MDI221
Statistiques élémentaires

Pascal Bianchi, Philippe Ciblat

28 novembre 2022

Table des matières

1 Bases mathématiques

1.1 Algèbre linéaire et optimisation

1.1.1 Espace vectoriel

On appelle espace vectoriel un ensemble $\mathcal{E}$ vérifiant les propriétés suivantes :

- il existe une opération, notée $+$, qui est interne (deux éléments additionnés de $\mathcal{E}$ restent dans $\mathcal{E}$), associative, possède un élément neutre, noté 0, et est symétrique (si $A + B = 0$ alors $B + A = 0$ aussi)
- il existe une opération, notée ., qui est externe. Soit un élément A de $\mathcal{E}$ et un élément λ d'un autre ensemble $\mathcal{K}$ (possédant certaines propriétés que nous ne mentionnerons pas ici), alors $\lambda.A$ est aussi dans $\mathcal{E}$. Et il y a une distributivité entre les opérations interne et externe, c-à-d, $\lambda(A + B) = \lambda A + \lambda B$ et $(\lambda + \mu)A = \lambda A + \mu A$.

L'exemple le plus courant est $\mathbb{R}^m$ pour lequel quand on additionne deux éléments de $\mathbb{R}^m$ on reste dans $\mathbb{R}^m$ et quand on multiple un vecteur par un réel (ici, on a $\mathcal{K} = \mathbb{R}$), on reste dans $\mathbb{R}^m$ aussi. De plus les lois de multiplication et d'addition ont bien les propriétés voulues, comme celle de la distributivité.

De plus on peut définir un sous-espace vectoriel à $\mathcal{E}$, noté $\mathcal{F}$, dont tous les éléments sont dans $\mathcal{E}$ mais quand on additionne deux éléments de $\mathcal{F}$, on reste dans $\mathcal{F}$ et quand on multiplie un élément de $\mathcal{F}$ par un point de $\mathcal{K}$ on reste dans $\mathcal{F}$.

Comme exemple, on peut penser à $\mathcal{F} = \mathbb{R}.\mathbf{1}_m$ avec $\mathcal{E} = \mathbb{R}^m$ où $\mathbf{1}_m$ est le vecteur de taille m composé uniquement de 1.

Dans la suite de ce polycopié, nous ne travaillerons qu'avec $\mathcal{E} = \mathbb{R}^m$.

1.1.2 Algèbre linéaire

Matrice

On considère un système de m équations avec d inconnues réelles à résoudre.

$$\begin{cases} a_{1,1}x_1 + \cdots + a_{1,d}x_d & = & b_1 \\ \ldots & = & \ldots \\ a_{m,1}x_1 + \cdots + a_{m,d}x_d & = & b_m \end{cases}$$

On appelle matrice A, le tableau suivant

$$A = \begin{pmatrix} a_{1,1} \cdots\cdots a_{1,d} \\ \vdots \\ a_{m,1} \cdots\cdots a_{m,d} \end{pmatrix}$$

et le système d'équations précédent s'écrit alors

$$A \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}.$$

Image d'une matrice

Une collection de vecteurs $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d$ de $\mathbb{R}^m$ forme une *famille libre* si la propriété suivante est vraie :

$$\forall x_1, \ldots, x_d,\ x_1\boldsymbol{a}_1 + \cdots + x_d\boldsymbol{a}_d = 0 \text{ implique que } x_1 = \cdots = x_d = 0\,.$$

Cela signifie que dans une famille libre, aucun vecteur ne peut s'écrire comme combinaison linéaire des autres. Si $\mathcal{E}$ est un sous-espace vectoriel de $\mathbb{R}^m$, on appelle *base de* $\mathcal{E}$ toute famille libre de $\mathcal{E}$ telle que, en outre, n'importe quel élément de $\mathcal{E}$ puisse s'écrire comme combinaison linéaire des vecteurs de la base. La *dimension* de l'espace vectoriel $\mathcal{E}$ est le nombre de vecteurs nécessaires pour constituer une base.

Soit A une matrice $m \times d$. L'image de A est l'ensemble :

$$\mathrm{Im}(A) = \{y \in \mathbb{R}^m \,:\, \exists x \in \mathbb{R}^d, y = Ax\}\,.$$

C'est un sous-espace vectoriel de $\mathbb{R}^m$. Plus exactement, si on note $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d$ les colonnes de la matrice A, soit :

$$A = \left(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d\right)\,, \tag{1.1}$$

alors $\mathrm{Im}(A)$ est l'ensemble des combinaison linéraires de ces vecteurs. Comme il est important d'avoir compris cela, nous démontrons ce point :

$$\begin{aligned} y \in \mathrm{Im}(A) \quad &\Leftrightarrow \quad \exists x,\ y = Ax \\ &\Leftrightarrow \quad \exists x,\ y = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d] \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \\ &\Leftrightarrow \quad \exists x,\ y = x_1\boldsymbol{a}_1 + \cdots + x_d\boldsymbol{a}_d \\ &\Leftrightarrow \quad y \text{ est une combinaison linéaire des colonnes de } A\,. \end{aligned}$$

L'image $\mathrm{Im}(A)$ d'une matrice A est donc le *sous-espace-vectoriel engendré par les colonnes de A*. Le *rang* de A, noté $\mathrm{rang}(A)$ est la dimension de $\mathrm{Im}(A)$. Dans de nombreux cas pratiques, les d

colonnes de A forment une famille libre, et par conséquent, forment une base de $\mathrm{Im}(A)$. Dans ce cas, $\mathrm{rang}(A) = d$. Mais dans d'autres cas, il peut y avoir de la redondance dans les colonnes de A (par exemple, une colonne est répétée, ou une colonne est la somme de deux autres). Dans ce cas de figure, il faudra moins de d vecteurs pour constituer une base de $\mathrm{Im}(A)$, et ainsi le rang de A sera plus petit que d.

Le noyau d'une matrice A est l'ensemble :

$$\ker(A) = \{x \in \mathbb{R}^d \ : \ Ax = 0\}\,.$$

C'est un sous-espace vectoriel de $\mathbb{R}^d$. Le théorème du rang stipule que :

$$\mathrm{rang}(A) + \dim(\ker(A)) = d\,.$$

En particulier, on a l'équivalence suivante :

$$\mathrm{rang}(A) = d \ \Leftrightarrow \ \ker(A) = \{0\} \ \Leftrightarrow \ \boldsymbol{a}_1, \ldots, \boldsymbol{a}_d \text{ est une famille libre.}$$

On rappelle que la transposée d'une matrice A est notée A^T. La transposition consiste à transformer les colonnes de A en lignes de A^T. On rappelle la règle : $(AB)^T = B^T A^T$. La transposée d'un vecteur colonne est un vecteur ligne.

On appelle $\mathrm{trace}(A)$ la somme des éléments diagonaux de A.

Matrices carrées

On dit qu'une matrice carrée $d \times d$ est

- inversible s'il existe une matrice, notée A^{-1} (et forcément unique), telle que $AA^{-1} = A^{-1}A = I_d$, où I_d est la matrice identité de taille $d \times d$. Cela revient à dire que $\ker(A) = \{0\}$, ou de manière équivalente que $\mathrm{rang}(A) = d$, où de manière encore équivalente, que $\det(A) \neq 0$, où $\det(A)$ est le *déterminant* de A.
- orthogonale si en outre $A^T = A^{-1}$.
- admet une valeurs propre et un vecteur propre si il existe un réel λ (valeur propre) et un vecteur v de $\mathbb{R}^d$ tels que
$$Av = \lambda v.$$
En général, une matrice admet un ensemble de valeurs propres et de vecteurs propres.
- symétrique si $A = A^T$.

Quand une matrice est symétrique, nous avons quelques définitions supplémentaires et propriétés intéressantes.

- A est symétrique semi-définie positive, si en outre $x^T Ax \geq 0$ pour tout vecteur colonne x de taille d.
- A symétrique définie positive, si en outre $x^T Ax > 0$ dès que $x \neq 0$.
- Soient deux matrices A et B symétriques. On dira que A est supérieure à B, et on notera $A \succeq B$ si la matrice $(A - B)$ qui est symétrique est aussi semi-définie positive, c-à-d, si $x^T(A - B)x \geq 0 \Leftrightarrow x^T Ax \geq x^T Bx$ pour tout vecteur colonne x de taille d.
Faire l'exercice 1.2.

Un des résultats les plus importants d'algèbre linéaire est le suivant.

Toute matrice symétrique est diagonalisable dans une base orthogonale

c'est à dire que si A est symétrique, alors il existe une matrice orthogonale P et une matrice diagonale Λ, telles que

$$A = P\Lambda P^T\,. \tag{1.2}$$

Dans ce cas, les coefficients de la diagonale de Λ coïncident avec les valeurs propres de A. Les colonnes de P sont constituées des vecteurs propres. De plus

— Si A est symétrique inversible, alors toutes ces valeurs propres sont non-nulles et la matrice inverse vaut

$$A^{-1} = P\Lambda^{-1}P^T.$$

— Si A est symétrique semi-définie positive, alors les valeurs propres (les coefficients sur la diagonale de Λ) sont positives ou nulles.

— Si A est définie positive, ces valeurs propres sont strictement positives, et dans ce cas la matrice A est inversible.

Produit scalaire

Soient deux vecteurs x, y dans $\mathbb{R}^m$. On supposera toujours que les vecteurs sont des vecteurs-colonne, et on notera x_i, y_i leurs composantes respectives. Le produit scalaire de x et y est défini par :

$$\begin{aligned}\langle x, y\rangle &= \sum_{i=1}^{m} x_i y_i \\ &= x^T y\,.\end{aligned}$$

La norme (euclidienne) de x est définie par :

$$\|x\| = \sqrt{x^T x}\,.$$

Faire l'exercice 1.3.

Faire l'exercice 1.4.

Faire l'exercice 1.5.

1.1.3 Optimisation

Si $f : \mathbb{R}^d \to \mathbb{R}$ est une fonction, on dit que x est un *minimiseur* de f si $f(y) \geq f(x)$ pour tout y. On note $\arg\min f$ l'ensemble des minimiseurs de f (c'est un ensemble, mais lorsque f admet

un unique minimiseur, $\arg\min f$ est simplement un point de $\mathbb{R}^d$). Le gradient d'une fonction f en un point $x \in \mathbb{R}^d$ est le vecteur :

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{pmatrix} .$$

Dans le cas unidimensionnel $d = 1$, le gradient est simplement la dérivée $f'(x)$ de la fonction. On peut aller un cran plus loin, et définir la *matrice hessienne* de f au point x. Il s'agit de la matrice $d \times d$ définie par :

$$\mathrm{Hess}(f) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_d} \\ \vdots & & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_d \partial x_1} & \frac{\partial^2 f(x)}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_d^2} \end{pmatrix}$$

c'est à dire que le coefficient (i, j) de la matrice est $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$. Cette matrice est symétrique, car on peut permuter l'ordre de dérivation entre x_i et x_j. Dans le cas unidimensionnel $d = 1$, le Hessien est simplement la dérivée-seconde $f''(x)$ de la fonction.

Un point x est un *point critique* de f si $\nabla f(x) = 0$. Tout minimiseur de f est un point critique. La réciproque n'est pas vraie en générale, mais elle est vraie si f est une fonction convexe. Une fonction f est dite *convexe* si la propriété suivante est satisfaite :

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

pour tout $t \in [0, 1]$ et tous points x et y dans $\mathbb{R}^d$. On notera que l'ensemble $\{tf(x)+(1-t)f(y), t \in [0, 1]\}$ est le segment de droite reliant $f(x)$ à $f(y)$.

Théorème 1.1. *Dans le cas où f est convexe, on a effectivement l'équivalence :*

$$x \in \arg\min f \;\Leftrightarrow\; \nabla f(x) = 0 \,.$$

Cela signifie que pour trouver un minimiseur d'une fonction il suffit de chercher un point qui annule le gradient. De ce point de vue, les fonctions convexes sont sympathiques, car on a alors un outil simple pour rechercher ces minimiseurs. En pratique, il faudra savoir repérer qu'une fonction est convexe. Dans le cas unidimensionnel $d = 1$ (f est une fonction de $\mathbb{R} \to \mathbb{R}$), on se convainc facilement que f est convexe si et seulement sa dérivée est croissante (faire un dessin pour s'en convaincre). Autrement dit, f est convexe si et seulement si sa dérivée-seconde est partout positive ou nulle :

$$f : \mathbb{R} \to \mathbb{R} \text{ est convexe } \Leftrightarrow\; \forall x \in \mathbb{R},\; f''(x) \geq 0 \,.$$

Ce résultat intuitif admet une généralisation dans $\mathbb{R}^d$, donnée par le théorème suivant :

Théorème 1.2. *Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est convexe si et seulement si* $\mathrm{Hess}(f)$ *est semi-définie positive.*

1.1.4 Projection orthogonale sur un sous-espace

Soit A la matrice $m \times d$ dont les colonnes sont données par $\boldsymbol{a}_1, \dots \boldsymbol{a}_d$, comme dans l'équation (1.1). Noter que ces colonnes sont des vecteurs de $\mathbb{R}^m$.

Soit $z \in \mathbb{R}^m$ un point quelconque. On appelle *projeté de z sur* $\mathrm{Im}(A)$ le point, noté $\Pi_A(z)$, qui appartient à $\mathrm{Im}(A)$, et qui est le plus proche de z parmi tous les points de $\mathrm{Im}(A)$. En clair :

$$\Pi_A(z) = \arg \min_{y \in \mathrm{Im}(A)} \|y - z\| \,.$$

Théorème 1.3. *Supposons que* $\mathrm{rang}(A) = d$. *Alors :*

$$\Pi_A(z) = A(A^T A)^{-1} A^T z \,.$$

La matrice $\Pi_A = A(A^T A)^{-1} A^T$ *est appelée le projecteur orthogonal sur* $\mathrm{Im}(A)$.

Démonstration. Avant toute chose, on remarque que la matrice $A^T A$ est bien inversible. En effet, sous l'hypothèse que $\mathrm{rang}(A) = d$, le théorème du rang implique que $\ker(A) = \{0\}$. Donc, par l'exercice 1.1, $\ker(A^T A) = \{0\}$, ce qui implique que $A^T A$ est bien inversible. Passons maintenant au reste de la preuve.

Appelons temporairement Π la matrice $A(A^T A)^{-1} A^T$. Soit $y \in \mathrm{Im}(A)$ un élément quelconque de l'image de A. Par définition, il existe un certain $x \in \mathbb{R}^d$ tel que $y = Ax$. Un petit calcul montre que :

$$\Pi y = \Pi A x = A(A^T A)^{-1} A^T A x = Ax = y \,.$$

On peut donc écrire que pour tout $y \in \mathrm{Im}(A)$,

$$z - y = \Pi(z - y) + (I - \Pi)z \,.$$

où I est ici la matrice identité de taille $m \times m$. Ainsi, d'après l'exercice 1.5,

$$\|z - y\|^2 = \|\Pi(z - y)\|^2 + 2\langle \Pi(z - y), (I - \Pi)z \rangle + \|(I - \Pi)z\|^2 \,.$$

En fait, le lecteur peut très facilement vérifier que

$$\Pi^2 = \Pi = \Pi^T \,.$$

Par conséquent, on montre facilement que le produit scalaire est nul :

$$\langle \Pi(z - y), (I - \Pi)z \rangle = (z - y)^T \Pi (I - \Pi) z = (z - y)^T (\Pi - \Pi^2) z = 0 \,.$$

Donc, on peut simplifier :

$$\|z - y\|^2 = \|\Pi(z - y)\|^2 + \|(I - \Pi)z\|^2 \,.$$

On a montré que pour tout $y \in \mathrm{Im}\, A$, $\|z - y\| \geq \|(I - \Pi)z\|$. Or il existe un (et un seul) point $y \in \mathrm{Im}(A)$ qui atteint cette borne : il s'agit du point $y = \Pi z$. En effet :

$$\|z - \Pi z\|^2 = \|\Pi(z - \Pi z))\|^2 + \|(I - \Pi)z\|^2 = \|(I - \Pi)z\|^2 \,,$$

en utilisant à nouveau le fait que $\Pi^2 = \Pi$. Cela signifie que le point $y = \Pi z$ est le point de $\mathrm{Im}(A)$ qui est le plus proche de z. □

Dans le cadre du Théorème 1.3, on notera que la matrice $A^{\#} := (A^T A)^{-1} A^T$ vérifie la propriété suivante

$$A^{\#} A = I$$

et cette matrice $A^{\#}$ est dite pseudo-inverse à gauche.

Faire l'exercice 1.6.

Faire l'exercice 1.7.

1.2 Probabilités

On se place sur un espace Ω arbitraire appelé l'univers. Un élément $\omega \in \Omega$ s'appelle une issue, c'est à dire, un résultat possible d'une certaine expérience aléatoire. Un ensemble $A \subset \Omega$ s'appelle un événement. Pour une issue $\omega \in A$, on dit que A est réalisé si $\omega \in A$. On munit l'univers Ω d'une probabilité $\mathbb{P}$. Formellement, $\mathbb{P}$ est une application qui à tout événement A associe un nombre $\mathbb{P}(A)$ dans $[0, 1]$.

Une variable aléatoire (v.a.) est une fonction $X : \Omega \to \mathbb{R}$, qui à toute issue ω associe une valeur $X(\omega)$. Ainsi, X est une grandeur dont la valeur dépend du résultat de l'expérience aléatoire. On peut généraliser à des variables aléatoires $X : \Omega \to \mathbb{R}^d$, où $\mathbb{R}^d$ est l'espace euclidien de dimension d, que l'on appelle parfois des *vecteurs aléatoires*. Quand $d = 1$, on parle de variable aléatoire réelle (v.a.r.).

1.2.1 Variables aléatoires discrètes

Le cas le plus simple est le cas où $X(\omega)$ prend ses valeurs dans $\mathbb{N}$ (ou éventuellement, dans un sous-ensemble de $\mathbb{N}$, comme par exemple $\{1, 2, \ldots, 6\}$ pour un lancer de dé. L'ensemble des valeurs possibles est alors appelé le *support*). On parle de *variables discrètes*. La *loi* d'une telle v.a. est donnée par les coefficients :

$$\mathbb{P}(X = k)$$

pour tout $k \in \mathbb{N}$. C'est à dire, pour être formel, la probabilités de l'événement $\{X = k\}$ = l'ensemble des ω pour lesquels $X(\omega) = k$. L'*espérance* de X est définie par

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k\, \mathbb{P}(X = k)\,,$$

et plus généralement, si g est une fonction à valeurs réelles, on a :

$$\mathbb{E}(g(X)) = \sum_{k=0}^{\infty} g(k)\, \mathbb{P}(X = k)\,.$$

Le moment d'ordre p est $\mathbb{E}(X^p)$. On va tout particulièrement s'intéresser au moment d'ordre deux, $\mathbb{E}(X^2)$, et à la variance définie par

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)\,.$$

Variable	Support	$\mathbb{P}(X=k)$	$\mathbb{E}(X)$	$\mathrm{Var}(X)$
Bernoulli $\mathcal{B}(p)$	$\{0,1\}$	$p^k(1-p)^{1-k}$	p	$p(1-p)$
Uniforme $\mathcal{U}(N)$	$\{0,1,\dots,N\}$	$\frac{1}{N}$	$*$	$*$
Binomiale $\mathcal{B}(p,n)$	$\{0,1,\dots,n\}$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Poisson $\mathcal{P}(\lambda)$	$\mathbb{N}$	$\frac{\lambda^k}{k!}e^{-\lambda}$	λ	λ
Géométrique $\mathcal{G}(p)$	$\mathbb{N}^*$	$p(1-p)^{k-1}$	$1/p$	$*$

TABLE 1.1 – Principales variables discrètes

On a aussi que $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. L'écart-type est défini par :

$$\sigma_X = \sqrt{\mathrm{Var}(X)}\,.$$

Rappelons les exemples principaux. Si on se donne maintenant deux v.a. discrètes X, Y sur $\mathbb{N}$, on peut considérer une nouvelle v.a. formée par le couple (X,Y). C'est une v.a. sur $\mathbb{N}\times\mathbb{N}$, c'est à dire que ce vecteur aléatoire a pour valeurs possibles les couples de la forme (k,ℓ) où k,ℓ sont des entiers. On appelle *loi jointe* les coefficients

$$\mathbb{P}(X=k, Y=\ell)$$

pour tous les entiers possibles k,ℓ. Les lois de X et Y (c'est à dire les coefficients $\mathbb{P}(X=k)$ et $\mathbb{P}(Y=\ell)$ respectivement) sont appelées les *lois marginales* de X et de Y. Si g est une fonction quelconque à valeurs réelles, on a :

$$\mathbb{E}(g(X,Y)) = \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} g(k,\ell)\mathbb{P}(X=k,Y=\ell)\,.$$

Les v.a. X et Y sont dites *indépendantes* (noté $X \perp\!\!\!\perp Y$) si

$$\mathbb{P}(X=k,Y=\ell) = \mathbb{P}(X=k)\mathbb{P}(Y=\ell)$$

pour tout couple (k,ℓ). Si X et Y sont indépendantes, la propriété suivante est vraie pour toutes fonctions h et g :

$$\mathbb{E}(h(X)g(Y)) = \mathbb{E}(h(X))\mathbb{E}(g(Y))\,. \tag{1.3}$$

Applications :

- *calculer les termes $*$ dans le tableau 1.1*
- *Prouver l'éq. (1.3)*

1.2.2 Variables aléatoires à densité

On parle aussi parfois de variables aléatoires "continues", même si cette terminologie est impropre. On dit qu'une v.a. $X:\Omega\to\mathbb{R}$ admet une *densité de probabilité*, s'il existe une fonction $f_X:\mathbb{R}\to[0,+\infty)$ telle que :

$$\mathbb{P}(X\le x) = \int_{-\infty}^{x} f_X(t)dt\,.$$

Le membre de gauche, en tant que fonction de x, est appelée la *fonction de répartition* de x, on la note souvent $F_X(x)=\mathbb{P}(X\le x)$. La fonction f_X est la densité de probabilité de X (ou juste, la

Variable	Support	Densité	Espérance	Variance
Uniforme $\mathcal{U}([a,b])$	$[a,b]$	$(b-a)^{-1}\mathbf{1}_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{2}$
Exponentielle $\mathcal{E}(\lambda)$	$[0,+\infty)$	$\lambda e^{-\lambda x}\mathbf{1}_{[0,\infty)}(x)$	λ^{-1}	λ^{-2}
Gaussienne $\mathcal{N}(m,\sigma^2)$	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-m)^2}{2\sigma^2}}$	m	σ^2
Chi-deux $\chi^2(k)$	$[0,+\infty)$	$\frac{1}{2^{k/2}\Gamma(k/2)}\ x^{k/2-1}e^{-x/2}$	k	$2k$
Student $\mathcal{T}(k)$	$\mathbb{R}$	$\frac{1}{\sqrt{k\pi}}\frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}$	0 (si $k>1$)	$\frac{k}{k-2}$ (si $k>2$)

TABLE 1.2 – Principales variables à densité. La loi $\mathcal{N}(0,1)$ s'appelle la loi gaussienne (ou normale) centrée réduite. La lettre Γ représente la fonction Gamma d'Euler. La loi $\chi^2(k)$, ou loi du chi-deux à k degrés de liberté ($k \geq 1$), est la loi d'une somme de la somme $X_1^2 + \cdots + X_k^2$, où $X_1, \ldots, X_k$ sont iid gaussiennes centrées réduites. La loi $\mathcal{T}(k)$, ou loi de Student à k degrés de liberté, est la loi du rapport $\frac{Z}{\sqrt{U/k}}$, où $Z \sim \mathcal{N}(0,1)$, $U \sim \chi^2(k)$ et U, Z sont indépendantes.

densité, pour faire court). Par construction, $\int f_X(x)dx = 1$. Ainsi, la fonction de répartition est l'intégrale de la densité et, réciproquement, la densité de probabilité est la dérivée de la fonction de répartition. L'espérance de X est définie par :

$$\mathbb{E}(X) = \int x f_X(x)dx\,,$$

et plus généralement, si g est une fonction réelle,

$$\mathbb{E}(g(X)) = \int g(x) f_X(x)dx\,.$$

La variance et l'écart-type sont définis comme dans le cas discret. Deux v.a.r. X et Y définissent un vecteur aléatoire (X,Y) sur $\mathbb{R}^2$. La fonction de répartition $F_{X,Y}$ de ce vecteur est définie par :

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)\,.$$

De même, on appelle *densité jointe* du couple (X,Y), si elle existe, la fonction positive $f_{X,Y}$ telle que

$$\forall x,y, \quad F_{X,Y}(x,y) = \int_\infty^x \int_\infty^y f_{X,Y}(s,t)dsdt,$$

ou, autrement dit,

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}\,.$$

Les densités respectives de X et Y, appelées *densités marginales* sont alors liées à la densité jointe par :

$$f_X(x) = \int f_{X,Y}(x,y)dy \quad \text{et} \quad f_Y(y) = \int f_{X,Y}(x,y)dx\,.$$

Si $g(x,y)$ est une fonction réelle de deux variables, on a :

$$\mathbb{E}(g(X,Y)) = \int\int g(x,y) f_{X,Y}(x,y)dxdy\,.$$

En particulier, dans le cas où $g(x,y) = \mathbf{1}_H(x,y)$ est l'indicatrice d'un ensemble $H \subset \mathbb{R}^2$ quelconque, l'espérance du membre de gauche se ramène à une probabilité, et on peut écrire :

$$\mathbb{P}((X,Y) \in H) = \int\int f_{X,Y}(x,y)\mathbf{1}_H(x,y)dxdy$$

c'est à dire que la probabilité qu'un vecteur aléatoire appartienne à une certaine région $H \subset \mathbb{R}^2$ est l'intégrale de la densité, sur cette région H.

Dans le cas particulier où $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, on dit que les deux v.a.r. X et Y sont *indépendantes*. Dans ce cas, l'égalité (1.3) est satisfaite.

Faire l'exercice 1.10.

Toutes ces notions et ces résultats se généralisent évidemment au cas de non plus deux, mais d variables aléatoires.

Noter également qu'une collection de d variables aléatoires peut être vue comme une variable aléatoire d-dimensionnelle et donc comme un vecteur aléatoire. Nous reviendrons sur ce point plus tard.

1.2.3 Inégalités

Soient X, Y des v.a.r. On rappelle l'inégalité de Cauchy-Schwarz.

Proposition 1.4. *On a :*

$$\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}\,.$$

De plus, le cas d'égalité $\mathbb{E}(XY) = \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$ *signifie soit que* Y *est la variable nulle, soit que* $\exists \lambda \geq 0,\ X = \lambda Y$.

Démonstration. Soit $t \in \mathbb{R}$. On a :

$$0 \leq \mathbb{E}((X+tY)^2) = \mathbb{E}(X^2) + 2t\mathbb{E}(XY) + t^2\mathbb{E}(Y^2)\,.$$

Donc le discriminant Δ du trinôme du second degré $P(t) = \mathbb{E}(X^2) + 2t\mathbb{E}(XY) + t^2\mathbb{E}(Y^2)$ est négatif ou nul. Ce discriminant vaut $\Delta = 4(\mathbb{E}(XY)^2 - \mathbb{E}(X^2)\mathbb{E}(Y^2))$, l'inégalité de Cauchy-Schwarz est donc démontrée.

Dans le cas d'égalité, $P(t) = \mathbb{E}(X^2) + 2t\sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)} + t^2\mathbb{E}(Y^2) = (\sqrt{\mathbb{E}(X^2)} + t\sqrt{\mathbb{E}(Y^2)})^2$. Si Y n'est pas la variable nulle, on peut définir, $\bar{t} = -\sqrt{\mathbb{E}(X^2)}/\sqrt{\mathbb{E}(Y^2)}$. On a alors $P(\bar{t}) = 0$. En reprenant la définition de départ de $P(t)$, cela implique que $0 = P(\bar{t}) = \mathbb{E}((X+\bar{t}Y)^2)$. Donc $X + \bar{t}Y$ est la variable nulle. On a montré que $X = -\bar{t}Y$, soit $X = \sqrt{\mathbb{E}(X^2)}Y/\sqrt{\mathbb{E}(Y^2)}$. □

Soit $\epsilon > 0$. L'inégalité de Bienaymé-Tchebychev, que nous ne redémontrons pas, est donnée par :

$$\mathbb{P}(X > \epsilon) \leq \frac{\mathbb{E}(X^2)}{\epsilon^2}\,. \tag{1.4}$$

1.2.4 Quantiles

Soit X une variable aléatoire à densité. Pour tout α entre 0 et 1, on appelle *quantile de niveau* α, le nombre $q_\alpha \in \mathbb{R}$ tel que :

$$\mathbb{P}(X \leq q_\alpha) = \alpha \,.$$

On suppose ici implicitement que ce nombre existe et est défini de façon unique, ce qui est vrai dans les tous cas d'usage. Autrement dit, si F_X désigne la fonction de répartition de X, cela se lit : $F_X(q_\alpha) = \alpha$, soit :

$$q_\alpha = F_X^{-1}(\alpha) \,,$$

où F_X^{-1} est l'inverse de la fonction de répartition.

Le quantile de niveau 0.5 s'appelle la médiane. Le quantile de niveau 0.25 s'appelle le premier quartile. Le quantile de niveau 0.75 s'appelle le troisième quartile. Le quantile de niveau 0.1 s'appelle le premier décile. Le quantile de niveau 0.2 s'appelle le deuxième décile. Et ainsi de suite.

1.2.5 Covariance, corrélation

Le coefficient suivant est appelé la *covariance* de X et Y :

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \,.$$

On introduit le *coefficient de corrélation* entre X et Y, défini par :

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \,.$$

Le coefficient de corrélation est une version renormalisée de la covariance, comme l'indique la proposition suivante :

Proposition 1.5. *On a :*

$$-1 \leq \rho_{X,Y} \leq 1 \,.$$

On suppose que les v.a. X, Y ne sont pas constantes. Alors :

- *si $\rho_{XY} = 1$, alors il existe $a > 0$ et $b \in \mathbb{R}$, tels que $Y = aX + b$.*
- *si $\rho_{XY} = -1$, alors il existe $a < 0$ et $b \in \mathbb{R}$, tels que $Y = aX + b$.*

Démonstration. C'est une conséquence immédiate de l'inégalité de Cauchy-Schwarz. □

On dit que deux v.a. sont *décorrélées* si $\mathrm{Cov}(X, Y) = 0$, soit de manière équivalente $\rho_{X,Y} = 0$. D'après l'équation (1.3), l'indépendance de deux v.a. implique la décorrélation (et la réciproque est fausse en général, hormis dans le cas notable des vecteurs gaussiens que nous verrons plus bas). Donc, si deux variables aléatoires sont corrélées, cela implique qu'elles sont dépendantes. Les deux cas extrêmes sont donnés par $\rho_{XY} = \pm 1$. Dans ce cas, la dépendance est totale car on peut même affirmer que Y s'écrit comme une fonction (affine) de X. Cela signifie en particulier que la donnée de X *détermine* entièrement la valeur de Y, au travers d'une relation affine. En

ce sens, le coefficient de corrélation ρ_{XY} peut être interprété comme une mesure de dépendance (affine) entre X et Y.

La corrélation $\rho_{X,Y}$ peut être positive ou négative. Une corrélation positive tend à indiquer que de grandes valeurs de X vont de pair avec de grandes valeurs de Y. Une corrélation négative tend à indiquer que de grandes valeurs de X vont de pair avec de faibles valeurs de Y.

1.2.6 Matrice de covariance

La donnée de d v.a.r. $X_1, X_2, \ldots, X_d$ défini un vecteur aléatoire, que nous supposerons dorénavant être un vecteur-colonne $d \times 1$, noté $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} . \tag{1.5}$$

On notera sa densité jointe par $f_{\boldsymbol{X}}(x_1, \ldots, x_d)$. On définira l'espérance d'un vecteur aléatoire comme étant le vecteur des espérances :

$$\mathbb{E}(\boldsymbol{X}) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_d) \end{pmatrix} .$$

De même, on définira la *matrice de covariance* $\mathrm{Cov}(\boldsymbol{X})$ du vecteur aléatoire $\boldsymbol{X}$ comme étant la matrice $d \times d$ formé par l'ensemble des coefficients $\mathrm{Cov}(X_i, X_j)$ pour i, j décrivant $\{1, \ldots, d\}$, soit :

$$\mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_d) \\ \vdots & & & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \cdots & \mathrm{Cov}(X_d, X_d) \end{pmatrix} .$$

On rappelle que :

- La diagonale de $\mathrm{Cov}(\boldsymbol{X})$ correspond aux variances $\mathrm{Var}(X_1), \ldots, \mathrm{Var}(X_d)$, car $\mathrm{Cov}(X_1, X_1) = \mathrm{Var}(X_1)$
- $\mathrm{Cov}(\boldsymbol{X})$ est une matrice symétrique car $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$.
- $\mathrm{Cov}(\boldsymbol{X})$ est une matrice semi-définie positive, en ce sens que pour tout vecteur-colonne $x \in \mathbb{R}^d$, $x^T \mathrm{Cov}(X) x \geq 0$.

Faire les exercices 1.12 et 1.13.

1.2.7 Vecteurs gaussiens

Soit $\boldsymbol{X} : \Omega \to \mathbb{R}^d$ un vecteur-colonne, de la forme (1.5). On dit que $\boldsymbol{X}$ est un vecteur gaussien si, pour tout vecteur $x \in \mathbb{R}^d$, le produit scalaire $x^T X$ est une variable gaussienne (c'est à dire, qui suit la loi $\mathcal{N}(m, \sigma^2)$ pour un certain m et un certain σ^2). Si $\boldsymbol{X}$ est un vecteur gaussien, alors $X_1, \ldots, X_d$ sont des v.a. gaussiennes. Attention, la réciproque n'est pas toujours vraie, mais elle est vraie au moins dans le cas donné par l'exemple suivant.

Exemple 1.6. Si $X_1, \dots, X_d$ sont des v.a. gaussiennes indépendantes, alors le vecteur (1.5) est un vecteur gaussien. En outre, sa matrice de covariance est diagonale.

La première propriété fondamentale des vecteurs gaussiens est la suivante :

Toute transformation affine d'un vecteur gaussien est un vecteur gaussien.

Autrement dit, si $\boldsymbol{X}$ est un vecteur gaussien, alors tout vecteur aléatoire de la forme $A\boldsymbol{X}+b$ est un vecteur gaussien (où A est une matrice, et b un vecteur, tous deux déterministes).

La seconde propriété fondamentale des vecteurs gaussiens est la suivante :

Si la matrice de covariance d'un vecteur gaussien est diagonale, alors les composantes de ce vecteur sont indépendantes.

Par exemple, on sait qu'en général, si X et Y sont deux v.a.r. décorrélées, elles ne sont pas nécessairement indépendantes. Par contre, si on sait en outre que $(X, Y)^T$ est un vecteur gaussien, alors décorrélation vaut indépendance.

On utilise la notation

$$\boldsymbol{X} \sim \mathcal{N}(m, \Sigma)$$

pour écrire que $\boldsymbol{X}$ est un vecteur gaussien d'espérance m et de matrice de covariance Σ. Dans le cas où Σ est inversible, alors $\boldsymbol{X}$ admet une densité, qui est donnée par la formule suivante, pour tout $x = (x_1, \dots, x_d)^T$:

$$f_{\boldsymbol{X}}(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} e^{-\frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)}.$$

1.3 Exercices

Algèbre linéaire

Exercice 1.1. Montrer que $\ker A = \ker(A^T A)$.

Exercice 1.2. Soient deux matrices symétriques A et B Montrer que si $A \succeq B$ alors $a_{\ell,\ell} \geq b_{\ell,\ell}$ pour tout $\ell = 1, \cdots, d$ et donc $\mathrm{trace}(A) \geq \mathrm{trace}(B)$.

Exercice 1.3. Soit x un vecteur-(colonne) de $\mathbb{R}^m$. Démontrer que $x^T.x = \mathrm{trace}(x.x^T)$.

Exercice 1.4. Soit une matrice A carrée symétrique semi-définie positive. Démontrer qu'il existe une matrice carrée de même taille, notée Γ, telle que, $A = \Gamma^T.\Gamma$. En déduire que $x^T A x = \|\Gamma x\|^2$.

Exercice 1.5. Démontrer l'identité $\|a+b\|^2 = \|a\|^2 + 2\langle a, b\rangle + \|b\|^2$.

Exercice 1.6. Soit Π_A le projecteur sur $\mathrm{Im}(A)$ où A est une matrice $m \times d$. Montrer l'identité de Pythagore :

$$\|z\|^2 = \|\Pi_A z\|^2 + \|(I - \Pi_A)z\|^2.$$

Exercice 1.7. Dans le cadre du Théorème 1.3, si $m > s$, montrer que $AA^{\#}$ ne peut être égal à l'identité. Si $m = d$, que se passe t-il pour le projecteur ?

Optimisation

Exercice 1.8. Soit $f : \mathbb{R}^d \to \mathbb{R}$. Calculer $\nabla f(x)$ et $\mathrm{Hess}(f)$ dans les cas suivants :

— $f(x) = \frac{1}{2}\|x\|^2$.

— $f(x) = \frac{1}{2}\|Ax - b\|^2$.

Ces fonctions sont-elles convexes ? Caractériser $\arg\min f$.

Probabilités

Exercice 1.9. Un ingénieur d'une entreprise de micro-électronique affirme après avoir évalué la robustesse d'un composant A : "63% des composants A ont une durée de vie inférieure à la durée de vie moyenne du composant". Ne devrait-ce pas être par définition 50% ?

Exercice 1.10. Soient X, Y deux v.a. iid (indépendantes identiquement distribuées) de loi uniforme sur $[0, 1]$. Calculer la probabilité que $X^2 + Y^2 \leq 1$.

Exercice 1.11. (Inégalité de Chebychev-Cantelli) L'inégalité de Chebychev-Cantelli est un raffinement utile de l'inégalité de Bienaymé-Chebychev :

$$\mathbb{P}(X - \mathbb{E}(X) \geq \epsilon) \leq \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + \epsilon^2}. \tag{1.6}$$

1. On peut supposer sans perte de généralité que $\mathbb{E}(X) = 0$. Justifier que pour tout $\epsilon > 0$,
$$\epsilon = \mathbb{E}[(\epsilon - X)] \leq \mathbb{E}[(\epsilon - X)\mathbf{1}(X \leq \epsilon)].$$
2. Justifier que
$$\epsilon^2 \leq \mathbb{E}[(\epsilon - X)^2]\mathbb{P}(X \leq \epsilon) = (\epsilon^2 + \mathrm{Var}(X))\mathbb{P}(X \leq \epsilon),$$
3. Conclure.

Exercice 1.12. Soit $\boldsymbol{X}$ un vecteur aléatoire (colonne) sur $\mathbb{R}^d$, d'espérance nulle. Montrer que $\mathrm{Cov}(\boldsymbol{X}) = \mathbb{E}(XX^T)$. En déduire que $\mathrm{Cov}(\boldsymbol{X})$ est une matrice symétrique semi-définie positive.

Exercice 1.13. Soit $\boldsymbol{X}$ un vecteur aléatoire sur $\mathbb{R}^d$. Soit A une matrice $m \times d$. Montrer que :

$$\mathrm{Cov}(A\boldsymbol{X}) = A\,\mathrm{Cov}(\boldsymbol{X})A^T\,.$$

Exercice 1.14. Soient X, Y deux v.a. gaussiennes centrées réduites, indépendantes. On pose :

$$\boldsymbol{W} = \begin{pmatrix} 1 & 0 \\ 1 & -1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}.$$

Montrer que $\boldsymbol{W}$ est un vecteur gaussien. Calculer son espérance et sa matrice de covariance.

2 Régression linéaire

Avant de décrire le problème de la régression linéaire, nous introduisons, en guise de préliminaire, une quantité courante en statistique, le coefficient de corrélation de Pearson.

2.1 Coefficient de corrélation de Pearson

2.1.1 Définition

Soit $\{(x_1, y_1), \dots, (x_N, y_N)\}$ un ensemble de N couples sur $\mathbb{R}\times\mathbb{R}$.

Définition 2.1. On appelle *moyenne empirique* des points $\{x_1, \dots, x_N\}$ la quantité :

$$\bar{x} := \frac{1}{N}\sum_{i=1}^{N} x_i \,,$$

et $\bar{y}$ est définie de façon similaire.

Définition 2.2. On appelle *coefficient de corrélation empirique* des couples $\{(x_1, y_1), \dots, (x_N, y_N)\}$, ou *coefficient de corrélation de Pearson*, la quantité :

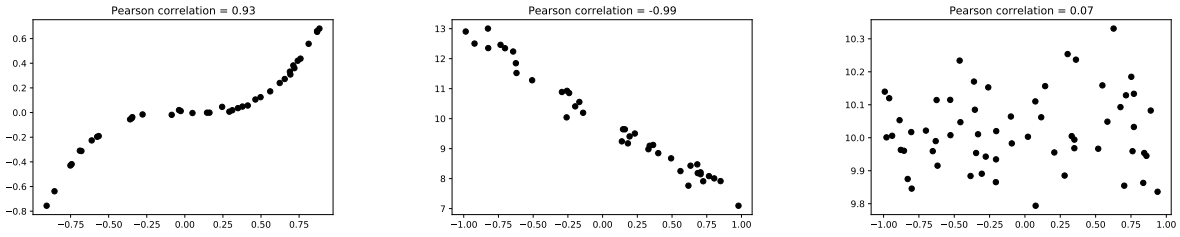
$$r_{x,y} := \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \,.$$

La définition fait sens dès que les x_i ne sont pas tous égaux à $\bar{x}$ (et idem pour y_i), ce que nous supposerons toujours implicitement.

Dans la définition ci-dessus, bien qu'on parle de corrélation, nous n'avons pas défini de variables aléatoires. Il n'y a au départ qu'un simple nuage de points, appelé un N-échantillon. Pourtant, il existe évidemment un lien avec le coefficient de corrélation $\rho_{X,Y}$ défini au premier chapitre dans un cadre probabiliste. Par exemple, si on introduit un vecteur aléatoire (X, Y), suivant la loi uniforme sur l'ensemble $\{(x_1, y_1), \dots, (x_N, y_N)\}$, il apparaît immédiatement que

$$\bar{x} = \mathbb{E}(X) \text{ et } r_{x,y} = \rho_{X,Y} \,.$$

Par conséquent, la proposition 1.5 s'applique. En particulier, le coefficient de Pearson appartient à l'intervalle $[-1, 1]$. Un coefficient de Pearson positif signifie que les valeurs de x_i et y_i ont tendance à être simultanément fortes ou simultanément faibles, par rapport à leur moyenne. Et dans le cas extrême où $r_{x,y} = 1$, les points x_i, y_i sont même alignés sur une droite de pente positive : $y_i = ax_i + b$ pour certains coefficients $a > 0$ et b. Inversement, si $r_{x,y} = -1$, les points x_i, y_i sont alignés sur une droite, mais dont la pente est cette fois négative. La figure 2.1 fournit quelques exemples.



(a) (b) (c)

FIGURE 2.1 – Coefficient de Pearson pour différents ensembles de points. On a les valeurs suivantes : (a) $r_{x,y} = 0.93$. (b) $r_{x,y} = -0.99$. (c) $r_{x,y} = 0.07$.

2.1.2 Corrélation versus causalité

Exemple 2.3. Deux économistes renommés, C. Reinhart et K. Rogoff, suggèrent dans un article que, dans les pays où la dette publique est supérieure à 90% du produit intérieur brut (PIB), la croissance économique est plus lente. L'article conclut par la recommandation suivante : puisqu'une dette élevée cause un ralentissement de la croissance, il convient de mettre en place des politiques d'austérité, destinées à faire passer le ratio de dette publique sous les 90% du PIB. Quelle erreur fondamentale de raisonnement est commise par les deux économistes ?

De votre environnement professionnel jusqu'à votre sphère privée, vous saurez résister aux interprétations hâtives : corrélation n'implique pas causalité.

Alors comment s'assurer de l'effet causal d'une action sur une quantité observée ? La bonne méthode (à condition qu'on puisse l'appliquer) est l'A/B test. Sans doute connaissez-vous déjà son principe, mais nous l'expliquons avec un exemple.

Exemple 2.4. Une entreprise de e-commerce propose des recommandations personnalisées aux clients qui se rendent sur son site internet. L'équipe datascience de l'entreprise a mis au point une nouvelle IA, dont elle prétend que les recommandations sont mieux adaptées aux besoins des clients. L'entreprise, qui déteste les risques, doit être certaine que ce nouvel algorithme a un impact positif sur le volume de vente, avant de remiser définitivement son ancien algorithme. L'A/B test consiste à diviser les clients en deux groupes A et B : le groupe A est constitué selon un tirage aléatoire uniforme de N individus, dans toute la population de clients. Idem pour le groupe B. La nouvelle IA est testée sur le groupe A, alors que les clients du groupe B utilisent toujours l'ancienne. On compare alors le volume de vente moyen par client dans chacun des deux groupes.

Remarque 2.1. Il n'est pas toujours possible de mettre en œuvre des A/B-tests pour vérifier un lien causal (pour l'exemple 2.3, cela n'aurait d'ailleurs guère de sens). Il faut alors recourir à d'autres méthodes statistiques (on parle d'*inférence causale*), qui dépassent le cadre de ce cours.

2.2 Régression linéaire simple

2.2.1 Etude de cas

L'Enquête canadienne sur les mesures de la santé (ECMS) a permis de mesurer la tension artérielle moyenne dans cinq groupes de femmes d'âges différent. Les données sont fournies dans le tableau suivant.

Tension	Âge
102	25
103	35
109	45
119	55
122	65
126	75

TABLE 2.1 – Tension artérielle moyenne de cinq groupes de femmes (source : ECMS)

> **Mise en situation** : En tant qu'ingénieur, vous représentez un cabinet d'étude responsable de l'interprétation des données recueillies par l'ECMS. Vous devez dans quelques jours présenter vos conclusions devant une assemblée de médecins. A quelles questions d'ordre médical allez vous chercher à répondre ?

Nous allons construire un modèle mathématique qui permet de lier la tension d'une femme à son âge. En représentant ces points sur une courbe, il semble que ces points soient approximativement situés sur une droite. Il est donc naturel de chercher une relation du type :

$$\text{tension} \simeq \beta_1 \times \text{âge} + \beta_0$$

où β_0 est l'ordonnée à l'origine, et β_1 la pente de la droite en question. La présence du signe $\simeq$ provient du fait que les points ne sont pas *exactement* situés sur une droite. L'objectif est de trouver la droite (c'est à dire les coefficients β_0, β_1) qui passe "au plus proche" des points (âge, tension). La première étape est donc de nous donner un *critère* qui quantifie à quel point une droite approxime bien ou mal notre nuage de points.

Remarque 2.2. L'âge est ici la *variable explicative* (en anglais : feature, ou input), et la tension est la *variable à expliquer* ou la *réponse* (en anglais : label, ou output).

2.2.2 Critère des moindres carrés

On notera N le nombre de points (x_i, y_i) disponibles, où x_i représente la $i$ème variable explicative et y_i la $i$ème variable à expliquer. Dans notre cas, $N = 6$, et les couples (x_i, y_i) sont donnés par l'ensemble :

$$\{(25, 102), (35, 103), (45, 109), (55, 119), (65, 122), (75, 126)\}\,.$$

Cet ensemble est le jeu de données (dataset) – on parle aussi de N-échantillon. Chaque point (x_i, y_i) du jeu de données est appelé un échantillon, ou un exemple (en anglais : sample).

Définition 2.5. On appelle *critère des moindres carrés* la fonction

$$J(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2 \,.$$

Cette fonction mesure la somme des écarts quadratiques entre la réponse d'un modèle $\beta_1 x_i + \beta_0$ et la variable y_i effectivement observée.

Remarque 2.3. On peut se demander pourquoi avoir élevé l'écart au carré, et pourquoi ne pas choisir, par exemple, la valeur absolue de l'écart. En fait, la valeur absolue aurait été, elle aussi, tout à fait intéressante. Les deux critères pénalisent les erreurs de façon différente (l'erreur quadratique utilisée ici pénalise davantage les grands écarts). Mais au delà de ça, ce qui fait le succès du critère des moindres carrés, c'est surtout la simplicité de sa mise en œuvre numérique, ainsi que l'interprétabilité des résultats.

Proposition 2.6. *Supposons que les x_i ne sont pas tous égaux. On pose $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, et $\bar{y}$ défini de même comme la moyenne des y_i. Alors, le critère des moindre carrés admet pour minimiseur le point $(\hat{\beta}_0, \hat{\beta}_1)$ défini par :*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

et $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Le point $(\hat{\beta}_0, \hat{\beta}_1)$ est appelé l'estimateur des moindres carrés.

Démonstration. En cours. □

> **Mise en situation** : Lors de votre présentation sur les liens entre âge et tension, devant l'assemblée de médecins, vous expliquez : "Nous avons optimisé un critère des moindres carrés, et notre conclusion est que $\hat{\beta}_1 = 0.53$ et $\hat{\beta}_0 = 86.79$." Quel est selon-vous le point de vue des médecins sur votre présentation ? Est-ce une conclusion suffisante de votre étude ?

On remarquera le lien naturel qui existe entre la pente $\hat{\beta}_1$ du modèle obtenu et le coefficient de corrélation de Pearson $r_{x,y}$. Une fois l'estimateur calculé, nous avons notre modèle définitif. Dans le cas de la Table 2.1, il s'agit de :

$$\hat{y}(x) = 0.53x + 86.79 \,,$$

où x représente l'âge et $\hat{y}(x)$ est la tension prédite. La fonction $\hat{y}(\cdot)$ est appelée le *prédicteur*. Dans le contexte du machine learning, la prédiction est une tâche très courante : l'objectif est souvent de prédire une réponse y lorsque l'on observe seulement une certaine variable explicative x. Pour cela, on apprend un modèle à partir d'un *jeu de données labellisé*, c'est à dire d'un N-échantillon constitué de couple (x_i, y_i) pour lesquels la réponse y_i a bien été mesurée. Nous verrons des exemples, plus tard dans ce cours.

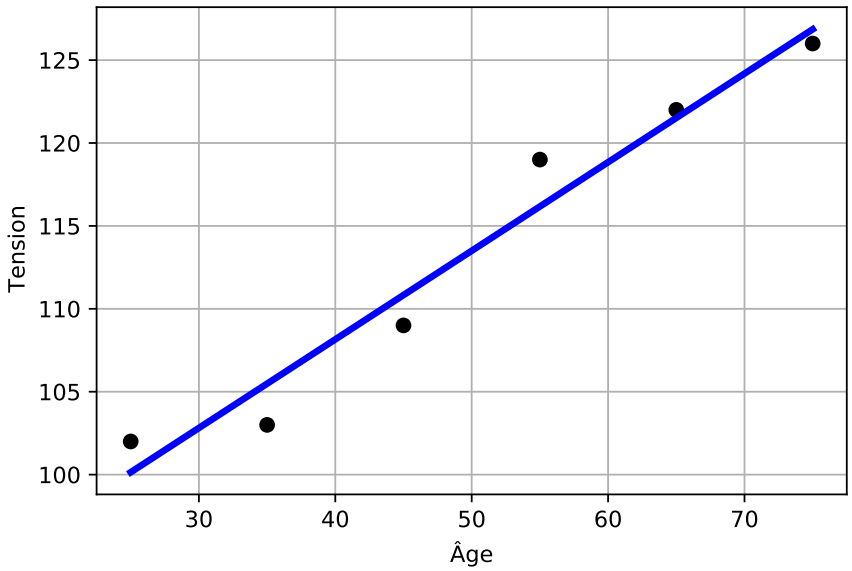


FIGURE 2.2 – Tension en fonction de l'âge chez les femmes. Les points sont les données de l'ECMS. La droite est le prédicteur.

Mise en situation : Après votre affirmation que $\hat{\beta}_1 = 0.53$, vous commencez à détecter l'ennui dans le regard des médecins qui vous écoutent. Ils semblent avoir du mal à comprendre votre langage. Vous tâchez alors d'être plus concret. Vous expliquez aux médecins que votre modèle vous permet de prédire la tension, et ce même pour des âges non présents dans le jeu de données. Par exemple, vous êtes capables de prédire qu'une femme de 50 ans aura une tension de 114. Avez vous regagné l'intérêt des médecins ?

2.2.3 Analyse de la variance

On se pose la question suivante : y a-t-il un effet significatif de l'âge x sur la tension y ? Les données semblent indiquer que la tension a tendance à augmenter avec l'âge. Et d'ailleurs, notre estimateur des moindres carrés conduit à $\hat{\beta}_1 = 0.53$, autrement dit, le prédicteur $\hat{y}(x)$ est une fonction croissante de x. Mais attention aux conclusions hâtives. Cela n'est-il pas simplement lié à l'aléa inhérent aux données ?

Exemple 2.7. Prenons un exemple absurde. Remplaçons les tensions de la table 2.1, par des nombres artificiels, choisis aléatoirement selon une loi uniforme sur l'intervalle $[100, 130]$, et sans aucun lien avec l'âge :

$$(107, 117, 115, 121, 127, 117)\,.$$

Ce sont nos nouveaux y_i, qui n'ont aucun sens, ne représentent rien. L'estimateur des moindres carrés conduit à $\hat{\beta}_1 = 0.25$ et $\hat{\beta}_0 = 105$. Puisque $\hat{\beta}_1 \neq 0$, on pourrait être tenté de conclure que les réponses sont effectivement liées à l'âge. C'est faux, c'est seulement une conséquence de l'aléa dans les données.

Le SS_M

Ce terme va permettre d'analyser une distance des N-échantillons estimés linéairement à la moyenne empirique.

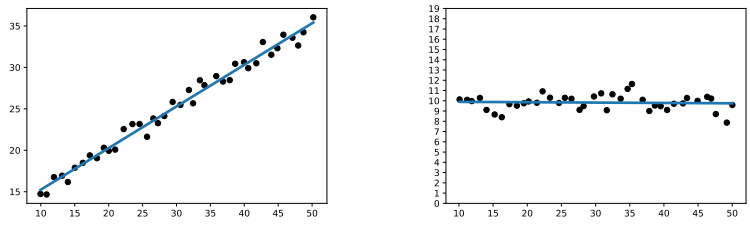
On appelle

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$$

la réponse prédite pour le $i$ème échantillon. On définit le SS_M (*Model Sum of Squares*) par la formule :

$$SS_M = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 \,,$$

où on rappelle que $\bar{y} = \frac{1}{N}\sum_i y_i$ est la valeur moyenne des réponses. Un faible SS_M signifie que les



(a) (b)

FIGURE 2.3 – Exemples de jeux de données générés aléatoirement. Dans la figure (a), les réponses simulées sont une fonction affine croissante de la variable explicative, à laquelle on a ajouté un bruit gaussien. Dans la figure (b), les réponses sont générées indépendamment de x, de manière iid. Le SS_M est plus grand dans la figure (a) que dans la figure (b). Donc le SS_M ne suffit pas pour savoir si les données collent au modèle.

prédictions $\hat{y}_i$ sont toutes proches de la moyenne $\bar{y}$. C'est typiquement le cas lorsque la variable x a peu d'influence sur y. Un fort SS_M signifie que les prédictions $\hat{y}_i$ varient beaucoup. C'est typiquement le cas lorsque la variable x influe beaucoup sur la valeur de y.

La remarque précédente a ses limites : la notion de "fort/faible" n'a pas de sens mathématique précis. Par exemple, si on s'intéresse non plus à la variable y mais à la variable $2y$, le SS_M obtenu après régression linéaire sera multiplié par quatre. Or l'application d'un facteur d'échelle ne devrait pas affecter notre conclusion. Autrement dit, il convient de *normaliser* le SS_M. La normalisation adéquate que nous décrirons à la fin de ce paragraphe donnera la quantité R^2 du statisticien Ronald Fisher. Nous devons auparavant introduire quelques autres notions.

Le SS_R

Ce terme va permettre d'analyser une distance des N-échantillons à leur estimée linéaire et donc une certaine distance au modèle linéaire.

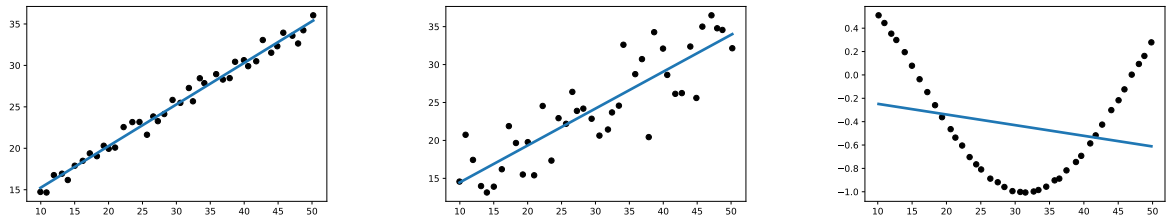
On appelle

$$e_i = y_i - \hat{y}_i$$

le $i$ème *résidu*. Si la prédiction était parfaite, on aurait $e_i = 0$ pour tout i. En pratique, ce n'est jamais le cas, et on quantifie l'erreur de prédiction par la quantité SS_R (*Residual Sum of Square*) :

$$SS_R = \sum_{i=1}^{N} e_i^2 \,,$$

c'est à dire la somme des carrés des résidus. Un SS_R faible indique une prédiction précise ; un SS_R fort indique de forts écarts entre la prédiction et la réponse. Cela peut provenir du fait que l'hypothèse du modèle linéaire est inadaptée aux données. Cela peut aussi provenir du fait que, bien que le modèle linéaire soit pertinent, il y a de fortes fluctuations des données.



(a) (b) (c)

FIGURE 2.4 – Exemples de jeux de données. (a) Faible SS_R : le modèle linéaire est adéquat et la prédiction est fiable. (b) Fort SS_R : le modèle linéaire est adequat mais il y a de fortes fluctuations de la réponse. (c) Fort SS_R : le modèle linéaire est inadéquat.

Le SS_T et la décomposition de la variance

Ce terme va permettre d'analyser une distance des N-échantillons à la moyenne empirique.

La quantité SS_T (*Total Sum of Square*) est définie par :

$$SS_T = \sum_{i=1}^{N} (y_i - \bar{y})^2 \,.$$

La valeur SS_T quantifie les fluctuations de la réponse y_i autour de la moyenne.

Proposition 2.8.

$$SS_T = SS_R + SS_M$$

Démonstration. En classe. □

La proposition 2.8 s'appelle la décomposition de la variance.

Remarque 2.4. Le terme de variance est utilisé ici, alors que nous n'avons fait aucune hypothèse probabiliste : cela provient du fait que, à un facteur $1/n$ près, SS_T peut-être interprétée comme la variance *empirique* des données. Mais nous insistons sur le fait que, jusqu'ici, nous ne faisons pas de probabilités !

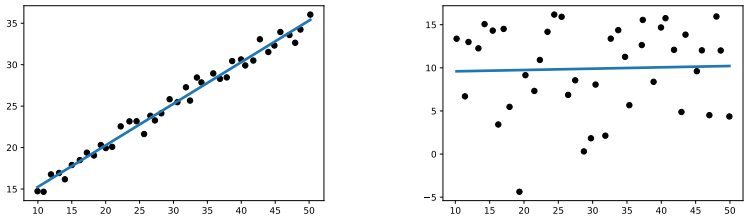
La proposition 2.8 justifie la dénomination suivante : SS_M est la *part de la variance expliquée par le modèle (écart à la moyenne)*, alors que SS_R est la *part de la variance expliquée par les résidus (écart entre les données et leurs estimées)*.

Le coefficient R^2

Définition 2.9. Le coefficient R^2, ou *coefficient de détermination*, est le rapport :

$$R^2 = \frac{SS_M}{SS_T}$$

D'après la proposition 2.8, on a $0 \leq R^2 \leq 1$. Il s'agit donc d'une version renormalisée du SS_M. Si R^2 est proche de un, toute la variance est expliquée par le modèle, c'est à dire par les variations de la prédiction en fonction de la variable explicative. D'ailleurs, le cas limite $R^2 = 1$ signifie que les résidus sont nuls, autrement dit, que le modèle prédit les réponses de façon exacte : la variable y est complètement déterminée par x, au travers de la relation $y = \hat{y}(x)$.



(a) (b)

FIGURE 2.5 – Exemples de deux jeux de données ayant une valeur de SS_T similaire. (a) $SS_R \simeq 0$ et $SS_T \simeq SS_M$. Toute la variance est expliquée par le modèle. On a $R^2 \simeq 1$. (b) $SS_M \simeq 0$ et $SS_T \simeq SS_R$. Toute la variance est expliquée par les résidus. On a $R^2 \simeq 0$.

Au contraire, si R^2 est proche de zéro, toute la variance est expliquée par les résidus. Cela signifie que le fait d'observer x ne fournit aucune information utile sur y. Le modèle n'explique aucunement les fluctuations de y.

Mise en situation :

— Après calcul, le coefficient de détermination correspondant aux données de la table 2.1 est $R^2 = 0.96$. Quelle conclusion tirez-vous ?

— Sur la figure 2.2.3, des mesures ont été effectuées pour plusieurs stations de base (les couleurs) de la consommation électrique en fonction du trafic. Les différents points de

même couleur sont des mesures effectuées à des instants différents pour la même station de base. Le R^2 a été calculé pour les différentes mesures et est compris entre $0,85$ et $0,95$ et donc les auteurs en ont déduit que la régression linéaire était satisfaisante et l'ont utilisée dans le reste de leur article.

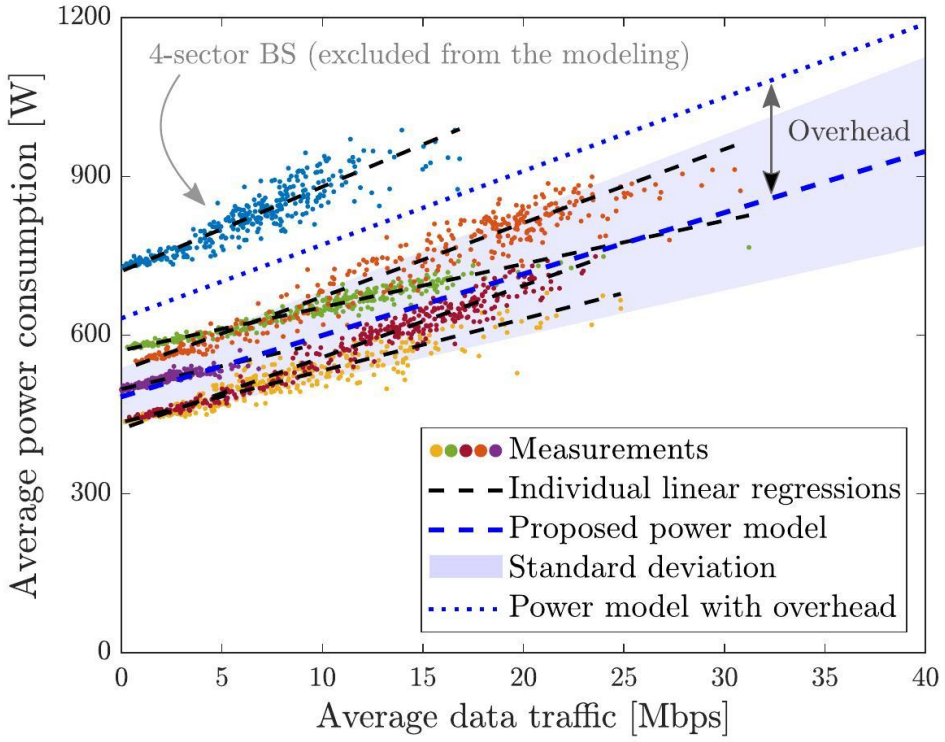


FIGURE 2.6 – Consommation électrique d'une station de base 4G en fonction de son débit. Figure provenant du papier suivant : L. Golard et al., "Evaluation and projection of 4G and 5G RAN energy footprints : the case of Belgium for 2020-2025", Annals of Telecommunications, 2023.

2.2.4 Data cleansing et bonnes pratiques

Si on suit ce polycopié à la lettre, les choses paraissent un peu trop simples : on calcule l'estimateur des moindres carrés, on en déduit le prédicteur, le coefficient de détermination, et voilà ?

Dans la pratique, un jeu de données s'appréhende toujours avec précaution.

— La première étape consiste toujours à visualiser les données. On trace les points du jeu de données. Cela permet de s'assurer que le modèle linéaire est une hypothèse raisonnable. Si ce n'est pas le cas, il faudra envisager d'autres modèles (polynômes ou autres). Certes il existe des méthodes permettant de quantifier rigoureusement l'adéquation des données au modèle linéaire, mais en première approche, votre oeil est déjà une garantie indispensable.

— Vous devez également "faire connaissance" avec vos données. Comment les variables explicatives sont-elles réparties ? Quelle est leur moyenne empirique, leur écart-type ? Souvent, les jeux de données comportent des *valeurs aberrantes* ou des *valeurs manquantes*, liées par

exemples à de mauvaises saisies. Les valeurs aberrantes auront un effet dramatique sur la qualité de votre estimateur des moindres carrés, et sur votre prédicteur. Il convient donc de les détecter, puis de les éliminer. Pour cela, il existe des méthodes statistiques, mais dans la plupart des cas, votre oeil sera là encore suffisant.

Cette étape d'inspection, de visualisation et de nettoyage des données se nomme le *data cleansing*. Elle est indispensable.

— Une fois l'estimateur des moindres carrés calculé, il faut visualiser vos résidus. Vous pouvez calculer le SS_R pour avoir une idée de leur amplitude, mais là encore, rien ne remplace un graphique. Vous pourrez ainsi contrôler visuellement la pertinence de votre prédicteur.

Ce qui vient d'être dit au-dessus est valable pour tout type de données : par exemple aussi pour des graphes de connexion dans un réseau social, des séries d'images, etc.

2.3 Régression linéaire multiple

2.3.1 Modèle

On cherche à expliquer une réponse $y \in \mathbb{R}$ en fonction d'un vecteur $\boldsymbol{x} \in \mathbb{R}^d$, de la forme $\boldsymbol{x} = (x_1, \ldots, x_d)^T$. Les variables $x_1, \ldots, x_d$ s'appellent les variables explicatives, ou les régresseurs (*features*, en anglais).

Exemple 2.10. On cherche à prédire la valeur y d'une action, en fonction des variables explicatives suivantes :

— x_1 = le taux d'intérêt de la Banque Centrale Européenne

— x_2 = le coût du baril de pétrole brut

— x_3 = l'indice boursier S&P-500.

On cherche une relation affine du type :

$$y \simeq \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d \,.$$

Pour cela, on dispose d'un N-échantillon :

$$\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$$

où chaque couple $(\boldsymbol{x}_i, y_i)$ est un élément de $\mathbb{R}^d \times \mathbb{R}$. Chaque vecteur $\boldsymbol{x}_i$ est caractérisé par ses composantes, que nous notons $x_{i,1}, \ldots, x_{i,d}$, soit :

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,d} \end{pmatrix} .$$

Remarque 2.5. S'il y a d variables explicatives, le modèle comporte donc $d+1$ paramètres. Cela est dû à la présence de l'ordonnée à l'origine β_0.

2.3.2 Critère des moindres carrés

On pose

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}.$$

Le vecteur $\boldsymbol{\beta}$ est de dimension $d+1$. Il contient les paramètres inconnus, que nous devons déterminer à l'aide du N-échantillon. Pour cela, nous minimisons le critère des moindres carrés, défini ci-dessous comme la somme des erreurs quadratiques entre la réponse prédite par le modèle et la variable à expliquer.

Définition 2.11. On appelle critère des moindres carrés la fonction $J : \mathbb{R}^{d+1} \to \mathbb{R}$ définie par :

$$J(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_{i,1} + \cdots + \beta_d x_{i,d})^2 .$$

A partir de maintenant, nous allons utiliser des notations matricielles pour simplifier les écritures. Notamment, on remarque que :

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d = (1, x_{i,1}, \ldots, x_{i,d})\, \boldsymbol{\beta} .$$

Ainsi, on obtient que

$$J(\boldsymbol{\beta}) = \|\boldsymbol{y} - \Phi\boldsymbol{\beta}\|^2$$

où $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ et où Φ est la matrice $N \times (d+1)$ donnée par :

$$\Phi = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,d} \\ \vdots & & & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,d} \end{pmatrix} \tag{2.1}$$

Théorème 2.12. *Supposons que* $\operatorname{rang}(\Phi) = d+1$. *Alors J admet un unique minimiseur donné par :*

$$\hat{\beta} = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{y} \ = \Phi^{\#}\boldsymbol{y}.$$

Le point $\hat{\beta}$ est appelé l'estimateur des moindres carrés.

Démonstration. C'est une conséquence de l'exercice 1.8. □

Dans le cas $d = 1$, on retrouve bien l'estimateur des moindres carrés donné par la proposition 2.6 (voir exercice 2.1)

Etant donné une nouvelle variable $\boldsymbol{x} \in \mathbb{R}^d$, le prédicteur est :

$$\hat{y}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_d x_d .$$

2.3.3 Et ensuite ?

Le travail du statisticien ne s'arrête pas au calcul de l'estimateur des moindres carrés. On doit répondre aux questions suivantes.

1. **Le modèle linéaire est-il bien fondé ?** Il existe trois façons de répondre à cette question. La plus simple est d'utiliser l'oeil : on trace les nuages de points pour vérifier le comportement linéaire de la réponse en fonction des variables explicatives, on trace les résidus pour s'assurer de leur faible amplitude. La seconde est de calculer R^2 (un R^2 proche de 1 suggère que le modèle est pertinent, mais attention, un R^2 faible ne signifie pas forcément que le modèle est à jeter, il peut simplement indiquer que les données fluctuent beaucoup). La troisième est d'essayer d'autres modèles, et de comparer les erreurs de prédiction.
2. **Y a-t-il des données aberrantes ?** Ces données correspondent à des points pour lesquels le résidu $e_i = y_i - \hat{y}_i$ est anormalement élevé. On peut tracer les résidus pour détecter les valeurs aberrantes, ou utiliser un test permettant de les détecter automatiquement.
3. **Peut-on fournir des intervalles de confiance ?** Dans l'étude de cas de la table 2.1 (ECMS), nous avons calculé $\hat{\beta}_1 = 0.53$, ce qui indique une croissance de la tension en fonction de l'âge avec une pente de 0.53. Mais quelle confiance a-t-on en ce 0.53 ? Peut-être que si l'enquête avait été menée auprès de plus de sujets, avec plus d'âges différents, nous aurions trouvé un autre résultat, comme $\hat{\beta}_1 = 0.6$? Ce qui va compter pour le statisticien, ce n'est pas la valeur de $\hat{\beta}_1$, c'est de pouvoir affirmer que, avec une forte probabilité, la "vraie" pente se trouve dans un certain intervalle $[a, b]$. Un tel intervalle s'appelle un intervalle de confiance. Son calcul nécessite de placer le problème de l'estimation dans un cadre probabiliste, ce que nous n'avons pas fait jusqu'ici. Nous y reviendrons donc plus tard dans ce polycopié, et la notion d'intervalle de confiance deviendra plus claire à ce moment là.
4. **Quelles sont les variables explicatives qui expliquent "réellement" la réponse ?** On a souvent de nombreuses variables explicatives, et toutes ne sont pas forcément utiles. Naïvement, on pourrait dire que si $\hat{\beta}_i$ est proche de zéro, c'est que la $i$ème variable explicative a peu d'influence. Malheureusement, ce raisonnement est trop simpliste car pour répondre proprement, il nous faut aussi la confiance que l'on a dans la valeur du $\hat{\beta}_i$. En fait, formaliser correctement ce problème revient à l'écrire sous la forme d'un test d'hypothèse, permettant de valider ou d'invalider la dépendance en la $i$ème variable. Nous y reviendrons plus tard dans ce cours. Un problème très lié à la question ci-dessus est celui de la *sélection de modèle*. Sur les d variables explicatives de départ, le statisticien cherche souvent à ne conserver qu'un plus petit nombre de variables importantes. Le modèle obtenu est ainsi plus simple, réduit aux variables qui portent réellement l'information utile. En outre, comme vous le verrez en machine learning, les modèles plus simples présentent en général de meilleures propriétés de généralisation, c'est-à-dire que le modèle produira des erreurs de prédiction plus faible quand il sera utilisé sur de nouvelles données.

Dans votre vie professionnelle, lorsque vous aurez à analyser des données, vous devrez vous poser les quatre questions ci-dessus.

2.4 Exercices

Exercice 2.1. Montrer que, dans le cas $d = 1$, l'estimateur des moindres carrés $\hat{\beta} = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{y}$ donné par le théorème 2.12 est bien identique à l'estimateur des moindres carrés donné par la

proposition 2.6.

Exercice 2.2. Dans le chapitre sur la régression linéaire multiple, nous avons supposé que la matrice Φ est de rang $d+1$. Qu'est-ce que cela sous-entend sur le rapport entre le nombre de régresseurs et le nombre N d'observations disponibles ?

Exercice 2.3. Posons $\hat{y}_i = \hat{y}(\boldsymbol{x}_i)$ la prédiction du modèle appliqué au $i$ème échantillon. Appelons $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_d)^T$ le vecteur des prédictions.

1. Montrer que
$$\hat{\boldsymbol{y}} = \Phi\hat{\beta}$$
2. En déduire, en utilisant le théorème 1.3, que $\hat{\boldsymbol{y}} = \Pi_\Phi \boldsymbol{y}$, c'est à dire que $\boldsymbol{y}$ est le projecteur sur $\mathrm{Im}(\Phi)$.
3. En utilisant les propriété de la projection, démontrer l'identité
$$SS_T = SS_M + SS_R\,,$$
où les quantités SS_T, SS_R, SS_M sont définies au paragraphe 2.2.3.
4. (Bonus) On définit le coefficient de détermination par $R^2 = SS_M/SS_T$. C'est la quantité proposée par le statisticien R. Fisher pour quantifier à quel point le modèle explique bien les données observées. Montrer que, si on ajoute un nouveau régresseur au modèle, disons x_{d+1}, le coefficient R^2 va mécaniquement se rapprocher de 1, et ce même si ce nouveau régresseur est sans lien avec la réponse. Quelle mise en garde faut il en déduire ?

Exercice 2.4. On considère le modèle de Volterra suivant :
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$
On considère un N-échantillon (avec $N \geq 3$)

1. Ecrire le modèle sous forme matricielle.
2. Résoudre les moindres carrés. Quelle propriété sur le N-échantillon assure la pseudo-inversion de la matrice. (Indication : regarder les propriétés d'une matrice de Vandermonde).

3 Modèle paramétrique

Dans ce chapitre, nous faisons une avancée majeure : nous allons nous donner un cadre probabiliste pour décrire les données. Dans le chapitre précédent, les données n'étaient que des points. Dans ce chapitre, ce seront des réalisations de certaines variables aléatoires. Cela va ouvrir beaucoup de possibilités, comme, entre autres, le calcul d'intervalles de confiance et le calcul d'erreur d'estimation.

Nous allons commencer par décrire un cadre abstrait, qui fixe le contexte de ce qu'on appelle l'estimation statistique. Nous verrons très vite des exemples qui rendront les choses plus concrètes, mais il faut en passer d'abord par cette abstraction : elle est nécessaire afin de bien comprendre ce que l'on fait, et de ne pas commettre d'erreur profonde de raisonnement. Ce cadre abstrait se nomme le *modèle paramétrique*.

Nous vous conseillons de bien lire le cadre formel. Les exemples qui suivent éclairciront les choses, et vous pourrez relire alors le cadre formel qui vous semblera alors bien moins abstrait.

3.1 Cadre formel

On considère N variables aléatoires réelles $Y_1, \dots, Y_N$, sur un univers Ω, muni d'une probabilité $\mathbb{P}$. L'utilisation d'une lettre majuscule indique bien que ce sont des variables aléatoires (nous utiliserons toujours la convention : majuscule = variable aléatoire, minuscule = variable déterministe). On note :

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}$$

le vecteur aléatoire sur $\mathbb{R}^N$.

On considère un observateur : c'est vous, le statisticien. L'observateur a accès à une *réalisation* $\boldsymbol{Y}(\omega)$ pour une certaine issue ω de l'expérience aléatoire. La loi de $\boldsymbol{Y}$ est inconnue de l'observateur. Toutefois, l'observateur fait une *hypothèse* sur cette loi. Il suppose que la loi fait partie d'une certaine famille, indexée par un certain vecteur $\boldsymbol{\theta} \in \mathbb{R}^d$, que l'on appelle le *paramètre*. Cette famille de lois s'appelle le *modèle* : elle relève du choix du statisticien. L'objectif de l'observateur est alors de déterminer la valeur du paramètre $\boldsymbol{\theta}$ qui explique le mieux possible la réalisation $\boldsymbol{Y}(\omega)$. Cette valeur est de la forme :

$$\hat{\boldsymbol{\theta}}(\omega) = \Theta(\boldsymbol{Y}(\omega)) \,.$$

La variable aléatoire $\hat{\boldsymbol{\theta}}$ s'appelle l'estimée. C'est une v.a. sur $\mathbb{R}^d$. La fonction $\Theta : \mathbb{R}^N \to \mathbb{R}^d$ s'appelle l'estimateur. Cette fonction représente la méthode ou l'algorithme utilisé par le statisticien pour produire l'estimée à partir de son observation $\boldsymbol{Y}(\omega)$.

Cas de variables à densité

Le modèle $\mathcal{P}$ choisi par le statisticien consiste en une famille de densités :

$$\mathcal{P} = \{p_{\boldsymbol{\theta}} \,:\, \boldsymbol{\theta} \in \mathbb{R}^d\}$$

où pour chaque $\boldsymbol{\theta}$ possible, $p_{\boldsymbol{\theta}}$ est une densité de probabilité sur $\mathbb{R}^N$. La plupart du temps (en tout cas, ce sera toujours le cas dans ce cours), le statisticien fait l'hypothèse que les observations $Y_1, \dots, Y_N$ sont indépendantes. Autrement dit, chaque densité $p_{\boldsymbol{\theta}}$ sera toujours supposée s'écrire comme le produit des densités marginales :

$$\forall (y_1, \dots, y_N) \in \mathbb{R}^N,\ p_{\boldsymbol{\theta}}(y_1, \dots, y_N) = p_{1,\boldsymbol{\theta}}(y_1) \times \cdots \times p_{N,\boldsymbol{\theta}}(y_N)\,,$$

où on a appelé $p_{1,\boldsymbol{\theta}}, \dots, p_{N,\boldsymbol{\theta}}$ les densités marginales (supposées !) de $Y_1, \dots, Y_N$ respectivement. Choisir un modèle $\mathcal{P}$ revient donc à choisir N densités de probabilités $p_{1,\boldsymbol{\theta}}, \dots, p_{N,\boldsymbol{\theta}}$, qui dépendent d'un certain paramètre $\boldsymbol{\theta}$.

Cas de variables discrètes

Si les variables aléatoires $Y_1, \dots, Y_N$ sont à valeurs dans $\mathbb{N}$, le statisticien se donne pour modèle une certaine famille de lois sur $\mathbb{N}^N$:

$$\mathcal{P} = \{p_{\boldsymbol{\theta}} \,:\, \boldsymbol{\theta} \in \mathbb{R}^d\}$$

où cette fois $p_{\boldsymbol{\theta}}$ n'est plus une densité de probabilité, mais une loi discrète de la forme :

$$\forall (k_1, \dots, k_N) \in \mathbb{N}^N,\ p_{\boldsymbol{\theta}}(k_1, \dots, k_N) = p_{1,\boldsymbol{\theta}}(k_1) \times \cdots \times p_{N,\boldsymbol{\theta}}(k_N)\,,$$

où $p_{1,\boldsymbol{\theta}}, \dots, p_{N,\boldsymbol{\theta}}$ sont des lois sur $\mathbb{N}$. On a raccourci la notation $p_{\boldsymbol{\theta}}(y_1 = k_1, \dots, y_N = k_N)$ en $p_{\boldsymbol{\theta}}(k_1, \dots, k_N)$ avec les k_n appartenant à l'ensemble discret des valeurs possibles pour bien montrer son aspect discret.

Remarque 3.1. Soyons rigoureux avec les notations ! Ne confondons pas :

- $\boldsymbol{Y} = (Y_1, \dots, Y_N)^T$ la variable aléatoire ;
- $\boldsymbol{Y}(\omega) = (Y_1(\omega) \dots, Y_N(\omega))^T$ la réalisation qui est effectivement observée par le statisticien lors d'une certaine expérience aléatoire ;
- $\boldsymbol{y} = (y_1, \dots, y_N)^T$ qui est une variable muette, un point de $\mathbb{R}^N$.

Remarque 3.2. Nous utilisons la même notation $p_{\boldsymbol{\theta}}$ dans le cas à densité et dans le cas discret. Mais il faut bien comprendre que dans le premier cas, les fonctions $p_{1,\boldsymbol{\theta}}(y), \dots, p_{N,\boldsymbol{\theta}}(y)$ sont des densités de probabilité sur $\mathbb{R}$, comme par exemple des densités gaussiennes. Alors que dans le second cas, les fonctions $p_{1,\boldsymbol{\theta}}(k), \dots, p_{N,\boldsymbol{\theta}}(k)$ sont des lois sur $\mathbb{N}$, comme par exemple des lois de Bernoulli ou de Poisson.

Dans certains cas de figures, on peut imposer une contrainte sur le paramètre $\boldsymbol{\theta}$, en stipulant que $\boldsymbol{\theta}$ est seulement autorisé à vivre dans une certaine région $D \subset \mathbb{R}^d$. Par exemple, $D = [0, +\infty)^d$ si on a une contrainte de positivité. En toute généralité, on a donc finalement la définition suivante.

Définition 3.1 (Modèle paramétrique)**.** Un modèle paramétrique est une famille

$$\mathcal{P} = \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in D\}$$

où D est un sous-ensemble de $\mathbb{R}^d$, et où pour tout $\boldsymbol{\theta} \in D$, $p_{\boldsymbol{\theta}}$ est une densité sur $\mathbb{R}^N$ (dans le cas à densité) ou une loi discrète sur $\mathbb{N}^N$ (dans le cas discret). L'entier d est la *dimension* du paramètre $\boldsymbol{\theta}$.

Pour chaque valeur de $\boldsymbol{\theta}$, on introduit une probabilité $\mathbb{P}_{\boldsymbol{\theta}}$ sur l'univers Ω telle que, sous la probabilité $\mathbb{P}_{\boldsymbol{\theta}}$, le vecteur aléatoire $\boldsymbol{Y}$ est de loi $p_{\boldsymbol{\theta}}$. Par exemple, dans le cas à densité, pour tout $H \subset \mathbb{R}^N$,

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} \in H) = \int \cdots \int_H p_{\boldsymbol{\theta}}(y_1, \ldots, y_N) dy_1 \ldots dy_N \,.$$

On note de même $\mathbb{E}_{\boldsymbol{\theta}}$ l'espérance associée, c'est à dire que pour une fonction $g : \mathbb{R}^N \to \mathbb{R}$ quelconque :

$$\mathbb{E}_{\boldsymbol{\theta}}(g(\boldsymbol{Y})) = \int \cdots \int g(y_1, \ldots, y_N) p_{\boldsymbol{\theta}}(y_1, \ldots, y_N) dy_1 \ldots dy_N \,.$$

Remarque 3.3. La probabilité $\mathbb{P}$ que nous nous sommes fixés au début, détermine la "vraie" loi des observations $\boldsymbol{Y}$. Cette vraie loi est inconnue (on ne sait rien de $\mathbb{P}$, la seule chose que nous connaissons, c'est le N-échantillon). En se donnant un modèle paramétrique, le statisticien se donne une infinité de mesures possibles, toutes de la forme $\mathbb{P}_{\boldsymbol{\theta}}$, où $\boldsymbol{\theta}$ décrit un ensemble D. Il va chercher la "vraie" loi au sein de la famille de lois qu'il s'est donné. Ainsi, si le statisticien choisit un certain modèle paramétrique, c'est parce qu'il a des raisons de penser que la "vraie" loi des observations $\boldsymbol{Y}$ appartient à la famille $\mathcal{P}$, ou en tout cas en est proche. Ces raisons proviennent de son inspection préalable des données, de son travail de visualisation des données. Si le statisticien a choisi un modèle paramétrique raisonnable, on peut alors supposer qu'il existe une certaine valeur $\boldsymbol{\theta}^*$, inconnue, telle que $\mathbb{P} = \mathbb{P}_{\boldsymbol{\theta}^*}$, c'est à dire que $\boldsymbol{Y}$ suit la loi $p_{\boldsymbol{\theta}^*}$.

Dans la suite, nous allons étudier deux cas de modèles très répandus.

3.2 Modèle de Bernoulli

3.2.1 Etude de cas

Afin d'estimer la part de fumeurs dans la population française, le ministère de la santé effectue un sondage auprès d'un échantillon de $N = 500$ personnes (voir la table 3.1).

3.2.2 Modèle paramétrique

Le statisticien formalise le problème de la manière suivante. La population française est divisée en deux classes, la classe "fumeur" des individus qui se disent fumeurs, et la classe "non-fumeur".

Indice	Réponse
1	0
2	0
3	0
4	1
5	0
⋮	⋮
499	0
500	1

TABLE 3.1 – Résultat d'un sondage. L'indice représente la personne interrogée, la valeur 0 indique la réponse "non" à la question "Etes-vous fumeur/fumeuse ?", la valeur 1 indique la réponse "oui".

Appelons θ^* le rapport entre le nombre de fumeurs divisé par la taille N_{pop} de la population. L'objectif est de déterminer θ^*.

Un sondage consiste à choisir N individus au sein de la population, et à mesurer leur réponse $0/1$ à la question posée. Le statisticien fait l'hypothèse que l'institut de sondage a choisi ces individus de manière aléatoire, indépendante et uniforme au sein de la population.

Remarque 3.4. L'hypothèse d'un échantillonnage iid uniforme pourrait être discutée. Imaginons par exemple que l'institut de sondage parisien ait, par facilité, interrogé des personnes de la région parisienne ? Quel problème cela peut-il poser ?

Le statisticien modélise les réponses des sondés comme étant une réalisation de N variables aléatoires $Y_1, \dots, Y_N$. Si l'hypothèse d'un échantillonnage iid uniforme est satisfaite, la probabilité que la réponse Y_i du $i$ème individu soit 1 ("oui") est égale à la probabilité que cet individu ait été choisi au sein de la classe "fumeur", soit θ^*.

Le modèle paramétrique naturel est donc une famille de lois de Bernoulli de paramètre $\theta \in [0,1]$. Autrement dit :

$$p_{1,\theta}(k) = \cdots = p_{N,\theta}(k) = \begin{cases} \theta & \text{si } k = 1 \\ 1-\theta & \text{si } k = 0\,. \end{cases} \tag{3.1}$$

Le modèle paramétrique est donc :

$$\mathcal{P} = \{k \mapsto \theta^k (1-\theta)^{1-k} \ : \ \theta \in [0,1]\}\,.$$

Le paramètre θ est ici un scalaire, c'est pourquoi nous ne l'écrivons pas en gras.

3.2.3 Estimateur de la moyenne empirique

Etant donné l'observation des variables $Y_1, \dots, Y_N$ donnée par la table 3.1, le statisticien définit naturellement l'estimée du paramètre θ par :

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} Y_i\,.$$

Afin d'alléger les notations, nous ne mettons pas la dépendance à l'issue ω lorsqu'il n'y pas d'équivoque. Le paramètre θ représente la probabilité qu'une personne réponde positivement au sondage, l'estimée $\hat{\theta}$ est le nombre moyen de personnes ayant répondu positivement, calculée sur le N-échantillon. Autrement dit, l'estimée est la fréquence de réponses positives au sondage.

Pourquoi cet estimateur plutôt qu'un autre ? Certes, nous pourrions justifier théoriquement que la moyenne empirique est un bon estimateur, et même qu'il est optimal, au sens d'un certain critère d'erreur quadratique que nous définirons plus bas. Mais à notre stade, une telle justification est superfétatoire. Un tel estimateur provient tout simplement du bon sens. Après tout, que pourrait on raisonnablement faire d'autre ?

Biais

Le biais de l'estimateur est défini, pour chaque valeur de θ, par :

$$b(\theta) = \mathbb{E}_\theta(\hat{\theta}) - \theta\,.$$

L'espérance $\mathbb{E}_\theta(\hat{\theta})$ est l'espérance de l'estimée $\hat{\theta}$, sous l'hypothèse que les observations sont distribuées selon la loi $p_{\boldsymbol{\theta}}$. Dans le cas présent :

$$\begin{aligned}\mathbb{E}_\theta(\hat{\theta}) &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\theta(Y_i)\\ &= \mathbb{E}_\theta(Y_1)\\ &= \theta\,,\end{aligned}$$

car, sous $\mathbb{P}_\theta$, les v.a. Y_i sont identiquement distribuées selon la loi de Bernoulli de paramètre θ. On remarque donc que :

$$b(\theta) = 0$$

quelque soit la valeur de θ. On dit que l'estimateur est *sans biais*. Cela veut juste dire que si on traçait la densité de probabilité de $\hat{\theta}$ (ce qui n'est pas possible car il faudrait une infinité d'observations ; en pratique on n'a accès qu'à une estimée de cette densité via un histogramme), elle serait centrée autour de la vraie valeur recherchée.

Variance et erreur quadratique moyenne

La variance de l'estimateur sous $\mathbb{P}_\theta$ est donnée par :

$$\mathrm{Var}_\theta(\hat{\theta}) = \mathbb{E}_\theta((\hat{\theta} - \mathbb{E}_\theta(\hat{\theta}))^2)\,.$$

Sous $\mathbb{P}_\theta$, la v.a. $\sum_{i=1}^{N} Y_i$ suit une loi binomiale de paramètres (θ, N). La variance est donnée par $N\theta(1-\theta)$. Ainsi la variance de l'estimateur est $\mathrm{Var}_\theta(\hat{\theta}) = \theta(1-\theta)/N$.

L'*erreur quadratique moyenne* (EQM) de l'estimateur est définie par

$$EQM_\theta = \mathbb{E}_\theta((\hat{\theta} - \theta)^2) = \mathrm{Var}_\theta(\hat{\theta}) + b(\theta)^2\,.$$

Naturellement, lorsque l'estimateur est non-biaisé (comme c'est le cas ici), l'erreur quadratique moyenne coïncide avec la variance de l'estimateur. On a donc :

$$EQM_\theta = \frac{\theta(1-\theta)}{N}\,.$$

L'EQM est un critère de performance de l'estimateur. Surtout, l'EQM permet de comparer deux estimateurs entre eux : on préférera utiliser l'estimateur dont l'EQM est plus petite pour tout θ.

3.2.4 Intervalle de confiance

A partir du sondage, on ne peut évidemment pas répondre à la question "Quelle est la part θ^* de fumeurs dans la population ?". On peut toutefois fournir une réponse statistique :

Grâce au sondage, on peut affirmer qu'avec une probabilité de 0.95, que la part de θ^* de fumeurs est comprise entre 30.5 et 31.5 pourcents.

Autrement dit, on répond à la question en donnant non pas une estimée $\hat\theta$ de θ^*, mais un *intervalle* qui contient θ^*, avec forte probabilité, choisie arbitrairement comme étant 0.95 (les valeurs les plus courantes sont 0.95 et 0.99). Cet intervalle $[30.5, 31.5]$ s'appelle un *intervalle de confiance de niveau 95%.*

Une astuce pour construire un intervalle de confiance à 95%, est de considérer l'estimateur $\hat\theta$. En utilisant l'inégalité de Bienaymé-Chebychev (1.4), on a pour tout $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}_\theta(|\hat\theta - \theta| > \epsilon) &\leq \frac{\mathrm{Var}_\theta(\hat\theta)}{\epsilon^2}\\ &= \frac{\theta(1-\theta)}{N\epsilon^2}\\ &\leq \frac{1}{4N\epsilon^2}\end{aligned}$$

où on a utilisé que $\theta(1-\theta) \leq 1/4$ pour tout $\theta \in [0,1]$. On choisit ϵ de telle sorte que $\frac{1}{4N\epsilon^2} = 0.05$, soit $\epsilon = \frac{1}{\sqrt{0.2N}}$. On obtient :

$$\mathbb{P}_\theta(|\hat\theta - \theta| \leq \frac{1}{\sqrt{0.2N}}) \geq 0.95\,.$$

En particulier, pour $\theta = \theta^*$ (la vraie part de fumeurs dans la population), on obtient :

$$\mathbb{P}_{\theta^*}\left(\theta^* \in \left[\hat\theta - \frac{1}{\sqrt{0.2N}}, \hat\theta + \frac{1}{\sqrt{0.2N}}\right]\right) \geq 0.95\,. \tag{3.2}$$

L'intervalle $\left[\hat\theta - \frac{1}{\sqrt{0.2N}}, \hat\theta + \frac{1}{\sqrt{0.2N}}\right]$ est un intervalle aléatoire. Ses bornes dépendent des observations $Y_1, \ldots, Y_N$. On l'appelle intervalle de confiance de niveau 0.95%. Autrement, sous l'hypothèse que la vraie loi des données est bien p_{θ^*} (c'est à dire sous l'hypothèse que le choix des sondés est iid uniforme dans la population), on peut affirmer que θ^* est dans l'intervalle en question, avec une probabilité de se tromper inférieure à 0.05.

x_1	x_2	x_3	x_4	x_5	Y
1.1250	0.2320	7.1600	0.0859	8.9050	0.0016
0.9200	0.2680	8.8040	0.0865	7.3880	0.0009
0.8350	0.2710	8.1080	0.0852	5.3480	0.0007
1.0000	0.2370	6.3700	0.0838	8.0560	0.0007
1.1500	0.1920	6.4410	0.0821	6.9600	0.0003
0.9900	0.2020	5.1540	0.0792	5.6900	0.0004
0.8400	0.1840	5.8960	0.0812	6.9320	0.0001
0.6500	0.2000	5.3360	0.0806	5.4000	0.0001
0.6400	0.1800	5.0410	0.0784	3.1770	-0.0002
0.5830	0.1650	5.0120	0.0793	4.4610	-0.0002
0.5700	0.1510	4.8250	0.0787	3.9010	0
0.5700	0.1710	4.3910	0.0780	5.0020	0
0.5100	0.2430	4.3200	0.0723	4.6650	-0.0001
0.5550	0.1470	3.7090	0.0749	4.6420	-0.0002
0.4600	0.2860	3.9690	0.0744	4.8400	-0.0004
0.2750	0.1980	3.5580	0.0725	4.4790	-0.0002
0.5100	0.1960	4.3610	0.0577	4.2000	-0.0002
0.1650	0.2100	3.3010	0.0718	3.4100	-0.0004
0.2440	0.3270	2.9640	0.0725	3.3600	-0.0005
0.0790	0.3340	2.7770	0.0719	2.5990	-0.0000

TABLE 3.2 – Données de Moore

Quand le statisticien fournit un intervalle de confiance, il cherche naturellement l'intervalle le plus court possible (on dit "le plus exact"). Or l'inégalité de Bienaymé-Chebychev que nous avons utilisée pour calculer l'intervalle de confiance est assez grossière. Les exercices 3.1 et 3.2 montrent que l'on peut calculer des intervalles de confiance plus exacts, en utilisant des inégalités plus fines.

3.3 Modèle linéaire gaussien

3.3.1 Etude de cas

Nous utilisons ici des résultats d'une expérience menée par Moore (1975) et analysée par Chatterjee et Hadi (1986). Ces données ont été collectées dans un bio-réacteur, pendant une période de 220 jours. Les données sont reproduites dans le tableau 3.2. Les variables mesurées sont : Y = log(demande d'oxygène) (g/min) ; x_1 = demande d'oxygène biologique (g/litre). x_2 = quantité totale d'azote, g/litre ; x_3 = quantité totale de matière solide, g/litre, x_4 = quantité totale de solides volatils, g/litre ; et x_5 =, demande chimique d'oxygène, g/litre. L'objectif est comme d'habitude d'expliquer la réponse en fonction des *régresseurs* $x_1, \ldots, x_5$. Pour cela, nous allons effectuer une régression linéaire multiple, comme nous avons appris à le faire au chapitre précédent.

Mais, afin d'aller plus loin dans l'interprétation des résultats, nous allons faire une hypothèse probabiliste sur les données. Autrement dit, nous allons commencer par nous fixer un modèle

paramétrique $\mathcal{P}$.

Afin de nous placer dans le cadre formel défini au paragraphe 3.1, nous allons donc supposer que le vecteur des réponses observées est une *réalisation* d'un vecteur aléatoire $\boldsymbol{Y} = (Y_1, \dots, Y_N)^T$, soit :

$$\boldsymbol{Y}(\omega) = \begin{pmatrix} Y_1(\omega) \\ Y_2(\omega) \\ \vdots \\ Y_{20}(\omega) \end{pmatrix} = \begin{pmatrix} 0.0016 \\ 0.0009 \\ \vdots \\ -0.0000 \end{pmatrix} .$$

Mais comme on le voit dans la table 3.2, l'observation ne se limite pas aux réponses. On observe également, pour chacune des vingt mesures, les valeurs de 5 régresseurs. Nous posons :

$$\begin{aligned} \boldsymbol{x}_1^T &= (1.1250, 0.2320, 7.1600, 0.0859, 8.9050) \\ &\vdots \\ \boldsymbol{x}_{20}^T &= (0.0790, 0.3340, 2.7770, 0.0719, 2.5990) . \end{aligned}$$

Le jeu de données de la table 3.2 est donc compris par le statisticien comme étant de la forme :

$$\{(\boldsymbol{x}_1, Y_1(\omega)), \dots, (\boldsymbol{x}_N, Y_N(\omega))\} ,$$

où ω est l'issue de l'expérience aléatoire, où pour tout i, $\boldsymbol{x}_i$ est un vecteur déterministe de $\mathbb{R}^5$, et Y_i est un variable aléatoire sur $\mathbb{R}$, et où $N = 20$.

Remarque 3.5. Insistons sur les notations. La notation $\boldsymbol{x}_i$ est en gras, donc il s'agit d'un vecteur, et en minuscule, donc il s'agit d'une quantité déterministe, qui ne dépend pas de ω. La notation Y_i est en majuscules, c'est donc une variable aléatoire, et n'est pas en gras, donc les valeurs sont scalaires.

Remarque 3.6. Ainsi, la différence majeure de ce chapitre par rapport au chapitre 2 est que les réponses sont supposées être des réalisations de variables aléatoires. Par comparaison au chapitre 2, ce choix de modèle apporte de nouvelles perspectives en termes de construction d'estimateurs et d'interprétation des résultats.

Il est temps de fixer notre choix de modèle paramétrique $\mathcal{P}$.

3.3.2 Modèle homoscédastique

Considérons un ensemble de couples $(\boldsymbol{x}_i, Y_i)$ pour $i = 1, \dots, N$, où $\boldsymbol{x}_i \in \mathbb{R}^d$ et où Y_i est un v.a.r. Les entrées de chaque vecteur $\boldsymbol{x}_i$ seront notées, comme au chapitre précédent, $\boldsymbol{x}_i = (x_{i,1}, \dots, x_{i,d})^T$. On suppose dorénavant que $N > d + 1$.

Le modèle homoscédastique consiste à faire l'hypothèse que les réponses Y_i observées correspondent à une fonction affine du vecteur $\boldsymbol{x}_i$, à laquelle s'ajoute une perturbation gaussienne. Autrement dit :

$$\forall i,\ Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_d x_{i,d} + \varepsilon_i , \tag{3.3}$$

où $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ est une v.a. gaussienne centrée d'espérance nulle et de variance σ^2. On suppose en outre que les variables $\varepsilon_1, \dots, \varepsilon_N$ sont indépendantes. Dans ce modèle, les v.a. $Y_1, \dots, Y_N$ sont

donc indépendantes, mais pas identiquement distribuées. Le vecteur $\boldsymbol{Y} = (Y_1, \dots, Y_N)^T$ s'écrit ainsi :

$$\boldsymbol{Y} = \Phi\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,,$$

où $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_d)^T$, où $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_N)^T$ et où Φ est la matrice définie par (2.1), et dont nous rappelons l'expression :

$$\Phi = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,d} \\ \vdots & & & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,d} \end{pmatrix}.$$

Par conséquent, $\boldsymbol{Y}$ est un vecteur gaussien d'espérance $\Phi\boldsymbol{\beta}$ et de matrice de covariance $\sigma^2 I_N$, où I_N est l'identité de taille N :

$$\boldsymbol{Y} \sim \mathcal{N}(\Phi\boldsymbol{\beta}, \sigma^2 I_N)\,.$$

Dans cette modélisation, la loi de $\boldsymbol{Y}$ dépend de $d+2$ paramètres : $\beta_0, \dots, \beta_d$ et la variance σ^2. Le vecteur final constituant l'ensemble des paramètres scalaires inconnus est donc :

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \\ \sigma^2 \end{pmatrix}.$$

La densité de probabilité de $\boldsymbol{Y}$ est donnée, pour tout $\boldsymbol{y} \in \mathbb{R}^N$, par :

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}) = \frac{1}{(\sqrt{2\pi\sigma^2})^N} \exp\left(-\tfrac{1}{2\sigma^2}\|\boldsymbol{y} - \Phi\boldsymbol{\beta}\|^2\right). \tag{3.4}$$

Le paramètre $\boldsymbol{\theta}$ décrit l'ensemble $D = \mathbb{R}^{d+1} \times (0, +\infty)$. Le modèle paramétrique est donc finalement donné par :

$$\mathcal{P} = \{p_{\boldsymbol{\theta}} \,:\, \boldsymbol{\theta} \in D\}\,.$$

Remarque 3.7. La variable ε_i est une perturbation stochastique (un peu abusivement, on parle parfois de *bruit*) qui caractérise le fait que les réponses ne s'écrivent pas exactement comme une fonction affine des régresseurs. Le terme homoscédastique traduit le fait que les ϵ_i sont supposés avoir tous la même variance, contrairement à un modèle dit hétéroscédastique, où les variances pourraient dépendre de i.

3.3.3 Estimateur

Le vecteur de paramètres à estimer se décompose en $\boldsymbol{\beta}$ et σ^2. Le paramètre d'intérêt consiste surtout en les coefficients $\beta_1, \dots, \beta_d$ qui vont révéler l'influence des différents régresseurs sur la réponse. Nous choisissons naturellement l'estimateur des moindres carrés :

$$\hat{\boldsymbol{\beta}} = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{Y}\,.$$

Théorème 3.2 (Gauss-Markov)**.** *Pour tout $\boldsymbol{\beta}, \sigma^2$, dans le cadre du modèle homoscédastique, on a*

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \frac{\sigma^2}{N}(\Phi^T\Phi)^{-1}).$$

En particulier, l'estimateur des moindres carrés est non biaisé, c'est à dire que $\mathbb{E}_{\boldsymbol{\beta},\sigma^2}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$*, et l'erreur quadratique moyenne sur le paramètre* β_k *est :*

$$\mathbb{E}_{\boldsymbol{\beta},\sigma^2}[(\hat{\beta}_k - \beta_k)^2] = \frac{\sigma^2}{N}\left[(\Phi^T\Phi)^{-1}\right]_{k,k}.$$

Enfin, si $\tilde{\boldsymbol{\beta}}$ *est un autre estimateur sans biais de* $\boldsymbol{\beta}$*, alors on a forcément :*

$$\mathbb{E}_{\boldsymbol{\beta},\sigma^2}[(\tilde{\beta}_k - \beta_k)^2] \geq \mathbb{E}_{\boldsymbol{\beta},\sigma^2}[(\hat{\beta}_k - \beta_k)^2]$$

pour tout $k = 0, \ldots, d$*. Autrement dit, l'estimateur des moindres carrés est celui qui minimise l'erreur quadratique moyenne, parmi tous les estimateurs non biaisés.*

Démonstration. En classe. □

3.3.4 Intervalle de confiance

Soit $k = 0, \ldots, d$ fixé. Nous voulons fournir un intervalle de confiance à 95% sur le paramètre β_k. Pour rappel, il s'agit d'un intervalle $I(\boldsymbol{Y})$ dont les extrémités sont des variables aléatoires, qui dépendent des observations, et qui sous $\mathbb{P}_{\boldsymbol{\beta},\sigma^2}$, contient β_k avec une probabilité au moins égale à 0.95.

Première approche

Nous allons commencer par une approche simple qui permet de comprendre le mécanisme de la construction d'un intervalle de confiance. Nous affinerons cette approche dans un second temps.

La technique pour déterminer un tel intervalle consiste à inspecter la loi de l'estimateur $\hat{\beta}_k$. Plaçons-nous sous la loi $\mathbb{P}_{\beta,\sigma^2}$ qui est gaussienne en raison du modèle homoscédastique. D'après le théorème 3.2,

$$\hat{\beta}_k \sim \mathcal{N}(\beta_k, \frac{\sigma^2 s_k}{N})$$

où s_k est défini comme le $k$ème coefficient de la diagonale de $(\Phi^T\Phi)^{-1}$. On peut se ramener à une loi normale centrée réduite par :

$$\sqrt{\frac{N}{\sigma^2 s_k}}(\beta_k - \hat{\beta}_k) \sim \mathcal{N}(0,1). \tag{3.5}$$

Définissons par $F(x)$ la fonction de répartition de la loi $\mathcal{N}(0,1)$, soit :

$$F(x) := \int_{-\infty}^{x} \tfrac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

D'après l'équation (3.5), on a pour tout $a > 0$,

$$\begin{aligned} F(a) - F(-a) &= \mathbb{P}_{\boldsymbol{\beta},\sigma^2}\left(-a \leq \sqrt{\frac{N}{\sigma^2 s_k}}(\beta_k - \hat{\beta}_k) \leq a\right) \\ &= \mathbb{P}_{\boldsymbol{\beta},\sigma^2}\left(\hat{\beta}_k - a\sqrt{\frac{\sigma^2 s_k}{N}} \leq \beta_k \leq \hat{\beta}_k + a\sqrt{\frac{\sigma^2 s_k}{N}}\right) \end{aligned}$$

Choisissons a pour que $F(a) - F(-a) = 0.95$. Si Z est une gaussienne centrée réduite, rappelons que

$$F(a) = \mathbb{P}(Z \leq a) = \mathbb{P}(-Z \leq a) = \mathbb{P}(Z \geq -a) = 1 - F(-a)\,.$$

Par conséquence, on doit choisir a pour que $F(a) = 0.975$. Autrement dit, a est le quantile de niveau 0.975 de la gaussienne centrée réduite. Tout statisticien aguerri sait que ce quantile vaut $a = 1.96$. Nous avons donc démontré que :

$$\mathbb{P}_{\boldsymbol{\beta},\sigma^2}\left(\beta_k \in \hat{\beta}_k \pm 1.96\sqrt{\frac{\sigma^2 s_k}{N}}\right) = 0.95\,.$$

L'intervalle $\hat{\beta}_k \pm 1.96\sqrt{\frac{\sigma^2 s_k}{N}}$ contient donc le paramètre inconnu β_k avec probabilité 0.95. Cet intervalle dépend des données au travers de l'estimée $\hat{\beta}_k$ et au travers du coefficient s_k.

Malheureusement, cet intervalle ne peut pas être considéré comme un intervalle de confiance, car il dépend du paramètre inconnu σ^2. Tout espoir n'est pas perdu : il nous suffit d'estimer σ^2.

Deuxième approche

On introduit l'estimateur :

$$\hat{\sigma}^2 = \frac{\|Y - \Phi\hat{\boldsymbol{\beta}}\|^2}{N - d - 1}\,.$$

On se souvient que le vecteur $\Phi\hat{\boldsymbol{\beta}}$ est le vecteur des prédictions, et donc la différence $Y - \Phi\hat{\boldsymbol{\beta}}$ est le vecteur des résidus. La norme au carré $\|\boldsymbol{Y} - \Phi\hat{\boldsymbol{\beta}}\|^2$ est donc égale au SS_R. Enfin, la division par $N - d - 1$ plutôt que par N assure que $\hat{\sigma}^2$ est un estimateur non biaisé, comme le montre le résultat suivant, c'est à dire que $\mathbb{E}_{\boldsymbol{\beta},\sigma^2}[\hat{\sigma}^2] = \sigma^2$ pour tout $\boldsymbol{\beta}, \sigma^2$.

On rappelle les définitions de la table 1.2. La loi du chi-deux à k degrés de libertés, notée $\chi^2(k)$, est la loi de la somme des carrés de k variables iid gaussiennes centrées réduites. La loi de Student à k degrés de libertés, notée $\mathcal{T}(k)$ est la loi du rapport $\frac{Z}{\sqrt{U/k}}$ où Z est une gaussienne centrée réduite, U suit un chi-deux à k degrés de libertés, indépendante de Z. Le résultat suivant est admis.

Lemme 3.3 (Lemme de Cochran)**.** *Sous $\mathbb{P}_{\boldsymbol{\beta},\sigma^2}$ du modèle homoscédastique, la v.a. $\hat{\sigma}^2$ est indépendante de $\hat{\boldsymbol{\beta}}$, et sa loi est caractérisée par :*

$$(N - d - 1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(N - d - 1)$$

En outre, pour tout $k = 0, \ldots, d$,

$$\sqrt{\frac{N}{s_k\hat{\sigma}^2}}(\hat{\beta}_k - \beta_k) \sim \mathcal{T}(N - d - 1)\,.$$

Ainsi, on peut reprendre le même raisonnement qu'au début du paragraphe, d'une part en remplaçant la variance σ^2 inconnue par son estimée $\hat{\sigma}^2$, et d'autre part en remplaçant la loi $\mathcal{N}(0,1)$ par la loi $\mathcal{T}(N - d - 1)$. On a donc démontré le résultat suivant.

Théorème 3.4 (Intervalle de confiance). *Soit $k = 0, \dots, d$. Soit $\alpha \in (0,1)$. Soit q_α le quantile de niveau $(1-\alpha/2)$ de la loi de Student $\mathcal{T}(N-d-1)$. Alors, sous $\mathbb{P}_{\boldsymbol{\beta},\sigma^2}$ du modèle homoscédastique, l'intervalle*

$$\hat{\beta}_k \pm q_\alpha \sqrt{\frac{\hat{\sigma}^2 s_k}{N}}$$

est un intervalle de confiance sur β_k de niveau $1-\alpha$.

3.3.5 Interprétation du modèle

Une fois les intervalles de confiances calculés, la question que l'on se pose est la suivante :

Le kème régresseur x_k contribue-t-il à expliquer la réponse Y ?

Cette question est essentielle pour l'*interprétation* du modèle. En supposant que les données suivent le modèle homoscédastique (3.3) pour certains paramètres inconnus $\boldsymbol{\beta}, \sigma^2$, la question une fois formalisée devient :

Est-ce que le coefficient β_k est non-nul ?

La réponse du statisticien ne sera pas "oui" ou "non" : il répondra en terme de niveau de confiance.

Exemple 3.5. On suppose que l'estimateur des moindres carrés appliqué à un certain modèle conduit aux résultats de la figure 3.3.

k	$\hat{\beta}_k$	Intervalle de confiance à 95%
1	0.01	[0.099, 0.011]
2	32.5	[-127, 188]

TABLE 3.3 – Estimées moindres-carrés et intervalles de confiances (exemple hypothétique).

On observe que $\hat{\beta}_1 = 0.01$ est petit, et que $\hat{\beta}_2 = 32.5$ est plus grand. Ces seules valeurs ne permettent pas de conclure : elles sont peut-être très incertaines ; elles peuvent aussi être dues à la nature des régresseurs x_1 et x_2 qui peuvent être d'amplitudes très différentes, si les régresseurs n'ont pas été préalablement normalisés. Une meilleure approche consiste à inspecter les intervalles de confiance. Avec probabilité 95%, on sait que $\beta_1 \in [0.099, 0.011]$, or cet intervalle ne contient pas zéro. On peut donc conclure avec au moins 95% de chance d'avoir raison, que β_1 n'est pas nul. Il n'en va pas de même pour le second régresseur : l'intervalle de confiance contient l'origine, donc on ne peut pas conclure à la significativité du second régresseur dans l'explication de la réponse.

p-valeur

Sous l'hypothèse que $\beta_k = 0$, le lemme 3.3 implique que $\sqrt{\frac{N}{s_k \hat{\sigma}^2}}\hat{\beta}_k$ suit une loi de Student à $N-d-1$ degrés de liberté. Le carré d'une v.a. de Student suit ce que l'on appelle une loi de

Fisher $\mathcal{F}(1, N-d-1)$ à $N-d-1$ degrés de liberté. La densité de Fisher est donnée dans la table 1.2, mais nous nous intéressons surtout à sa fonction de répartition complémentaire que l'on note

$$\bar{F}(x) = \mathbb{P}(F > x) \quad \text{où } F \sim \mathcal{F}(1, N-d-1)\,.$$

L'expression exacte de $\bar{F}$ est un peu alambiquée, inutile donc de l'écrire, il faut simplement retenir que cette fonction $\bar{F}$ existe, et qu'elle est disponible dans tout bon logiciel ou librairie de statistique.

Ainsi, si nous posons :

$$F_k := \frac{N}{s_k \hat{\sigma}^2} \hat{\beta}_k^2\,,$$

nous pouvons affirmer que, sous l'hypothèse que $\beta_k = 0$, F_k suit la loi $\mathcal{F}(1, N-d-1)$. Pour savoir si l'hypothèse $\beta_k = 0$ est plausible ou si elle ne l'est pas, il suffit de comparer la valeur de F_k effectivement calculée à la distribution de Fisher. Est-il vraisemblable que ce F_k soit une réalisation de la loi $\mathcal{F}(1, N-d-1)$? La quantité

$$p_k := \bar{F}\left(F_k\right)$$

est appelée la *p-valeur associée à l'hypothèse* $\beta_k = 0$. Il s'agit de la probabilité qu'une variable suivant une loi de Fisher soit au moins aussi grande que la valeur F_k observée. Par exemple, si p_k vaut 0.5, cela signifie qu'en tirant une variable selon $\mathcal{F}(1, N-d-1)$, on a une chance sur deux d'observer un résultat au moins aussi grand que F_k. Dans ce cas, il est tout à fait plausible que $\beta_k = 0$, cette hypothèse n'est nullement contredite par la valeur F_k observée. Si au contraire p_k vaut 0.001, cela signifie qu'en tirant une variable selon $\mathcal{F}(1, N-d-1)$, nous aurions une chance sur mille d'obtenir un résultat aussi grand que F_k. Dans ce cas, l'hypothèse $\beta_k = 0$ est très improbable.

En conclusion, la p-valeur quantifie notre croyance en le fait que le $k$ème régresseur contribue ou non à expliquer la réponse. Plus précisément, une p-valeur faible donne confiance en l'hypothèse $\beta_k \neq 0$, alors qu'une p-valeur de l'ordre de 0.5 donne confiance en l'hypothèse $\beta_k = 0$.

3.4 Modèle général

Nous venons de voir deux modèles pour lesquels nous avons donné un estimateur raisonnable (et même optimal dans un certain sens dans le cas linéaire gaussien). Dans un cadre général, pouvons-nous définir des estimateurs raisonnables et savons-nous décrire les performances (au sens de la variance d'erreur) indépassables comme énoncé dans le théorème de Gauss-Markov ?

3.4.1 Estimateur du maximum de vraisemblance

On définit l'estimateur du *maximum de vraisemblance* (Maximum Likelihood -ML-, en anglais) comme suit

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{y}).$$

On peut avoir l'intuitif que cet estimateur est raisonnable car il cherche le paramètre rendant l'observation disponible la plus vraisemblable.

Nous allons que cet estimateur nous permet de (re)-construire les estimateurs vus dans les deux modèles précédents et nous permet aussi d'en construire d'autres pour des modèles plus complexes.

Modèle Bernoulli

Nous avons vu via l'Eq. (3.1) que

$$p_\theta(y(n)) = \theta^{y(n)}(1-\theta)^{1-y(n)}$$

avec $y(n) \in \{0,1\}$ pour $n = 1, \cdots, N$.

Si on suppose que la collection d'observations est iid, alors

$$p_\theta(\boldsymbol{y}) = \prod_{n=1}^{N} \theta^{y(n)}(1-\theta)^{1-y(n)} = \theta^{\sum_{n=1}^{N} y(n)}(1-\theta)^{N-\sum_{n=1}^{N} y(n)}$$

On a tracé sur la figure 3.4.1 la fonction $\theta \mapsto p_\theta(\boldsymbol{y})$ avec $N = 30$, $\sum_{n=1}^{N} y(n) = 10$. La fonction n'est concave sur $[0,1]$ (qui est notre intervalle de recherche car le paramètre recherché est une probabilité) mais admet un unique maximum (une analyse formelle de son tableau de variation permettra de le montrer) et donc le maximum est atteint pour la dérivée-nulle.

Ainsi

$$p'_\theta(\boldsymbol{y}) = \theta^s\theta^{N-s}.[s(1-\theta) - (N-s)\theta]$$

avec $s = \sum_{n=1}^{N} y(n)$. Par conséquent, on obtient

$$\hat{\theta}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} y(n)$$

ce qui est l'estimateur empirique de la Section 3.2.3.

Modèle linéaire gaussien

Etant donné l'Eq. (3.4), il est facile de voir que l'estimateur des moindres carrés est en fait aussi l'estimateur du maximum de vraisemblance.

Modèle de régression logistique

Dans de nombreux problèmes de classification (on doit décider une action par exemple, ou on doit décider si une image contient un objet ou pas), la variable recherchée y est dans $\{0,1\}$

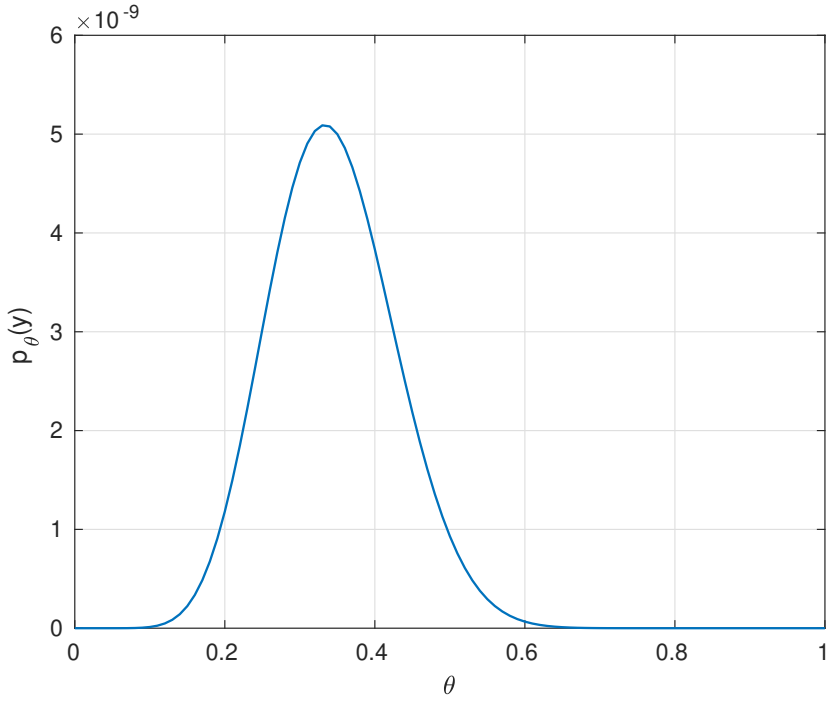


FIGURE 3.1 – Fonction $\theta \mapsto p_\theta(\boldsymbol{y})$ dans le cas Bernoulli

et les données sont les variables explicatives $\boldsymbol{x}$ qui permettent de prendre la décision. Néanmoins dans de très nombreux cas (il suffit de voir toutes les situations de ce type induites par l'apprentissage-machine) le lien entre la décision et les données n'est pas simple et ne peut être exhiber facilement.

Dans cette section, on propose une méthode pour trouver ce lien qui n'utilise pas a priori les réseaux de neurones. Cette méthode est dite la *régression logistique*. Nous verrons néanmoins que le résultat se met sous forme de réseaux de neurones et que c'est une manière assez simple de justifier l'intérêt des réseaux de neurones.

Il est naturel de vouloir de vouloir prendre la décision $y = 1$ si la probabilité $p_{\boldsymbol{x}}(Y = 1)$ est plus grande que celle de l'autre décision $p_{\boldsymbol{x}}(Y = 0)$ pour la valeur $\boldsymbol{x}$ des variables explicatives. Ici, on voit qu'on paramétrise la probabilité par les variables explicatives. La difficulté est que, pour la plupart des problèmes pratiques, nous n'avons pas accès à ces valeurs, autrement dit, nous ne savons pas les écrire en fonction de $\boldsymbol{x}$.

Par conséquent, on prendra la décision $y = 1$ si et seulement si

$$\begin{aligned} p_{\boldsymbol{x}}(Y=1) &\geq p_{\boldsymbol{x}}(Y=0) \\ \frac{p_{\boldsymbol{x}}(Y=1)}{p_{\boldsymbol{x}}(Y=0)} &\geq 1 \\ f(\boldsymbol{x}) := \ln\left(\frac{p_{\boldsymbol{x}}(Y=1)}{p_{\boldsymbol{x}}(Y=0)}\right) &\geq 0 \end{aligned}$$

Bref l'objectif maintenant est d'approximer cette fonction f (et donc de faire une régression).

Nous la modéliserons par une fonction linéaire, d'où,

$$f(\boldsymbol{x}) = \sum_{d=1}^{D} \theta_d x_d = \boldsymbol{x}^T \boldsymbol{\theta}$$

avec $\boldsymbol{x} = [x_1, \cdots, x_D]^T$ et $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_D]^T$.

Ainsi on cherche $\boldsymbol{\theta}$ tel que

$$\begin{aligned}
\ln\left(\frac{p_{\boldsymbol{x}}(Y=1)}{p_{\boldsymbol{x}}(Y=0)}\right) &\approx \boldsymbol{x}^T\boldsymbol{\theta} \\
\frac{p_{\boldsymbol{x}}(Y=1)}{1-p_{\boldsymbol{x}}(Y=1)} &\approx e^{\boldsymbol{x}^T\boldsymbol{\theta}} \\
p_{\boldsymbol{x}}(Y=1) &\approx \frac{e^{\boldsymbol{x}^T\boldsymbol{\theta}}}{1+e^{\boldsymbol{x}^T\boldsymbol{\theta}}} \\
p_{\boldsymbol{x}}(Y=0) &\approx \frac{1}{1+e^{\boldsymbol{x}^T\boldsymbol{\theta}}}
\end{aligned}$$

car $p_{\boldsymbol{x}}(Y=0) = 1 - p_{\boldsymbol{x}}(Y=1)$. Le terme *logistique* provient du fait que la dernière fonction est de type sigmoïde $x \mapsto 1/(1+e^x)$ qui est reliée à la fonction de répartition de la loi logistique.

Le but est maintenant de trouver le paramètre $\boldsymbol{\theta}$ en ayant à disposition des valeurs des variables explicatives et des décisions associées (on pourrait appeler cela une *phase d'apprentissage*, donc learning phase en anglais) . On possède ainsi un N-échantillon $\{(\boldsymbol{x}_i, y_i)\}_{n=1}^N$.

On ne peut procéder par moindres carrés car pour cela, il nous faudrait la valeur de $z_i = \ln\left(\frac{p_{\boldsymbol{x}_i}(Y=1)}{p_{\boldsymbol{x}_i}(Y=0)}\right)$. Cette valeur n'est pas disponible ou alors en la forçant à $+\infty$ si $y_i = 1$ et $-\infty$ si $y_i = 0$. Ces valeurs non finies de z_i empêchent une régression linéaire. C'est pourquoi nous allons procéder autrement.

Comme dans le cas Bernoulli , on peut écrire que

$$\begin{aligned}
p_{\boldsymbol{x}_i}(Y=y_i) &= p_{\boldsymbol{x}_i}(Y=1)^{y_i}.(1-p_{\boldsymbol{x}_i}(Y=1))^{1-y_i} \\
&= \left(\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\theta}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\theta}}}\right)^{y_i} \cdot \left(\frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\theta}}}\right)^{1-y_i} \\
&= \frac{e^{y_i\boldsymbol{x}_i^T\boldsymbol{\theta}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\theta}}}.
\end{aligned}$$

Si on suppose l'indépendance entre les y_i, on obtient la vraisemblance (qui dépend maintenant aussi de $\boldsymbol{\theta}$, d'où, l'écriture de cette dépendance) suivante

$$\begin{aligned}
\boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta},\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N}(y_1,\cdots,y_N) &= \prod_{i=1}^{N} \frac{e^{y_i\boldsymbol{x}_i^T\boldsymbol{\theta}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\theta}}} \\
&= \frac{e^{\sum_{i=1}^N y_i\boldsymbol{x}_i^T\boldsymbol{\theta}}}{\prod_{i=1}^N(1+e^{\boldsymbol{x}_i^T\boldsymbol{\theta}})}.
\end{aligned}$$

Et maintenant le résultat de notre régression est

$$\hat{\boldsymbol{\theta}}_{\text{Logistique}} = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta},\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N}(y_1,\cdots,y_N).$$

Pour obtenir, le maximum de cette fonction, c'est très difficile (car non concave a priori). C'est pour cela nous allons utiliser un algorithme de Gradient (qui permet d'arriver à un maximum local uniquement). C'est un algorithme itératif. A la $j$ème itération, nous avons

$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j + \mu \frac{\partial L}{\partial \boldsymbol{\theta}}_{|\boldsymbol{\theta}=\boldsymbol{\theta}_j}$$

où $\boldsymbol{\theta} \mapsto L(\boldsymbol{\theta}) := p_{\boldsymbol{\theta},\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N}(y_1,\cdots,y_N)$ et μ est le *taux d'apprentissage* (learning rate, en anglais).

On a finalement la fonction f qui permet de prendre la décision (en appliquant un seuil dessus) qui s'écrit

$$f(\boldsymbol{x}) = \sigma\left(\hat{\boldsymbol{\theta}}^T_{\text{Logistique}}\boldsymbol{x}\right).$$

C'est un réseau de neurones à une couche avec comme poids les $\hat{\boldsymbol{\theta}}_{\text{Logistique}}$ et comme fonction d'activation la sigmoïde $x \mapsto \sigma(x) := \frac{1}{1+e^x}$!

3.4.2 Borne de Cramer-Rao

Dans cette section, on va montrer que pour une certaine classe d'estimateurs, l'erreur quadratique moyenne de tout estimateur de cette classe est bornée inférieurement par une certaine valeur, appelée *borne de Cramer-Rao*, que l'on est capable de caractériser et parfois de calculer analytiquement.

3.4.3 Cas simple

On considère l'ensemble des estimateurs $\hat{\theta}$ non biaisé de θ (avec θ un paramètre scalaire à estimer). Ce cas simple est aussi dit cas scalaire ou cas mono-varié.

Sous certaines conditions techniques sur la fonction de vraisemblance (qu'on introduira au cours de la démonstration), on a que pour tout $\hat{\theta}$ non biaisé,

$$\mathbb{E}_\theta[(\hat{\theta}-\theta)^2] \geq \text{BCR}(\theta)$$

avec

— la BCR qui est donnée par

$$\text{BCR}(\theta) = \frac{1}{F(\theta)}$$

où $F(\theta)$ est l'information de Fisher décrite par

$$F(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial \ln p_\theta(\boldsymbol{y})}{\partial\theta}\right)^2\right]$$

et $p_\theta(\boldsymbol{y})$ est la vraisemblance des données $\boldsymbol{y}$ paramétrée par θ.

— L'espérance mathématique est à prendre sur toutes les variables aléatoires du problème.

Démonstration. On pose

$$\begin{aligned}
\mathbb{E}_\theta\left[(\hat{\theta}-\theta).\frac{\partial \ln p_\theta(\boldsymbol{y})}{\partial\theta}\right] &= \int (\hat{\theta}-\theta).\frac{\partial \ln p_\theta(\boldsymbol{y})}{\partial\theta} p_\theta(\boldsymbol{y})d\boldsymbol{y} \\
&\overset{(a)}{=} \int (\hat{\theta}-\theta).\frac{\partial p_\theta(\boldsymbol{y})}{\partial\theta} d\boldsymbol{y} \\
&= \int \hat{\theta}\frac{\partial p_\theta(\boldsymbol{y})}{\partial\theta} d\boldsymbol{y} - \int \theta\frac{\partial p_\theta(\boldsymbol{y})}{\partial\theta} d\boldsymbol{y} \\
&\overset{(b)}{=} \frac{\partial}{\partial\theta}\int \hat{\theta} p_\theta(\boldsymbol{y})d\boldsymbol{y} - \theta\frac{\partial}{\partial\theta}\int p_\theta(\boldsymbol{y})d\boldsymbol{y} \\
&\overset{(c)}{=} \frac{\partial}{\partial\theta}\mathbb{E}_\theta[\hat{\theta}] - \theta\frac{\partial 1}{\partial\theta} \\
&\overset{(d)}{=} \frac{\partial\theta}{\partial\theta} \\
&= 1
\end{aligned}$$

L'égalité (a) vient de la propriété de dérivation du ln. Pour le premier terme dans (b), on utilise le fait que $\hat{\theta}$ ne dépend que des données (et non de θ) et qu'on suppose qu'on peut sortir la dérivation de l'intégrale. Pour la seconde partie de (b), on sort le θ de l'intégrale car on intègre sur $\boldsymbol{y}$ et de nouveau on suppose que la dérivation peut être sortie de l'intégrale. Pour (c), on utilise la définition de l'espérance (et ceci marche car le p_θ est calculé sur la vraie valeur de paramètre) et on utilise aussi le fait qu'une densité de probabilité intègre à 1. Enfin (d) est obtenu grâce à l'hypothèse de l'estimateur non-biaisé (l'hypothèse forte ne sert que là !).

Et en appliquant Cauchy-Schwarz, on a bien

$$\mathbb{E}_\theta\left[(\hat{\theta}-\theta)^2\right].\mathbb{E}_\theta\left[\left(\frac{\partial \ln p_\theta(\boldsymbol{y})}{\partial\theta}\right)^2\right] \geq 1$$

ce qui conclut la preuve.□

Exemple 3.6. On considère le cas gaussien linéaire simple. On a un N-échantillon avec

$$y_i = \theta x_i + \varepsilon_i$$

avec ε_i iid gaussien de moyenne nulle et de variance σ^2 connue.

On calcule donc la log-vraisemblance. On a

$$\begin{aligned}
\ln p_\theta(\boldsymbol{y}) &= \ln\prod_{i=1}^{N}\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-x_i\theta)^2}{2\sigma^2}}\right) \\
&= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{N}(y_i-x_i\theta)^2}{2\sigma^2}.
\end{aligned}$$

En dérivant, on a

$$\begin{aligned}
\frac{\partial \ln p_\theta(\boldsymbol{y})}{\partial\theta} &= \frac{\sum_{i=1}^{N}(y_i-x_i\theta)x_i}{\sigma^2} \\
&\overset{(a)}{=} \frac{\sum_{i=1}^{N}\varepsilon_i x_i}{\sigma^2}.
\end{aligned}$$

Dans (a), on remplace y_i par son modèle.

Pour l'information de Fisher, on a

$$\begin{aligned} F(\theta) &= \mathbb{E}_\theta \left[\left(\frac{\sum_{i=1}^N \varepsilon_i x_i}{\sigma^2}. \right)^2 \right] \\ &= \frac{1}{\sigma^4} \sum_{i,j=1}^N x_i x_j \mathbb{E}[\varepsilon_i \varepsilon_j] \\ &\overset{(a)}{=} \frac{1}{\sigma^4} \sum_{i=1}^N x_i^2 \mathbb{E}[\varepsilon_i^2] \\ &= \frac{\sum_{i=1}^N x_i^2}{\sigma^2}. \end{aligned}$$

L'égalité (a) est obtenue en observant que $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$ pour $i \neq j$.

Finalement cela conduit à

$$\mathrm{BCR}(\theta) = \frac{\sigma^2}{\sum_{i=1}^N x_i^2}.$$

On voit que la BCR diminue quand le bruit diminue et aussi quand la taille du N-échantillon augmente. Mais pouvez-vous dire comment il évolue en fonction de N si $x_i \in \{-1, 1\}$? Pensez-vous que cela est un comportement généralisable ?

3.4.4 Cas multiple

On considère maintenant un paramètre vectoriel $\boldsymbol{\theta}$. Ce cas est dit aussi multi-varié.

Toujours sous certaines conditions techniques (que l'on précisera au cours de la preuve), on a que la matrice des erreurs quadratiques est inférieure (au sens de l'ordre partiel des matrices semi-définies positives) à l'inverse de la matrice d'information de Fisher. Ainsi

$$\mathbb{E}_{\boldsymbol{\theta}} \left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right)^T \right] \succeq \mathbf{F}(\boldsymbol{\theta})^{-1} \tag{3.6}$$

où

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left[\left(\frac{\partial \ln p_{\boldsymbol{\theta}}(\boldsymbol{y})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \ln p_{\boldsymbol{\theta}}(\boldsymbol{y})}{\partial \boldsymbol{\theta}} \right)^T \right].$$

La matrice $\mathbf{F}$ est dite matrice d'information de Fisher et son inverse est dite matrice de la borne de Cramer-Rao.

Démonstration : en classe.

Notez que l'erreur quadratique est obtenue en appliquant la trace de part et d'autre de l'Eq. (3.6) et l'erreur quadratique de chaque composante de $\boldsymbol{\theta}$ en prenant le terme diagonal correspondant de la matrice de BCR. Ainsi, on a

$$\mathbb{E}_{\boldsymbol{\theta}} \left[(\hat{\theta}_d - \theta_d)^2 \right] \geq \left[\mathbf{F}(\theta)^{-1} \right]_{d,d} \tag{3.7}$$

et

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|^2\right] \geq \operatorname{trace}\left(\mathbf{F}(\theta)^{-1}\right). \tag{3.8}$$

Exemple 3.7. Prenons le cas gaussien avec la moyenne à calculer et la variance à calculer. Ainsi, on considère que

$$y_i = m + \varepsilon_i$$

avec ε_i un processus iid gaussien de moyenne nulle et de variance inconnue σ^2.

On a donc $\boldsymbol{\theta} = [m, \sigma^2]^T$.

On obtient que

$$\mathbf{F}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}.$$

Essayez de trouver le résultat suivant par vous-même et en déduire la BCR pour chaque composante de $\boldsymbol{\theta}$.

3.4.5 Sélection de modèle

Dans le cadre de la sélection de modèle, on peut aussi utiliser le principe de maximum de vraisemblance.

Imaginons de nouveau un N-échantillon dont nous cherchons un modèle approprié entre les y_i et les $\boldsymbol{x}_i$ avec une taille D pour les vecteurs $\boldsymbol{x}_i$. On appelle $\mathcal{M}_D$ le modèle dépendant de D variables explicatives dépendant donc d'un paramètre multi-varié $\boldsymbol{\theta}$ de taille D.

La vraisemblance des $\boldsymbol{y}$ dépend donc du modèle choisi et du paramètre associé. On la notera

$$p_{\mathcal{M}_D,\boldsymbol{\theta}}(\boldsymbol{y}).$$

La vraisemblance du modèle sera obtenue en remplaçant le paramètre $\boldsymbol{\theta}$ inconnu par un de ces estimateurs, typiquement celui du maximum de vraisemblance. Par conséquent la vraisemblance des données par rapport au modèle $\mathcal{M}_D$ est noté et vaut respectivement

$$p_{\mathcal{M}_D}(\boldsymbol{y}) = p_{\mathcal{M}_D,\hat{\boldsymbol{\theta}}_{\mathrm{ML}}}(\boldsymbol{y}).$$

Néanmoins si les modèles $\{\mathcal{M}_D\}_{D\in\mathcal{D}}$ sont emboîtés ce qui signifie qu'en forçant une composante de $\boldsymbol{\theta}$ à zéro dans le modèle $\mathcal{M}_D$, on obtient le modèle $\mathcal{M}_{D-1}$, il est clair qu'on choisira le modèle avec le plus de paramètres ce qui n'est pas ce qu'on souhaite non plus car on veut conserver un modèle simple. C'est pourquoi, on va pénaliser les modèles ayant trop de paramètres via la fonction $\mathrm{Pen}(\mathcal{M}_D)$. Ainsi

$$\hat{\mathcal{M}} = \arg \min_{\{\mathcal{M}_D\}_{D\in\mathcal{D}}} -\ln\left(p_{\mathcal{M}_D}(\boldsymbol{y})\right) + \mathrm{Pen}(\mathcal{M}_D).$$

avec, par exemple,

- $\mathrm{Pen}(\mathcal{M}_D) = D$ pour le critère d'information d'Akaike (AIC)
- $\mathrm{Pen}(\mathcal{M}_D) = \ln(N).D$ pour le critère d'information bayésienne (BIC)

3.5 Exercices

Exercice 3.1. En utilisant l'inégalité de Chebychev-Cantelli (1.6) (voir l'exercice 1.11), fournir un intervalle de confiance pour le modèle de Bernoulli. Est-il plus exact que celui de l'équation (3.2) ?

Exercice 3.2. L'*inégalité de Hoeffding*[1] stipule que, si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes, telles que pour tout i, $a_i \leq X_i \leq b_i$. Soit $S_n = X_1 + \cdots + X_n$. Alors pour tout $\epsilon > 0$,

$$\mathbb{P}\left(S_n - \mathbb{E}(S_n) > \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

En utilisant cette inégalité, fournir un intervalle de confiance pour le modèle de Bernoulli. Est-il plus exact que celui de l'équation (3.2) ?

Exercice 3.3. On souhaite estimer une fréquence f_0 dépendant de cette manière des observations

$$y(n) = e^{2i\pi f_0 n} + \varepsilon(n)$$

avec i le nombre complexe tel que $i^2 = -1$ et $\varepsilon(n)$ un processus gaussien iid de moyenne nulle et de variance σ^2 (et de parties réelles et imaginaires indépendantes entre elles).

Donner la formule de l'estimateur du maximum de vraisemblance pour f_0. Obtient-on un résultat intuitif ? Exprimer analytiquement la borne de Cramer-Rao du problème d'estimation de la fréquence ? Tracer cette borne en fonction de N pour un niveau de bruit pré-déterminé par vous-même.

Exercice 3.4. Dans le cas de la régression logistique avec $D = 1$, calculer L' qui est la dérivée de L. Astuce : calculer d'abord la dérivée de $\ln(L)$.

Exercice 3.5. On considère un modèle à deux paramètres θ_1 et θ_2 tel que

$$y_i = \theta_1 x_{i,1} + \theta_2 x_{i,2} + \varepsilon_i$$

avec ε_i iid gaussien de moyenne nulle et de variance σ_2^2.

1. Calculer la borne de Cramer-Rao sur le paramètre θ_1.
2. On suppose maintenant qu'on était pas trop sûr de la dépendance du modèle en fonction de θ_2. On considère donc maintenant que

   $$y_i = \theta_1 x_{i,1} + \varepsilon'_i, \quad i = 1, \cdots, N$$

   avec ε'_i iid gaussien de moyenne nulle et de variance σ_1^2. On suppose $N \geq D$.
   Calculer la borne de Cramer-Rao pour ce modèle et donner la valeur de σ_1^2 qu'on peut s'autoriser pour avoir la même erreur d'estimation qu'en 1. Commentez.

Exercice 3.6. On considère un modèle à D données explicatives avec bruit gaussien iid de moyenne nulle et de variance (connue) σ^2. On considère une collection de N paires $(y_i, \boldsymbol{x}_i)$ avec $\boldsymbol{x}_i$ de taille D. Le paramètre à estimer de taille D s'écrit $\boldsymbol{\theta}$.

Montrer que si $\boldsymbol{y} = [y_1, \cdots, y_N]^T$, alors on a

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

avec une matrice $\boldsymbol{X} = (x_{i,d})_{i=1,\cdots,N;d=1,\cdots,D}$ de taille $N \times D$ et $\boldsymbol{\varepsilon} = [\varepsilon_1, \cdots, \varepsilon_N]^T$.

Calculer la matrice d'information de Fisher (à quelle condition minimale sur N et D, cette matrice est-elle inversible ?). Montrer qu'on retrouve le résultat de l'exemple 3.6 avec $D = 1$.

1. Ceux que cela intéresse iront visiter la page Wikipedia correspondante, pour la preuve de cette inégalité.

4 Tests d'hypothèses

4.1 Introduction

Dans nombreuses situations pratiques, on aimerait savoir si une hypothèse est plus juste qu'une autre. On l'a vu dans le chapitre précédent avec un test de dépendance d'un modèle par rapport à une variable explicative ou bien dans la sélection de modèle (dans le cas alors de deux modèles). Historiquement, ce problème s'est d'abord rencontré dans des contextes militaires où une hypothèse (dite $\mathcal{H}_1$)) correspond à la détection d'un missile ou d'un avion et l'autre hypothèse (dite $\mathcal{H}_0$) correspond à l'espace sûr. Dans ce dernier exemple, on voit clairement que les deux hypothèses ne sont pas mises au même niveau ce qui induira une différence d'analyse de performance entre les deux.

D'abord, on considère que la loi des données $\boldsymbol{y}$ est indexée par l'hypothèse et est donc différente d'une hypothèse à l'autre. Ainsi on a

$$\begin{cases} \text{Hypothèse } \mathcal{H}_0 : & \boldsymbol{y} & \sim & p_{\mathcal{H}_0}(\boldsymbol{y}) \\ \text{Hypothèse } \mathcal{H}_1 : & \boldsymbol{y} & \sim & p_{\mathcal{H}_1}(\boldsymbol{y}) \end{cases}$$

On notera $\hat{\mathcal{H}}$ l'hypothèse détectée dans toute la suite.

On peut définir trois types usuels de performance comme dans le Tableau suivant.

Hyp. détectée \ Vraie hyp.	$\mathcal{H}_0$	$\mathcal{H}_1$
$\hat{\mathcal{H}} = \mathcal{H}_0$	$\times$	$P_M = \Pr_{\mathcal{H}_1}\{\mathcal{H}_0\}$ Probabilité de mauvaise détection Probabilité d'erreur de type II
$\hat{\mathcal{H}} = \mathcal{H}_1$	$P_{FA} := \Pr_{\mathcal{H}_0}\{\mathcal{H}_1\}$ Probabilité de fausse alarme Probabilité d'erreur de type I	$P_D = \Pr_{\mathcal{H}_1}\{\mathcal{H}_1\} = 1 - P_M$ Puissance de bonne détection Puissance du test

TABLE 4.1 – Différentes types de performance et leurs relations

Dans le cadre historique, il apparaît raisonnable de vouloir maximiser P_D ce qui est très facile à faire en choisissant toujours $\mathcal{H}_1$ et donc il faut interdire ce choix de test en bornant aussi la P_{FA} (qui serait aussi égale à 1 avec le test trivial précédent). Ceci va conduire au test suivant décrit dans la prochaine section.

4.2 Test optimal

Théorème 4.1 (Test de Neyman-Pearson)**.** *Maximiser la probabilité de bonne détection sous la condition que la probabilité de fausse alarme soit en-dessous d'un certain niveau (noté P_{FA}^t) conduit au test de Neyman-Pearson suivant, également, appelé test du Rapport de Vraisemblance (Likelihood Ratio Test-LRT , en anglais),*

$$T(\boldsymbol{y}) = \ln\left(\frac{p_{\mathcal{H}_1}(\boldsymbol{y})}{p_{\mathcal{H}_0}(\boldsymbol{y})}\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \mu,$$

avec

- *T le Log Likelihood Ratio-LLR*
- *μ le seuil du test qui permet de satisfaire le niveau-cible de probabilité de fausse alarme.*

Démonstration : en classe.

Evidemment pour mettre en place ce test en pratique, il faut être capable de trouver une forme analytique pour $p_{\mathcal{H}_0}$ et $p_{\mathcal{H}_1}$ et idéalement d'ajuster le seuil μ aussi mathématiquement.

Noter que le test trivial décrit à la section précédente s'écrit $T(\boldsymbol{y}) = 1, \forall \boldsymbol{y}$, et alors $P_D = 1$ et $P_{FA} = 1$ ce qui évidemment dans un contexte militaire est absurde car cela reviendrait à détecter un missile ou un avion ennemi à tous les coups et réduire les stocks à une vitesse grand V.

Il est clair aussi que pour un test non-trivial, P_D va fortement dépendre de P_{FA} notamment en jouant sur le seuil μ. Par conséquent, il serait intéressant de tracer P_D en fonction de P_{FA}.

Définition 4.2. Pour une configuration donnée (Rapport Signal-à-Bruit donné, nombre d'échantillons N donné, etc), la fonction $P_{FA} \mapsto P_D$ est appelée courbe *Receiver Operating Characteristics-ROC.*

Comment la trace-t-on ? Typiquement comme une courbe paramétrée par μ en reliant les points $(P_{FA}(\mu), P_D(\mu))$ dans l'ordre des μ.

Exemple 4.3.

$$\left\{ \begin{array}{ll} \mathcal{H}_0 & : y_n = w_n \\ \mathcal{H}_1 & : y_n = x_n + w_n \end{array} \right. , n = 1, \cdots, N$$

avec

- w_n suite gaussienne iid de moyenne nulle et de variance (connue) $\sigma_w^2 = \mathbb{E}[w_n^2]$,
- x_n également une suite gaussienne iid de moyenne nulle et de variance (connue) $\sigma_x^2 = \mathbb{E}[x_n^2]$.

On a

$$\left\{ \begin{array}{lcl} p_{\mathcal{H}_0}(\boldsymbol{y}) & = & \prod_{n=1}^N p_{\mathcal{H}_0}(y_n) \text{ avec } p_{\mathcal{H}_0}(y_n) = \frac{1}{(2\pi\sigma_w^2)^{1/2}} e^{-\frac{y_n^2}{2\sigma_w^2}} \\ p_{\mathcal{H}_1}(\boldsymbol{y}) & = & \prod_{n=1}^N p_{\mathcal{H}_1}(y_n) \text{ avec } p_{\mathcal{H}_1}(y_n) = \frac{1}{(2\pi(\sigma_x^2+\sigma_w^2))^{1/2}} e^{-\frac{y_n^2}{2(\sigma_x^2+\sigma_w^2)}} \end{array} \right. .$$

d'où

$$\begin{aligned} T(\boldsymbol{y}) &= \ln\left(\frac{\frac{1}{(2\pi)^{N/2}(\sigma_x^2+\sigma_w^2)^{N/2}}e^{-\frac{\sum_{n=1}^N y_n^2}{2(\sigma_x^2+\sigma_w^2)}}}{\frac{1}{(2\pi)^{N/2}\sigma_w^N}e^{-\frac{\sum_{n=1}^N y_n^2}{2\sigma_w^2}}}\right) \\ &= \ln\left(\left(\frac{\sigma_w^2}{\sigma_x^2+\sigma_w^2}\right)^{N/2} e^{-(\frac{1}{2(\sigma_x^2+\sigma_w^2)}-\frac{1}{2\sigma_w^2})\sum_{n=1}^N y_n^2}\right) \\ &= \text{constante positive} \times \sum_{n=1}^N y_n^2 + \text{constante} \end{aligned}$$

Le test LRT est donc un test d'énergie. On peut en fait choisir les constantes à notre guise. Et donc on fera en sorte que le test final soit le suivant.

$$T(\boldsymbol{y}) = \frac{1}{\sigma_x^2+\sigma_w^2}\sum_{n=1}^N y_n^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta.$$

— Sous $\mathcal{H}_1$, $T(\boldsymbol{y})$ suit une loi du χ_2 à N degrés de liberté

$$p_{\chi_2,N}(x) = \frac{1}{\Gamma_c(N/2)}x^{N/2-1}e^{-x},\ x \geq 0$$

— Sous $\mathcal{H}_0$, $T(\boldsymbol{y})$ suit une loi du χ_2 à N degrés de liberté

$$p_{\chi_2,N}(x) = \frac{1}{(\sigma_w^2/(\sigma_x^2+\sigma_w^2))^{N/2}\Gamma_c(N/2)}x^{N/2-1}e^{-\frac{(\sigma_x^2+\sigma_w^2)x}{\sigma_w^2}},\ x \geq 0$$

avec les fonctions Gamma complète et incomplète suivante

$$\Gamma_c(s) = \int_0^\infty x^{s-1}e^{-x}dx$$

et

$$\Gamma_{\text{inc}}(s,u) = \int_u^\infty x^{s-1}e^{-x}dx.$$

Par conséquent, on obtient

$$\begin{aligned} P_{FA} &= \Pr_{\mathcal{H}_0}(T(\boldsymbol{y}) > \eta) \\ &= \int_\eta^\infty \frac{1}{(\sigma_w^2/(\sigma_x^2+\sigma_w^2))^{N/2}\Gamma_c(N/2)}x^{N/2-1}e^{-\frac{(\sigma_x^2+\sigma_w^2)x}{\sigma_w^2}}dx \\ &= \frac{1}{\Gamma_c(N/2)}\cdot\frac{1}{(\sigma_w^2/(\sigma_x^2+\sigma_w^2))^{N/2}}\cdot\int_\eta^\infty x^{N/2-1}e^{-\frac{(\sigma_x^2+\sigma_w^2)x}{\sigma_w^2}}dx \\ &= \frac{\Gamma_{\text{inc}}\left(N/2, \eta\frac{\sigma_x^2+\sigma_w^2}{\sigma_w^2}\right)}{\Gamma_c(N/2)}. \end{aligned}$$

De manière similaire, on a

$$P_D = \frac{\Gamma_{\text{inc}}(N/2,\eta)}{\Gamma_c(N/2)}.$$

Sur la figure 4.1, on trace pour le test précédent la courbe ROC. On souhaite évidemment que la courbe se rapproche le plus rapidement possible du coin Nord-Ouest.

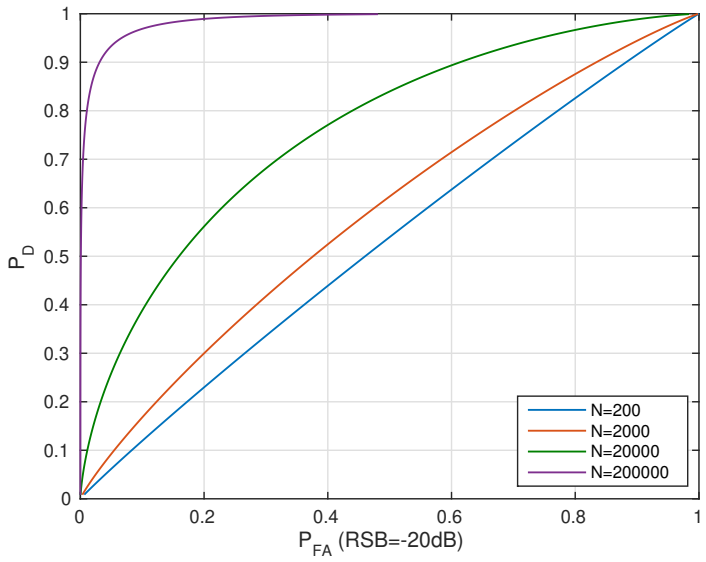


FIGURE 4.1 – Courbe ROC pour le test d'hypothèse LRT (avec RSB= -20dB)

4.3 Lien entre intervalle de confiance et test

On peut voir un intervalle de confiance comme un test qu'un certain paramètre θ recherché appartienne à cet intervalle avec une certaine probabilité.

Plus précisément, quand on veut savoir si le paramètre recherché vaut 0 (et donc sera associé à une variable non explicative), on associera cette hypothèse à $\mathcal{H}_0$. Alors le test détectera correctement $\mathcal{H}_0$ (avec une certaine probabilité qui sera alors la confiance $(1-\alpha)$ que l'on a sur le fait que la variable est non explicative) si le test $T(\boldsymbol{y})$ est inférieur à un seuil. Ce test peut être par exemple un estimateur de θ que l'on note $\hat{\theta}(\boldsymbol{y})$. Par conséquent

$$1 - P_{FA} = \Pr_{\mathcal{H}_0}(\hat{\theta}(\boldsymbol{y}) \leq \mu) = 1 - \alpha.$$

4.4 Exercices

Exercice 4.1.

- On considère des échantillons bi-dimensionnel (bref, $y \in \mathbb{R}^2$).
- On considère que l'hypothèse 0 correspond à une gaussienne (bi-dimensionnelle) de moyenne $\boldsymbol{m}_0$ et de variance σ_0^2 (chaque composante de y est indépendante et de même variance σ_0^2).
- On considère que l'hypothèse 1 correspond à une gaussienne (bi-dimensionnelle) de moyenne $\boldsymbol{m}_1$ et de variance σ_1^2 (chaque composante de y est indépendante et de même variance σ_1^2).

En appliquant le test LRT (avec $\mu = 0$), trouve la règle de décision entre les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$. Montrer que ce test s'écrit sous la forme d'un réseau de neurones à une couche avec l'échelon d'Heavyside comme a fonction d'activation quand $\sigma_0^2 = \sigma_1^2$. Est-ce encore vrai si $\sigma_0^2 \neq \sigma_1^2$?

Exercice 4.2. On considère le même modèle que l'exemple 4.3 mais maintenant le signal x_n appartient à une M-PAM (cf. cours de COM105) d'amplitude A. Ecrire le test de Neyman-Pearson correspondant.