I-1.4 Les compléments

Les contenus de cette partie ne sont pas exigibles ; ils enrichissent la culture statistique et préparent certains chapitres ultérieurs.

Exemples introductifs supplémentaires

Avant de procéder à la construction mathématique d'un modèle statistique, il est utile d'avoir en tête quelques exemples de problèmes concrets de statistique. Ceux qui suivent complètent les exemples de la partie principale (Section I-1.1.2, Section I-1.1.2 et Section I-1.1.2).

Exemple I-1.4.1: Sondage

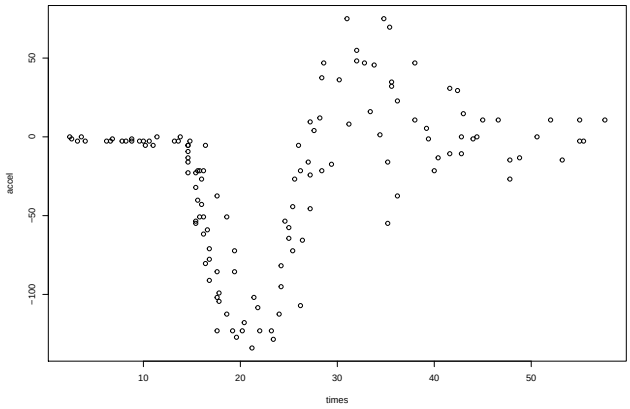
Une élection entre deux candidats A et B a lieu : on effectue un sondage à la sortie des urnes. On interroge n votants, le nombre n étant considéré comme petit devant le nombre total de votants, et on obtient ainsi les nombres n_A et n_B de voix pour les candidats A et B respectivement ($n_A + n_B = n$, en ne tenant pas compte des votes blancs ou nuls pour simplifier). Les questions naturelles auxquelles nous tâcherons de répondre sont typiquement les suivantes.

- Quelle est la proportion d'électeurs ayant voté pour le candidat A ?
- Peut-on affirmer que A ou B a gagné au vu de n_A et n_B seulement ?
- Si l'on décide d'annoncer A (ou B) vainqueur, comment quantifier l'erreur de décision ?

Le modèle statistique associé à ce sondage, dans le cas d'un tirage avec remise, est construit dans l'Section I-1.1.2 ; le cas d'un tirage sans remise est traité en Section I-1.4.

Exemple I-1.4.2: Reconstruction d'un signal

On souhaite apprendre une fonction $t \mapsto f(t)$, et pour ce faire on dispose d'observations affectées par une erreur de mesure de la fonction inconnue f aux instants multiples de T_e, sur un intervalle de temps $[0, T]$: les relevés $y_k, k = 1, \dots, n$ dont on dispose sont donc des approximations des quantités inconnues $f(kT_e)$, $k = 1, \dots, n$, où $n := \lfloor T/T_e \rfloor$ - la partie entière inférieure de T/T_e.
Dans ce problème, on est typiquement intéressé par "estimer" la fonction f. Ceci signifie que nous allons chercher à construire une fonction $t \mapsto \widehat{f}\big(t;\ \{y_k\}_{k=1}^n\big)$ ne dépendant que de l'échantillon mesuré $\{y_k\}_{k=1}^n$, qui soit en un certain sens "proche" de la fonction inconnue f. La difficulté ici va dépendre de la période d'échantillonnage, du rapport entre l'amplitude du signal à reconstruire f et celle des erreurs de mesure $\{y_k - f(kT_e)\}_{k=1}^n$, et bien sûr de la "complexité" de la fonction f. Intuitivement, une fonction constante $f(t) = \mu$ ou ayant une forme prescrite, par exemple $f(t) = \beta_0 + \beta_1 t$ sera plus facile à reconstruire qu'une fonction très irrégulière.



Accélération de la tête en fonction du temps suite à un impact.

On peut s'intéresser au même problème en dimension supérieure, par exemple en dimension 2 où f peut représenter une image. La fonction f est alors définie sur le pavé $[0,1]\times[0,1]$. En pratique, cette image est discrétisée en pixels et la valeur de f en ce pixel quantifie un niveau de gris dans $\{0,\dots,M-1\}$. On collecte donc $n=N^2$ données $y_{k,\ell}$ pour $1\le k,\ell\le N$, qui sont des mesures bruitées de la quantité inconnue $f(k/N,\ell/N)$. On cherche à reconstruire l'image f ou bien à décider si une certaine caractéristique est présente dans l'image.

Exemple I-1.4.3: Modèle gaussien pour la reconstruction de signal

Dans le problème de reconstruction d'un signal (Section I-1.4), les observations sont clairement à valeurs dans $\mathsf{Z}:=\mathbb{R}^n$ avec $n:=\lfloor T/T_e\rfloor$. Il est naturel de munir cet ensemble de sa tribu borélienne (voir **??**) : $\mathcal{Z}:=\mathcal{B}(\mathbb{R}^n)$. Pour construire un modèle statistique, il faut se donner un ensemble $\mathcal{C}$ de lois possibles pour les observations. Supposons que :

— la fonction f est modélisée par une combinaison linéaire de fonctions de base connues $f(t)=\sum_{k=1}^p \beta_k\phi_k(t)\in\mathbb{R}$,

— les erreurs de mesure sont modélisées comme la réalisation de variables aléatoires gaussiennes centrées et de même variance σ^2.

L'ensemble des paramètres inconnus est donc ici $\theta=(\beta_1,\dots,\beta_p,\sigma^2)$ où $(\beta_1,\dots,\beta_p)\in\mathbb{R}^p$ et $\sigma^2\in\mathbb{R}_+$. Nous poserons donc $\Theta:=\mathbb{R}^p\times\mathbb{R}_+^*$. Par suite, pour tout $\theta\in\Theta$, dans le modèle $\mathbb{P}_\theta$, la loi du vecteur d'observations est la loi Gaussienne sur $\mathbb{R}^n$, d'espérance $(\mu_1(\theta),\mu_2(\theta),\dots,\mu_n(\theta))$ avec $\mu_i=\sum_{\ell=1}^p\beta_\ell\phi_\ell(iT_e)$ et de matrice de covariance $\sigma^2\mathrm{I}_{n\times n}$. Nous écrivons le modèle statistique suivant :

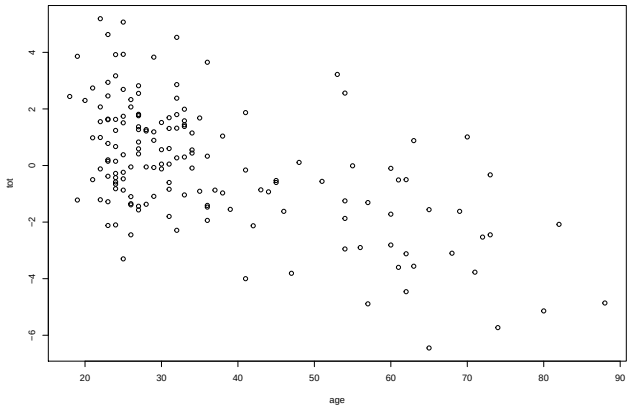
$$\left(\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n),\left\{p_\theta\cdot\mathrm{Leb}^{\otimes n}\,,\ \theta\in\Theta:=\mathbb{R}^p\times\mathbb{R}_+^*\right\}\right),$$

où

$$p_\theta(x_1,\dots,x_n):=(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n\left(x_i-\sum_{k=1}^p\beta_k\phi_k(iT_e)\right)^2\right)\ .$$

Exemple I-1.4.4: Influence d'une variable sur une autre

On s'intéresse à étudier le lien entre la qualité d'un greffon de rein (caractérisée par une note agrégeant un certain nombre de caractéristiques) et l'âge du donneur. On considère une population de n donneurs potentiels. Pour chaque donneur $i\in\{1,\dots,n\}$, on note x_i son âge et y_i la qualité du greffon.



Qualité d'un greffon de rein en fonction de l'âge du donneur (données du laboratoire de néphrologie du Dr. B. Myers, Université de Stanford).

Il est bien entendu irréaliste de postuler l'existence d'une fonction $f : \mathbb{R} \to \mathbb{R}$ telle que $y = f(x)$. Toutefois, il est raisonnable d'introduire un modèle statistique exprimant que y_i est en partie expliquée par x_i ; ce qui peut être fait en exprimant que chaque mesure y_i est une observation dans un bruit de mesure de $f(x_i)$ où f est une fonction inconnue. Le problème du statisticien sera ici de reconstruire la fonction de régression f et de caractériser l'erreur de modélisation.
Le problème est proche de celui de l'section I-1.4, à ceci près que les points kT_e sont remplacés par les âges x_i. Les variables explicatives x_i ne sont pas nécessairement scalaires, on peut ainsi remplacer x_i par un vecteur $\mathbf{x}_i \in \mathbb{R}^k$ qui collecte un ensemble de variables explicatives possibles. Dans ce cas, f est une fonction $\mathbb{R}^k \to \mathbb{R}$.
On peut, si cette information est disponible, incorporer une information a priori sur la fonction f : par exemple, postuler que la fonction f est de la forme $f(\mathbf{x}) = t(\boldsymbol{\theta}^\top \mathbf{x})$, où $t : \mathbb{R} \to \mathbb{R}$ est une fonction connue et $\theta = (\theta_1, \ldots, \theta_k)^T \in \mathbb{R}^k$ sont des paramètres inconnus. Dans les cas les plus élémentaires, $t(z) = z$, et on parle de *régression linéaire*. Il est bien entendu possible de considérer des modèles de régression plus généraux incorporant explicitement des non-linéarités et des interactions entre les différentes variables explicatives. La nature des informations exploitables pour construire un tel modèle dépend naturellement fortement des applications considérées. Nous reviendrons dans la suite sur l'importance centrale des modèles en statistique et des méthodes pour construire de tels modèles.
Il existe aussi des situations où y_i est une variable *qualitative*, c'est-à-dire ne prenant qu'un nombre fini de valeurs (instances). On peut penser que le risque d'être victime d'un infarctus est influencé par un ensemble de facteurs : consommation de tabac, d'alcool, taux de cholestérol, indice de masse corporelle, âge, terrain familial, pression artérielle, ces différents facteurs étant eux-mêmes souvent liés.

Sondage sans remise et loi hypergéométrique

Reprenons l'Section I-1.4 et le modèle de l'Section I-1.1.2, et supposons maintenant que les tirages sont sans replacement. On dispose d'une population de N individus (N est typiquement très grand). On veut introduire un modèle qui exprime que dans la population, $N\theta$ individus votent pour le candidat A, où $\theta \in \Theta := \{k/N\,,\ k \in \{0, \ldots, N\}\}$. L'espace des observations est $\mathsf{Z} := \{0, \ldots, n\}$ et $\mathcal{Z} := \mathcal{P}(\mathsf{Z})$. La loi des observations est cette fois donnée par

$$\mathbb{P}_\theta(\{k\}) = \begin{cases} \frac{\binom{N\theta}{k}\binom{N-N\theta}{n-k}}{\binom{N}{n}}, & \text{si } k \in \{\max(n - N(1-\theta), 0), \ldots, \min(N\theta, n)\}, \\ 0, & \text{si } k \notin \{\max(n - N(1-\theta), 0), \ldots, \min(N\theta, n)\}. \end{cases}$$

C'est le nombre de façons de choisir k individus dans une population de taille $N\theta$, puis indépendamment $(n-k)$ individus dans une population de taille $N(1-\theta)$ divisé par le nombre total de façons de choisir n individus parmi N. La loi $\mathbb{P}_\theta$ définie ci-dessus est appelée *hypergéométrique*, notée $\text{Hyper}(N\theta, N, n)$.

Données censurées

Exemple I-1.4.5: Contrôle de qualité, données censurées

On cherche – en laboratoire – à tester la fiabilité d'un dispositif. On fait fonctionner en parallèle n appareils jusqu'à ce qu'ils tombent tous en panne. On note $x_1, \dots, x_n$ les instants de panne observés. On cherche alors à obtenir des informations sur la fiabilité du dispositif, par exemple

- garantir qu'avec une probabilité donnée, le système fonctionnera plus qu'un temps T donné.
- donner une valeur "représentative" de la durée de vie du dispositif.

Répondre à ces deux questions nécessite bien entendu de formaliser ce qu'on entend ici par probabilité ou "valeur représentative". Nous nous y attacherons dans la suite de l'exposé.
Si les appareils sont fiables et que le nombre n d'appareils testés est grand, attendre que tous les dispositifs soient tombés en panne peut s'avérer impossible. Une idée fréquemment utilisée en fiabilité est de fixer a priori un temps terminal τ et d'observer les temps de défaillance des systèmes apparaissant avant l'horizon τ, ce qui revient à observer $x_i^\star = \min\{x_i, \tau\}$, pour $i \in \{1, \dots, n\}$. Une question importante consiste à quantifier la perte d'information associée à l'horizon τ dans cette seconde expérience plus réaliste.

Exemple I-1.4.6: Modèle exponentiel censuré

Pour l'section I-1.4 du contrôle de qualité, un modèle classique de durée de vie est fourni par la famille de lois exponentielles de paramètre $\theta \in \mathbb{R}_+^*$, de densité par rapport à la mesure de Lebesgue Leb donnée par

$$q_\theta(x) := \theta \mathrm{e}^{-\theta x} \mathbb{1}_{\mathbb{R}_+}(x) \, . \tag{I-1.2}$$

Le n-échantillon du modèle statistique $\left(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \left\{q_\theta \cdot \text{Leb}\,,\ \theta \in \Theta := \mathbb{R}_+^*\right\}\right)$ est

$$\left(\mathbb{R}_+^n, \mathcal{B}(\mathbb{R}_+^n), \left\{q_\theta^{\otimes n} \cdot \text{Leb}^{\otimes n}\,,\ \theta \in \Theta\right\}\right) \, ,$$

où $\text{Leb}^{\otimes n}$ est la mesure de Lebesgue sur $\mathbb{R}^n$ et

$$q_\theta^{\otimes n}(x_1, \dots, x_n) = \prod_{i=1}^n q_\theta(x_i).$$

Pour tout $\theta \in \Theta$, sous $\mathbb{P}_\theta$, les observations $(X_1, \dots, X_n)$, définies par $X_i(x_1, \dots, x_n) = x_i$ pour tout $i \in \{1, \dots, n\}$ et $(x_1, \dots, x_n) \in \mathbb{R}_+^n$ sont indépendantes. Pour tout i, le modèle statistique induit par X_i est

$$\left(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \left\{q_\theta \cdot \text{Leb}\,,\ \theta \in \Theta := \mathbb{R}_+^*\right\}\right) \, .$$

Supposons que les observations $(X_1, \dots, X_n)$ soient censurées à un instant terminal τ connu. Pour $i \in \{1, \dots, n\}$, nous considérons les statistiques

$$X_i^* := \min(X_i, \tau) \, , \quad i \in \{1, \dots, n\} \, .$$

Pour $\theta \in \Theta$, notons $\mathbb{Q}_\theta^*$ la loi image $\mathbb{Q}_\theta^{X_i^*}$. Le modèle statistique induit par la variable X_i^* est donc

$$\left([0, \tau], \mathcal{B}([0, \tau]), \{\mathbb{Q}_\theta^*\,,\ \theta \in \Theta\}\right) \, .$$

Notons que la loi $\mathbb{Q}_\theta^*$ n'est pas absolument continue par rapport à la mesure de Lebesgue. En revanche, la famille $\{\mathbb{Q}_\theta^*\,,\ \theta\in\Theta\}$ est dominée par $\mu = \mathrm{Leb} + \delta_\tau$, où δ_τ désigne la mesure de Dirac en τ. Pour tout $\theta\in\Theta$, la densité de $\mathbb{Q}_\theta^*$ par rapport à μ est donnée par

$$q_\theta^*(x) := \theta \mathrm{e}^{-\theta x}\mathbb{1}_{\{x<\tau\}} + c(\theta)\mathbb{1}_{\{x=\tau\}}\ , \quad \text{avec } c(\theta) := \int_\tau^{+\infty} \theta \mathrm{e}^{-\theta t}\mathrm{d}t = \mathrm{e}^{-\theta\tau}\ .$$

Le modèle statistique induit par $(X_1^*,\dots,X_n^*)$ est donc donné par

$$\big([0,\tau]^n\,, \mathcal{B}([0,\tau]^n), \big\{(q_\theta^*)^{\otimes n}\cdot\mu^{\otimes n}\,,\ \theta\in\Theta\big\}\big)\ .$$

C'est un n-échantillon du modèle statistique $([0,\tau]\,, \mathcal{B}([0,\tau]), \{q_\theta^*\cdot\mu\,,\ \theta\in\Theta\})$.

Illustration / animation : L'atome en τ grossit à mesure que la censure devient plus sévère

On illustre ici l'histogramme des durées observées $\min(X_i,\tau)$, qui garde la forme exponentielle en dessous de τ et concentre une masse croissante en τ à mesure que τ diminue.

link

Régression logistique

Exemple I-1.4.7: Régression logistique

Une nouvelle molécule anti-cancéreuse est en cours de développement. Pour déterminer la dose à administrer aux futurs patients, un "bio-essai" est mené. Celui-ci consiste à injecter des doses croissantes $x_1 < \cdots < x_q$ de cette molécule à des populations de $n_1, n_2, \dots, n_q$ souris et à mesurer dans chaque population le nombre de souris y_i qui décèdent. On dispose donc de q-couples de données $\{(x_i,y_i)\}_{i=1}^q$. Rappelons tout d'abord que la loi binomiale $\mathrm{Bin}(n,\pi)$, $\pi\in]0,1[$, est la loi sur $\{0,\dots,n\}$ de densité par rapport à la mesure de comptage donnée par

$$k\mapsto \binom{n}{k}\pi^k(1-\pi)^{n-k}\ , \quad k\in\{0,\dots,n\}\ .$$

L'application

$$\pi\mapsto \mathrm{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

est appelée "application logit". C'est une fonction croissante, $\lim_{\pi\downarrow 0^+}\mathrm{logit}(\pi) = -\infty$ et $\lim_{\pi\uparrow 1^-}\mathrm{logit}(\pi) = \infty$. Les chercheurs postulent un lien linéaire entre la dose x et le taux de mortalité. Pour chaque population individuelle $i\in\{1,\dots,q\}$, nous postulons donc le modèle statistique binomial

$$\mathcal{E}_i := \big(\{0,\dots,n_i\}, \mathcal{P}(\{0,\dots,n_i\}), \big\{\mathrm{Bin}(n_i,\pi(\theta,x_i))\,,\ \theta=(\theta_0,\theta_1)\in\Theta=\mathbb{R}^2\big\}\big)$$

où

$$\pi(\theta,x) := \frac{\exp(\theta_0+\theta_1 x)}{1+\exp(\theta_0+\theta_1 x)}.$$

On vérifie aisément que

$$\mathrm{logit}(\pi(\theta,x)) = \theta_0 + \theta_1 x\ .$$

Considérons le modèle statistique

$$\left(\prod_{i=1}^{q}\{0,\dots,n_i\}, \mathcal{P}\left(\prod_{i=1}^{q}\{0,\dots,n_i\}\right), \left\{\bigotimes_{i=1}^{q}\mathrm{Bin}(n_i,\pi(\theta,x_i))\,,\ \theta\in\Theta\right\}\right)\ .$$

Pour $i\in\{1,\dots,q\}$ on note Y_i la i-ème observation qui est la statistique définie, pour tout $z=(y_1,\dots,y_q)\in\mathsf{Z}$ par $Y_i(z)=y_i$. Pour tout $\theta\in\Theta$, sous $\mathbb{P}_\theta$, les observations $(Y_1,\dots,Y_q)$ sont indépendantes et le modèle induit par chaque observation est le modèle binomial $\mathcal{E}_i$. Par abus de langage, nous dirons que "les observations $(Y_1,\dots,Y_q)$ sont indépendantes et pour $i\in\{1,\dots,q\}$, Y_i est distribuée suivant le modèle binomial de paramètre $\pi(\theta,x_i)$".

Exemples détaillés développés en cours (annexe des slides)

Les slides du cours 1 se terminent par trois exemples développés en annexe : un modèle de survie, un modèle de régression pour des données de température et une formalisation statistique des systèmes de recommandation. Nous les reprenons ici avec le même niveau de détail que les slides, en corrigeant au passage quelques coquilles de ces dernières. Ces exemples mettent en œuvre les notions du chapitre (modèle dominé, n-échantillon, modèle induit, identifiabilité) et anticipent sur les méthodes d'estimation du cours 2.

Modèles de survie.

Exemple I-1.4.8: Modèle de survie

Un modèle classique de durée de vie est fourni par la famille des lois exponentielles de paramètre $\theta\in\mathbb{R}_+^*$, de densité par rapport à la mesure de Lebesgue Leb donnée par

$$q_\theta(x) := \theta \mathrm{e}^{-\theta x}\mathbb{1}_{\mathbb{R}_+}(x)\ .$$

Le n-échantillon du modèle statistique $\left(\mathbb{R}_+,\mathcal{B}(\mathbb{R}_+),\left\{q_\theta\cdot\mathrm{Leb}\,,\ \theta\in\Theta:=\mathbb{R}_+^*\right\}\right)$ est

$$\left(\mathbb{R}_+^n,\mathcal{B}(\mathbb{R}_+^n),\left\{\mathbb{P}_\theta := q_\theta^{\otimes n}\cdot\mathrm{Leb}^{\otimes n}\,,\ \theta\in\Theta\right\}\right)\ ,$$

où $\mathrm{Leb}^{\otimes n}$ est la mesure de Lebesgue sur $\mathbb{R}^n$ et

$$q_\theta^{\otimes n}(x_1,\dots,x_n) := \prod_{i=1}^{n} q_\theta(x_i) = \theta^n \exp\Big(-\theta\sum_{i=1}^{n}x_i\Big)\,\mathbb{1}_{\mathbb{R}_+^n}(x_1,\dots,x_n)\ .$$

L'observation $(X_1,\dots,X_n)$ est définie par $X_i(x_1,\dots,x_n):=x_i$ pour tout $i\in\{1,\dots,n\}$ et $(x_1,\dots,x_n)\in\mathbb{R}_+^n$. D'après le Section I-1.1.7, pour tout $\theta\in\Theta$, les statistiques X_i et X_j sont indépendantes sous $\mathbb{P}_\theta$ pour $i\neq j$, et le modèle statistique induit par X_i est

$$(\mathbb{R}_+,\mathcal{B}(\mathbb{R}_+),\{q_\theta\cdot\mathrm{Leb}\,,\ \theta\in\Theta\})\ .$$

Absence de mémoire. Pour tous $a,b>0$ et $\theta\in\Theta$, comme $\mathbb{P}_\theta(X_1\geq a)=\int_a^{+\infty}\theta\mathrm{e}^{-\theta x}\mathrm{d}x=\mathrm{e}^{-a\theta}$,

$$\mathbb{P}_\theta\left(X_1\geq a+b\,|\,X_1\geq a\right)=\frac{\mathbb{P}_\theta(X_1\geq a+b)}{\mathbb{P}_\theta(X_1\geq a)}=\frac{\mathrm{e}^{-(a+b)\theta}}{\mathrm{e}^{-a\theta}}=\mathrm{e}^{-b\theta}=\mathbb{P}_\theta(X_1\geq b)\ .$$

Un dispositif qui a déjà fonctionné pendant une durée a a donc la même loi de durée de vie résiduelle qu'un dispositif neuf : le modèle exponentiel ne rend pas compte de l'usure.

Lemme I-1.4.1: Moments de la loi exponentielle

Pour tout $\theta \in \mathbb{R}_+^*$ et tout $k \in \mathbb{N}^*$, $\mathbb{E}_\theta[X_1^k] = k!/\theta^k$.

Démonstration

La *fonction génératrice des moments* de X_1 est finie pour $|s| < \theta$ et vaut

$$\begin{aligned}\mathbb{E}_\theta[\mathrm{e}^{sX_1}] &= \int_0^{+\infty} \mathrm{e}^{sx}\theta\mathrm{e}^{-\theta x}\mathrm{d}x = \theta\int_0^{+\infty}\mathrm{e}^{-(\theta-s)x}\mathrm{d}x \\ &= \frac{\theta}{\theta-s} = \frac{1}{1-s/\theta} = \sum_{k=0}^{+\infty}\frac{s^k}{\theta^k}\,.\end{aligned}$$

D'autre part, en développant l'exponentielle en série entière, $\mathrm{e}^{sX_1} = \sum_{k\geq 0} s^k X_1^k/k!$, et l'interversion de la somme et de l'espérance est licite puisque $\sum_{k\geq 0}|s|^k X_1^k/k! = \mathrm{e}^{|s|X_1}$ est $\mathbb{P}_\theta$-intégrable pour $|s| < \theta$. Ainsi

$$\mathbb{E}_\theta[\mathrm{e}^{sX_1}] = \sum_{k=0}^{+\infty}\frac{s^k}{k!}\,\mathbb{E}_\theta[X_1^k]\,.$$

En identifiant les coefficients des deux développements en série entière, valables sur $]-\theta,\theta[$, on obtient directement $\mathbb{E}_\theta[X_1^k] = k!/\theta^k$ pour tout $k \in \mathbb{N}^*$. □

Construction d'estimateurs par la méthode des moments. Considérons les fonctions $T(x) = x$ et $\tilde{T}(x) = x^2$. D'après le lemme précédent,

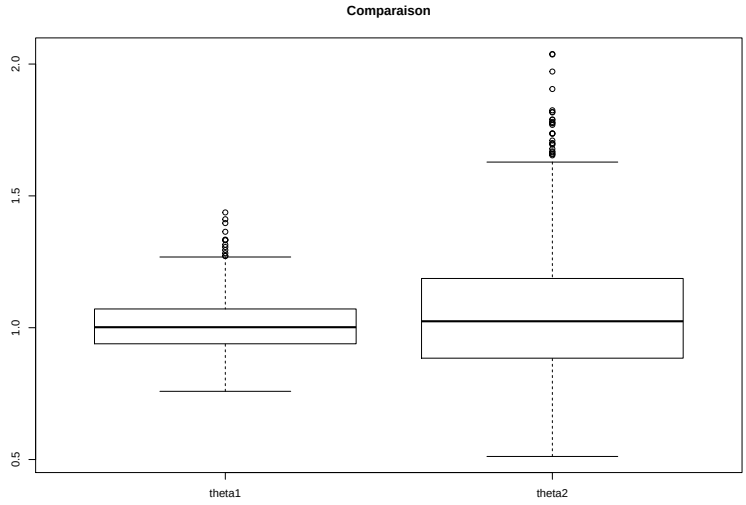
$$\begin{aligned}e(\theta) &:= \int T(x)\,\mathbb{Q}_\theta(\mathrm{d}x) = \int_0^{+\infty} x\,\theta\exp(-\theta x)\,\mathrm{d}x = \frac{1}{\theta}\,, \\ \tilde{e}(\theta) &:= \int \tilde{T}(x)\,\mathbb{Q}_\theta(\mathrm{d}x) = \int_0^{+\infty} x^2\,\theta\exp(-\theta x)\,\mathrm{d}x = \frac{2}{\theta^2}\,.\end{aligned}$$

Les *estimateurs des moments* associés sont les solutions en θ des équations obtenues en remplaçant les moments théoriques $e(\theta)$ et $\tilde{e}(\theta)$ par leurs contreparties empiriques :

$$e(\theta) = \frac{1}{\theta} = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{et} \quad \tilde{e}(\theta) = \frac{2}{\theta^2} = \frac{1}{n}\sum_{i=1}^n X_i^2\,.$$

Bien entendu, chacun des moments utilisés donne dans ce contexte un estimateur différent. Ces équations ont des solutions uniques dans $\mathbb{R}_+^*$ (dès que les observations ne sont pas toutes nulles, ce qui est le cas $\mathbb{P}_\theta$-presque sûrement) :

$$\hat{\theta}_{n,1} := \frac{1}{\frac{1}{n}\sum_{i=1}^n X_i} \quad \text{et} \quad \hat{\theta}_{n,2} := \left(\frac{2}{\frac{1}{n}\sum_{i=1}^n X_i^2}\right)^{1/2}\,.$$



Comparaison des estimateurs $\hat{\theta}_{n,1} = 1/(n^{-1}\sum_{i=1}^n X_i)$ *et* $\hat{\theta}_{n,2} = \left(2/(n^{-1}\sum_{i=1}^n X_i^2)\right)^{1/2}$ *: boîtes à moustaches des valeurs prises par chacun des deux estimateurs sur un grand nombre d'échantillons simulés de taille* n *sous* $\mathbb{P}_\theta$*, pour* $\theta = 1$*. Les deux estimateurs se concentrent autour de la vraie valeur du paramètre, mais* $\hat{\theta}_{n,2}$ *est nettement plus dispersé que* $\hat{\theta}_{n,1}$*.*

Illustration / animation : Les deux estimateurs des moments se concentrent tous deux autour de θ, mais pas à la même vitesse

L'animation accessible par le QR code ci-contre permet de comparer, sur des échantillons répétés de taille n, les boîtes à moustaches de $\hat{\theta}_{n,1}$ et de $\hat{\theta}_{n,2}$, tous deux centrés sur la vraie valeur mais dispersés différemment.

link

Pour étudier ces estimateurs, nous utilisons les lois Gamma (voir aussi la Section I-1.2).

Définition I-1.4.1: Lois Gamma

(i) Pour tout réel $p > 0$, on appelle *loi Gamma réduite* de paramètre de forme p (et l'on note Gamma(p)) la loi sur $\mathbb{R}_+$ de densité

$$f_p(x) := \frac{1}{\Gamma(p)} \exp(-x)\, x^{p-1}, \quad x > 0\,,$$

où $\Gamma(p) := \int_0^{+\infty} \mathrm{e}^{-x} x^{p-1} \mathrm{d}x$ est la fonction Gamma d'Euler.

(ii) Pour $\lambda > 0$, on appelle Gamma(p, λ) la loi de $X = Z/\lambda$ où $Z \sim$ Gamma(p) ; λ est le *paramètre d'intensité* (l'échelle est $1/\lambda$) et p le *paramètre de forme*. La densité de la loi Gamma(p, λ) est donnée par

$$f_{p,\lambda}(x) = \frac{\lambda^p}{\Gamma(p)} \exp(-\lambda x)\, x^{p-1}\,, \quad x > 0\,.$$

La loi exponentielle d'intensité $\lambda > 0$ est la loi Gamma$(1, \lambda)$. Si $Z \sim$ Gamma(p), alors pour tout $r > -p$,

$$\mathbb{E}[Z^r] = \frac{1}{\Gamma(p)} \int_0^{+\infty} \mathrm{e}^{-x} x^{p+r-1} \mathrm{d}x = \frac{\Gamma(p+r)}{\Gamma(p)}\,, \tag{I-1.3}$$

l'intégrale étant divergente pour $r \leq -p$. Par changement d'échelle, si $X \sim \text{Gamma}(p, \lambda)$, alors $\mathbb{E}[X^r] = \lambda^{-r}\,\Gamma(p+r)/\Gamma(p)$ pour $r > -p$.

Lemme I-1.4.2: Convolution des lois Gamma

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes distribuées suivant des lois $\text{Gamma}(p_i, \lambda)$, avec $\lambda > 0$ et $p_i > 0$ pour $i \in \{1, \ldots, n\}$. Alors $\sum_{i=1}^n X_i$ est distribuée suivant la loi $\text{Gamma}(\sum_{i=1}^n p_i, \lambda)$.

Ce lemme se démontre par récurrence sur n, par exemple avec la méthode de la fonction muette (Section I-1.2). En particulier, la somme $X_1 + X_2$ de deux variables aléatoires exponentielles *indépendantes et de même paramètre* θ suit une loi $\text{Gamma}(2, \theta)$: en notant $\star$ le produit de convolution, $\mathcal{E}(\theta) \star \mathcal{E}(\theta) = \text{Gamma}(2, \theta)$.

Proposition I-1.4.1: Biais et erreur quadratique du premier estimateur des moments

Pour tout $\theta \in \Theta$, sous $\mathbb{P}_\theta$, $\sum_{i=1}^n X_i \sim \text{Gamma}(n, \theta)$. Par conséquent, pour $n \geq 2$,

$$\mathbb{E}_\theta[\hat{\theta}_{n,1}] = \frac{n}{n-1}\,\theta\,,$$

et, pour $n \geq 3$,

$$\mathbb{E}_\theta[(\hat{\theta}_{n,1} - \theta)^2] = \frac{n+2}{(n-1)(n-2)}\,\theta^2\,.$$

Démonstration

Sous $\mathbb{P}_\theta$, les X_i sont indépendantes de loi $\text{Gamma}(1, \theta)$; le lemme de convolution montre que $Y := \sum_{i=1}^n X_i$ suit la loi $\text{Gamma}(n, \theta)$, de densité $y \mapsto \frac{\theta^n}{(n-1)!}\mathrm{e}^{-\theta y} y^{n-1}$ sur $\mathbb{R}_+^*$ (rappelons que $\Gamma(n) = (n-1)!$). Comme $\hat{\theta}_{n,1} = n/Y$, la formule de transfert donne, pour $n \geq 2$,

$$\mathbb{E}_\theta[\hat{\theta}_{n,1}] = \mathbb{E}_\theta\left[\frac{n}{Y}\right] = \int_0^{+\infty} \frac{\theta^n}{(n-1)!}\,\mathrm{e}^{-\theta y}\,y^{n-1}\,\frac{n}{y}\,\mathrm{d}y = \frac{n\,\theta^n}{(n-1)!}\,\frac{(n-2)!}{\theta^{n-1}} = \frac{n}{n-1}\,\theta\,,$$

où l'on a utilisé $\int_0^{+\infty} \mathrm{e}^{-\theta y} y^{n-2}\mathrm{d}y = \Gamma(n-1)/\theta^{n-1}$; cette intégrale diverge pour $n = 1$, auquel cas $\mathbb{E}_\theta[\hat{\theta}_{1,1}] = \mathbb{E}_\theta[1/X_1] = +\infty$. De même, pour $n \geq 3$,

$$\mathbb{E}_\theta[\hat{\theta}_{n,1}^2] = \int_0^{+\infty} \frac{\theta^n}{(n-1)!}\,\mathrm{e}^{-\theta y}\,y^{n-1}\,\frac{n^2}{y^2}\,\mathrm{d}y = \frac{n^2\,\theta^n}{(n-1)!}\,\frac{(n-3)!}{\theta^{n-2}} = \frac{n^2}{(n-1)(n-2)}\,\theta^2\,.$$

Il est important de remarquer que ces calculs reviennent à utiliser des renormalisations de lois Gamma : d'après (I-1.3), pour $Y \sim \text{Gamma}(n, \theta)$ et $k \in \mathbb{N}$, $\mathbb{E}_\theta[Y^{-k}] = \theta^k\,\Gamma(n-k)/\Gamma(n)$ est fini si et seulement si $n > k$. En développant le carré,

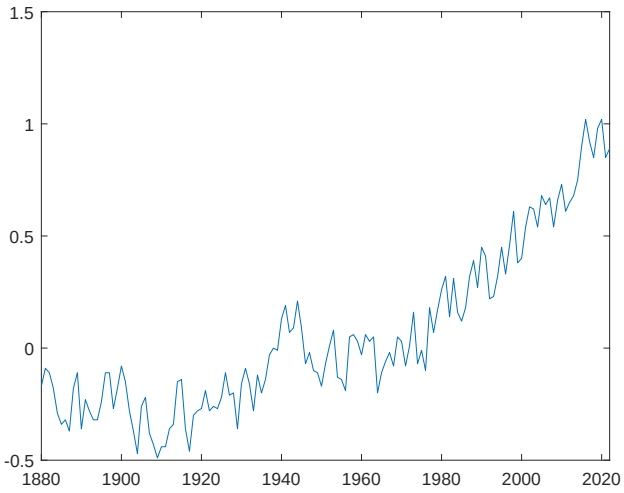
$$\begin{aligned}
\mathbb{E}_\theta[(\hat{\theta}_{n,1} - \theta)^2] &= \mathbb{E}_\theta[\hat{\theta}_{n,1}^2] - 2\theta\,\mathbb{E}_\theta[\hat{\theta}_{n,1}] + \theta^2 \\
&= \theta^2\,\frac{n^2}{(n-1)(n-2)} - 2\theta^2\,\frac{n}{n-1} + \theta^2 \\
&= \theta^2\,\frac{n^2 - 2n(n-2) + (n-1)(n-2)}{(n-1)(n-2)} = \frac{n+2}{(n-1)(n-2)}\,\theta^2\,.
\end{aligned}$$

□

Remarque. L'estimateur $\hat{\theta}_{n,1}$ est donc *biaisé* : son biais $\mathbb{E}_\theta[\hat{\theta}_{n,1}] - \theta = \theta/(n-1)$ est strictement positif, mais tend vers 0 quand $n \to +\infty$; l'estimateur $(n-1)/\sum_{i=1}^n X_i$ est, lui, sans biais. Le

carré de l'erreur en moyenne, $\frac{n+2}{(n-1)(n-2)}\theta^2$, est d'autant plus grand que θ est grand ; l'erreur relative $\mathbb{E}_\theta[(\hat{\theta}_{n,1}/\theta - 1)^2] = \frac{n+2}{(n-1)(n-2)}$ ne dépend pas, elle, de θ. On pourrait mener les mêmes calculs pour le second estimateur des moments $\hat{\theta}_{n,2}$; la figure ci-dessus suggère qu'il est moins précis.

Modèle de régression. Le Goddard Institute for Space Studies (GISS) de la NASA publie chaque année l'anomalie de température moyenne à l'échelle du globe, c'est-à-dire l'écart entre la température moyenne de l'année et celle de la période de référence 1951–1980. La série de 1880 à 2022 est représentée ci-dessous.



Anomalie de température moyenne à l'échelle du globe (en degrés Celsius), année par année de 1880 à 2022, par rapport à la période 1951–1980. Source : NASA, Goddard Institute for Space Studies (GISTEMP).

Exemple I-1.4.9: Construction du modèle

— *Observations* : les déviations des températures annuelles par rapport à la moyenne 1951–1980, $(x_1, \dots, x_n) \in \mathbb{R}^n$, où x_i est l'anomalie de l'année d'indice i ; on note $t(i)$ l'année correspondante (l'index de temps).

— *Modèle statistique* paramétrique, dominé par la mesure de Lebesgue $\mathrm{Leb}^{\otimes n}$, de densité

$$p_{\boldsymbol{\beta},\sigma^2}(x_1, \dots, x_n) := \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\{x_i - m(\boldsymbol{\beta}, i)\}^2\right) ,$$

où $\boldsymbol{\beta} \in \mathbb{R}^p$ est un paramètre et, pour tout $\boldsymbol{\beta} \in \mathbb{R}^p$, $i \mapsto m(\boldsymbol{\beta}, i)$ est une fonction donnée (la *tendance*).

— *Paramètre* : $\theta := (\boldsymbol{\beta}, \sigma^2) \in \Theta := \mathbb{R}^p \times \mathbb{R}_+^*$.

Le modèle statistique est donc

$$\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \left\{p_\theta \cdot \mathrm{Leb}^{\otimes n}\,,\ \theta \in \Theta\right\}\right) .$$

Hypothèses du modèle. Ici $\mathsf{Z} = \mathbb{R}^n$ et $Z = (X_1, \dots, X_n)$ avec $X_i(x_1, \dots, x_n) = x_i$ pour $i \in \{1, \dots, n\}$. La forme produit de la densité traduit trois hypothèses :

(i) pour tout $\theta \in \Theta$, les statistiques X_i et X_j, $i \neq j$, sont indépendantes sous $\mathbb{P}_\theta$;

(ii) sous $\mathbb{P}_\theta$, pour tout $i \in \{1, \dots, n\}$, la loi de X_i est gaussienne de variance σ^2 ;
(iii) sous $\mathbb{P}_\theta$, pour tout $i \in \{1, \dots, n\}$, la moyenne de X_i est $m(\boldsymbol{\beta}, i)$.

Autrement dit, $X_i = m(\boldsymbol{\beta}, i) + \varepsilon_i$ où les erreurs $\varepsilon_1, \dots, \varepsilon_n$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Ce sont des hypothèses de modélisation, qui peuvent (et doivent) être discutées :

— est-il raisonnable de supposer que la variance est constante au cours du temps (modèle *homoscédastique*) ?
— a-t-on une bonne raison de penser que la distribution des erreurs est gaussienne ?
— les observations sont-elles indépendantes ? Pour une série chronologique, deux années consécutives peuvent être liées par des phénomènes climatiques persistants.

Exemple I-1.4.10: Tendance linéaire

Hypothèse 1 : tendance linéaire sur l'ensemble de la période,

$$m(\boldsymbol{\beta}, i) := \beta_0 + \beta_1 t(i) \,, \qquad \boldsymbol{\beta} = (\beta_0, \beta_1) \in \mathbb{R}^2 \,,$$

où $t(i)$ est l'index de temps (ici $p = 2$). On construit un estimateur de $\boldsymbol{\beta}$ en minimisant l'erreur quadratique (*moindres carrés*), puis on estime la variance à partir des résidus :

$$\hat{\boldsymbol{\beta}} := \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^2} \sum_{i=1}^n (X_i - \beta_0 - \beta_1 t(i))^2 \,,$$
$$\hat{\sigma}^2 := (n-2)^{-1} \sum_{i=1}^n (X_i - \hat{\beta}_0 - \hat{\beta}_1 t(i))^2 \,.$$

Dans le modèle gaussien, minimiser l'erreur quadratique en $\boldsymbol{\beta}$ revient à *maximiser la vraisemblance des observations*. En effet,

$$\log p_{\boldsymbol{\beta}, \sigma^2}(X_1, \dots, X_n) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \beta_0 - \beta_1 t(i))^2 \,,$$

de sorte que, quelle que soit la valeur de σ^2, le maximum en $\boldsymbol{\beta}$ est atteint en $\hat{\boldsymbol{\beta}}$; en maximisant ensuite en σ^2, on trouve

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_{\mathrm{MV}}) = \operatorname*{arg\,max}_{(\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^2 \times \mathbb{R}_+^*} p_{\boldsymbol{\beta}, \sigma^2}(X_1, \dots, X_n) \,, \qquad \text{avec } \hat{\sigma}^2_{\mathrm{MV}} := n^{-1} \sum_{i=1}^n (X_i - \hat{\beta}_0 - \hat{\beta}_1 t(i))^2 \,.$$

L'estimateur du maximum de vraisemblance de σ^2 est donc $\hat{\sigma}^2_{\mathrm{MV}} = \frac{n-2}{n} \hat{\sigma}^2$; la normalisation par $n-2$ (nombre d'observations moins nombre de coefficients estimés) fait de $\hat{\sigma}^2$ un estimateur sans biais de σ^2, comme on le verra ci-dessous.

Proposition I-1.4.2: Solution explicite des moindres carrés

Notons $\bar{t} := n^{-1} \sum_{i=1}^n t(i)$ et $\bar{X}_n := n^{-1} \sum_{i=1}^n X_i$. Si les dates $t(1), \dots, t(n)$ ne sont pas toutes égales (au moins deux dates distinctes), la solution des moindres carrés est unique et explicite :

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (t(i) - \bar{t})(X_i - \bar{X}_n)}{\sum_{i=1}^n (t(i) - \bar{t})^2} \,, \qquad \widehat{\beta}_0 = \bar{X}_n - \widehat{\beta}_1 \, \bar{t} \,,$$

ou, de façon plus compacte,

$$\widehat{\boldsymbol{\beta}} = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{X} \quad \text{avec} \quad \Phi := \begin{bmatrix} 1 & t(1) \\ \vdots & \vdots \\ 1 & t(n) \end{bmatrix} \quad \text{et} \quad \mathbf{X} := \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \,.$$

La condition sur les dates équivaut à ce que Φ soit de rang 2, c'est-à-dire à l'inversibilité de $\Phi^\top\Phi$. L'estimateur de la variance s'écrit

$$\hat\sigma^2 = (n-2)^{-1}\sum_{i=1}^{n}\{X_i - \widehat\beta_0 - \widehat\beta_1 t(i)\}^2 = (n-2)^{-1}\left\|\mathbf{X} - \Phi\hat{\boldsymbol\beta}\right\|^2 .$$

Démonstration

La fonction $\boldsymbol\beta \mapsto \|\mathbf{X} - \Phi\boldsymbol\beta\|^2$ est une forme quadratique convexe, de gradient $-2\Phi^\top(\mathbf{X} - \Phi\boldsymbol\beta)$; elle est strictement convexe dès que $\Phi^\top\Phi$ est inversible, et son unique minimum est alors caractérisé par les *équations normales* $\Phi^\top\Phi\,\boldsymbol\beta = \Phi^\top\mathbf{X}$. Explicitement, ces équations s'écrivent $n\beta_0 + \beta_1\sum_i t(i) = \sum_i X_i$ et $\beta_0\sum_i t(i) + \beta_1\sum_i t(i)^2 = \sum_i t(i)X_i$; la première donne $\beta_0 = \bar X_n - \beta_1\bar t$ et, en reportant dans la seconde, $\beta_1\sum_i(t(i)-\bar t)^2 = \sum_i(t(i)-\bar t)(X_i - \bar X_n)$. □

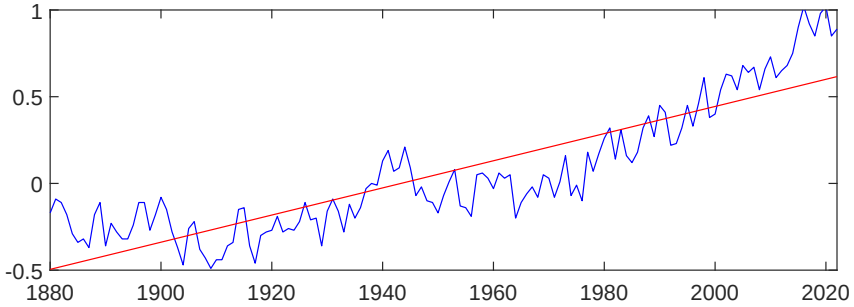
Proposition I-1.4.3: Lois des estimateurs

Pour tout $(\boldsymbol\beta, \sigma^2) \in \mathbb{R}^2 \times \mathbb{R}_+^*$, sous $\mathbb{P}_{\boldsymbol\beta,\sigma^2}$,

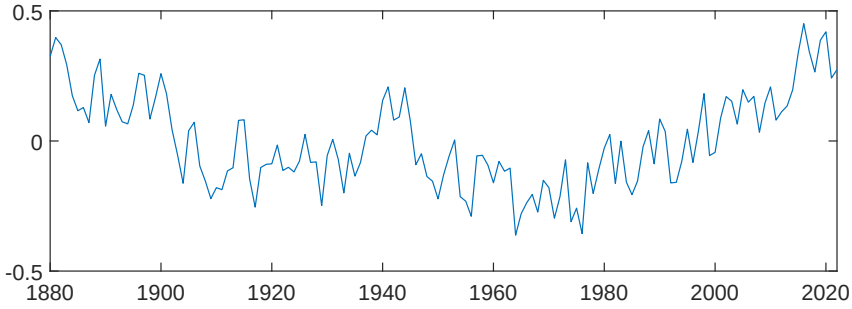
$$\widehat{\boldsymbol\beta} \sim \mathcal{N}\big(\boldsymbol\beta, \sigma^2(\Phi^\top\Phi)^{-1}\big) , \qquad \frac{(n-2)\,\hat\sigma^2}{\sigma^2} \sim \chi^2(n-2) ,$$

et les statistiques $\widehat{\boldsymbol\beta}$ et $\hat\sigma^2$ sont indépendantes. En particulier, $\mathbb{E}_{\boldsymbol\beta,\sigma^2}[\widehat{\boldsymbol\beta}] = \boldsymbol\beta$ et $\mathbb{E}_{\boldsymbol\beta,\sigma^2}[\hat\sigma^2] = \sigma^2$.

Ces résultats découlent du théorème de Cochran, énoncé dans le cours principal (voir l'énoncé), que nous ne redémontrons pas ici : sous $\mathbb{P}_{\boldsymbol\beta,\sigma^2}$, $\mathbf{X} \sim \mathcal{N}(\Phi\boldsymbol\beta, \sigma^2\mathrm{I}_n)$, $\Phi\widehat{\boldsymbol\beta}$ est la projection orthogonale de $\mathbf{X}$ sur l'image de Φ, sous-espace de dimension 2, et $\mathbf{X} - \Phi\widehat{\boldsymbol\beta}$ est sa projection sur l'orthogonal de ce sous-espace, de dimension $n-2$. Le modèle de régression linéaire gaussienne est l'objet de l'exercice 4 de la PC1, poursuivi en PC 2.



Anomalie de température moyenne à l'échelle du globe par rapport à la période 1951–1980 (NASA, GISS), en bleu, et droite $t \mapsto \widehat\beta_0 + \widehat\beta_1 t$ ajustée par moindres carrés, en rouge. La pente estimée est $\widehat\beta_1 = 0{,}0072\,^\circ\mathrm{C}$ par an, avec un écart-type estimé $\mathrm{se}_{\beta_1} = 0{,}0008\,^\circ\mathrm{C}$ par an ; l'écriture $\widehat\beta_1 \pm \mathrm{se}_{\beta_1}$ est une estimation accompagnée de son écart-type estimé et non un intervalle de confiance (un intervalle de confiance à 95 % serait $\widehat\beta_1 \pm 1{,}96\,\mathrm{se}_{\beta_1}$, soit environ de $0{,}0056$ à $0{,}0088\,^\circ\mathrm{C}$ par an).



Résidus de l'ajustement linéaire, $\hat{R}_i := X_i - \hat{\beta}_0 - \hat{\beta}_1 t(i)$, en fonction de l'année. Si le modèle à tendance linéaire était adéquat, les résidus devraient se comporter comme un bruit sans structure ; on observe au contraire des résidus négatifs sur de longues périodes (1905–1930, 1950–1980) et positifs aux deux extrémités de la série et autour de 1940, ce qui suggère que la tendance n'est pas linéaire sur l'ensemble de la période.

Exemple I-1.4.11: Modèle avec rupture de pente

Hypothèse 2 : rupture de pente à une date inconnue. On pose

$$m(\boldsymbol{\beta}, i) := \beta_0 + \beta_1 t(i) + \beta_2 \big(t(i) - t(\beta_3)\big)\, \mathbb{1}_{\{i > \beta_3\}} , \qquad \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) \in \mathbb{R}^3 \times \{1, \dots, n\} .$$

La tendance est ainsi affine par morceaux et continue : de pente β_1 jusqu'à la date $t(\beta_3)$, puis de pente $\beta_1 + \beta_2$ ensuite ; le paramètre β_3 est l'indice de la date de rupture. On construit les estimateurs en minimisant l'erreur quadratique,

$$\hat{\boldsymbol{\beta}} := \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^3 \times \{1,\dots,n\}} \sum_{i=1}^{n} (X_i - m(\boldsymbol{\beta}, i))^2 ,$$

$$\hat{\sigma}^2 := n^{-1} \sum_{i=1}^{n} (X_i - m(\hat{\boldsymbol{\beta}}, i))^2 ,$$

ce qui fournit, par le même argument que pour la tendance linéaire, l'estimateur du maximum de vraisemblance

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \operatorname*{arg\,max}_{\boldsymbol{\beta}, \sigma^2} p_{\boldsymbol{\beta}, \sigma^2}(X_1, \dots, X_n) .$$

Le paramètre β_3 étant discret, la minimisation se fait en deux temps.

(i) Pour β_3 – la position de la rupture de pente – fixé, le critère $(\beta_0, \beta_1, \beta_2) \mapsto \sum_{i=1}^{n} (X_i - m(\boldsymbol{\beta}, i))^2$ est à nouveau un critère de moindres carrés linéaire, et les paramètres $(\beta_0, \beta_1, \beta_2)$ peuvent être calculés de façon explicite :

$$\hat{\boldsymbol{\beta}}(\beta_3) := \begin{bmatrix} \hat{\beta}_0(\beta_3) \\ \hat{\beta}_1(\beta_3) \\ \hat{\beta}_2(\beta_3) \end{bmatrix} = \big(\Phi(\beta_3)^\top \Phi(\beta_3)\big)^{-1} \Phi(\beta_3)^\top \mathbf{X} ,$$

où

$$\Phi(\beta_3) := \begin{bmatrix} 1 & t(1) & 0 \\ \vdots & \vdots & \vdots \\ 1 & t(\beta_3) & 0 \\ 1 & t(\beta_3 + 1) & t(\beta_3 + 1) - t(\beta_3) \\ \vdots & \vdots & \vdots \\ 1 & t(n) & t(n) - t(\beta_3) \end{bmatrix} ,$$

pourvu que $\Phi(\beta_3)$ soit de rang 3, ce qui impose au moins deux dates distinctes de part et d'autre de la rupture ; en pratique, on restreint donc β_3 à un ensemble de valeurs admissibles.

(ii) On estime ensuite β_3 en maximisant la *vraisemblance profilée* $\beta_3 \mapsto \max_{(\beta_0,\beta_1,\beta_2,\sigma^2)} p_{\boldsymbol{\beta},\sigma^2}(X_1,\dots,X_n)$. À β_3 fixé, ce maximum vaut $\big(2\pi \mathrm{e}\,\mathrm{RSS}(\beta_3)/n\big)^{-n/2}$ avec $\mathrm{RSS}(\beta_3) := \big\|\mathbf{X} - \Phi(\beta_3)\,\hat{\boldsymbol{\beta}}(\beta_3)\big\|^2$; c'est une fonction décroissante de la somme des carrés des résidus, de sorte que maximiser la vraisemblance profilée revient à minimiser cette somme :

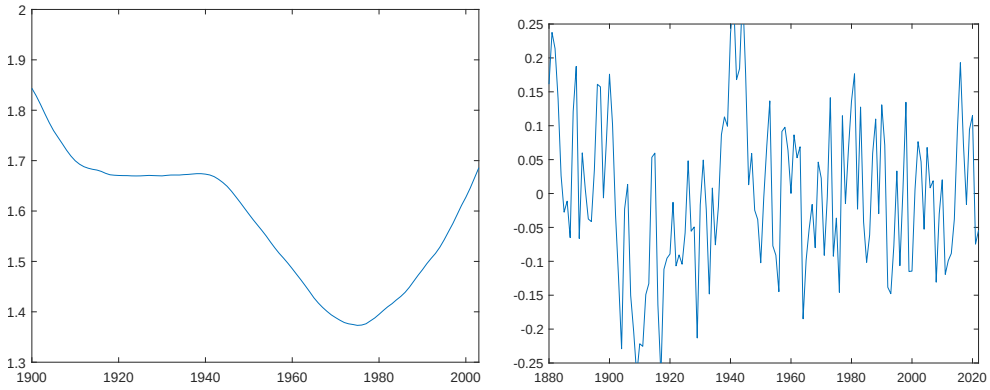
$$\hat{\beta}_3 := \operatorname*{arg\,min}_{\beta_3 \in \{1,\dots,n\}} \big\|\mathbf{X} - \Phi(\beta_3)\,\hat{\boldsymbol{\beta}}(\beta_3)\big\|^2 ,$$

puis $\hat{\boldsymbol{\beta}} = \big(\hat{\boldsymbol{\beta}}(\hat{\beta}_3), \hat{\beta}_3\big)$. Cette dernière minimisation se fait par simple énumération des valeurs admissibles de β_3.

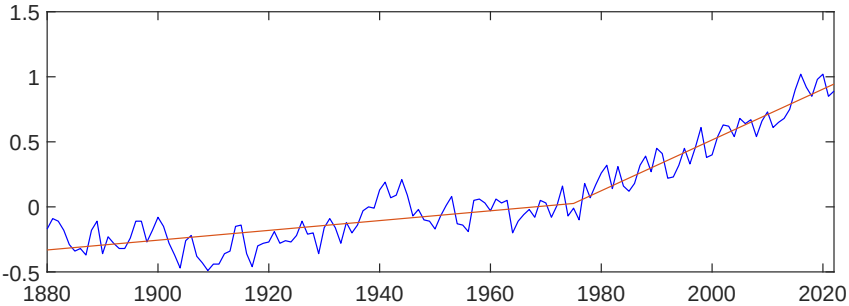
Illustration / animation : La vraisemblance profilée désigne la position de rupture qui minimise les résidus

Vous pouvez retester ici, pour chaque position candidate de rupture, l'ajustement obtenu et la vraisemblance profilée associée, dont le maximum désigne la position de rupture retenue.

link



À gauche : somme des carrés des résidus $\mathrm{RSS}(\beta_3) = \big\|\mathbf{X} - \Phi(\beta_3)\hat{\boldsymbol{\beta}}(\beta_3)\big\|^2$ *en fonction de l'année de rupture* $t(\beta_3)$*, pour les années de rupture admissibles de 1900 à 2003 ; la vraisemblance profilée en est une fonction décroissante, et son maximum, atteint vers 1975, correspond au minimum de la courbe. À droite : résidus* $X_i - m(\hat{\boldsymbol{\beta}}, i)$ *du modèle avec rupture de pente, qui ne présentent plus la structure observée pour la tendance linéaire.*



Anomalie de température moyenne à l'échelle du globe par rapport à la période 1951–1980 (NASA, GISS), en bleu, et tendance ajustée $i \mapsto m(\hat{\boldsymbol{\beta}}, i)$ *du modèle avec rupture de pente, en rouge : une pente faible jusqu'au milieu des années 1970, puis nettement plus forte.*

Systèmes de recommandation. Les plateformes de diffusion de films ou de musique et les sites de vente en ligne cherchent à prédire l'intérêt d'un utilisateur pour un produit qu'il n'a pas encore noté, à partir des notes déjà attribuées par l'ensemble des utilisateurs.

				
Claire	4	?	3	...
Nial	?	4	?	...
Brendon	?	5	4	...
Andrew	?	4	?	...
Adrian	1	?	?	...
Damien	?	1	?	...
⋮	⋮	⋮	⋮	⋱

Extrait d'une matrice de notes : chaque ligne correspond à un utilisateur et chaque colonne à un film. Les notes observées vont de 1 à 5 ; les points d'interrogation signalent les notes manquantes, que l'on souhaite prédire pour recommander à chaque utilisateur les films susceptibles de lui plaire.

Exemple I-1.4.12: Une formalisation statistique possible

- m_1 : le nombre de sujets (utilisateurs) ;
- m_2 : le nombre d'objets (films, produits, etc.) ;
- Y : la matrice $m_1 \times m_2$ dont l'entrée $Y_{i,j}$ est la note du sujet i au produit j ;
- *données* : les notes $Y_{i,j}$ pour $(i,j) \in I$, où $I \subset \{1,\dots,m_1\} \times \{1,\dots,m_2\}$ est l'ensemble des couples pour lesquels une note est observée, et

$$n = \text{ nombre d'observations } = \mathrm{card}(I) \ll m_1 m_2 \ .$$

Les notes prennent leurs valeurs dans $\{1,\dots,K\}$ ($K = 5$ dans l'illustration). On suppose que les notes $Y_{i,j}$, $(i,j) \in I$, sont indépendantes et que $Y_{i,j}$ suit une loi catégorielle (loi multinomiale d'ordre 1) sur $\{1,\dots,K\}$, de probabilités

$$\mathbb{P}_\theta(Y_{i,j} = \ell) = \frac{\exp(\theta_{i,j,\ell})}{\sum_{k=1}^K \exp(\theta_{i,j,k})} \ , \qquad \ell \in \{1,\dots,K\} \ .$$

Par convention, on fixe $\theta_{i,j,K} = 0$: autrement le modèle n'est pas *identifiable*, puisque remplacer $(\theta_{i,j,1},\dots,\theta_{i,j,K})$ par $(\theta_{i,j,1}+c,\dots,\theta_{i,j,K}+c)$ ne change pas ces probabilités. Le paramètre est $\theta = (\theta_{i,j,\ell})$ pour $(i,j) \in \{1,\dots,m_1\} \times \{1,\dots,m_2\}$ et $\ell \in \{1,\dots,K-1\}$. L'espace des observations est $\mathsf{Z} := \{1,\dots,K\}^I$, muni de la tribu de ses parties, et la densité de l'observation par rapport à la mesure de comptage μ sur Z est

$$p_\theta\big(\{y_{i,j}\}_{(i,j)\in I}\big) = \prod_{(i,j)\in I} \prod_{\ell=1}^K \left(\frac{\exp(\theta_{i,j,\ell})}{\sum_{k=1}^K \exp(\theta_{i,j,k})} \right)^{\mathbb{1}_{\{y_{i,j}=\ell\}}} \ .$$

Le modèle statistique est $\big(\mathsf{Z}, \mathcal{P}(\mathsf{Z}), \{p_\theta \cdot \mu \ , \ \theta \in \Theta\}\big)$ avec $\Theta = \mathbb{R}^{m_1 \times m_2 \times (K-1)}$.

Une hypothèse sur la structure de θ. Tel quel, le modèle comporte $(K-1)\,m_1 m_2$ paramètres réels pour seulement $n \ll m_1 m_2$ observations, et les paramètres $\theta_{i,j,\cdot}$ des couples $(i,j) \notin I$ n'apparaissent même pas dans la densité : sans hypothèse supplémentaire, les données n'apportent aucune information sur les notes manquantes. Il est donc nécessaire de faire une hypothèse sur la structure de θ, qui soit raisonnable pour l'application considérée, et plusieurs possibilités existent. Considérons pour simplifier des notes binaires, $K = 2$ (par exemple « j'aime » ou « je n'aime pas ») : avec la convention $\theta_{i,j,2} = 0$, il reste un seul paramètre $\theta_{i,j} := \theta_{i,j,1}$ par couple (i,j),

$$\mathbb{P}_\theta(Y_{i,j} = 1) = \frac{\exp(\theta_{i,j})}{1+\exp(\theta_{i,j})} \,,$$

et $\theta = (\theta_{i,j})$ est une matrice $m_1 \times m_2$. L'*hypothèse de rang faible* exprime que les préférences des utilisateurs sont des mélanges d'un petit nombre q de comportements « prototypes », c'est-à-dire que la matrice θ est de rang au plus q :

$$\theta = UV^\top \,, \qquad U \in \mathbb{R}^{m_1 \times q} \,, \quad V \in \mathbb{R}^{m_2 \times q} \,,$$

soit $\theta_{i,j} = \langle u_i, v_j \rangle$, où $u_i \in \mathbb{R}^q$ (la i-ème ligne de U) décrit l'utilisateur i et $v_j \in \mathbb{R}^q$ (la j-ème ligne de V) décrit le produit j dans un même espace de « facteurs latents ». Le nombre de paramètres est alors $q(m_1+m_2)$, à q^2 près : la factorisation n'est pas unique, puisque $UV^\top = (UA)\big(V(A^{-1})^\top\big)^\top$ pour toute matrice inversible A de taille $q \times q$, et l'ensemble des matrices de rang q est de dimension $q(m_1+m_2) - q^2$. Le rang q est souvent choisi de telle sorte que

$$q(m_1+m_2) \ll n \ll m_1 m_2 \,.$$

Seule la matrice θ, et non le couple (U,V), est identifiable ; c'est bien elle, c'est-à-dire l'ensemble des probabilités $\mathbb{P}_\theta(Y_{i,j}=1)$ pour tous les couples (i,j), qui importe pour la recommandation.

Recommandation.

- Estimer les matrices U et V, ou plutôt leur produit $\theta = UV^\top$, à partir des notes observées. Dans ce cas, comme dans la plupart des applications « réelles », il n'y a pas de méthode d'estimation « élémentaire » : les méthodes générales (maximum de vraisemblance, méthode des moments) sont l'objet du cours 2.
- Imputer les données manquantes : pour $(i,j) \notin I$, la note $Y_{i,j}$ n'a pas été observée, mais l'estimation $\hat\theta_{i,j} = \langle \hat u_i, \hat v_j \rangle$ fournit une prédiction de la probabilité que le sujet i apprécie le produit j ; on lui recommande alors les produits pour lesquels cette probabilité est la plus élevée.
- C'est la base des systèmes de recommandation ou de *filtrage collaboratif* : les notes des autres utilisateurs, à travers les facteurs latents partagés, renseignent sur les goûts de chacun. Dans un système opérationnel, il y a bien entendu un certain nombre de raffinements à apporter, mais le principe est celui-là.

Pour aller plus loin : mesures de domination et modèles non dominés

La mesure de domination peut être choisie de multiples façons. Soient μ et ν deux mesures σ-finies distinctes sur $(\mathsf{Z}, \mathcal{Z})$. Supposons que le modèle statistique paramétrique $(\mathsf{Z}, \mathcal{Z}, \mathcal{C}) = (\mathsf{Z}, \mathcal{Z}, \{\mathbb{P}_\theta \,,\ \theta \in \Theta\})$ est dominé par rapport à μ et à ν. Cette hypothèse implique que, pour tout $\theta \in \Theta$, nous pouvons associer deux fonctions mesurables positives, p_θ et q_θ telles que

$$\mathbb{P}_\theta = p_\theta \cdot \mu = q_\theta \cdot \nu \,.$$

La fonction p_θ est la densité de $\mathbb{P}_\theta$ par rapport à μ et q_θ est la densité de $\mathbb{P}_\theta$ par rapport à ν. Il est facile de montrer que ces deux densités diffèrent d'un facteur multiplicatif. Le raisonnement élémentaire est le suivant. Notons tout d'abord que les mesures σ-finies μ et ν sont absolument continues par rapport à $\lambda := \mu + \nu$. En effet, pour tout $B \in \mathcal{Z}$, $\lambda(B) = \mu(B) + \nu(B) = 0$ implique que $\mu(B) = 0$ et $\nu(B) = 0$. En appliquant le **??**, il existe donc des fonctions mesurables positives

m et n, uniques à une λ-équivalence près, telles que $\mu = m \cdot \lambda$ et $\nu = n \cdot \lambda$. Nous avons donc, pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta = p_\theta \cdot \mu = p_\theta m \cdot \lambda = q_\theta \cdot \nu = q_\theta n \cdot \lambda \, ,$$

et donc pour tout $\theta \in \Theta$, $p_\theta m = q_\theta n \quad \lambda$-p.p. Posons $\mathsf{Z}_0 = \{z \in \mathsf{Z} \, , \; m(z) > 0\}$. Notons que

$$\mu(\mathsf{Z}_0^c) = \int_{\mathsf{Z}_0^c} m(z) \lambda(\mathrm{d}z) = 0 \, ,$$

et donc, pour tout $\theta \in \Theta$, $\mathbb{P}_\theta(\mathsf{Z}_0^c) = 0$. Pour tout $z \in \mathsf{Z}_0$ et $\theta \in \Theta$, nous avons

$$p_\theta(z) = q_\theta(z) \frac{n(z)}{m(z)} \, , \quad \lambda - \text{p.p.}$$

Remarque. L'identité précédente ne vaut que λ-presque partout : une densité n'est définie qu'à un ensemble négligeable près, et l'argument montre que le facteur n/m ne dépend pas de θ. L'invariance des points de maximum de la vraisemblance, invoquée dans les fondamentaux et au chapitre 2, vaut donc pour des versions fixées des densités. La réserve n'est pas gratuite : la vraisemblance s'évalue aux points observés, qui forment eux-mêmes un ensemble négligeable pour les lois à densité, et modifier chaque p_θ sur un ensemble μ-négligeable *qui dépend de* θ produit une autre famille de versions, tout aussi légitime, dont les points de maximum peuvent être différents. Y. Baraud et L. Birgé (*Rho-estimators revisited : general theory and applications*, Annals of Statistics, vol. 46, 2018, proposition 1) construisent ainsi, dans le modèle $\{\mathcal{N}(\theta, 1) \, , \; \theta \in \mathbb{R}\}$ dominé par $\mu = \mathcal{N}(0, 1)$, des versions des densités modifiées au seul point $z = \theta$ pour lesquelles l'estimateur du maximum de vraisemblance vaut $\max(X_1, \ldots, X_n)$ avec une probabilité qui tend vers un, quelle que soit la vraie valeur du paramètre : la consistance est perdue. Les versions continues des densités, utilisées dans tout ce cours, écartent cette pathologie. Nous remercions Matthieu Lerasle de nous avoir signalé ce point et cette référence.

Les deux exemples qui suivent sont plus avancés et peuvent être omis en première lecture.

Exemple I-1.4.13

Un exemple où il n'existe pas de mesure dominante est la famille paramétrique $\{\mathbb{P}_\theta = \delta_\theta, \theta \in \Theta = \mathbb{R}\}$, où δ_θ est la mesure de Dirac au point θ. Cet exemple correspond à l'"expérience parfaite" où une seule réalisation de la loi permet de déterminer sans erreur la valeur du paramètre. Montrons que ce modèle n'est pas dominé. Nous procédons par contradiction et supposons que ce modèle est dominé par une mesure σ-finie μ. Alors, nous avons $\mu(\{\theta\}) > 0$ pour tout $\theta \in \Theta$ (car $\mathbb{P}_\theta \ll \mu$ et $\mathbb{P}_\theta(\{\theta\}) > 0$), ce qui est impossible car une mesure σ-finie a au plus un nombre dénombrable d'atomes.
Nous allons prouver cette dernière propriété par contradiction. Comme μ est σ-finie, il existe une suite $\{F_k, \; k \in \mathbb{N}\}$ telle que $\mu(F_k) < \infty$ pour tout $k \in \mathbb{N}$, et $\mathbb{R} = \bigcup_{k=1}^\infty F_k$. On note $I = \{\theta \in \mathbb{R} \, : \, \mu(\{\theta\}) > 0\}$ et on rappelle que, par hypothèse, cet ensemble n'est pas dénombrable. Comme $I = \bigcup_{k=1}^\infty F_k \cap I$, il existe $n \in \mathbb{N}$ tel que $I \cap F_n$ est non dénombrable. Pour tout $p \in \mathbb{N}$, notons alors $I_p = \{\theta \in I \cap F_n \, : \, \mu(\{\theta\}) > 1/p\}$. Comme $I \cap F_n = \bigcup_{p=1}^\infty I_p$, il existe $q \in \mathbb{N}$ tel que I_q est non dénombrable. En particulier, I_q est donc infini. Pour tout sous-ensemble J de cardinal N de I_q, nous avons

$$\mu(F_n) \geq \mu(I_q) \geq \mu(J) = \sum_{\theta \in J} \mu(\{\theta\}) \geq \frac{N}{p} \, .$$

Comme p est fixé et que N peut être choisi arbitrairement grand, nous avons nécessairement $\mu(F_n) = \infty$, ce qui conduit à une contradiction.

Exemple I-1.4.14

Un exemple plus subtil est donné par le modèle statistique dont la famille de lois est donnée par

$$P_\theta = \sum_{k=1}^{\infty} \mathrm{e}^{-k} \delta_{\theta k} \ , \quad \text{où } \theta \in \Theta = \mathbb{R}_+ \setminus \{0\} \text{ est le paramètre.}$$

Dans ce cas, la loi de l'observation est non dégénérée (réduite à un point). Le modèle n'est pas dominé car là aussi, toute mesure de domination aurait alors nécessairement une infinité non dénombrable d'atomes.