Modèles de durée

Motivation:

* $X \geq 0$ p.s
* observations incomplètes
  - tronquées à gauche
    $X \mid X > e$
  - censure droite
    $\begin{cases} X \wedge C = T \\ \mathbb{1}_{X<C} \end{cases}$

On a $(X_1, \ldots, X_n)$ iid
↳ informations sur F_X.

I / Introduction : Outils

1- Outils de base

$F(x) = \mathbb{P}(X \leq x)$
$f(x) = \frac{d}{dx} F(x)$

$S(x) = \mathbb{P}(X > x) = 1 - F(x) \longrightarrow \mathbb{E}[X] = \int_0^\infty S(x)\,dx$

↳ $\mathbb{E}[X] = \int_0^\infty x\,dF(x) = -\int_0^\infty x\,dS(x)$

$-\int_0^u x\,dS(x) = uS(u) - \int_0^u S(x)\,dx \xrightarrow[u \to \infty]{} \int_0^\infty S(x)\,dx$

MARKOV $\to uS(u) \leq \mathbb{E}[X]$
$\Rightarrow$ borné $\Rightarrow$ CV

↳ $\int_0^\infty S(x)\,dx = \int_0^\infty \mathbb{P}(X > x)\,dx = \int_0^\infty \mathbb{E}[\mathbb{1}_{X>x}]\,dx$
$= \mathbb{E}[\underbrace{\int_0^\infty \mathbb{1}_{X>x}\,dx}_{=X}]$

Exercice : Mq $\mathbb{E}[X^2] = 2\int_0^\infty x S(x)\,dx$

$\mathbb{E}[X^2] = \int_0^\infty x^2 f(x)\,dx = \int_0^\infty x^2 \frac{d}{dx} F(x)\,dx = \int_0^\infty x^2\,dF(x) = -\int_0^\infty x^2\,dS(x)$

$u = x^2 \quad u' = 2x$
$v' = dS(x) \quad v = S(x)$

$= -[x^2 S(x)]_0^\infty + \int_0^\infty 2u\,S(u)\,du$
$= 0 + \int_0^\infty 2x\,S(x)\,dx$

D'où $\mathbb{E}[X^2] = 2\int_0^\infty x S(x)\,dx$.

$\int uv' = [uv] - \int u'v$

Survie conditionnelle

$$\mathbb{P}(X > x+u \mid X>u) = \frac{\mathbb{P}(X>x+u)}{\mathbb{P}(X>u)} = \frac{S(u+x)}{S(u)} = S_u(x)$$

$$X_u \overset{\mathcal{L}}{=} X \mid X>u$$

Fonction de Hasard :

$$\mathbb{P}(X \leq x+1 \mid X>x) = q_x = \frac{S(x) - S(x+1)}{S(x)}$$

$$\mathbb{P}(X \leq x+u \mid X>x) = \frac{S(x)-S(x+u)}{u\,S(x)} \xrightarrow[u\to 0^+]{} \frac{-S'(x)}{S(x)}$$

$$\lim_{u\to 0^+} \mathbb{P}(X \leq x+u \mid X>x) = h(x) = -\frac{S'(x)}{S(x)}$$

Si u est "petit", $\mathbb{P}(X\leq x+u \mid X>x) \approx u\,h(x)$

$$\longrightarrow \boxed{S(x) = \exp\left(-\int_0^x h(t)\,dt\right)}$$

$$X_u \overset{\mathcal{L}}{=} X \mid X>u$$

$$S_u(x) = \frac{S(u+x)}{S(u)} = \frac{\exp\left(-\int_0^{u+x} h(t)\,dt\right)}{\exp\left(-\int_0^u h(t)\,dt\right)} = \exp\left(-\int_u^{u+x} h(v)\,dv\right)$$

$$h_u(x) = -\frac{d}{dx}\ln\left(\frac{S(u+x)}{S(u)}\right) = h(u+x)$$

$$S(x) = e^{-\int_0^x h(t)\,dt} = e^{-H(x)}$$

Loi de $H(X)$? $Y = H(X)$

$$\mathbb{P}(Y>x) = \mathbb{P}(H(X)>x) = \mathbb{P}(X > H^{-1}(x))$$
$$= e^{-H(H^{-1}(x))} = e^{-x}$$

$$H(t) = \int_0^t h(u)\,du$$
$$h(x) = \lambda \quad \forall x>0$$
$$S(x) = e^{-\lambda x}, \leadsto \exp(\lambda x)$$

Espérance de vie conditionnelle :

$$e(x) = \mathbb{E}[X \mid X>x]$$

$$X_x \overset{\mathcal{L}}{=} X \mid X>x \qquad S_x(u) = \frac{S(x+u)}{S(x)}$$

$$e(x) = \int_0^\infty S_x(u)\,du = \frac{1}{S(x)}\int_x^\infty S(u)\,du$$

$$\tilde{e}(u) = \frac{1}{\ell_x}\sum_{y \geq x} \ell_y$$

$$e(x) = \frac{1}{S(x)} \sum_{l \geq 0} \underbrace{\int_{x+l}^{x+l+1} S(u)\,du}_{\substack{\approx S(x+l+1) \\ \approx S(x+l)}}$$

$$e(x) = \frac{1}{S(x)} \int_x^{\infty} S(u)\,du$$

$$e'(x) = \frac{1}{S(x)^2}\left[-S(x)^2 - \int_x^{\infty} S(u)\,du\, S'(x)\right]$$

$$= -1 - \underbrace{\frac{1}{S(x)} \int_x^{\infty} S(u)\,du}_{e(x)}\ \underbrace{\frac{S'(x)}{S(x)}}_{-h(x)}$$

$$\longrightarrow \boxed{e'(x) = -1 + h(x)\,e(x)}$$

$$h(x) = \frac{1+e'(x)}{e(x)}$$

<u>Exercice</u>:

* Si h est petit, $e' \to -1$ => espérance de vie perd 1 an par an...

→ Modèles de durée à espérance de vie affine

i.e modèles tq $e(x) = ax + b$.

$$\Rightarrow h(x) = \frac{1+a}{ax+b} = \frac{1+a}{a}\,\frac{a}{ax+b} = \frac{1+a}{a}\,\frac{d}{dx} \ln\left(\frac{ax+b}{b}\right)$$

$$\Rightarrow S(x) = \left(1 + \frac{ax}{b}\right)^{-(1+1/a)}$$

* <u>Cas particulier</u>: $S_u(x) = \dfrac{S(x+u)}{S(u)}$

$S(x+u) = S(x)S(u)$ Modèles de durée ?

$$S_u(x) = \frac{S(x+u) - S(u)}{S(u)} = \frac{S(x)S(u) - S(u)}{S(u)} = S(x) - 1$$

$$\frac{S(x+u) - S(x)}{u} = \frac{S(x)S(u) - S(x)}{u} = S(x)\,\frac{S(u)-1}{S(u)}$$

$\lambda = -S'(0)$

$$\Rightarrow S'(x) = -S(x)\lambda$$

$$\to S(x) = e^{-\lambda x}$$

$t_0 = 0$, t_1, t_2, $t = t_m$

$$\mathbb{P}(X \leq t_i \mid X > t_{i-1}) = \frac{S(t_{i-1}) - S(t_i)}{S(t_{i-1})}$$

$$Z_m(t) = \sum_{i=1}^{m} \mathbb{P}(X \leq t_i \mid X > t_{i-1})$$

$$\lim_{m \to 0} Z_m(t) = \lim \sum_{i=1}^{m} -\frac{S(t_i) - S(t_{i-1})}{S(t_{i-1})} = -\int_0^t \frac{dS(u)}{S(u)} = H(t)$$

* $Z = X_1 \wedge X_2$

$$\mathbb{P}(Z > x) = \mathbb{P}(X_1 \wedge X_2 > x)$$

$$= \mathbb{P}(X_1 > x \cap X_2 > x)$$

$$\underset{\text{Supposé } \perp\!\!\!\perp}{=} \mathbb{P}(X_1 > x)\,\mathbb{P}(X_2 > x)$$

$\longrightarrow S_Z(x) = S_{X_1}(x) S_{X_2}(x)$

$h(x) = \lambda x^{\alpha - 1}$

$X_1, \ldots, X_n$ iid $\sim G$

$X_{(1)} = \min(X_1, \ldots, X_n)$

Loi de $X_{(1)}$ lorsque $n \to \infty$?

Supposons que $\lim_{x \to 0^+} \frac{G(x)}{\lambda x^\alpha} = 1 \qquad (G(x) \underset{x \to 0^+}{\sim} \lambda x^\alpha)$

$$\begin{aligned}
\mathbb{P}(n^{1/\alpha} X_{(1)} > x) &= \mathbb{P}(X_{(1)} > x n^{-1/\alpha}) \\
&= \mathbb{P}(X_1 > x n^{-1/\alpha})^n \\
&= (1 - G(x n^{-1/\alpha}))^n \\
&\simeq \left(1 - \lambda \frac{x^\alpha}{n}\right)^n \longrightarrow e^{-\lambda x^\alpha} \sim \mathcal{W}(\lambda, \alpha)
\end{aligned}$$

Loi de Makeham / Thatcher

$h(x) = \alpha + \beta e^{\gamma x}$ (Makeham, 1850)

$X = X_1 \wedge X_2 \qquad \begin{cases} h_{X_1}(x) = \alpha \\ h_{X_2}(x) = \beta e^{\gamma x} \end{cases}$ (Gompertz, 1825)

$h(x) = \alpha + \frac{\beta e^{\gamma x}}{1 + \beta e^{\gamma x}}$ (Thatcher, 2000).

Censure aléatoire droite / troncature gauche

$(X_1, \ldots, X_n)$ iid

$\ln \mathcal{L}_\theta(x_1, \ldots, x_n) = \sum_{i=1}^n \ln(f_\theta(x_i))$

Censure $\longrightarrow$ on n'observe pas X_i mais on observe $Y_i \begin{cases} T_i = X_i \wedge C_i \\ D_i = \mathbb{1}_{X_i \leq C_i} \end{cases}$ $\quad (C_1, \ldots, C_n) \perp (X_1, \ldots, X_n)$.

$(Y_1, \ldots, Y_n)$ iid

$t \leq c < \infty$

$t \leq u \leq c$

$$\begin{aligned}
\mathbb{P}(T > t, D = 1) &= \mathbb{P}(X \wedge C > t, X \leq C) \\
&= \mathbb{P}(X > t, X \leq C) \\
&= \mathbb{P}(t < X \leq C) \\
&= \mathbb{E}[\mathbb{P}(t < X \leq c \mid C = c)] \\
&= \int_t^\infty \int_t^c f_X(u)\, du\, f_C(c)\, dc \overset{\text{Fubini}}{=} \int_t^\infty \int_u^\infty f_C(c) f_X(u)\, dc\, du \\
&= \int_t^\infty f_X(u) S_C(u)\, du = S_X(t) S_C(t)
\end{aligned}$$

$\longrightarrow -\frac{d}{dt} \mathbb{P}(T > t, D = 1) = f_X(t) S_C(t)$

$\mathbb{P}(T > t, D = 0) = f_C(t) S_X(t)$

$$\mathcal{L}(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} [f_X(t_i) S_C(t_i)]^{d_i} [f_C(t_i) S_X(t_i)]^{1-d_i}$$

$$\boxed{\mathcal{L} = \prod_{i=1}^{n} h_X(t_i)^{d_i} S_X(t_i) \prod_{i=1}^{n} S_C(t_i) h_C(t_i)^{1-d_i}}$$

Hypothèse : "Censure non informative"
$\Leftrightarrow$ Les lois de X et C n'ont pas de paramètres commun.

$$\ln(\mathcal{L}_\theta(Y_1, \ldots, Y_n)) = \sum_{i=1}^{n} [d_i \ln(h_\theta(t_i)) + \ln(S_\theta(t_i))].$$

Exemple de censure informative

$X \perp\!\!\!\perp C$

$S_C(x) = S_X(x)^\beta$

$h_C(x) = \beta h_X(x).$

Troncature gauche E.

$X_i \longrightarrow X_i \mid X_i > E_i$

$$S_\theta(t_i) \longrightarrow \frac{S_\theta(t_i)}{S_\theta(e_i)}$$

$h_\theta(t_i) \longrightarrow h_\theta(t_i)$

$$\ln(\mathcal{L}_\theta(x_1, \ldots, x_n)) = \sum \ln(f_\theta(x_i)) = \sum (\ln(h_\theta(x_i)) + \ln(S_\theta(x_i))) \quad f = hS.$$

$$\ln(\mathcal{L}_\theta(y_1, \ldots, y_n)) = \sum (d_i \ln(h_\theta(t_i)) + \ln(S_\theta(t_i)))$$

Exemple : $X \sim \exp(\theta)$

$f = hS$

$S = e^{-\theta t}$
$f = \theta e^{-\theta t}$
$\Rightarrow h = \theta$

① Modèle non censuré :

$$\ln(\mathcal{L}_\theta) = \sum_{i=1}^{n} (\ln(\theta) - \theta x_i)$$

$$\hookrightarrow \boxed{\hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i}}$$

② Modèle censuré et tronqué

$$\ln(\mathcal{L}_\theta) = \sum_{i=1}^{n} (d_i \ln(\theta) - \theta(t_i - e_i))$$

$$\hookrightarrow \boxed{\hat{\theta} = \frac{d}{\sum (t_i - e_i)}} \quad d = \sum d_i$$

$h_\theta(x_i) = \lambda \alpha x^{\alpha-1}$

$S_\theta(x) = e^{-\lambda x^\alpha}$

$\ln \mathcal{L}_\theta = \sum_{i=1}^{n} d_i \ln(\lambda \alpha t_i^{\alpha-1}) - \lambda(t_i^\alpha - e_i^\alpha)$

$\ln \mathcal{L}_\theta = \ln(\lambda) d + \ln(\alpha) d + \sum_{i=1}^{n} (d_i (\alpha-1) \ln(t_i) - \lambda(t_i^\alpha - e_i^\alpha))$

$$\hat{\lambda} = \frac{d}{\sum (t_i^{\hat\alpha} - e_i^{\hat\alpha})} \qquad \hat{\alpha} = ?$$

Algorithme Newton-Raphson

$$\frac{\partial}{\partial \theta} \ln(\mathcal{L}_\theta) = 0$$

Rappel: $f(x) = 0$

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \qquad \text{car } f(x_{i+1}) = f(x_i) + (x_{i+1} - x_i)\, f'(x_i) + \dots$$

$$\theta_{i+1} = \theta_i - \left[\frac{\partial^2 \ln(\theta)}{\partial \theta_i \partial \theta_j}\right]^{-1} \frac{\partial \ln(\theta_i)}{\partial \theta}$$

Rappels séance précédente:

$$(X_1, \dots, X_n)$$
$$\downarrow$$
$$Y_i = (T_i, D_i, E_i)_{1 \le i \le m}$$

$$\mathcal{L}_\theta(y_1, \dots, y_n) = \sum_{i=1}^{m} \left[d_i \ln(h_\theta(t_i)) + \ln(S_\theta(t_i)) - \ln(S_\theta(e_i))\right]$$

① Données requises

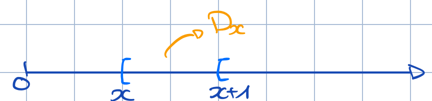
* Structure des données ind. ⊕ période d'observation.

a) Décès

- identifiant
- Date de naissance ← survenance
- Date d'entrée dans le risque] pour calculer E ← début d'indemnisation
- Date de sortie de risque ← fin d'indemnisation
-] pour calculer T
- Motif de sortie — pour calculer D.
- Fin d'observation : information à jour D_f
- Début d'observation: toutes les D_0 données soient connues.

$$D'_{E_i} = D_0 \vee D_{E_i}$$
$$D'_{T_i} = D_f \wedge D_{T_i}$$

Approche alternative:



$$q_x = 1 - \exp\left(-\int_x^{x+1} h_\theta(u)\, du\right)$$

$$D_x \sim \mathcal{B}(N_x, q_x) \qquad \mathbb{E}[D_x] = N_x q_x \qquad V(D_x) = \underbrace{N_x \hat{q}_x (1 - \hat{q}_x)}_{\hat{\sigma}_x^2}$$

$$P(D_x = d_x) = \frac{\sqrt{N_x}}{\sqrt{2\pi}\,\hat{\sigma}_x} \exp\left(-\frac{N_x}{2}\left(\frac{q_x - \hat{q}_x}{\hat{\sigma}_x}\right)^2\right)$$

$$\ln(\mathcal{L}_\theta) = \sum_{x = x_m}^{x_N} \left[\ln\left(\frac{\sqrt{N_x}}{\sqrt{2\pi}\,\hat{\sigma}_x}\right) - \frac{N_x}{2}\left(\frac{q_x - \hat{q}_x}{\hat{\sigma}_x}\right)^2\right]$$

$$\text{Max}(\ln \mathcal{L}_\theta) \iff \text{Min } \varphi(\theta)$$

$$\boxed{\varphi(\theta) = \sum_{x = x_m}^{x_N} \frac{N_x}{\hat{q}_x (1 - \hat{q}_x)} (\hat{q}_x - q_x(\theta))^2}$$

$N_x = ?$

$\mathbb{E}[D_x] = N_x q_x$

$\mathbb{E}[D_x] = ?$

$\mathbb{P}(D_x^* = 1) = \mathbb{P}(X \leq C)$

pour 1 individu

$= \mathbb{E}_c(\mathbb{P}(X \leq C | C))$

$= \int_e^{x+1} \left(1 - \frac{S_X(c)}{S_X(e_x)}\right) F_c(d_c)$

$\mathbb{P}_c(X \leq c) = 1 - S_{X|c}(c) = 1 - \frac{S_X(c)}{S_X(e_x)}$

<u>Hyp 1</u>: la fonction de hasard est $\simeq$ cte sur $[x, x+1[$, $1 - \frac{S_X(c)}{S_X(e_x)} = 1 - e^{-\mu_x (c - e_x)}$

<u>Hyp2</u>: μ_x "petit"

$1 - e^{-\mu_x (c - e_x)} \simeq \mu_x (c - e_x)$

$\longrightarrow \mathbb{P}(D_x^* = 1) \simeq \mu_x \int_e^{x+1} (c - e_x) F_c(d_c)$

$\mathbb{E}[D_x] = \mu_x \sum_{i=1}^{m} \int_{e_i(x)}^{x+1} (c - e_i(x)) F_c(d_c)$

$\Longrightarrow N_x = \sum_{i=1}^{m} \underbrace{\int_{e_i(x)}^{x+1} (c - e_i(x)) F_c(dc)}_{t_i(x) - e_i(x)}$

estimé par $t_i(x) - e_i(x)$

$N_x \simeq E_x = \sum_{i=1}^{m} [t_i(x) - e_i(x)]$

Sortie de $[x, x+1[$; entrée dans $[x, x+1[$

$$\boxed{\varphi(\theta) = \sum_{x = x_m}^{x_N} \frac{E_x}{\hat{q}_x(1 - \hat{q}_x)} (\hat{q}_x - \hat{q}_x(\theta))^2}$$

avec $\boxed{\hat{q}_x = \frac{D_x}{E_x}}$ Estimateur de Hoem

<u>Rq</u> : $\ln(\mathcal{L}_\theta) = \sum_{i=1}^{m} [d_i \ln(\mu) - \mu(t_i - e_i)]$

$= D \ln(\mu) - \mu E \qquad \hat{\mu} = \frac{D}{E}$

Si μ est cte, $q = 1 - e^{-\mu} \qquad \hat{q} = 1 - e^{-\frac{D}{E}}$

$D \sim \mathcal{P}(E\mu)$

$\mathbb{P}(D = d) = \frac{(E\mu)^d}{d!} e^{-E\mu}$

$\ln(\mathbb{P}(D = d)) = d \ln(E\mu) - \ln(d!) - E\mu$. inutile pour maximisat° de μ

$= \cancel{d \ln(E)} + d \ln(\mu) - \cancel{\ln(d!)} - E\mu$.

$\simeq d \ln(\mu) - E\mu$. à cte près

Méthode ① EMV $\Rightarrow \hat{\theta}_{EMV}$

② PMV $\Rightarrow \hat{\theta}_{PMV}^{x+1}$

PMV : PSEUDO MAX de VRAISEMBLANCE

Généralisation semi-paramétrique (modèle de Brass, 1971)

$$\varphi(\theta) = \sum_{x=x_m}^{x_N} \frac{E_x}{\hat{q}_x(1-\hat{q}_x)} (\hat{q}_x - q_x(\theta))^2$$

On dispose de $(q_x^{Ref})_{x=x_1,\dots,x_N}$

$$q_x(\theta) = a\, q_x^{Ref} + b \quad \text{par certain} \in [0,1]$$

$$f(x) = \ln\left(\frac{x}{1-x}\right) \qquad]0,1[\longrightarrow \mathbb{R}$$

$$\ln\left(\frac{q_x(\theta)}{1-q_x(\theta)}\right) = a \ln\left(\frac{q_x^{Ref}}{1-q_x^{Ref}}\right) + b$$

Comportement de $f(x)$:

$$f(x) = \ln\left(\frac{x}{1-x}\right) \qquad y = \ln\left(\frac{x}{1-x}\right) \Leftrightarrow e^y = \frac{x}{1-x} \Leftrightarrow \boxed{x = \frac{e^y}{1+e^y}}$$

$$f'(x) = \frac{1}{x} + \frac{1}{1-x} = \frac{1}{x(1-x)} > 0$$

$$f''(x) = \frac{2x-1}{x^2(1-x)^2} \qquad f''(x) \begin{cases} <0 & \text{si } x < 1/2 \\ >0 & \text{si } x > 1/2 \end{cases}$$

$$\hat{q} = \frac{D}{E} \qquad \mathbb{E}[\hat{q}] = q$$

$$\mathbb{E}[f(\hat{q})] \leq f(\bar{q}) = f(q)$$

$\Rightarrow$ sur la plage $q < 1/2$, l'ajustement par le modèle de Brass sous-estime la proba cond. de sortie.

$$h(x) \longrightarrow e(x) = \frac{1}{S(x)} \int_x^{\infty} S(u)\,du$$

$$S(x) = \exp\left(-\int_0^x h(u)\,du\right)$$

$$= \exp(-H(x)) \qquad f: x \mapsto e^{-x} \text{ est cvx donc } \mathbb{E}[\underbrace{e^{-\hat{H}(x)}}_{\hat{S}(x)}] \geq e^{-\mathbb{E}[\hat{H}(x)]} = e^{-H(x)} = S(x)$$

$$\Rightarrow \mathbb{E}[\hat{S}(x)] \geq S(x)$$

$$X \sim S_\theta$$

① Estimation MV

$$\mathcal{L}_\theta = \sum_{i=1}^{m} [d_i \ln(h_\theta(t_i)) + \ln(S_\theta(t_i)) - \ln(S_\theta(e_i))]$$

② Estimation "discrète":

$$\varphi(\theta) = \sum_{x=x_m}^{x_N} \frac{E_x}{\hat{q}_x(1-\hat{q}_x)} (\hat{q}_x - q_x(\theta))^2$$

(①, ②) Paramétriques

③ Semi-paramétrique (Brass)

$$f(q_x(\theta)) = a\, f(q_x^{Ref}) + b$$

$$f(x) = \frac{x}{1-x}$$

④ non-paramétrique $\quad \varphi(\theta) = \sum_x \frac{E_x}{\hat{q}_x(1-\hat{q}_x)} (\hat{q}_x - q_x)^2 + \mathrm{Pen}(q_x)$

Exercice : modèle à risques concurrents

$T = X \wedge C$

I : temps d'entrée en dépendance

D : temps de décès.

$Z = I \wedge D$.

$\varphi(x,y) = \mathbb{P}(x < I < D \wedge y)$

$S_Z(x) = S_I(x) S_D(x) \quad (\text{si } I \perp\!\!\!\perp D)$

$$\varphi(x,y) = \mathbb{E}_D[\mathbb{P}(x < I < D \wedge y \mid D)]$$

$$= \int_x^{\infty} (S_I(x) - S_I(y \wedge u)) f_D(u)\,du = S_I(x) S_D(x) - \Big(\int_x^y S_I(u) f_D(u)\,du + \underbrace{\int_y^{\infty} S_I(y) f_D(u)\,du}_{S_I(y) S_D(y)} \Big)$$

$$= \int_x^y (S_I(x) - S_I(u)) f_D(u)\,du + \int_y^{\infty} (S_I(x) - S_I(y)) f_D(u)\,du$$

$$\varphi(x,y) = S_I(x) S_D(x) - S_I(y) S_D(y) - \int_x^y S_I(u) f_D(u)\,du$$

$S_I(x) = e^{-\lambda_I x}$ $\quad \lambda = \lambda_D + \lambda_I$

$S_D(x) = e^{-\lambda_D x}$

$$\varphi(x,y) = e^{-\lambda x} - e^{-\lambda y} - \underbrace{\int_x^y e^{-\lambda_I u} \lambda_D e^{-\lambda_D u}\,du}_{\lambda_D \int_x^y e^{-\lambda u} du = \frac{\lambda_D}{\lambda}(e^{-\lambda x} - e^{-\lambda y})}$$

$$\varphi(x,y) = (e^{-\lambda x} - e^{-\lambda y}) \underbrace{\left[1 - \frac{\lambda_D}{\lambda}\right]}_{\frac{\lambda_I}{\lambda_I + \lambda_D}}$$

$$\pi(x) = \frac{\mathbb{P}(x < I < D \wedge (x+1))}{\mathbb{P}(I \wedge D > x)}$$

$$= \frac{e^{-\lambda x} - e^{-\lambda(x+1)}}{e^{-\lambda x}} \frac{\lambda_I}{\lambda_I + \lambda_D}$$

constance de $\lambda_D(x)$ et $\lambda_I(x)$ sur $[x, x+1[$

$q_D(x) = 1 - e^{-\lambda_D(x)}$

$\lambda_D(x) = \ln(1 - q_D(x))$

$$\Rightarrow \boxed{\pi(x) = (1 - e^{-\lambda_x}) \frac{\lambda_I(x)}{\lambda_x}}$$

Censure au $r^{\text{ième}}$ décès.

$C_i = X_{(r)}$

$$\mathcal{L}_\theta = \frac{m!}{(m-r)!} \left(\prod_{i=1}^{r} f_\theta(x_{(i)}) \right) S_\theta(x_{(r)})^{m-r}$$

$$= \frac{m!}{(m-r)!} \prod_{i=1}^{r} f_\theta(t_i)^{d_i} S_\theta(x_{(r)})^{m-r}$$

$$= \frac{m!}{(m-r)!} \prod_{i=1}^{m} h_\theta(t_i)^{d_i} S_\theta(t_i)$$

$$\ln \mathcal{L}_\theta = \sum_{i=1}^{m} (d_i \ln(h_\theta(t_i)) + \ln(S_\theta(t_i)))$$

$h_\theta(t) = \theta \qquad \ln \mathcal{L}_\theta = n \ln \theta - \theta \sum_{i=1}^{m} t_i$

$$\boxed{\hat{\theta} = \frac{n}{\sum_{i=1}^{m} t_i} = \frac{n}{\mathcal{E}}} \qquad \hat{\theta} = \frac{D}{\mathcal{E}}$$

Censure informative.

$$\begin{cases} T = X \wedge C \\ D = \mathbb{1}_{X \leq C} \end{cases}$$

$S_C(x) = S_X(x)^\beta$

$h_C(x) = \beta h_X(x).$

$\mathbb{P}(D=1) = \mathbb{P}(X \leq C)$

$= \int_0^\infty S_C(x) f_X(x) dx$

$u = S_X(x)$

$du = -f_X(x)dx$

$= \int_0^\infty S_X(x)^\beta f_X(x) dx = \int_0^1 u^\beta \, du = \frac{1}{\beta+1}$

$$\boxed{\mathbb{P}(D=1) = \frac{1}{\beta+1}} \qquad \boxed{\hat{\beta} = 1 + \frac{1}{\bar{D}}} \qquad \text{avec } \bar{D} = \frac{1}{m}\sum_{i=1}^{m} D_i.$$

Vraisemblance :

$\mathcal{L} = \prod_{i=1}^{m} \left(f_X(t_i) S_C(t_i)\right)^{d_i} \prod_{i=1}^{m} \left(f_C(t_i) S_X(t_i)\right)^{1-d_i} = \prod h_X(t_i)^{d_i} S_X(t_i) h_C(t_i)^{1-d_i} S_C(t_i)$

$S_\theta(t_i) = e^{-\theta t_i}$

$S_C(t_i) = e^{-\beta\theta t_i}$

$\ln \mathcal{L} = \sum_{i=1}^{m} \left[d_i \ln(\theta) - \theta t_i + (1-d_i)\ln(\beta\theta) - \beta\theta t_i\right]$

$\ln \mathcal{L} = D\ln(\theta) - \theta\mathcal{E} + (m-D)\ln\beta\theta - \beta\theta\mathcal{E}$

$= D\ln(\theta) + (m-D)\ln(\beta) + (m-D)\ln(\theta) - \theta(1+\beta)\mathcal{E}$

$\frac{\partial}{\partial\beta} \ln \mathcal{L} = \frac{m-D}{\beta} - \theta\mathcal{E}$

$\frac{\partial}{\partial\theta} \ln \mathcal{L} = \frac{m}{\theta} - (1+\beta)$

$$\begin{cases} \frac{m-D}{\mathcal{E}} = \theta\beta \\ \frac{m}{\mathcal{E}} = \theta(1+\beta) \end{cases} \qquad \frac{\beta}{1+\beta} = \frac{m-D}{m} \longrightarrow \begin{cases} \hat{\beta} = 1 - \frac{1}{D} \\ \hat{\theta} = \dots \end{cases}$$

09/10/23

On donne une table de mortalité $q_x \quad x = x_m, \ldots, x_M$

3 méthodes :

① - EMV : $h_\theta(x) = \alpha + \frac{\beta e^{\gamma x}}{1+\beta e^{\gamma x}}$ (Thatcher)

$\ln(\mathcal{L}_\theta) = \sum_{i=1}^{m} \left(d_i \ln(h_\theta(t_i)) + \ln(S_\theta(t_i))\right)$

Rappel : $(X_1, \dots, X_n)$ nos observat°
$(Y_1, \dots, Y_n)$

Censure non informative $Y_i \begin{cases} T_i = X_i \wedge C_i \\ D_i = \mathbb{1}_{X_i \wedge C_i} \end{cases}$

$X \perp C$ → $S_T(x) = \mathbb{P}(T > x) = \mathbb{P}(X \wedge C > x) = S_X(x) S_C(x)$

$h_T(x) = h_X(x) + h_T(x)$

Si $S_C(x) = S_X(x)^\beta$

La loi de Y indép de S_X et S_C
"La loi de Y et celle de C n'ont pas de paramètre commun"

② PMV ($\hat\theta_{PMV}$)

e_i, D_x^i, x, $x+1$

$E_x = \sum_{i=1}^{n} [t_i \wedge (x+1) - e_i \vee x]^+$

temps passé par i dans $[x, x+1[$

$$\varphi(\theta) = \sum_{x=x_m}^{x_M} \frac{E_x}{\hat q_x(1-\hat q_x)} \left(\hat q_x - q_x(\theta)\right)^2$$

$$\hat q_x = \frac{D_x}{E_x} \qquad q_x(\theta) = 1 - e^{-\int_x^{x+1} h_\theta(u)\,du}$$

③ Brass (semi-paramétrique)

$$\text{où} \quad \ln\left(\frac{q_x(\theta)}{1-q_x(\theta)}\right) = a \ln\left(\frac{q_x^{Ref}}{1-q_x^{Ref}}\right) + b$$

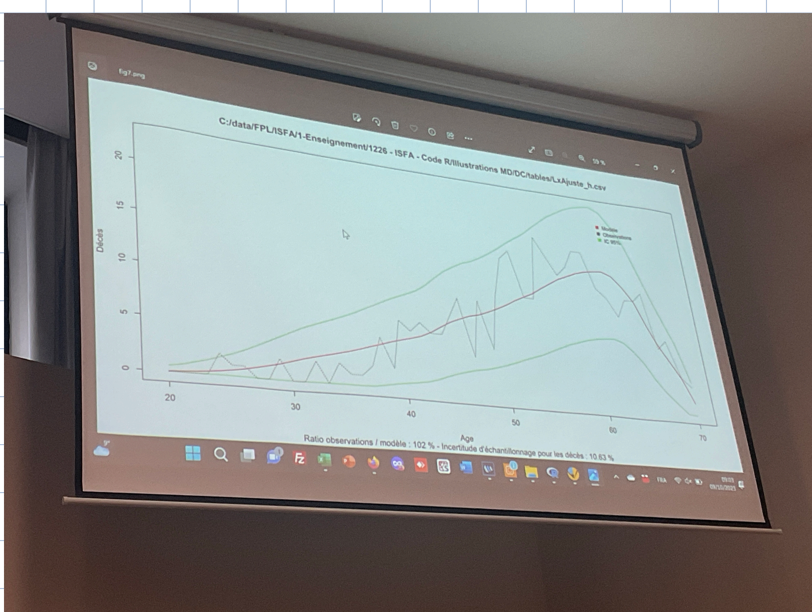
$(\theta = (a,b))$

$f(x) = \frac{x}{1-x} \iff g(y) = \frac{e^y}{1+e^y}$

NB : biais de sous-estimation des taux ajustés si $q_x < \frac{1}{2}$ (Jensen).

↑ Connaître ces 3 points c'est connaître 90% du cours jusqu'à maintenant

→ Voir l'application R. Le 2e paramètre dans le code pour les exponentielles correspond au γ de l'EMV.



- D_x $\quad x = x_m, \dots, x_M$
- théorique → $D_x^{th} = D_x(\hat\theta) = E_x\, q_x(\hat\theta)$

$E[D_x] = E_x \mu_x \approx E_x q_x$

$Var(\hat q) = \frac{1}{E} q(1-q)$

$Var(\hat D) = E\, q(1-q)$

$D \sim B(E, q)$

- $D_x^{+/-} = D_x^{th} \pm 1{,}96 \sqrt{E_x q_x(\hat\theta)(1-q_x(\hat\theta))}$

interv. de conf. $\approx D_x^{th} \pm 2 D_x^{th}$

Le 102% représente le $SMR = \frac{\sum D_x}{\sum D_x^{th}}$ ← aussi appelé Best Estimate

$$D^{th} = \sum_x D_x^{th} = \sum_x E_x q_x(\hat\theta)$$

$$D^{+/-} = D^{th} \pm 1{,}96 \sqrt{E\, \bar q(\hat\theta)(1-\bar q(\hat\theta))}$$

$E = \sum E_x \qquad \bar q(\hat\theta) = \frac{1}{E} \sum_x E_x q_x(\hat\theta)$

Le 10,63% représente le $\frac{D^+ - D^{th}}{D^{th}}$ demi-largeur relative de l'intervalle de conf.

→ Diapo sur mortalité.

Weibull

$h_\theta(t) = \theta\, h_0(t)$ où $\theta > 0$ PH (Proportional Hasard)

$h_\theta(t) = h_0(\theta t)$ AFT (Accelerated Failure Time)

Quels sont les fonctions h_0 tq PH et AFT ? Pas forcément les même θ.

Si $h_0(t) = \lambda \alpha t^{\alpha-1}$

$h_0(\theta t) = \lambda \alpha \theta^{\alpha-1} t^{\alpha-1}$

$= \theta^{\alpha-1} h_0(t)$

donc $W(\lambda, \alpha)$ est PH et AFT

Réciproquement,

$h_0(\theta t) = \ell(\theta)\, h_0(t)$

$h_0(t) = k(\theta)\, h_0(\theta t)$

$0 = k'(\theta)\, h_0(\theta t) + t\, k(\theta)\, h_0'(\theta t)$

$\Rightarrow -\frac{k'(\theta)}{k(\theta)} = \frac{t\, h_0'(\theta t)}{h_0(\theta t)}$

$\ln(h_0(u)) = -\lambda u + cte \longrightarrow h_0(u) = c e^{-\lambda u}$

$u \frac{h_0'(u)}{h_0(u)} = -\lambda \qquad h_0(u) = e^{-\lambda u}$

$\ln(h_0(u)) = -\lambda \ln(u) + cte$

$h_0(u) = c u^{-\lambda} \longrightarrow$ Weibull.

Lien entre vraisemblances latente et observable

modèle complet $(X_1, \dots, X_m)$ $\ell_\theta^*(x)$

censuré $(Y_1, \dots, Y_m)$ $\ell_\theta(y)$ $\qquad Y = X \wedge C$

notation du support de cours

$\mathcal{L}_\theta(x) = \ell_\theta^*(x) = \ell_\theta(x \mid Y = y)\, \overbrace{\ell_\theta(y)}^{\mathcal{L}_\theta(y)}$

$\Rightarrow \ln(\ell_\theta^*(x)) = \ln(\ell_\theta(x \mid Y = y)) + \ln(\ell_\theta(y))$

$\Rightarrow \frac{\partial}{\partial\theta} \ln(\ell_\theta^*(x)) = \frac{\partial}{\partial\theta} \ln(\ell_\theta(x \mid Y = y)) + \frac{\partial}{\partial\theta} \ln(\ell_\theta(y))$

$\Rightarrow \mathbb{E}\left[\frac{\partial}{\partial\theta} \ln(\ell_\theta^*(x)) \mid Y = y\right] = \underbrace{\mathbb{E}\left[\frac{\partial}{\partial\theta} \ln(\ell_\theta(x \mid Y = y))\right]} + \frac{\partial}{\partial\theta} \ln(\ell_\theta(y))$

$\int \frac{\partial \ell_\theta(x \mid Y = y)}{\partial\theta} \frac{1}{\ell_\theta(x \mid Y = y)} \ell_\theta(x \mid Y = y)\, dy = \int \frac{\partial \ell_\theta(x \mid Y = y)}{\partial\theta}\, dy = 0$

$\Rightarrow \boxed{\frac{\partial}{\partial\theta} \ln(\ell_\theta(y)) = \mathbb{E}\left[\frac{\partial}{\partial\theta} \ln(\ell_\theta^*(x)) \mid Y = y\right]}$

Hétérogénéité facteur v $\quad S(x),\ h(x)$

$(X_1, \dots, X_m)$ $\quad \ln(\mathcal{L}_\theta(x)) = \sum_{i=1}^m \ln(\ell_\theta(x_i)) \longrightarrow \frac{\partial}{\partial\theta} \ln(\mathcal{L}_\theta(x)) = 0 \Rightarrow \hat\theta^*$

$\hookrightarrow (Y_1, \dots, Y_m)$ $\quad \ln(\mathcal{L}_\theta(y)) = \sum_{i=1}^m (d_i \ln(h_\theta(t_i)) + \ln(S_\theta(t_i))) \longrightarrow \frac{\partial}{\partial\theta} \ln(\mathcal{L}_\theta(y)) = 0 \Rightarrow \hat\theta^c$

répartit° d'hétérogénéité

$\pi(dv) \qquad \pi_t(dv) = \frac{S(t, v)}{S(t)} \pi(dv)$

$\left.\begin{array}{l} S(t,v) \\ h(t,v) \end{array}\right\} \Rightarrow \begin{array}{l} S(t) = \int S(t,v)\,\pi(dv) \\ h(t) = \int h(t,v)\,\pi_t(dv) \end{array}$

c'est $\pi(dv)$ car $S(t) = \mathbb{E}_v[\underbrace{\mathbb{P}(T>t\,|\,v)}_{S(t,v)}]$

$u^{-1}\mathbb{P}(T\le t+u\,|\,T>t) = \int u^{-1}\mathbb{P}(T\le t+u\,|\,T>t, v)\,\overbrace{\frac{S(t,v)}{S(t)}\pi(dv)}^{\pi_t(dv)} \underset{u\to 0}{\Longrightarrow} h(t) = \int h(t,v)\,\pi_t(dv)$

On suppose que $h(t,v) = \lambda(v)$

$h(t) = \int \lambda(v)\,\pi_t(dv)$

$\forall v$, $t \mapsto h(t,v)$ cte. Quid de $h(t)$?

$$\frac{d}{dt}h(t) = \int \lambda(v)\frac{\partial}{\partial t}\pi_t(dv) \qquad S(t,v) = e^{-\lambda(v)t}$$

$$\Rightarrow \frac{\partial}{\partial t}\left(\frac{S(t,v)}{S(t)}\right) = \frac{\frac{\partial}{\partial t}S(t,v)S(t) - S(t,v)S'(t)}{S(t)^2} = \frac{-\lambda(v)S(t)S(t,v) - S(t,v)S'(t)}{S(t)^2} = \frac{-\lambda(v)S(t,v)}{S(t)} + h(t)\frac{S(t,v)}{S(t)}$$

$$\Rightarrow \frac{\partial}{\partial t}\pi_t(dv) = -\lambda(v)\underbrace{\frac{S(t,v)}{S(t)}\pi(dv)}_{\pi_t(dv)} + h(t)\underbrace{\frac{S(t,v)}{S(t)}\pi(dv)}_{\pi_t(dv)}$$

$$\Rightarrow \frac{\partial}{\partial t}\left[\int \lambda(v)\pi_t(dv)\right] = -\int \lambda^2(v)\pi_t(dv) + \underbrace{\int \lambda(v)h(t)\pi_t(dv)}_{h(t)^2}$$

$$\Rightarrow \frac{\partial}{\partial t}h(t) = -\Big[\underbrace{\int \lambda^2(v)\pi_t(dv) - \Big(\int \lambda(v)\pi_t(dv)\Big)^2}_{\geq 0 \text{ (Ineg de Schwartz)}}\Big] \leq 0$$

Prise en compte de l'hétérogénéité

$h(t,v) = g(v)h_0(t)$ Modèle à hasard proportionnel, PH.

Rappel : $h_\theta(t) = \theta h_0(t) \Leftrightarrow S_\theta(t) = S_0(t)^\theta$

$S_\theta(t) = e^{-\int_0^t \theta h_0(u)du} = S(t)^\theta$

- hétérogénéité non observable (fragilité) $g(v) \sim \Gamma(\lambda, r)$ (voir fin du support d'intro) on prend souvent $\lambda = r$.

  $S(t) = \mathbb{E}_v[S(t,v)] = \mathbb{E}[S_0(t)^X]$ $X \sim \Gamma(\lambda, r)$

- hétérogénéité observable $V := Z = (z_1, \dots, z_p)$

  $h(t|z) = \theta(z)h(t) = e^{-z'\beta}h_0(t)$ $\theta(z) = \exp(-\sum_{j=1}^{p}\beta_j z_j)$

→ Cas n°1 : $h_0(t) = h_\theta(t)$ (modèle paramétrique).

déjà fait car

$$\ln(\mathcal{L}_{\theta,\beta}(y)) = \sum_{i=1}^{n}\left[d_i \ln(h_\theta(t_i|e_i, z_i)) + \ln(S_\theta(t_i|e_i, z_i))\right]$$

→ Cas n°2 : $h_0(t)$ connue cf supports → Peut faire objet d'un sujet d'exam ! À regarder

→ Cas n°3 : $h_0(t)$ inconnue et non spécifiée (Cox, 1972)

Estimation de β :

$$\mathcal{L}(\beta) = \prod_{i=1}^{n}\left(\frac{e^{-z_i'\beta}}{\sum_{T_j \geq T_i} e^{-z_j'\beta}}\right)^{d_i}$$

sortie ou non

Justification : $0 < t_1 \dots t_k$ k sorties observés $\sum_{i=1}^{n} d_i = k$

$E = \{1, \dots, n\}$ ← individus.

$C_i = \{$individus sorties par censure avant $t_i\}$

$D_i = \{$individu non censuré en $t_i\} = \{j_i\}$

$$\{1, \dots, n\} \to \overbrace{C_1 D_1 C_2 D_2 \dots C_k D_k}^{\text{OBS}} \to \{1, \dots, n\} \setminus \{\text{tous les censurés}\} \setminus \{\text{toutes sorties non censurées}\}$$

$$\mathbb{P}(OBS) = \mathbb{P}(C_1)\mathbb{P}(D_1|C_1)\mathbb{P}(C_2|C_1D_1)\mathbb{P}(D_2|C_1D_1D_2) - \mathbb{P}(D_k|C_1D_1 - D_{k-1}C_k)$$

$$\mathbb{P}(OBS) = \mathbb{P}(C_1)\mathbb{P}(C_2|C_1D_1) - \mathbb{P}(C_k|C_1 - C_{k-1}D_1 - D_k) \times \mathbb{P}(D_1|C_1)\mathbb{P}(D_2|C_1D_1) - \mathbb{P}(D_k|D_1 - D_{k-1}C_1 - C_k)$$

$$= \prod_{\ell=1}^{k} \underbrace{\mathbb{P}(C_\ell|C_1 - C_{\ell-1}D_1 - D_{\ell-1})}_{\text{censure}} \prod_{\ell=1}^{k} \underbrace{\mathbb{P}(D_\ell|C_1 - C_\ell D_1 - D_{\ell-1})}_{\text{non censure}}$$

$$\Rightarrow \text{On garde } \mathcal{L}(\beta) = \prod_{\ell=1}^{k} \mathbb{P}(D_\ell|\underbrace{C_1 - C_\ell D_1 - D_{\ell-1}}_{\overline{\overline{R_\ell}}}) = \prod_{\ell=1}^{k} \mathbb{P}(D_\ell|R_\ell) \quad \text{où } R_\ell = \{i \in \{1,-,n\} \text{ tq } T_i \geq T_{i_\ell}\}$$

individus restants

Les i_ℓ sont les numéros des ℓ^e individu parti censuré ou non censuré

Soit $i \in R_\ell$, $\mathbb{P}(\text{sortie en } t_\ell | \text{présent avant } t_\ell) = h(t_\ell|z_i)dt_\ell = \theta(z_i)h_0(t_\ell)dt_\ell$

$$= \mathbb{P}(t_\ell < T \leq t_\ell + dt_\ell | T > t_\ell)$$

$$\mathbb{P}(D_\ell|R_\ell) = \frac{\theta(z_{i_\ell})\cancel{h_0(t_\ell)dt_\ell}}{\sum_{T_i \geq T_\ell} \theta(z_i)\cancel{h_0(t_\ell)dt_\ell}} = \frac{e^{-z_i'\beta}}{\sum_{T_i \geq T_\ell} e^{-z_i'\beta}}$$

$$\Rightarrow \boxed{\mathbb{P}(D_\ell|R_\ell) = \frac{e^{-z_i'\beta}}{\sum_{T_i \geq T_\ell} e^{-z_i'\beta}}}$$

$$\Rightarrow \mathcal{L}(\beta) = \prod_{\ell=1}^{k} \frac{e^{-z_i'\beta}}{\sum_{T_i \geq T_\ell} e^{-z_i'\beta}} = \prod_{i=1}^{n} \left(\frac{e^{-z_i'\beta}}{\sum_{T_j \geq T_i} e^{-z_j'\beta}} \right)^{d_i}$$

Mise en oeuvre

- Estimer β
- Estimer h_0

Segmentation

- la + grande population $\rightarrow h_0$ (Celle avec le + de sortie)
- les $(p-1)$ qui restent avec Cox

Validation de l'hypothèse PH,

$$r_i = d_i\left(z_i - \frac{\sum_{T_j \geq T_i} e^{-z_j'\beta} z_j}{\sum_{T_j \geq T_i} e^{-z_j'\beta}}\right) = d_i\left(z_i - \sum_{T_j \geq T_i} w_j z_j\right) \qquad w_j = \frac{e^{-z_j'\beta}}{\sum_{T_k \geq T_i} e^{-z_k'\beta}}$$

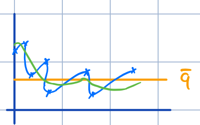
$\Rightarrow$ Test des résidus de Schoenfeld

Lissage de W-H (1922)

$$\varphi(\theta) = \sum_{x=x_m}^{x_M} \frac{E_x}{\hat{q}_x(1-\hat{q}_x)} (\hat{q}_x - q_x(\theta))^2 = \sum_x w_x (\hat{q}_x - q_x(\theta))^2$$

- 1ère extension: Brass
- 2ème extension: non-paramétrique

$$\varphi(q) = \sum_{i=1}^{p} w_i (\hat{q}_i - q_i)^2 + h \sum_{i=1}^{p-2} (\Delta^2 q_i)^2$$



$$\Delta q = \begin{pmatrix} q_2 - q_1 \\ | \\ q_p - q_{p-1} \end{pmatrix} \qquad \Delta^2 q = k_2 q$$

$$\varphi(q) = \underset{(1p)}{(q-\hat{q})'} \underset{(pp)}{W} \underset{(p1)}{(q-\hat{q})} + h \underset{(1p)}{q'} \underset{(pp)}{k_2'} \underset{(pp)}{k_2} \underset{(p1)}{q}$$

$$q^* = \text{argmin } \varphi(q)$$

$$\varphi'(q) = 2W(q-\hat{q}) + 2h k_2' k_2 q$$

$$\varphi'(q)=0 \iff Wq - W\hat{q} + h k_2' k_2 q = 0$$

$$\to W(q + h k_2' k_2 q) = W\hat{q}$$

$$\Rightarrow \boxed{q^* = (W + h k_2' k_2)^{-1} W \hat{q}'}$$

Voir documents de Gieseche

En dimension 2 :

Soit $(q_{ij})_{\substack{1\le i\le p\\ 1\le j\le q}}$

V : vertical
H : horizontal

$$\varphi(q) = \underbrace{\sum_{i,j} w_{ij}(\hat{q}_{ij} - q_{ij})^2}_{(q-\hat{q})' W (q-\hat{q})} + \underbrace{h_V \sum_{j=1}^{q} \sum_{i=1}^{p-z_V} (\Delta^{z_V} q_{ij})^2}_{h_V (q-\hat{q})' k_{z_V}' k_{z_V} (\hat{q}-q)} + \underbrace{h_H \sum_{i=1}^{p} \sum_{j=1}^{q-z_H} (\Delta^{z_H} q_{ij})^2}_{h_H (q-\hat{q})' k_{z_H}' k_{z_H} (q-\hat{q})}$$

18/10

Estimation non paramétrique

$$\varphi(\theta) = \sum_x \frac{E_x}{\hat{q}_x(1-\hat{q}_x)} (\underbrace{\hat{q}_x}_{=\frac{D_x}{E_x}} - q_x(\theta))^2$$

$\mu(t)$ de sur $[x, x+1]$ $\to$ $\hat{\mu}_x = \frac{D_x}{E_x}$

Si μ_x est "petit" alors $q_x = 1 - e^{-\mu_x}$

$\Rightarrow q_x \approx \mu_x$

Soit $(X_1, \dots, X_n)$ iid

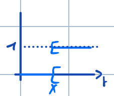
$\sqrt{n}(F_n - F) \xrightarrow{\mathcal{L}} W$

Soit $(Y_1, \dots, Y_n)$ iid $\quad Y_i = (T_i, D_i)$

$\to$ Trouver un $\hat{S}_n$ tq

$\sqrt{n}(\hat{S}_n - S) \xrightarrow{\mathcal{L}} W_{x,c}$

Soit $N(t) = \mathbb{1}_{X \le t}$ (càd)



$(T, D) \quad N'(t) = \mathbb{1}_{\{T \le t, D=1\}}$

Théorème de Donsker

$$X = A + M$$

A : processus prévisible, variat° bornées

M : martingale centrée

Décomposition de $N(t)$

$$N(t-) = \lim_{u \uparrow t} N(u)$$

$$dN(t) = N(t+dt) - N(t)$$

$$\mathbb{P}(dN(t) = 1 \mid N(t-)) = \begin{cases} 0 & \text{si } N(t-) = 1 \\ h(t)dt & \text{si } N(t-) = 0 \end{cases}$$

Car on a déjà sauté 1 fois

Comme un processus d'exposit°

Soit $Y(t) = \mathbb{1}_{X \ge t}$ $\quad \mathbb{P}(dN(t) \mid N(t-)) = Y(t)h(t)dt$

$Y(t) = 0 \iff N(t-) = 1 \quad \to \mathbb{E}[dN(t) \mid N(t-)] = Y(t)h(t)dt$

$\Rightarrow M(t) = N(t) - \int_0^t Y(t)h(t)\,dt$ est une martingale centrée.

$\Rightarrow N(t) = \int_0^t Y(t)h(t)\,dt + M(t)$

$$N'(t) = \int_0^t R(t)\,h(u)\,du + M'(t)$$

avec $R(t) = \mathbb{1}_{T \geq t}$ ← exposit° au risque.

Souvent on note $R(t)h(u) = \lambda(t)$ appelé intensité.

$\bar{N}'(t) = \sum_{i=1}^{m} N_i'(t)$ $\quad \bar{R}(t) = \sum_{i=1}^{m} R_i(t)$ $\quad t$ tq $\bar{R}(t) > 0$ $\quad \bar{R}(u) > 0 \ \forall u \leq t$

$$N_i'(t) = \int_0^t R_i(t)\,h(u)\,du + M_i'(t)$$

$$\Rightarrow \bar{N}'(t) = \int_0^t \bar{R}(u)\,h(u)\,du + \underbrace{\sum_{i=1}^{m} M_i'(t)}_{M'(t)}$$

← martingale centrée comme Σ de martingales centrées

$$\Rightarrow \bar{N}'(t) = \underbrace{\int_0^t \bar{R}(u)\,h(u)\,du}_{\text{Compensateur prévisible de } \bar{N}'(t)} + M'(t)$$

observat° ↑ $\quad$ modèle $\Lambda(t)$ ↑ $\quad$ erreur ↑

$\hat{\Lambda}(t) = \bar{N}'(t)$ $\qquad d\bar{N}'(t) = \bar{R}(t)h(t)dt + dM'(t)$

$$\int_0^t \frac{\mathbb{1}_{\bar{R}(u)>0}}{\bar{R}(u)}\,d\bar{N}'(u) = \int_0^t \mathbb{1}_{\bar{R}(u)>0}\,h(u)\,du + M(t)$$

← reste une martingale centrée car stable par Σ

Si $\bar{R}(t) > 0$, alors $\int_0^t \mathbb{1}_{\bar{R}(u)>0}\,h(u)\,du = H(t)$

$$\Rightarrow \boxed{\hat{H}(t) = \int_0^t \frac{\mathbb{1}_{\bar{R}(u)>0}}{\bar{R}(u)}\,d\bar{N}'(u)}$$

$\bar{N}'$: $\quad \bar{R}$:

$\hat{H}(t) = \sum_{T_i \geq t} \frac{D_i}{\bar{R}(T_i)}$ $\qquad D_i = \Delta\bar{N}'(T_i)$

Exercice Démontrer que $\mathbb{E}[\hat{H}(t)] \leq H(t)$

$$\mathbb{E}[\hat{H}(t)] = \mathbb{E}\left[\int_0^t \mathbb{1}_{\bar{R}(u)>0}\,h(u)\,du\right] + \underbrace{\mathbb{E}[M(t)]}_{=0}$$

$$= \int_0^t P(\bar{R}(u) > 0)\,h(u)\,du$$

$$\leq \int_0^t h(u)\,du = H(t)$$

$$\hat{S}(t) = e^{-\hat{H}_{NA}(t)}$$

$$\mathbb{E}[\hat{S}(t)] \underset{\text{Jensen}}{\geq} e^{-\mathbb{E}[\hat{H}_{NA}(t)]} \underset{x \mapsto e^{-x} \searrow}{\geq} e^{-\mathbb{E}[H(t)]} = S(t)$$

Estimateur de variance de $\hat{H}$

$$\frac{\bar{N}'(t+u) - \bar{N}'(t)}{\bar{R}(t)} \approx h(t)u$$

$\Rightarrow \bar{N}'(t+u) - \bar{N}'(t) \approx \bar{R}(t)\,h(t)\,u = \lambda$ $\quad \Sigma$ de bernoulli $\perp\!\!\!\perp$ $\Rightarrow$ Poisson

$$V\left(\frac{\bar{N}'(t+u) - \bar{N}'(t)}{\bar{R}(t)}\right) \approx \frac{h(t)u}{\bar{R}(t)}$$

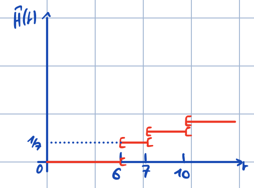
$$\rightarrow \hat{V}\left(\frac{\bar{N}'(t+dt) - \bar{N}'(t)}{\bar{R}(t)}\right) \approx \frac{h(t)}{\bar{R}(t)^2}\,\mathbb{1}_{\bar{R}(t)>0}\,dt$$

$$\rightarrow \hat{V}(\hat{H}(t)) = \int_0^t \frac{h(u)\mathbb{1}_{\bar{R}(u)>0}}{\bar{R}(u)^2}\,du \qquad \rightarrow \boxed{\hat{V}(\hat{H}(t)) = \sum_{T_i \geq t} \frac{D_i}{\bar{R}_i^2}}$$

Exemple : Données : $\overset{\text{Sortie en rémiss° censure}}{6, 6, 6, 6+}, 7, 9+, 10, 10+$ → Durée avant rémission

$T=6, D=1$ $\quad T=6, D=0 \ \dots$ $\quad$ 21 observat°

Sortie	t_i	r_i	d_i	$\frac{d_i}{r_i}$	$\frac{d_i}{r_i^2}$	$\hat{H}(t)$
1,2,3	6	21	3	1/7		1/7
5	7	17	1	1/17		1/7 + 1/17
7	10	15	1	1/15		...



→ But connaître $\hat{H}$

Introduction à l'estimateur KM (1958)

$$S(t) = 1 - \int_0^t S(u)\, h(u)\, du \qquad \int_0^t S(u) h(u) du = -\int_0^t S'(u) du = 1 - S(t)$$

avec $h(u) = -\frac{S'(u)}{S(u)}$

$$(*) \quad \hat{S}(t) = 1 - \int_0^t \hat{S}(u_-) \frac{\mathbb{1}_{\bar{R}(u)>0}}{\bar{R}(u)} d\bar{N}'(u)$$

Existe-t-il un $\hat{S}$ qui vérifie $(*)$? $\quad$ En +, mq que 2 estimateurs qui vérifie ça sont égaux

Construction directe

$$S(t) = \mathbb{P}(X>t) = \mathbb{P}(X>t \mid X>s)\, \mathbb{P}(X>s) = \mathbb{P}(X>t \mid X>s)\, S(s)$$

Soit $0 = t_0 < t_1 - < t_k = t$

$$\Rightarrow S(t) = \prod_{i=1}^{k} \underbrace{\mathbb{P}(X>t_i \mid X>t_{i-1})}_{1-\mathbb{P}(\text{Sortie entre } t_{i-1} \text{ et } t_i)}$$

$$\hat{S}_{HF}(t) = \prod_{T_i \le t} e^{-\frac{D_i}{R_i}} \qquad \boxed{\hat{S}_{KM}(t) = \prod_{T_i \le t} \left(1 - \frac{D_i}{R_i}\right)}$$

$$\hat{H}_{NA}(t) = \sum_{T_i \le t} \frac{D_i}{R_i}$$

$$\hat{S}_{HF}(t) = \prod_{T_i \le t} e^{-\frac{D_i}{R_i}} \quad \text{et} \quad \hat{S}_{KM}(t) = \prod_{T_i \le t} \left(1 - \frac{D_i}{R_i}\right)$$

$$\Rightarrow -\ln(\hat{S}_{HF}(t)) + \ln(\hat{S}_{KM}(t)) = \sum_{T_i \le t} \left(\frac{D_i}{R_i} + \ln\left(1 - \frac{D_i}{R_i}\right) \right)$$

$g : x \mapsto \ln(1-x) + x$

$g(0) = 0 \quad g'(x) = -\frac{1}{1-x} + 1 = \frac{-x}{1-x} \le 0$ sur $[0, 1[$ $\qquad \mathbb{E}[\hat{S}_{HF}(t)] \ge S(t)$

$$\Rightarrow \ln(\hat{S}_{HF}(t)) \ge \ln(\hat{S}_{KM}(t)) \Rightarrow \boxed{\hat{S}_{KM} \le \hat{S}_{HF}(t)}$$

HF : Harington-Flemming
NA : Nelson-Aalen
KM : Kaplan-Meier

$$\ln(\hat{S}_{KM}(t)) = \sum_{T_i \le t} \ln(1 - \hat{q}_i)$$

$$V(g(x)) \approx V(X)\, g'(\mathbb{E}[X])^2$$

$$\to V(\ln(1-\hat{q}_i)) = \frac{1}{(1-\hat{q}_i)^2} \underbrace{V(\hat{q}_i)}_{= \frac{1}{r_i} \hat{q}_i (1-\hat{q}_i)} = \frac{1}{(1-\hat{q}_i)^2} \frac{\hat{q}_i (1-\hat{q}_i)}{r_i} = \frac{\hat{q}_i}{r_i (1-\hat{q}_i)}$$

$$\Rightarrow \sum_{T_i \le t} \frac{\hat{q}_i}{r_i (1-\hat{q}_i)} = V(\ln(\hat{S}(t)))$$

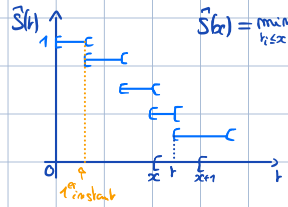
$$V(\hat{S}_{KM}(t)) = V(\ln(\hat{S}(t)))\, \hat{S}(t)^2$$

$$\hat{V}(\hat{S}_{KM}(t)) = \hat{S}^2(t) \times \sum_{T_i \le t} \frac{\hat{q}_i}{r_i (1-\hat{q}_i)}$$

$$\Rightarrow \boxed{\hat{S}^2(t) = \sum_{T_i \le t} \frac{d_i}{r_i (r_i - d_i)} = \hat{V}(\hat{S}_{KM}(t))}$$ Greenwood.

$$\hat{q}_{KM}(x) = 1 - \frac{\hat{S}_{KM}(x+1)}{\hat{S}_{KM}(x)}$$

$$\hat{q}_{HOGM}(x) = \frac{D_x}{E_x}$$



$$\hat{S}(x) = \min_{t_i \leq x} \hat{S}(t_i)$$

Présence d'hétérogénéité (Modèle d'Aalen)

$$h(t|\theta) = \theta h(t) \longrightarrow \text{Cox} \qquad \theta = e^{-z'\beta}$$

$$h(t) = X'(t)\beta(t) = \sum_{j=1}^{p} X_j(t)\beta_j(t)$$

$$(N_i'(t), R_i(t), X_i(t))_{1\leq i\leq m} \qquad \lambda_i(t) = R_i(t) X^i(t)'\beta(t)$$

$$\lambda(t) = (\lambda_i(t))'_{1\leq i\leq m} \qquad \in \mathcal{M}_{m,1}$$

$$N'(t) = (N_i'(t))'_{1\leq i\leq m} \qquad \in \mathcal{M}_{m,1}$$

$$X(t) = (R_i(t) X^i(t))'_{1\leq i\leq m} \quad \in \mathcal{M}_{m,p}$$

$$N'(t) = \int_0^t R_i(u) h(u)\, du + M_i(t)$$

$$\underset{(m,1)}{dN'(t)} = \underset{(m,p)}{X(t)}\,\underset{(p,1)}{\beta(t)}dt + \underset{(m,1)}{dM(t)}$$

$$Y = X\beta + \varepsilon \qquad X^-(t) = \begin{cases} [X'(t)X(t)]^{-1}X'(t) & \text{si } X'X \text{ inversible} \\ 0 & \text{Sinon} \end{cases}$$

$$\underset{(p,1)}{\hat{B}(t)} = \int_0^t \underset{(p\,m)}{X^-(t)}\, \underset{(m\,1)}{dN'(t)} \qquad B(t) = \int_0^t \beta(u)\, du$$

$$\boxed{\hat{B}_j(t) = \sum_{T_i \leq t} X_j^-(T_i)\, D_i}$$

① Régularisat°

→ W-H

$$\sum W_i(\hat{q}_i - q_i) + h\sum(\Delta^2 q_i)^2$$

$\hat{q}_{KM}(x_i)$

② Test de comparaison