Modèles linéaires

- la variable réponse : Y
- les variables explicatives : $X^{(1)}, \dots, X^{(p)}$

question : existe t'il une fonction f telle que :

$$Y \approx f(X^{(1)}, \dots, X^{(n)})$$

Pour déterminer f, on utilise le risque et un problème de minimisation

On cherche f tq

$$\mathbb{E}\left[\left(Y - f(X^{(1)}, \dots, X^{(T)})\right)^2\right] = \min_{g} \mathbb{E}\left[\left(Y - g(X^{(1)}, \dots, X^{(T)})\right)^2\right]$$

Plutôt que de faire la minimisation sur l'ensemble des f et g possibles, on se limite aux fonctions linéaires

$$\mathcal{F} = \left\{ g : \mathbb{R}^p \longrightarrow \mathbb{R} \text{ tq } g(x_1, \dots, x_p) = \beta_0 + \sum_{k=1}^{T} \beta_k x_k \right\}$$

pourquoi cette famille.

Car dans le cadre où $Y, X^{(1)}, \dots, X^{(p)}$ sont des variables quantitatives, en réalité, on connaît de façon théorique la meilleure fonction f il s'agit de

$$f(X^{(1)}, \dots, X^{(p)}) = \mathbb{E}\left[Y \mid X^{(1)}, \dots, X^{(p)} \right]$$

et si l'on est dans le cadre gaussien, à savoir que le modèle mathématique est $Y = f(X^{(1)}, \dots, X^{(p)}) + \varepsilon$ avec ε une variable gaussienne, alors l'espérance conditionnelle est une f^{ct} linéaire.

$\rightarrow$ cette espérance conditionnelle est bien souvent incalculable.

On remplace le risque par le risque empirique car à la base, on dispose d'observations.

échantillon d'apprentissage : $\mathcal{L} = \left\{ \left(x_i^{(1)}, \dots, x_i^{(p)}, y_i \right), i \in [\![1, n]\!] \right\}$

- y_i une observation de la variable aléatoire Y_i.
- $x_i^{(k)}$ une observation de la variable $X_i^{(k)}$ que nous supposerons non aléatoire

le modèle s'écrit : $\forall i \in [\![1, n]\!], \; Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_i^{(k)} + \varepsilon_i$ la variable de bruit

hypothèses classiques :

- $\forall i \in [\![1, n]\!], \quad \mathbb{E}[\varepsilon_i] = 0$
- $\forall i \in [\![1, n]\!], \quad \mathbb{V}[\varepsilon_i] = \sigma^2$ (homocédasticité)
- $\forall i \neq j, \quad \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$

bien souvent, on ajoute l'hypothèse de normalité du bruit :

$$U = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \sim \mathcal{N}(0_n, \sigma^2 I_n)$$

U : vecteur aléatoire

$\mathcal{N}(0_n, \sigma^2 I_n)$: vecteur gaussien

$\sigma^2 I_n$: matrice de variance

def vecteur gaussien

Soit $\overline{T} = \begin{pmatrix} \overline{T}_1 \\ \vdots \\ \overline{T}_n \end{pmatrix}$ un vecteur aléatoire ($\overline{T}_1, \overline{T}_n$ sont des variables aléatoires)

$\overline{T}$ est un vecteur gaussien si et seulement si :

$$\forall (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n, \quad \sum_{k=1}^{n} \lambda_k \overline{T}_k \text{ est une variable gaussienne}$$

Rq : Automatiquement $\forall i \in [\![1, n]\!]$, $\overline{T}_i$ est une variable gaussienne

def : matrice de variance d'un vecteur aléatoire $\overline{T}$:

$$V[\overline{T}] = \mathbb{E}\left[\left(\overline{T} - \mathbb{E}[\overline{T}] \right) \cdot \left(\overline{T} - \mathbb{E}[\overline{T}] \right)' \right]$$: matrice $n \times n$ positive et symétrique

def: espérance d'un vecteur aléatoire T

$$\mathbb{E}[T] := \begin{pmatrix} \mathbb{E}(T_1) \\ \vdots \\ \mathbb{E}[T_n] \end{pmatrix}$$

Rq: $$\mathbb{V}[T] := \left(\mathrm{cov}(T_i, T_j) \right)_{i,j \in [1,n]}$$

la théorie des modèles linéaires est plus vaste

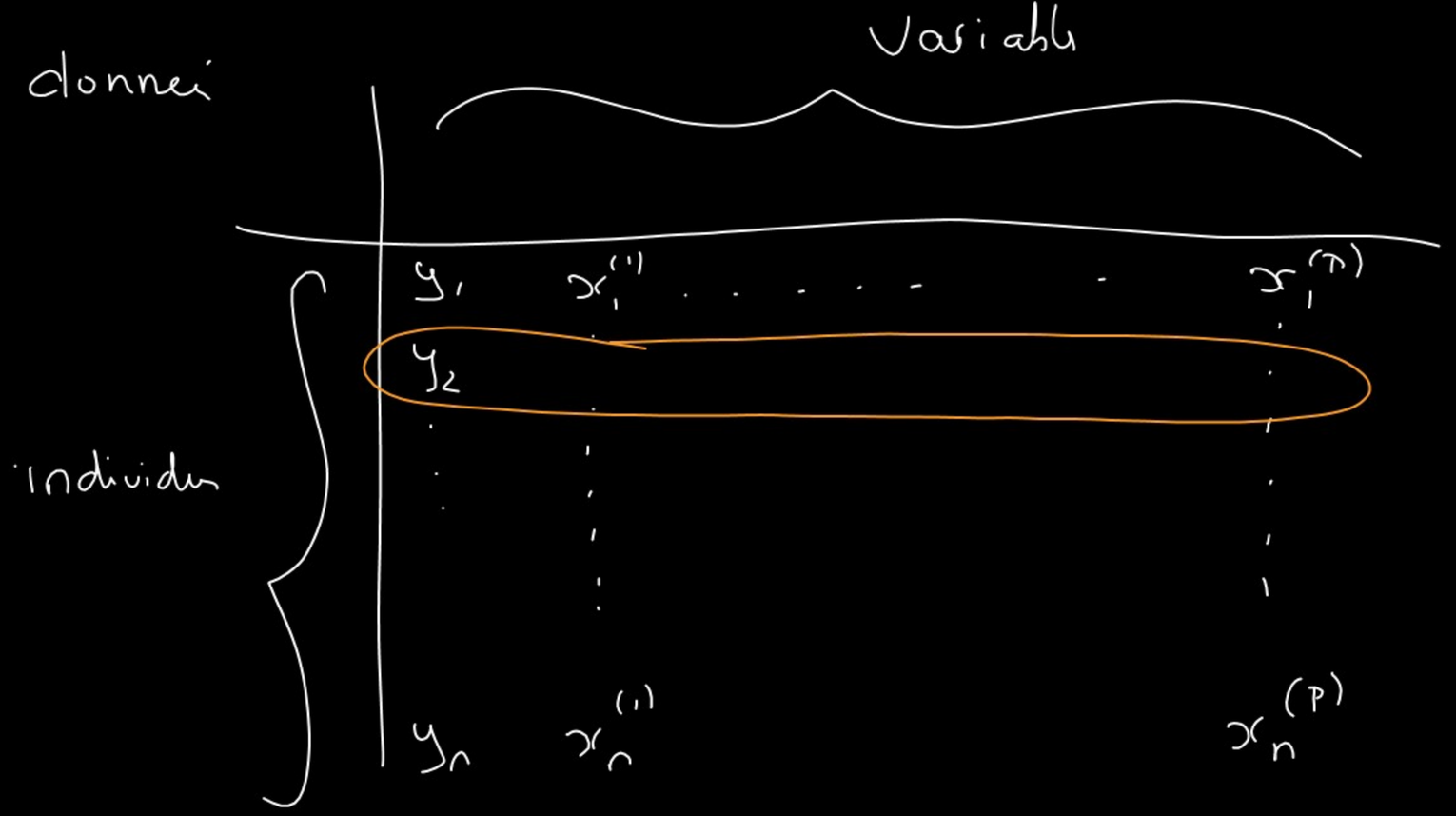
type variables explicatives / type de la variance réponse	quantitatives	qualitatives	mélange
quantitative	→ Modèle linéaire par moindres carrés (OLS)	Analyse de la variance (ANOVA)	Analyse de la covariance (ANCOVA)
qualitative	Régression Logistique		

⚠

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_i^{(k)} + \varepsilon_i$$

on a $\mathbb{E}[Y_i] = \beta_0 + \sum_{k=1}^{p} \beta_k x_i^{(k)}$

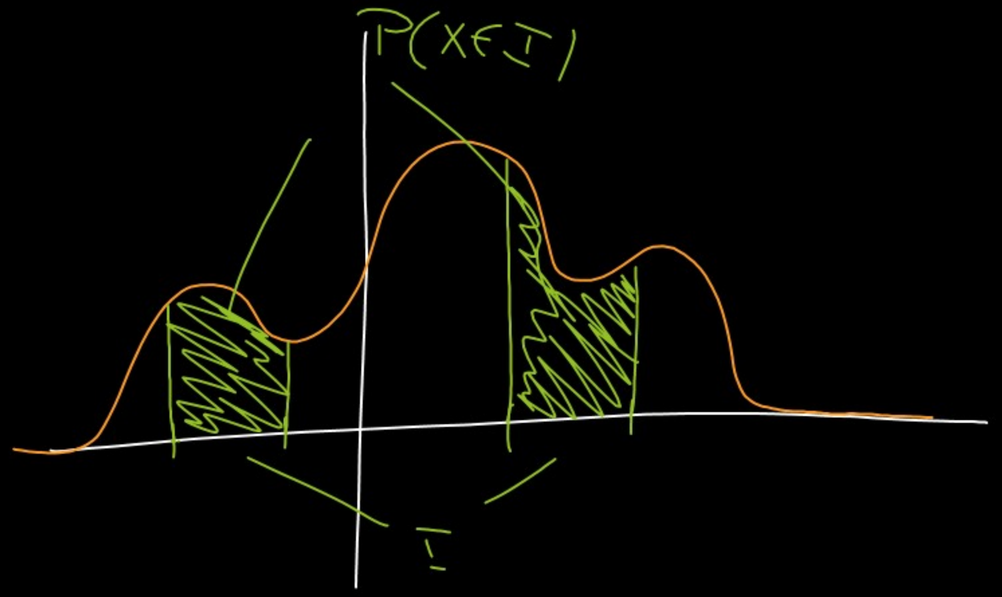
les Y_i ne sont pas identiquement distribuées car déjà pas toutes la même espérance !


Variable
données
individus
y_1
$x_1^{(1)}$
$x_1^{(p)}$
y_2
y_n
$x_n^{(1)}$
$x_n^{(p)}$

Rappel :

Fonction de densité f :

- $f(x) \geq 0 \quad \forall x \in \mathbb{R}$
- $\int_{\mathbb{R}} f(x)\,dx = 1$



$\Rightarrow$ Si X est une variable aléatoire admettant f pour fonction de densité

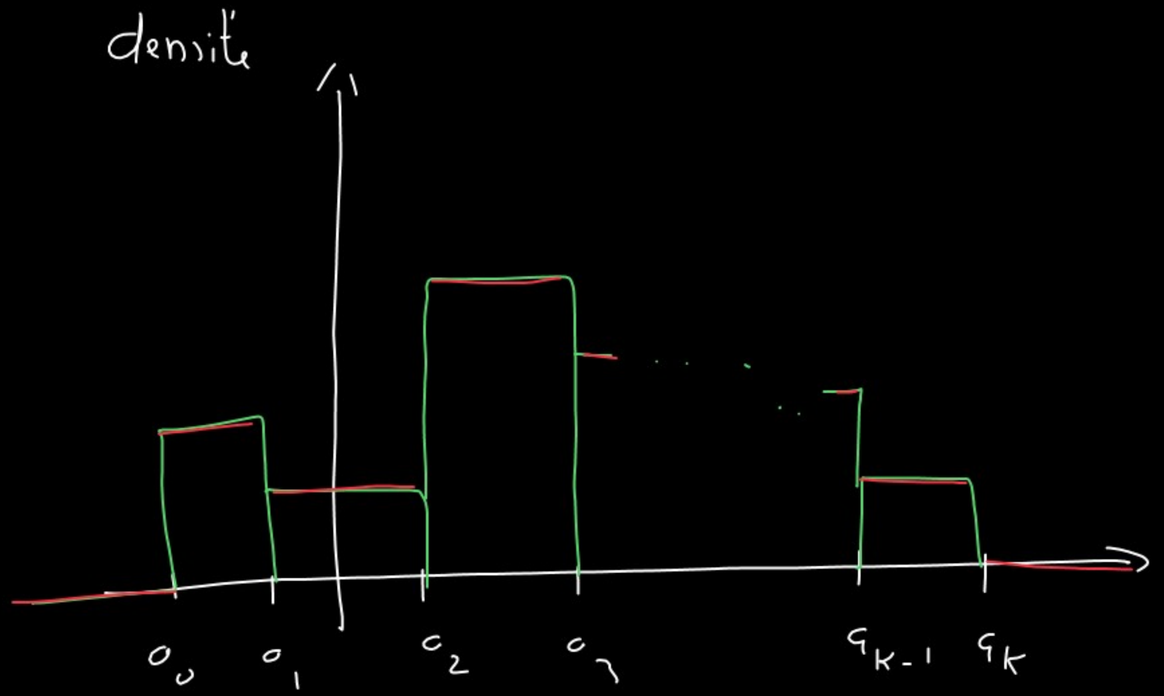
$$\forall I \subset \mathbb{R}, \quad P(X \in I) = \int_I f(x)\,dx$$

histogramme :

classes	effectifs	densité
$[a_0 ; a_1[$	n_1	d_1
$[a_1 ; a_2[$	n_2	d_2
$[a_2 ; a_3[$	n_3	d_3
$\vdots$	$\vdots$	$\vdots$
$[a_{K-1} ; a_K[$	n_K	d_K
	n	

n_i : nb d'observations qui sont dans $[a_{i-1} ; a_i[$

$$d_i = \frac{n_i}{n(a_i - a_{i-1})}$$



$$\int_{\mathbb{R}} \hat{f}(t)\,dt = \sum_{k=1}^{K} d_k (a_k - a_{k-1}) = \sum_{k=1}^{K} \frac{n_k}{n} = 1$$

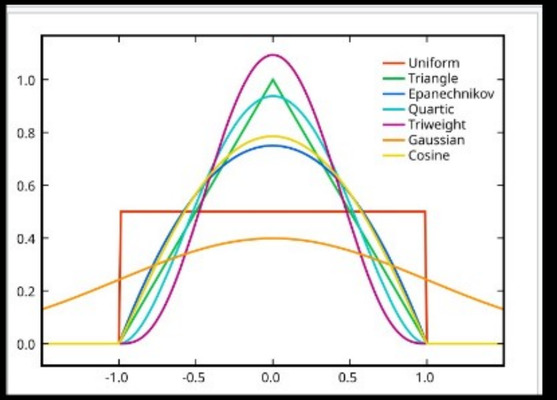
$$\hat{f}(t) = \begin{cases} d_k \text{ pour } t \in [a_{k-1}, a_k[\\ 0 \text{ sinon} \end{cases} \longrightarrow f^{\text{ct}} \text{ densité}$$

Estimation de la densité par la méthode des noyaux

def : Noyau (Kernel)

fonction réelle K telle que :

- $\forall x \in \mathbb{R}, \; K(x) \geq 0$
- $\forall x \in \mathbb{R}, \; K(-x) = K(x)$
- $\int_{\mathbb{R}} K(x)\,dx = 1$



Exemples : • $K(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$ • $K(x) = \frac{1}{2} \mathbb{1}_{[-1;1]}(x)$

Triangle [modifier | modifier le code]

La forme du noyau est une fonction triangulaire :

$$K(u) = (1 - |u|)\ \mathbf{1}_{(|u| \leq 1)}$$

La fonction estimée sera alors linéaire par morceaux.

Epanechnikov [modifier | modifier le code]

On parle aussi de noyau « parabolique ». Il porte le nom de V.A. Epanechnikov, 1969[1] :

$$K(u) = \frac{3}{4}(1 - u^2)\ \mathbf{1}_{(|u| \leq 1)}$$

Ce noyau permet d'avoir l'estimateur le plus efficace pour la densité.

Quartique [modifier | modifier le code]

$$K(u) = \frac{15}{16}(1 - u^2)^2\ \mathbf{1}_{(|u| \leq 1)}$$

Cubique [modifier | modifier le code]

$$K(u) = \frac{35}{32}(1 - u^2)^3\ \mathbf{1}_{(|u| \leq 1)}$$

Gaussien [modifier | modifier le code]

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

Circulaire [modifier | modifier le code]

autres exemples de noyaux

Estimateur par noyaux

$$\hat{f}_n(t) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

où les x_i sont les observations de la variable dont on cherche à estimer la densité

tuning parameter : bandwich (fenêtre)

↳ à calibrer

Rappel : loi du χ^2 (Chi-Deux)

Soit $T_1, \dots, T_d$ des variables i.i.d $\mathcal{N}(0;1)$

On pose :

$$C = \sum_{k=1}^{d} T_k^2$$

$$C \sim \chi^2(d)$$

C suit une loi du χ^2 à d degrés de liberté

Test d'adéquation du χ^2 :

→ on crée une distance entre les effectifs observés et les effectifs théoriques, les effectifs théoriques étant calculés sous l'hypothèse H_0 qui précise totalement la loi inférée pour la variable.

On peut alors définir la statistique du χ^2 :

$$T = \sum_{j=1}^{J} \frac{(N\hat{p}_j - Np_j)^2}{Np_j} = \sum_{j=1}^{J} \frac{(n_j - Np_j)^2}{Np_j} \quad \text{où } n_j = N\hat{p}_j = \sum_{i=1}^{N} 1_{y_i = v_j}$$

Résultat :

Sous H_0, $T \underset{\mathcal{L}}{\sim} \chi^2(J-1)$

La règle de décision est :

. si $T > s \Rightarrow$ on rejette H_0

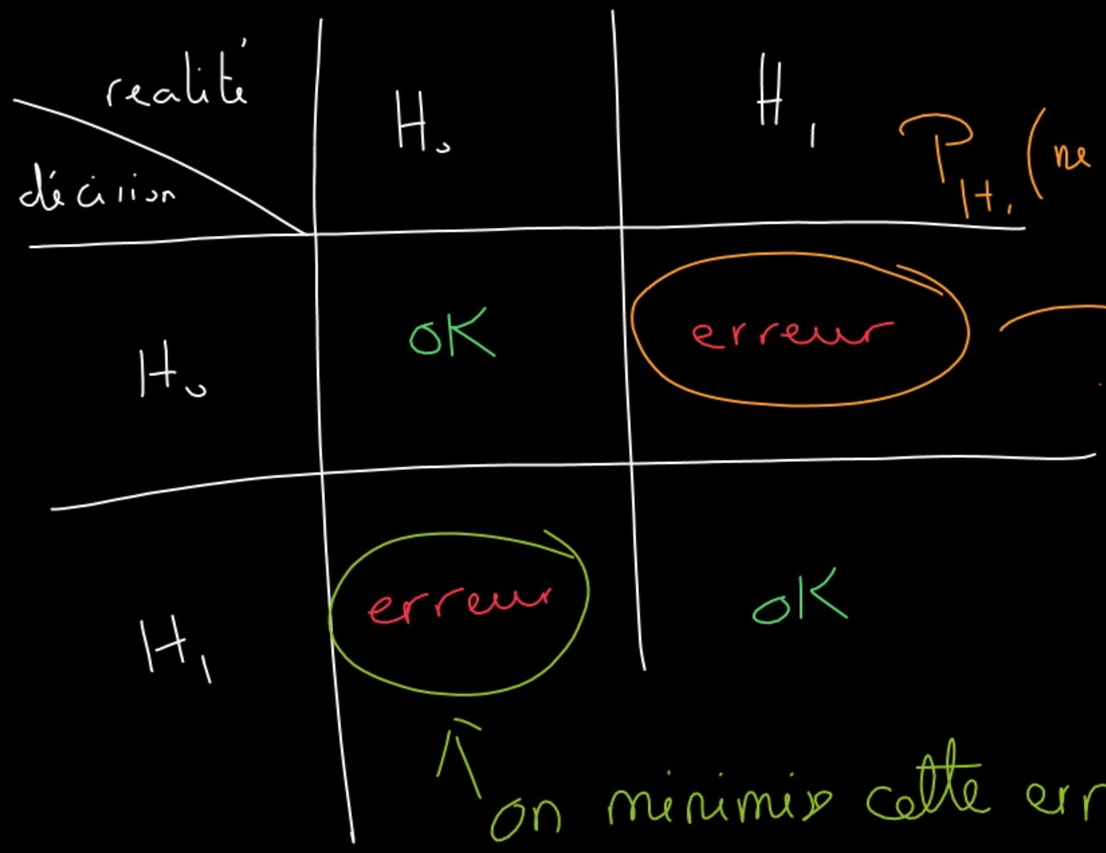
. si $T < s \Rightarrow$ on ne rejette pas H_0

avec $s = c_{1-\alpha ; J-1}$: le quantile d'ordre $1-\alpha$ pour $\chi^2(J-1)$

$$\left(P(\chi^2(J-1) > c_{1-\alpha ; J-1}) \leq \alpha \right.$$

α est ce que l'on appelle le niveau de test.

Niveau d'un test



P_{H_1}(ne pas rejeter H_0) = 1 − puissance

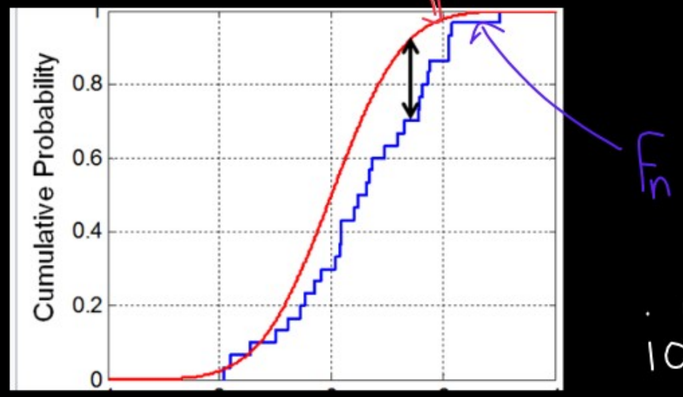
↑ on minimise cette erreur : $P_{H_0}(\text{décider } H_1) \leq \alpha$

test de Kolmogorov

On dispose d'observations $x_1, \dots, x_n$

H_0 : les observations sont associées à une variable aléatoire de f.d.r. F

H_1 : ce n'est pas le cas



On introduit la f.d.r. empirique

$$F_n(t) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{X_k \leq t}$$

idée : créer une distance entre F et F_n

la distance est :

$$K_n = \sup_{t \in \mathbb{R}} \left| F(t) - F_n(t) \right|$$

Sous H_0,

$\sqrt{n} K_n \rightsquigarrow K$

$\uparrow$ loi de Kolmogorov

Region de rejet :

$\{ \sqrt{n} K_n > s \}$ avec s tq

$P(K > s) \leq \varphi$

En pratique, on ne calcule pas α mais la p-valeur.

Pour le test de Kolmogorov, la p-valeur est donnée par :

$$\text{p-value} = P\left(K > \sqrt{n}\, K_{n,\text{obs}}\right)$$

$K_{n,\text{obs}}$: la valeur calculée de K_n sur le jeu de données

Règle de décision :

- si p-value $< \alpha \Rightarrow$ on décide H_1
- si p-value $> \alpha \Rightarrow$ on ne rejette pas H_0