

# MAP 433 : Introduction aux méthodes statistiques. Cours 5

25 Septembre 2015

# Aujourd'hui

## 1 Méthode d'estimation dans le modèle de régression

- Modèle de régression
- Régression à design déterministe
- La droite des moindres carrés
- Régression linéaire multiple
- Propriétés de l'estimateur des Moindres Carrés
- Modèle linéaire gaussien

## 2 Sélection de variables

## 3 Régression non-linéaire

# 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  $Y_i$  ».

# 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).

# 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  $\mathbf{X}$** .

# 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

# régresseurs aléatoires

## Definition

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

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

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

# Modèle de régression à design déterministe

## Definition

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

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

- $\mathbf{x}_i$  déterministes, donnés (ou choisis) : plan d'expérience, points du « design ».
- Hypothèses sur les  $\xi_i$  : à débattre. *Pour simplifier*, les variables  $\xi_i$  sont centrées,  $\mathbb{E}_\theta[\xi_i] = 0$ , décorrélées,  $\mathbb{E}_\theta[\xi_i \xi_j] = 0$  si  $i \neq j$  et de variance unité  $\mathbb{E}[\xi_i^2] = 1$  (*homoscédasticité*).
- **Attention !** Les  $Y_i$  ne sont *pas identiquement distribuées*.



# Régression gaussienne

- Modèle de régression à design déterministe :

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

- Supposons :  $\xi_i \sim \mathcal{N}(0, 1)$ , i.i.d.
- On a alors le modèle de **régression gaussienne**. Comment estimer  $\theta$  ? **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  $Y_i$  :

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

# EMV pour régression gaussienne

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

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

- La fonction de vraisemblance

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

# Estimateur des moindres carrés

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

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

## Definition

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

- L'EMC est un M-estimateur. Pour le modèle de régression gaussienne : EMV = EMC.
- Existence, unicité.

# Droite de régression

- Modèle le plus simple  $r(\beta, x) = \beta_0 + \beta_1 x$

$$Y_i = \beta_0 + \beta_1 x_i + \xi_i, \quad i = 1, \dots, n$$

avec  $\beta = (\beta_0, \beta_1)^T \in \mathbb{R}^2$  et les  $(x_1, \dots, x_n)$  données.

- L'estimateur des moindres carrés :

$$\hat{\beta}_n = (\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(b_0, b_1) \in \mathbb{R}^2} \sum_{i=1}^n (Y_i - b_0 - b_1 x_i)^2.$$

- Solution explicite

# Droite de régression

- Le minimum est caractérisé par les équations

$$\begin{cases} b_0 + b_1 n^{-1} \sum_{i=1}^n x_i &= n^{-1} \sum_{i=1}^n Y_i \\ b_0 n^{-1} \sum_{i=1}^n x_i + b_1 n^{-1} \sum_{i=1}^n x_i^2 &= n^{-1} \sum_{i=1}^n x_i Y_i. \end{cases}$$

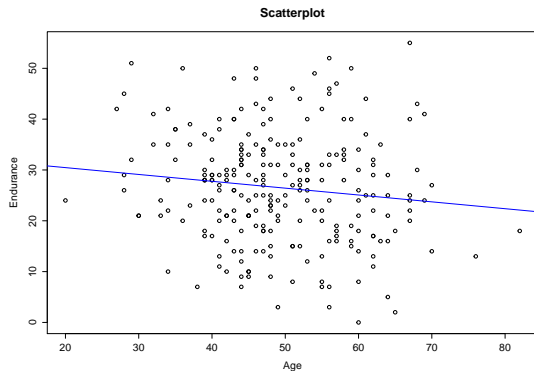
- Notons  $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$ . Si le déterminant  $\Delta_n \neq 0$  où

$$\Delta_n = \begin{vmatrix} 1 & n^{-1} \sum_{i=1}^n x_i \\ n^{-1} \sum_{i=1}^n x_i & n^{-1} \sum_{i=1}^n x_i^2 \end{vmatrix} = S_{xx} = n^{-1} \sum_{i=1}^n (x_i^2 - \bar{x}_n)^2, \quad ,$$

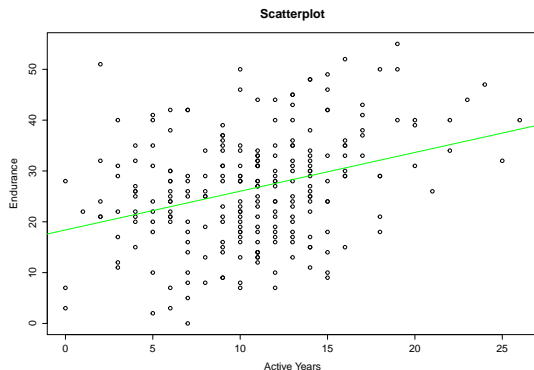
alors ce système d'équations a une solution unique :

$$\begin{cases} \hat{\beta}_{n0} &= \bar{Y}_n - \hat{\beta}_{n1} \bar{x}_n \\ \hat{\beta}_{n1} &= \frac{S_{xy}}{S_{xx}}, \quad S_{xy} = n^{-1} \sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n). \end{cases}$$

# Régression linéaire simple



# Régression linéaire simple



# Régression linéaire multiple (=Modèle linéaire)

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

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

avec

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

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

- Matriciellement

$$\mathbf{Y} = \mathbb{X}\beta + \sigma\xi$$

avec

- $\mathbf{Y} = (Y_1 \cdots Y_n)^T$ ,
- $\xi = (\xi_1 \cdots \xi_n)^T$
- $\mathbb{X}$  la matrice  $(n \times k)$  dont la  $i$ -ème ligne est  $\mathbb{X}_{i,\cdot} = \mathbf{x}_i^T$ .



# Régression linéaire multiple (=Modèle linéaire)

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

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

avec

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

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

- Matriciellement

$$\mathbf{Y} = \mathbb{X}\beta + \sigma \boldsymbol{\xi}$$

avec

- $\mathbf{Y} = (Y_1 \cdots Y_n)^T$ ,
- $\boldsymbol{\xi} = (\xi_1 \cdots \xi_n)^T$
- $\mathbb{X}$  la matrice  $(n \times k)$  dont la  $i$ -ème ligne est  $\mathbb{X}_{i,\cdot} = \mathbf{x}_i^T$ .

# Régression linéaire multiple (=Modèle linéaire)

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

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

avec

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

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

- Matriciellement

$$\mathbf{Y} = \mathbb{X}\beta + \sigma\xi$$

avec

- $\mathbf{Y} = (Y_1 \cdots Y_n)^T$ ,
- $\xi = (\xi_1 \cdots \xi_n)^T$
- $\mathbb{X}$  la matrice  $(n \times k)$  dont la  $i$ -ème ligne est  $\mathbb{X}_{i,\cdot} = \mathbf{x}_i^T$ .

# EMC en régression linéaire multiple

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

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

- En notation matricielle :

$$\begin{aligned} \|\mathbf{Y} - \mathbb{X} \hat{\beta}_n\|^2 &= \min_{\mathbf{b} \in \mathbb{R}^k} \|\mathbf{Y} - \mathbb{X} \mathbf{b}\|^2 \\ &= \min_{\mathbf{v} \in V} \|\mathbf{Y} - \mathbf{v}\|^2 \end{aligned}$$

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

# Géométrie de l'EMC

- L'EMC vérifie

$$\mathbb{X} \hat{\beta}_n = P_V \mathbf{Y}$$

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

- Comme  $\mathbf{Y} - P_V \mathbf{Y} \perp V$ , on en déduit les équations normales des moindres carrés :

$$\mathbb{X}^T \mathbb{X} \hat{\beta}_n = \mathbb{X}^T \mathbf{Y}.$$

- Remarques.

- L'EMC est un  $Z$ -estimateur.
- **unicité** de  $\hat{\beta}_n$  si la matrice de Gram  $\mathbb{X}^T \mathbb{X}$  est inversible (la matrice  $\mathbb{X}$  est de rang complet).

# Géométrie de l'EMC

## Proposition

Si  $\mathbb{X}^T \mathbb{X}$  (matrice  $k \times k$ ) inversible, alors  $\hat{\beta}_n$  est *unique* et

$$\boxed{\hat{\beta}_n = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y}} = \mathbb{X}^\# \mathbf{Y}$$

Contient le cas précédent de la droite de régression simple.

# Géométrie de l'EMC

## Proposition

Si  $\mathbb{X}^T \mathbb{X}$  (matrice  $k \times k$ ) inversible, alors  $\hat{\beta}_n$  est *unique* et

$$\boxed{\hat{\beta}_n = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y}} = \mathbb{X}^\# \mathbf{Y}$$

Résultat géométrique, *non stochastique*.  $\mathbb{X}^T \mathbb{X} \geq 0$ ;  $\mathbb{X}^T \mathbb{X}$  inversible  $\iff \mathbb{X}^T \mathbb{X} > 0$ ;

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

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

# Géométrie de l'EMC

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

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

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

## Proposition

Si  $\mathbb{X}^T \mathbb{X} > 0$ , alors  $A$  est le projecteur sur  $V$  :  $A = P_V$  et  $\text{rang}(A) = k$ .

## Démonstration.

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

# Géométrie de l'EMC

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

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

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

## Proposition

*Si  $\mathbb{X}^T \mathbb{X} > 0$ , alors  $A$  est le projecteur sur  $V$  :  $A = P_V$  et  $\text{rang}(A) = k$ .*

**Chapeau**, car  $A$  génère la prévision de  $\mathbb{X}\theta$  notée  $\hat{\mathbf{Y}}$  :

$$\hat{\mathbf{Y}} = \mathbb{X} \hat{\boldsymbol{\beta}}_n = A \mathbf{Y}.$$



# Pseudo-inverse de Moore-Penrose

Soit  $\mathbb{X}$  une matrice  $n \times p$  avec  $p \leq n$ . On suppose que  $\mathbb{X}$  est de rang  $p$ .

- $\mathbb{X}^\# = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T$  est la **pseudo-inverse** de Moore-Penrose.
- $\mathbb{X}^\# \mathbb{X} = Id_{p \times p}$  :  $\mathbb{X}$  est un inverse à gauche de la matrice  $\mathbb{X}$ .
- $\mathbb{X} \mathbb{X}^\# = \mathbb{X} (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T$  est le projecteur sur l'espace image (l'espace vectoriel engendré par les colonnes de  $\mathbb{X}$ ).
- Méthode de calcul : décomposition QR ou décomposition en valeurs singulières.

# Hypothèses

$$Y = X\beta + \sigma\xi$$

- 1  $X$  est de rang complet.
- 2  $\mathbb{E}_\theta[\xi] = 0$  pour tout  $\theta \in \Theta$  (les erreurs sont centrées)
- 3 La variance des erreurs est constante et les erreurs sont décorrélées  $\mathbb{E}_\theta[\xi\xi^T] = I$  (homoscédasticité)

# Hypothèses

$$Y = X\beta + \sigma\xi$$

- 1  $X$  est de rang complet.
- 2  $\mathbb{E}_\theta[\xi] = 0$  pour tout  $\theta \in \Theta$  (les erreurs sont centrées)
- 3 La variance des erreurs est constante et les erreurs sont décorrélées  $\mathbb{E}_\theta[\xi\xi^T] = I$  (homoscédasticité)

# Hypothèses

$$Y = X\beta + \sigma\xi$$

- 1  $X$  est de rang complet.
- 2  $\mathbb{E}_\theta[\xi] = 0$  pour tout  $\theta \in \Theta$  (les erreurs sont centrées)
- 3 La variance des erreurs est constante et les erreurs sont décorrélées  $\mathbb{E}_\theta[\xi\xi^T] = I$  (homoscédasticité)

# Estimateur sans biais

## Théorème

*L'estimateur  $\hat{\beta}_n$  est sans biais, i.e. pour tout  $\theta \in \Theta$ ,*

- $\mathbb{E}_\theta[\hat{\beta}_n] = \theta$
- $\text{Cov}_\theta(\hat{\beta}_n) = \sigma^2(\mathbb{X}^T \mathbb{X})^{-1}$ .

$$\hat{\beta}_n = \mathbb{X}^\# \mathbf{Y} = \beta + \mathbb{X}^\# \xi.$$

$$\begin{aligned}\mathbb{E}_\theta[\hat{\beta}_n] &= \beta + \mathbb{X}^\# \mathbb{E}_\theta[\xi] \\ &= \beta\end{aligned}$$

# Estimateur sans biais

## Théorème

*L'estimateur  $\hat{\beta}_n$  est sans biais, i.e. pour tout  $\theta \in \Theta$ ,*

- $\mathbb{E}_\theta[\hat{\beta}_n] = \theta$
- $\text{Cov}_\theta(\hat{\beta}_n) = \sigma^2(\mathbb{X}^T \mathbb{X})^{-1}$ .

$\hat{\beta}_n = \mathbb{X}^\# \mathbf{Y} = \beta + \mathbb{X}^\# \boldsymbol{\xi}$ .  $\hat{\beta}_n - \beta = \mathbb{X}^\# \boldsymbol{\xi}$  ce qui implique

$$\begin{aligned}\text{Cov}_\theta(\hat{\beta}_n) &= \mathbb{E}_\theta [\{\mathbb{X}^\# \boldsymbol{\xi}\} \{\mathbb{X}^\# \boldsymbol{\xi}\}^T] \\ &= \sigma^2 \{\mathbb{X}^\#\} \{\mathbb{X}^\#\}^T = \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1}.\end{aligned}$$

# Erreur de prédiction

- Erreur de prédiction :

$$\begin{aligned}\hat{\xi} &= \mathbf{Y} - \mathbb{X} \hat{\beta}_n = \mathbf{Y} - \mathbb{X}(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y} \\ &= (I - A) \mathbf{Y}\end{aligned}$$

- Sous  $\mathbb{P}_\theta$ ,  $\mathbf{Y} = \mathbb{X}\beta + \sigma\xi$ . Donc,

$$\begin{aligned}\hat{\xi} &= (I - A)\mathbb{X}\beta + \sigma(I - A)\xi \\ &= \sigma(I - A)\xi\end{aligned}$$

car  $A\mathbb{X} = \mathbb{X}$  ( $A$  is the orthogonal projector on the image of  $\mathbb{X}$ ).

# Résidus et variance résiduelle

## Theorem

Pour tout  $\theta \in \Theta$

- 1  $\mathbb{E}_\theta[\hat{\xi}] = 0.$
- 2  $\text{Cov}_\theta(\hat{\xi}) = \sigma^2(I - A).$
- 3  $\mathbb{E}_\theta[\hat{Y}] = X\beta.$
- 4  $\text{Cov}_\theta(\hat{\xi}, \hat{Y}) = 0.$

## Démonstration.

$$\begin{aligned}\mathbb{E}_\theta[\hat{\xi}] &= \sigma \mathbb{E}_\theta[(I - A)\xi] \\ &= \sigma(I - A) \mathbb{E}_\theta[\xi].\end{aligned}$$



# Résidus et variance résiduelle

## Theorem

Pour tout  $\theta \in \Theta$

- 1  $\mathbb{E}_\theta[\hat{\xi}] = 0.$
- 2  $\text{Cov}_\theta(\hat{\xi}) = \sigma^2(I - A).$
- 3  $\mathbb{E}_\theta[\hat{Y}] = X\beta.$
- 4  $\text{Cov}_\theta(\hat{\xi}, \hat{Y}) = 0.$

## Démonstration.

$$\begin{aligned}\text{Cov}_\theta(\hat{\xi}) &= \sigma^2(I - A) \mathbb{E}_\theta[\xi \xi'] (I - A) \\ &= \sigma^2(I - A).\end{aligned}$$

# Résidus et variance résiduelle

## Theorem

Pour tout  $\theta \in \Theta$

- 1  $\mathbb{E}_\theta[\hat{\xi}] = 0.$
- 2  $\text{Cov}_\theta(\hat{\xi}) = \sigma^2(I - A).$
- 3  $\mathbb{E}_\theta[\hat{\mathbf{Y}}] = \mathbb{X}\beta.$
- 4  $\text{Cov}_\theta(\hat{\xi}, \hat{\mathbf{Y}}) = 0.$

## Démonstration.

$$\begin{aligned}\mathbb{E}_\theta[\hat{\mathbf{Y}}] &= \mathbb{E}_\theta[A(\mathbb{X}\beta + \sigma\xi)] \\ &= A\mathbb{X}\beta + \sigma \mathbb{E}_\theta[\xi] \\ &= \mathbb{X}\beta.\end{aligned}$$

# Résidus et variance résiduelle

## Theorem

Pour tout  $\theta \in \Theta$

- 1  $\mathbb{E}_\theta[\hat{\xi}] = 0.$
- 2  $\text{Cov}_\theta(\hat{\xi}) = \sigma^2(I - A).$
- 3  $\mathbb{E}_\theta[\hat{Y}] = X\beta.$
- 4  $\text{Cov}_\theta(\hat{\xi}, \hat{Y}) = 0.$

## Démonstration.

On a  $\hat{Y} - \mathbb{E}_\theta[\hat{Y}] = \sigma A\xi$  et donc

$$\begin{aligned}\text{Cov}_\theta(\hat{\xi}, \hat{Y}) &= \sigma^2 \mathbb{E}_\theta[(I - A)\xi\xi' A] \\ &= \sigma^2(I - A)A = 0.\end{aligned}$$

# Estimateur sans biais de la variance de l'erreur de prédiction

## Théorème

$\hat{\sigma}_n^2 = (n - p)^{-1} \|\hat{\xi}\|^2$  est un estimateur sans biais de la variance de l'erreur.

## Démonstration.

Comme  $(I - A)^2 = (I - A)$ , nous avons

$$\begin{aligned}\mathbb{E}_\theta[\hat{\sigma}_n^2] &= (n - p)^{-1} \mathbb{E}_\theta[\xi^T (I - A) \xi] \\ &= (n - p)^{-1} \mathbb{E}_\theta[\text{Tr}((I - A) \xi \xi)] \\ &= \sigma^2 (n - p)^{-1} \text{Tr}(I - A) = \sigma^2.\end{aligned}$$



# Coefficient de détermination

## ■ Pythagore

$$\begin{aligned}\| \mathbf{Y} \|^2 &= \| A \mathbf{Y} \|^2 + \| (I - A) \mathbf{Y} \|^2 \\ &= \| \hat{\mathbf{Y}} \|^2 + \| \hat{\boldsymbol{\xi}} \|^2\end{aligned}$$

## ■ Coefficient de détermination

$$R^2 = \frac{\| \mathbf{Y} \|^2}{\| \mathbf{Y} \|^2} = 1 - \frac{\| \hat{\boldsymbol{\xi}} \|^2}{\| \mathbf{Y} \|^2} = 1 - \frac{\text{SCR}}{\text{SCT}}$$

où **SCR** est la somme des carrés résiduels (RSS : residual sum of squares) et **SCT** est la somme des carrés totaux.

# Diagnostic de régression

```
> summary(model1)

Call:
lm(formula = endur$endurance ~ endur$age)

Residuals:
    Min       1Q   Median       3Q      Max
-25.0734  -7.6331   0.0974   6.7710  30.8696

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.15667    3.42033   9.694  <2e-16 ***
endur$age   -0.13472    0.06812  -1.978   0.0491 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.76 on 243 degrees of freedom
Multiple R-squared:  0.01584, Adjusted R-squared:  0.01179
F-statistic: 3.911 on 1 and 243 DF,  p-value: 0.04911
```

FIGURE – Régression à un facteur : endurance / âge

# Diagnostic de régression

```
> summary(model2)

Call:
lm(formula = endur$endurance ~ endur$activeyears)

Residuals:
    Min       1Q   Median       3Q      Max
-23.7296  -7.0671   0.5579   5.7454  31.0829

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    18.3921     1.5998  11.496 < 2e-16 ***
endur$activeyears  0.7625     0.1369   5.571 6.7e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.21 on 243 degrees of freedom
Multiple R-squared:  0.1133, Adjusted R-squared:  0.1096
F-statistic: 31.04 on 1 and 243 DF,  p-value: 6.697e-08
```

FIGURE – Régression à un facteur : endurance / nombre d'années de pratique

# Diagnostic de régression

```
> summary(model3)

Call:
lm(formula = endur$endurance ~ endur$age + endur$activeyears)

Residuals:
    Min       1Q   Median       3Q      Max
-21.7994  -6.9040   0.5701   5.6326  27.2279

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    29.3952     3.2054   9.171 < 2e-16 ***
endur$age      -0.2571     0.0655  -3.925 0.000113 ***
endur$activeyears  0.9163     0.1386   6.610 2.44e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.919 on 242 degrees of freedom
Multiple R-squared:  0.1663, Adjusted R-squared:  0.1594
F-statistic: 24.14 on 2 and 242 DF,  p-value: 2.754e-10
```

FIGURE – Régression à un deux facteurs : endurance / âge + nombre d'années de pratique



# Régression gaussienne

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

- On sait expliciter la loi **exacte** (non-asymptotique !) de  $(\hat{\beta}_n, \hat{\sigma}^2)$ .
- **Ingrédient** :
  - loi des vecteurs gaussiens sont caractérisés par leur moyenne et matrice de variance-covariance.
  - pour des vecteurs gaussiens, la décorrélation implique l'indépendance.

# Cadre gaussien : loi des estimateurs

## Proposition

- 1  $\hat{\beta}_n \sim \mathcal{N}(\beta, \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1})$
- 2  $\|(I - A) \mathbf{Y}\|^2 \sim \sigma^2 \chi^2(n - k)$  *loi du Chi 2 à  $n - k$  degrés de liberté*
- 3  $\hat{\beta}_n$  et  $(I - A) \mathbf{Y}$  sont indépendants.

# Théorème de Cochran

## Théorème

Soit  $\mathbf{Y} \sim N(\mu, \sigma^2 I_n)$ ,  $\mathcal{M}$  un sous espace de  $\mathbb{R}^n$  de dimension  $k$ ,  $\Pi$  la matrice de projection orthogonale sur  $\mathcal{M}$  et  $\Pi_{\perp} = I_n - \Pi$  la matrice de projection orthogonale sur  $\mathcal{M}^{\perp}$ . Nous avons

- 1  $\Pi \mathbf{Y} \sim \mathcal{N}(\Pi \mu, \sigma^2 \Pi)$ ,  $\Pi_{\perp} \mathbf{Y} \sim \mathcal{N}(\Pi_{\perp} \mu, \sigma^2 \Pi_{\perp})$
- 2 les vecteurs  $\Pi \mathbf{Y}$  et  $\Pi_{\perp} \mathbf{Y}$  sont indépendants
- 3  $\|\Pi(\mathbf{Y} - \mu)\|^2 / \sigma^2 \sim \chi_k^2$  et  $\|\Pi_{\perp}(\mathbf{Y} - \mu)\|^2 / \sigma^2 \sim \chi_{n-k}^2$ .

# Théorème de Cochran

## Théorème

Soit  $\mathbf{Y} \sim N(\mu, \sigma^2 I_n)$ ,  $\mathcal{M}$  un sous espace de  $\mathbb{R}^n$  de dimension  $k$ ,  $\Pi$  la matrice de projection orthogonale sur  $\mathcal{M}$  et  $\Pi_{\perp} = I_n - \Pi$  la matrice de projection orthogonale sur  $\mathcal{M}^{\perp}$ . Nous avons

- 1  $\Pi \mathbf{Y} \sim \mathcal{N}(\Pi \mu, \sigma^2 \Pi)$ ,  $\Pi_{\perp} \mathbf{Y} \sim \mathcal{N}(\Pi_{\perp} \mu, \sigma^2 \Pi_{\perp})$
- 2 les vecteurs  $\Pi \mathbf{Y}$  et  $\Pi_{\perp} \mathbf{Y}$  sont indépendants
- 3  $\|\Pi(\mathbf{Y} - \mu)\|^2 / \sigma^2 \sim \chi_k^2$  et  $\|\Pi_{\perp}(\mathbf{Y} - \mu)\|^2 / \sigma^2 \sim \chi_{n-k}^2$ .

# Théorème de Cochran

## Théorème

Soit  $\mathbf{Y} \sim N(\mu, \sigma^2 I_n)$ ,  $\mathcal{M}$  un sous espace de  $\mathbb{R}^n$  de dimension  $k$ ,  $\Pi$  la matrice de projection orthogonale sur  $\mathcal{M}$  et  $\Pi_{\perp} = I_n - \Pi$  la matrice de projection orthogonale sur  $\mathcal{M}^{\perp}$ . Nous avons

- 1  $\Pi \mathbf{Y} \sim \mathcal{N}(\Pi \mu, \sigma^2 \Pi)$ ,  $\Pi_{\perp} \mathbf{Y} \sim \mathcal{N}(\Pi_{\perp} \mu, \sigma^2 \Pi_{\perp})$
- 2 les vecteurs  $\Pi \mathbf{Y}$  et  $\Pi_{\perp} \mathbf{Y}$  sont indépendants
- 3  $\|\Pi(\mathbf{Y} - \mu)\|^2 / \sigma^2 \sim \chi_k^2$  et  $\|\Pi_{\perp}(\mathbf{Y} - \mu)\|^2 / \sigma^2 \sim \chi_{n-k}^2$ .

# Cadre gaussien : loi des estimateurs

## Proposition

- 1  $\hat{\beta}_n \sim \mathcal{N}(\beta, \sigma^2(\mathbb{X}^T \mathbb{X})^{-1})$
- 2  $\|(I - A) \mathbf{Y}\|^2 \sim \sigma^2 \chi^2(n - k)$  loi du Chi 2 à  $n - k$  degrés de liberté
- 3  $\hat{\beta}_n$  et  $(I - A) \mathbf{Y}$  sont indépendants.

- Définition :  $\hat{\beta}_n = \mathbb{X}^\# \mathbf{Y}$  et  $\mathbf{Y} \sim N(\mathbb{X}\beta, \sigma^2 I_n)$ .
- $\hat{\beta}_n \sim N(\mathbb{X}^\# \mathbb{X} \beta, \sigma^2 \{\mathbb{X}^\#\} \{\mathbb{X}^\#\}^T)$
- On conclut en remarquant  $\mathbb{X}^\# \mathbb{X} = I_k$  et  $\{\mathbb{X}^\#\} \{\mathbb{X}^\#\}^T = (\mathbb{X}^T \mathbb{X})^{-1}$

# Cadre gaussien : loi des estimateurs

## Proposition

- 1  $\hat{\beta}_n \sim \mathcal{N}(\beta, \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1})$
- 2  $\|(I - A) \mathbf{Y}\|^2 \sim \sigma^2 \chi^2(n - k)$  loi du Chi 2 à  $n - k$  degrés de liberté
- 3  $\hat{\beta}_n$  et  $(I - A) \mathbf{Y}$  sont indépendants.

Application directe de Cochran en remarquant que

$$(I - A) \mathbb{E}_\theta[\mathbf{Y}] = (I - A) \mathbf{X} \beta = 0.$$

# Cadre gaussien : loi des estimateurs

## Proposition

- 1  $\hat{\beta}_n \sim \mathcal{N}(\beta, \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1})$
- 2  $\|(I - A) \mathbf{Y}\|^2 \sim \sigma^2 \chi^2(n - k)$  loi du Chi 2 à  $n - k$  degrés de liberté
- 3  $\hat{\beta}_n$  et  $(I - A) \mathbf{Y}$  sont indépendants.

- Le théorème de Cochran montre que  $A \mathbf{Y}$  et  $(I - A) \mathbf{Y}$  sont indépendants.
- On conclut en remarquant que  $\hat{\beta}_n = \mathbb{X}^\# \mathbf{Y} = \mathbb{X}^\# A \mathbf{Y}$ .



# Estimateur de la variance $\sigma^2$ : cadre gaussien

$$\hat{\sigma}_n^2 = \frac{\|(I - A) \mathbf{Y}\|^2}{n - k} = \frac{1}{n - k} \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_n)^2$$

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 :

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

- $\hat{\sigma}_n^2$  est indépendant de  $\hat{\boldsymbol{\beta}}_n$ .

# Lois des coordonnées de $\hat{\beta}_n$ : cadre gaussien

$$\hat{\beta}_{nj} - \beta_j \sim \mathcal{N}(0, \sigma^2 b_j)$$

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

$$\frac{\hat{\beta}_{nj} - \beta_j}{\hat{\sigma}_n \sqrt{b_j}} \sim t_{n-k}$$

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  $\eta$ .

# Exemple de données de régression

## 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

# 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  $N = \{j : \theta_j \neq 0\}$ .

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

# RSS (Residual Sum of Squares)

Modèle de régression

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

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

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

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

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

- En régression **linéaire** :  $\boxed{\text{RSS} = \|\mathbf{Y} - \mathbb{X}\hat{\theta}_n\|^2.}$

# Prévision

Modèle de régression

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

Régression **linéaire** :  $r(\boldsymbol{\beta}, \mathbf{x}_i) = \boldsymbol{\beta}^T \mathbf{x}_i$ . 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  $\mathbf{x} \in \mathbb{R}^k$ . Donner la prévision de la valeur de fonction de régression  $r(\boldsymbol{\beta}, \mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x}_0$  (=état du patient dans 1 an).

- Soit  $\hat{\boldsymbol{\beta}}_n$  un estimateur de  $\boldsymbol{\beta}$ . **Prévision par substitution** :

$$\hat{Y}(\mathbf{x}) = r(\hat{\boldsymbol{\beta}}_n, \mathbf{x}).$$

- Question statistique : quelle est la qualité de la prévision ?  
**Intervalle de confiance** pour  $r(\hat{\boldsymbol{\beta}}_n, \mathbf{x})$  ?

# Moyenne et variance de la prévision

## Theorem

- $\mathbb{E}_\theta[\hat{Y}_n(\mathbf{x})] = \mathbf{x}^T \boldsymbol{\beta}$
- $\text{Var}_\theta(\hat{Y}_n(\mathbf{x})) = \sigma^2 \mathbf{x}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbf{x}$
- $\mathbb{E}_\theta[(Y(\mathbf{x}) - \hat{Y}_n(\mathbf{x}))^2] = \sigma^2(1 + \mathbf{x}'(\mathbb{X}^T \mathbb{X})^{-1} \mathbf{x})$

$$\hat{Y}_n(\mathbf{x}) = \mathbf{x}^T \hat{\boldsymbol{\beta}}_n \text{ et } \mathbb{E}_\theta[\hat{\boldsymbol{\beta}}_n] = \boldsymbol{\beta}$$



# Moyenne et variance de la prévision

## Theorem

- $\mathbb{E}_\theta[\hat{Y}_n(\mathbf{x})] = \mathbf{x}^T \boldsymbol{\beta}$
- $\text{Var}_\theta(\hat{Y}_n(\mathbf{x})) = \sigma^2 \mathbf{x}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbf{x}$
- $\mathbb{E}_\theta[(Y(\mathbf{x}) - \hat{Y}_n(\mathbf{x}))^2] = \sigma^2(1 + \mathbf{x}'(\mathbb{X}^T \mathbb{X})^{-1} \mathbf{x})$

$$\begin{aligned}\hat{Y}_n(\mathbf{x}) - \mathbf{x}^T \boldsymbol{\beta} &= \mathbf{x}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y} - \mathbf{x}^T \boldsymbol{\beta} \\ &= \mathbf{x}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T (\mathbb{X} \boldsymbol{\beta} + \sigma \boldsymbol{\xi}) - \mathbf{x}^T \boldsymbol{\beta} \\ &= \sigma \mathbf{x}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \boldsymbol{\xi}\end{aligned}$$

# Moyenne et variance de la prévision

## Theorem

- $\mathbb{E}_\theta[\hat{Y}_n(\mathbf{x})] = \mathbf{x}^T \boldsymbol{\beta}$
- $\text{Var}_\theta(\hat{Y}_n(\mathbf{x})) = \sigma^2 \mathbf{x}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbf{x}$
- $\mathbb{E}_\theta[(Y(\mathbf{x}) - \hat{Y}_n(\mathbf{x}))^2] = \sigma^2(1 + \mathbf{x}'(\mathbb{X}^T \mathbb{X})^{-1} \mathbf{x})$

$$\begin{aligned}\mathbb{E}_\theta[(Y(\mathbf{x}) - \hat{Y}_n(\mathbf{x}))^2] &= \mathbb{E}_\beta[(Y(\mathbf{x}) - \mathbb{E}_\beta[\hat{Y}_n(\mathbf{x})])^2] + \text{Var}_\theta(\hat{Y}_n(\mathbf{x})) \\ &= \mathbb{E}_\beta[(Y(\mathbf{x}) - \mathbf{x}^T \boldsymbol{\beta})^2] + \sigma^2 \mathbf{x}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbf{x}\end{aligned}$$

## Prévision : modèle linéaire gaussienne

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

### Proposition

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

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

# Prévision : modèle linéaire gaussienne

- D'après la Proposition,

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

- On replace  $\sigma^2$  inconnu par  $\hat{\sigma}_n^2 = \|(I - A) \mathbf{Y}\|^2 / (n - k)$ .
- $t$ -statistique :

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

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

## Prévision : intervalle de confiance

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

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

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

# 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**.

# Régression linéaire non-gaussienne

Modèle de régression linéaire

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

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

Proposition (Normalité asymptotique de l'EMC)

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

- A comparer avec le cadre gaussien :

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

# Régression non-linéaire

- On observe

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

où

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

avec

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

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

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

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

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



# Moindre carrés non-linéaires

## Definition

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

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

- *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**.

# 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

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

- Représentation

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

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

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

## Modèle à réponse discrète

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

Vraisemblance

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

→ méthodes de résolution numérique.

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

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

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

# 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,

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

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

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

■

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