92	57

$n=442, k=10$

bmi = Body Mass Index

map = Blood Pressure

tc, ldl, tch, ltg, glu = Blood Serum Measurements

Response  $Y$  = a quantitative measure of disease progression 1 year after baseline

MAP 433 :  
Introduction aux méthodes statistiques.  
Cours 5

Méthode d'estimation dans le modèle de régression

Modèle de régression, notion de « design »  
Régression à design déterministe

La droite des moindres carrés  
Régression linéaire multiple  
Le cas gaussien  
Modèle linéaire gaussien

Sélection de variables

Régression non-linéaire

Bilan provisoire

## Résultats de traitement statistique initial

	Estimate	Std. Error	t value	Pr(>  t )
(Intercept)	152.133	2.576	59.061	< 2e - 16 * **
age	-10.012	59.749	-0.168	0.867000
sex	-239.819	61.222	-3.917	0.000104 * **
bmi	519.840	66.534	7.813	4.30e - 14 * **
map	324.390	65.422	4.958	1.02e - 06 * **
tc	-792.184	416.684	-1.901	0.057947
ldl	476.746	339.035	1.406	0.160389
hdl	101.045	212.533	0.475	0.634721
tch	177.064	161.476	1.097	0.273456
ltg	751.279	171.902	4.370	1.56e - 05 * **
glu	67.625	65.984	1.025	0.305998

MAP 433 :  
Introduction aux méthodes statistiques.  
Cours 5

Méthode d'estimation dans le modèle de régression

Modèle de régression, notion de « design »  
Régression à design déterministe

La droite des moindres carrés  
Régression linéaire multiple  
Le cas gaussien  
Modèle linéaire gaussien

Sélection de variables

Régression non-linéaire

Bilan provisoire

## Questions statistiques

- Sélection de variables.** Lesquelles parmi les 10 variables :

age, sex, bmi, map, tc, ldl, hdl, tch, ltg, glu

sont significatives ? Formalisation mathématique : trouver (estimer) l'ensemble  $N = \{j : \vartheta_j \neq 0\}$ .

- Prévison.** Un nouveau patient arrive avec son vecteur des 10 variables  $\mathbf{x}_0 \in \mathbb{R}^{10}$ . Donner la prévison de la réponse  $Y$  = état du patient dans 1 an.

MAP 433 :  
Introduction aux méthodes statistiques.  
Cours 5

Méthode d'estimation dans le modèle de régression

Modèle de régression, notion de « design »  
Régression à design déterministe

La droite des moindres carrés  
Régression linéaire multiple  
Le cas gaussien  
Modèle linéaire gaussien

Sélection de variables

Régression non-linéaire

Bilan provisoire

## Sélection de variables : Backward Stepwise Regression

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables  
Backward  
Stepwise  
Regression  
LASSO

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

- On se donne un critère d'élimination de variables (**plusieurs choix de critère possibles...**).
- On élimine une variable, la moins significative du point de vue du critère choisi.
- On calcule l'EMC  $\hat{\vartheta}_{n,k-1}^{mc}$  dans le nouveau modèle, avec seulement les  $k - 1$  paramètres restants, ainsi que le RSS :
 
$$RSS_{k-1} = \|\mathbf{Y} - \mathbb{M}\hat{\vartheta}_{n,k-1}^{mc}\|^2.$$
- On continue à éliminer des variables, une par une, jusqu'à la **stabilisation de RSS** :  $RSS_m \approx RSS_{m-1}$ .

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

## Données de diabète : Backward Regression

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables  
Backward  
Stepwise  
Regression  
LASSO

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

### Backward Regression : Itération 2.

Critère d'élimination : plus grande valeur de  $\Pr(> |t|)$ .

	Estimate	Std. Error	t value	$\Pr(>  t )$
(Intercept)	152.133	2.573	59.128	$< 2e - 16$
sex	-240.835	60.853	-3.958	0.000104
bmi	519.905	64.156	5.024	$8.85e - 05$
map	322.306	65.422	4.958	$7.43e - 07$
tc	-790.896	416.144	-1.901	0.058
ldl	474.377	338.358	1.402	0.162
<b>hdl</b>	99.718	212.146	0.470	<b>0.639</b>
tch	177.458	161.277	1.100	0.272
ltg	749.506	171.383	4.373	$1.54e - 05$
glu	67.170	65.336	1.013	0.312

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

## RSS (Residual Sum of Squares)

Modèle de régression

$$Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i, \quad i = 1, \dots, n.$$

- **Résidu** : si  $\hat{\vartheta}_n$  est un estimateur de  $\vartheta$ ,

$$\hat{\xi}_i = Y_i - r(\hat{\vartheta}_n, \mathbf{x}_i) \text{ résidu au point } i.$$

- **RSS** : **Residual Sum of Squares**, somme résiduelle des carrés. Caractérise la qualité d'approximation.

$$RSS(= RSS_{\hat{\vartheta}_n}) = \|\hat{\xi}\|^2 = \sum_{i=1}^n (Y_i - r(\hat{\vartheta}_n, \mathbf{x}_i))^2.$$

- En régression **linéaire** :  $RSS = \|\mathbf{Y} - \mathbb{M}\hat{\vartheta}_n\|^2.$

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables  
Backward  
Stepwise  
Regression  
LASSO

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

## Données de diabète : Backward Regression

- **Sélection "naïve"** : {sex, bmi, map, ltg}
- **Sélection par Backward Regression** :  
Critère d'élimination : plus grande valeur de  $\Pr(> |t|)$ .

	Estimate	Std. Error	t value	$\Pr(>  t )$
(Intercept)	152.133	2.576	59.061	$< 2e - 16$ ***
<b>age</b>	-10.012	59.749	-0.168	<b>0.867000</b>
sex	-239.819	61.222	-3.917	0.000104 ***
bmi	519.840	66.534	7.813	$4.30e - 14$ ***
map	324.390	65.422	4.958	$1.02e - 06$ ***
tc	-792.184	416.684	-1.901	0.057947
ldl	476.746	339.035	1.406	0.160389
hdl	101.045	212.533	0.475	0.634721
tch	177.064	161.476	1.097	0.273456
ltg	751.279	171.902	4.370	$1.56e - 05$ ***
glu	67.625	65.984	1.025	0.305998

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables  
Backward  
Stepwise  
Regression  
LASSO

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

## Données de diabète : Backward Regression

### Backward Regression : Itération 5 (dernière).

Variables sélectionnées :  
 $\{\text{sex}, \text{bmi}, \text{map}, \text{tc}, \text{ldl}, \text{ltg}\}$

	Estimate	Std. Error	t value	Pr(>  t )
(Intercept)	152.133	2.572	59.159	< 2e - 16
sex	-226.511	59.857	-3.784	0.000176
bmi	529.873	65.620	8.075	6.69e - 15
map	327.220	62.693	5.219	2.79e - 07
tc	-757.938	160.435	-4.724	3.12e - 06
ldl	538.586	146.738	3.670	0.000272
ltg	804.192	80.173	10.031	< 2e - 16

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables  
Backward  
Stepwise  
Regression  
LASSO

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

◀ ▶ ⏪ ⏩ 🔍 🔄

## Sélection de variables : Backward Regression

### Discussion de Backward Regression :

- Méthode de sélection purement empirique, pas de justification théorique.
- Application d'autres critères d'élimination en Backward Regression peut amener aux résultats différents.  
Exemple. Critère  $C_p$  de Mallows-Akaike : on élimine la variable  $j$  qui réalise

$$\min_j \left( \text{RSS}_{m,(-j)} + 2\hat{\sigma}_n^2 m \right).$$

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables  
Backward  
Stepwise  
Regression  
LASSO

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

◀ ▶ ⏪ ⏩ 🔍 🔄

## Sélection de variables : LASSO

LASSO = Least Absolute Shrinkage and Selection Operator

- **Estimateur LASSO** : tout estimateur  $\hat{\vartheta}_n^L$  vérifiant

$$\hat{\vartheta}_n^L \in \arg \min_{\vartheta \in \mathbb{R}^k} \left( \sum_{i=1}^n (Y_i - \vartheta^T \mathbf{x}_i)^2 + \lambda \sum_{j=1}^k |\vartheta_j| \right) \text{ avec } \lambda > 0.$$

- Si  $\mathbb{M}^T \mathbb{M} > 0$ , l'estimateur LASSO  $\hat{\vartheta}_n^L$  est unique.
- Estimateur des moindres carrés **pénalisé**. Pénalisation par  $\sum_{j=1}^k |\vartheta_j|$ , la norme  $\ell_1$  de  $\vartheta$ .

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables  
Backward  
Stepwise  
Regression  
LASSO

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

◀ ▶ ⏪ ⏩ 🔍 🔄

## Sélection de variables : LASSO

- Deux utilisations de LASSO :
  - **Estimation de  $\vartheta$**  : alternative à  $\hat{\vartheta}_n^{\text{mc}}$  si  $k > n$ .
  - **Sélection de variables** : on ne retient que les variables qui correspondent aux coordonnées non-nulles du vecteur  $\hat{\vartheta}_n^L$ .
- LASSO admet une **justification théorique** : sous certaines hypothèses sur la matrice  $\mathbb{M}$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}\{\hat{N}_n = N\} = 1,$$

où  $N = \{j : \vartheta_j \neq 0\}$  et  $\hat{N}_n = \{j : \hat{\vartheta}_{n,j}^L \neq 0\}$ .

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables  
Backward  
Stepwise  
Regression  
LASSO

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

◀ ▶ ⏪ ⏩ 🔍 🔄

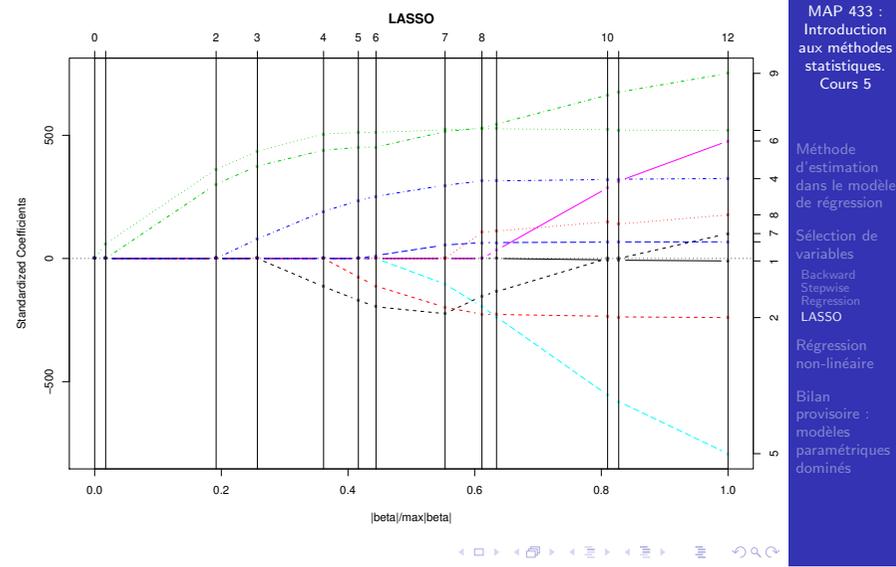
# Données de diabète : LASSO

Application aux données de diabète.

- L'ensemble de variables sélectionné par LASSO :
 
$$\{\text{sex, bmi, map, tc, hdl, ltg, glu}\}$$
- Backward Regression :
 
$$\{\text{sex, bmi, map, tc, ldl, ltg}\}$$
- Sélection naïve :
 
$$\{\text{sex, bmi, map, tc}\}$$

MAP 433 : Introduction aux méthodes statistiques. Cours 5  
Méthode d'estimation dans le modèle de régression  
Sélection de variables  
Backward Stepwise Regression LASSO  
Régression non-linéaire  
Bilan provisoire : modèles paramétriques dominés

# Application de LASSO : "regularization path"



MAP 433 : Introduction aux méthodes statistiques. Cours 5  
Méthode d'estimation dans le modèle de régression  
Sélection de variables  
Backward Stepwise Regression LASSO  
Régression non-linéaire  
Bilan provisoire : modèles paramétriques dominés

# Prévision

Modèle de régression

$$Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i, \quad i = 1, \dots, n.$$

Régression **linéaire** :  $r(\vartheta, \mathbf{x}_i) = \vartheta^T \mathbf{x}_i$ . Exemple :  $\mathbf{x}_i$  vecteur de 10 variables explicatives (age, sex, bmi, ...) pour patient  $i$ .

- **Problème de prévision** : Un nouveau patient arrive avec son vecteur des 10 variables  $\mathbf{x}_0 \in \mathbb{R}^{10}$ . Donner la prévision de la valeur de fonction de régression  $r(\vartheta, \mathbf{x}_0) = \vartheta^T \mathbf{x}_0$  (=état du patient dans 1 an).
- Soit  $\hat{\vartheta}_n$  un estimateur de  $\vartheta$ . **Prévision par substitution** :
 
$$\hat{Y} = r(\hat{\vartheta}_n, \mathbf{x}_0).$$
- **Question statistique** : quelle est la qualité de la prévision ? **Intervalle de confiance** pour  $r(\vartheta, \mathbf{x}_0)$  basé sur  $\hat{Y}$  ?

MAP 433 : Introduction aux méthodes statistiques. Cours 5  
Méthode d'estimation dans le modèle de régression  
Sélection de variables  
Backward Stepwise Regression LASSO  
Régression non-linéaire  
Bilan provisoire : modèles paramétriques dominés

# Prévision : modèle linéaire gaussienne

- Traitement sur l'exemple :  $r(\vartheta, \mathbf{x}) = \vartheta^T \mathbf{x}$ , régression **linéaire gaussienne** et  $\hat{\vartheta}_n = \hat{\vartheta}_n^{\text{mc}}$ .  $\implies \hat{Y} = \mathbf{x}_0^T \hat{\vartheta}_n^{\text{mc}}$
- **Hyp. 1** :  $\xi \sim \mathcal{N}(0, \sigma^2 \text{Id}_n)$ .
- **Hyp. 2** :  $\mathbb{M}^T \mathbb{M} > 0$ .

## Proposition

- (i)  $\hat{Y} \sim \mathcal{N}(\mathbf{x}_0^T \vartheta, \sigma^2 \mathbf{x}_0^T (\mathbb{M}^T \mathbb{M})^{-1} \mathbf{x}_0)$
- (ii)  $\hat{Y} - \mathbf{x}_0^T \vartheta$  et  $\mathbf{Y} - \mathbb{M} \hat{\vartheta}_n^{\text{mc}}$  sont indépendants.

Rappel :  $\|\mathbf{Y} - \mathbb{M} \hat{\vartheta}_n^{\text{mc}}\|^2 \sim \sigma^2 \chi^2(n - k)$  loi du Chi 2 à  $n - k$  degrés de liberté.

MAP 433 : Introduction aux méthodes statistiques. Cours 5  
Méthode d'estimation dans le modèle de régression  
Sélection de variables  
Backward Stepwise Regression LASSO  
Régression non-linéaire  
Bilan provisoire : modèles paramétriques dominés

## Prévision : modèle linéaire gaussienne

- D'après la Proposition,

$$\eta := \frac{\hat{Y} - \mathbf{x}_0^T \vartheta}{\sqrt{\sigma^2 \mathbf{x}_0^T (\mathbb{M}^T \mathbb{M})^{-1} \mathbf{x}_0}} \sim \mathcal{N}(0, 1).$$

- On remplace  $\sigma^2$  inconnu par  $\hat{\sigma}_n^2 = \|\mathbf{Y} - \mathbb{M} \hat{\vartheta}_n^{\text{mc}}\|^2 / (n - k)$ .
- **t-statistique** :

$$t := \frac{\hat{Y} - \mathbf{x}_0^T \vartheta}{\sqrt{\hat{\sigma}_n^2 \mathbf{x}_0^T (\mathbb{M}^T \mathbb{M})^{-1} \mathbf{x}_0}} = \frac{\eta}{\sqrt{\chi / (n - k)}} \sim t_{n-k},$$

loi de Student à  $n - k$  degrés de liberté, car  $\eta \sim \mathcal{N}(0, 1)$ ,  
 $\chi := \|\mathbf{Y} - \mathbb{M} \hat{\vartheta}_n^{\text{mc}}\|^2 / \sigma^2 \sim \chi^2(n - k)$  et  $\eta \perp \chi$ .

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables

Backward  
Stepwise  
Regression  
LASSO

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

## Prévision : intervalle de confiance

$$\begin{aligned} \mathbb{P}\left(-q_{1-\frac{\alpha}{2}}(t_{n-k}) \leq \frac{\hat{Y} - \mathbf{x}_0^T \vartheta}{\sqrt{\hat{\sigma}_n^2 \mathbf{x}_0^T (\mathbb{M}^T \mathbb{M})^{-1} \mathbf{x}_0}} \leq q_{1-\frac{\alpha}{2}}(t_{n-k})\right) \\ = \mathbb{P}(-q_{1-\frac{\alpha}{2}}(t_{n-k}) \leq t \leq q_{1-\frac{\alpha}{2}}(t_{n-k})) = 1 - \alpha. \end{aligned}$$

⇒ **intervalle de confiance** de niveau  $1 - \alpha$  pour  
 $r(\vartheta, \mathbf{x}_0) = \mathbf{x}_0^T \vartheta$  est  $[r_L, r_U]$ , où :

$$\begin{aligned} r_L &= \hat{Y} - q_{1-\frac{\alpha}{2}}(t_{n-k}) \sqrt{\hat{\sigma}_n^2 \mathbf{x}_0^T (\mathbb{M}^T \mathbb{M})^{-1} \mathbf{x}_0}, \\ r_U &= \hat{Y} + q_{1-\frac{\alpha}{2}}(t_{n-k}) \sqrt{\hat{\sigma}_n^2 \mathbf{x}_0^T (\mathbb{M}^T \mathbb{M})^{-1} \mathbf{x}_0}. \end{aligned}$$

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables

Backward  
Stepwise  
Regression  
LASSO

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

## Limites des moindres carrés et du cadre gaussien

- Calcul **explicite** (et efficace) de l'EMC limité à une fonction de régression **linéaire**.
- Modèle linéaire donne un cadre assez général :
  - Modèle polynomial,
  - **Modèles avec interactions...**
- **Hypothèse de gaussianité** = cadre asymptotique implicite.
- Besoin d'outils pour les modèles à réponse **Y discrète**.

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables

Backward  
Stepwise  
Regression  
LASSO

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

## Régression linéaire non-gaussienne

Modèle de régression linéaire

$$Y_i = \vartheta^T \mathbf{x}_i + \xi_i, \quad i = 1, \dots, n.$$

- **Hyp. 1'** :  $\xi_i$  i.i.d.,  $\mathbb{E}[\xi_i] = 0$ ,  $\mathbb{E}[\xi_i^2] = \sigma^2 > 0$ .
- **Hyp. 2'** :  $\mathbb{M}^T \mathbb{M} > 0$ ,  $\lim_n \max_{1 \leq i \leq n} \mathbf{x}_i^T (\mathbb{M}^T \mathbb{M})^{-1} \mathbf{x}_i = 0$ .

**Proposition (Normalité asymptotique de l'EMC)**

$$\sigma^{-1} (\mathbb{M}^T \mathbb{M})^{1/2} (\hat{\vartheta}_n^{\text{mc}} - \vartheta) \xrightarrow{d} \mathcal{N}(0, \text{Id}_k), \quad n \rightarrow \infty.$$

- A comparer avec le cadre gaussien :

$$\sigma^{-1} (\mathbb{M}^T \mathbb{M})^{1/2} (\hat{\vartheta}_n^{\text{mc}} - \vartheta) \sim \mathcal{N}(0, \text{Id}_k) \text{ pour tout } n.$$

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables

Backward  
Stepwise  
Regression  
LASSO

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

## Régression non-linéaire

- On observe

$$(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n),$$

où

$$Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i, \quad i = 1, \dots, n$$

avec

$$\mathbf{x}_i \in \mathbb{R}^k, \quad \text{et } \vartheta \in \Theta \subset \mathbb{R}^d.$$

- Si  $\xi_i \sim \text{i.i.d. } \mathcal{N}(0, \sigma^2)$ ,

$$\mathcal{L}_n(\vartheta, Y_1, \dots, Y_n) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - r(\vartheta, \mathbf{x}_i))^2\right)$$

et l'estimateur du **maximum de vraisemblance** est obtenu en minimisant la fonction

$$\vartheta \rightsquigarrow \sum_{i=1}^n (Y_i - r(\vartheta, \mathbf{x}_i))^2.$$

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

## Moindre carrés non-linéaires

### Définition

- $M$ -estimateur associé à la **fonction de contraste**  
 $\psi : \Theta \times \mathbb{R}^k \times \mathbb{R} \rightarrow \mathbb{R}$  : tout estimateur  $\hat{\vartheta}_n$  satisfaisant

$$\sum_{i=1}^n \psi(\hat{\vartheta}_n, \mathbf{x}_i, Y_i) = \max_{a \in \Theta} \sum_{i=1}^n \psi(a, \mathbf{x}_i, Y_i).$$

- Estimateur des **moindres carrés non-linéaires** : associé au contraste  $\psi(a, \mathbf{x}, y) = -(y - r(a, \mathbf{x}))^2$ .

- Extension** des résultats en densité  $\rightarrow$  théorèmes limites pour des sommes de v.a. indépendantes **non-équidistribuées**.

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

## Modèle à réponse binaire

- On observe

$$(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n), \quad Y_i \in \{0, 1\}, \quad \mathbf{x}_i \in \mathbb{R}^k.$$

- Modélisation **via la fonction de régression**

$$\mathbf{x} \rightsquigarrow p_{\mathbf{x}}(\vartheta) = \mathbb{E}_{\vartheta} [Y | \mathbf{X} = \mathbf{x}] = \mathbb{P}_{\vartheta} [Y = 1 | \mathbf{X} = \mathbf{x}]$$

- Représentation**

$$\begin{aligned} Y_i &= p_{\mathbf{x}_i}(\vartheta) + (Y_i - p_{\mathbf{x}_i}(\vartheta)) \\ &= r(\vartheta, \mathbf{x}_i) + \xi_i \end{aligned}$$

avec  $r(\vartheta, \mathbf{x}_i) = p_{\mathbf{x}_i}(\vartheta)$  et  $\xi_i = Y_i - p_{\mathbf{x}_i}(\vartheta)$ .

- $\mathbb{E}_{\vartheta} [\xi_i] = 0$  mais structure des  $\xi_i$  **compliquée** (dépendance en  $\vartheta$ ).

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

## Modèle à réponse discrète

- $Y_i$  v.a. de Bernoulli de paramètre  $p_{\mathbf{x}_i}(\vartheta)$ .  
**Vraisemblance**

$$\mathcal{L}_n(\vartheta, Y_1, \dots, Y_n) = \prod_{i=1}^n p_{\mathbf{x}_i}(\vartheta)^{Y_i} (1 - p_{\mathbf{x}_i}(\vartheta))^{1 - Y_i}$$

$\rightarrow$  méthodes de résolution numérique.

- Régression logistique** (très utile dans les applications)

$$p_{\mathbf{x}}(\vartheta) = \psi(\mathbf{x}^T \vartheta),$$

$$\psi(t) = \frac{e^t}{1 + e^t}, \quad t \in \mathbb{R} \quad \text{fonction logistique.}$$

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

## Régression logistique et modèles latents

- **Représentation équivalente de la régression logistique** : on observe

$$Y_i = 1_{\{Y_i^* > 0\}}, \quad i = 1, \dots, n$$

(les  $\mathbf{x}_i$  sont donnés), et  $Y_i^*$  est une **variable latente** ou cachée,

$$Y_i^* = \vartheta^T \mathbf{x}_i + U_i, \quad i = 1, \dots, n$$

avec  $U_i \sim \text{i.i.d. } F$ , où

$$F(t) = \frac{1}{1 + e^{-t}}, \quad t \in \mathbb{R}.$$

- $$\begin{aligned} \mathbb{P}_\vartheta [Y_i^* > 0] &= \mathbb{P}_\vartheta [\mathbf{x}_i^T \vartheta + U_i > 0] \\ &= 1 - \mathbb{P}_\vartheta [U_i \leq -\mathbf{x}_i^T \vartheta] \\ &= 1 - (1 + \exp(-\mathbf{x}_i^T \vartheta))^{-1} = \psi(\mathbf{x}_i^T \vartheta). \end{aligned}$$

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

## Bilan provisoire : modèles paramétriques dominés

- **Modèle de densité** : on observe

$$X_1, \dots, X_n \sim \text{i.i.d.} \mathbb{P}_\vartheta, \quad \vartheta \in \Theta \subset \mathbb{R}^d.$$

**Estimateurs** : moments, Z- et M-estimateurs, **EMV**.

- **Modèle de régression** : on observe

$$Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i, \quad i = 1, \dots, n, \quad \xi_i \text{ i.i.d.}, \quad \vartheta \in \Theta \subset \mathbb{R}^d.$$

**Estimateurs** :

- Si  $r(\vartheta, \mathbf{x}) = \vartheta^T \mathbf{x}$ , EMC (coincide avec l'**EMV** si les  $\xi_i$  gaussiens)
- Sinon, M-estimateurs, **EMV**...
- Autres méthodes selon des **hypothèses** sur le « design »...

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

## Bilan provisoire (cont.) : précision d'estimation

$\hat{\vartheta}_n$  estimateur de  $\vartheta$  : **précision, qualité** de  $\hat{\vartheta}_n$  ? Approche par **région-intervalle de confiance**

- Pour  $\alpha \in (0, 1)$ , on construit  $\mathcal{C}_{n,\alpha}(\hat{\vartheta}_n)$  **ne dépendant pas de  $\vartheta$**  (observable) tel que

$$\mathbb{P}_\vartheta [\vartheta \in \mathcal{C}_{n,\alpha}(\hat{\vartheta}_n)] \geq 1 - \alpha$$

asymptotiquement lorsque  $n \rightarrow \infty$ , uniformément en  $\vartheta$ ...

La **précision** de l'estimateur est le **diamètre** (moyen) de  $\mathcal{C}_{n,\alpha}(\hat{\vartheta}_n)$ .

- Par exemple :  $\mathcal{C}_{n,\alpha}(\hat{\vartheta}_n) =$  boule de centre  $\hat{\vartheta}_n$  et de rayon à **déterminer**.

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés

En pratique, une information **non-asymptotique** de type

$$\mathbb{E} [\|\hat{\vartheta}_n - \vartheta\|^2] \leq c_n(\vartheta)^2,$$

ou bien **asymptotique** de type

$$v_n(\hat{\vartheta}_n - \vartheta) \xrightarrow{d} Z_\vartheta, \quad n \rightarrow \infty$$

(avec  $v_n \rightarrow \infty$ ) permet « souvent » de construire un(e) **région-intervalle de confiance**.

MAP 433 :  
Introduction  
aux méthodes  
statistiques.  
Cours 5

Méthode  
d'estimation  
dans le modèle  
de régression

Sélection de  
variables

Régression  
non-linéaire

Bilan  
provisoire :  
modèles  
paramétriques  
dominés