

## TD 2: APPROCHE NON-PARAMÉTRIQUE DE LA SURVIE.

Modèle de durée M1 DUAS– Semestre 1  
P.-O. Goffard

---

1. Soit une variable aléatoire de fonction de hasard constante par morceau avec

$$h(t) = \sum_{l=1}^k \alpha_l \mathbb{I}_{[v_l, v_{l+1})}(t),$$

où  $0 = v_1 < v_2 \dots < v_{k+1} = \infty$ , et  $\alpha_1, \dots, \alpha_k \geq 0$ .

- (a) Si  $k = 1$ , quel est la loi de  $T$ .

**Solution:** Il s'agit de la loi exponentielle de paramètre  $1/\alpha_1$

- (b) Pour  $k > 1$ , donner l'expression de la fonction de survie .

**Solution:** On a

$$\begin{aligned} H(t) &= \int_0^t h(s) ds \\ &= \sum_{l=1}^k 0 \cdot \alpha_l \cdot \mathbb{I}_{t < v_l} + \alpha_l \cdot (t - v_l) \mathbb{I}_{t \in [v_l, v_{l+1})} + \alpha_l \cdot (v_{l+1} - v_l) \mathbb{I}_{t \geq v_{l+1}} \\ &= \text{Faire le graphique de } t \mapsto 0 \cdot \alpha_l \cdot \mathbb{I}_{t < v_l} + \alpha_l \cdot (t - v_l) \mathbb{I}_{t \in [v_l, v_{l+1})} + \alpha_l \cdot (v_{l+1} - v_l) \mathbb{I}_{t \geq v_{l+1}} \\ &= \sum_{l=1}^k \alpha_l (t \wedge v_{l+1} - v_l)_+ = \sum_{l=1}^k \alpha_l \max(t \wedge v_{l+1} - v_l, 0) \end{aligned}$$

puis  $S(t) = \exp[-H(t)]$ .

- (c) Soit un échantillon de  $n$  observations i.i.d. et censurée à droite (censure non informative)

$$\mathcal{D} = (x_i, \delta_i) = (t_i \wedge c_i, \mathbb{I}_{t_i \leq c_i}), \quad i = 1, \dots, n.$$

Donner l'estimateur du maximum de vraisemblance des paramètres  $\alpha_l$  pour  $l = 1, \dots, k$ . L'estimateur obtenu porte le nom d'estimateur de Hoem [1], il est très populaire en science actuarielle.

Indication: L'estimateur doit faire apparaître un nombre d'évènement dans le segment  $[v_l, v_{l+1})$

$$d(v_l) = \sum_{i=1}^n \delta_i \mathbb{I}_{[v_l, v_{l+1})}(x_i)$$

et une exposition au risque

$$e(v_l) = \sum_{i=1}^n (x_i \wedge v_{l+1} - v_l)_+ = \sum_{i=1}^n \max(x_i \wedge v_{l+1} - v_l, 0),$$

qui s'interprète ici comme la somme des temps passés par les individus sur le segment  $[v_l, v_{l+1})$ .

**Solution:** La vraisemblance s'écrit

$$\begin{aligned}\mathcal{L}(\mathcal{D}; \theta) &= \prod_{i=1}^n h(x_i)^{\delta_i} S(x_i) \\ &= \prod_{i=1}^n \prod_{l=1}^k \alpha_l^{\delta_i \mathbb{I}_{[v_l, v_{l+1})}(x_i)} \exp \{-\alpha_l (x_i \wedge v_{l+1} - v_l)_+\} \\ &= \prod_{l=1}^k \alpha_l^{d(v_l)} \exp \{-\alpha_l e(v_l)\}.\end{aligned}$$

La log-vraisemblance est donnée par

$$l(\mathcal{D}; \theta) = \sum_{l=1}^k \ln(\alpha_l) d(v_l) - \alpha_l e(v_l)$$

On en déduit l'estimateur du maximum de vraisemblance en résolvant les équations du score avec

$$\hat{\alpha}_l = \frac{d(v_l)}{e(v_l)}.$$

- (d) Donner une estimation de la matrice d'information de Fisher. En déduire une estimation de la variance asymptotique (valable pour un grand nombre d'observations) et un intervalle de confiance de niveau  $\epsilon$  pour les paramètres.

**Solution:** La matrice d'information de Fisher est estimée par

$$(I_n(\hat{\theta}))_{i,j} = -\frac{\partial}{\partial \alpha_i \partial \alpha_j} l(\mathcal{D}; \theta) = \begin{cases} e(v_i)^2 / d(v_i), & \text{si } i = j, \\ 0, & \text{sinon.} \end{cases}$$

On en déduit que

$$\mathbb{V}(\hat{\alpha}_l) = d(v_l) / e(v_l)^2$$

et

$$\alpha_l \in [\hat{\alpha}_l \pm z_{1-\epsilon/2} \mathbb{V}(\hat{\alpha}_l)]$$

2. Le modèle de Gompertz-Makeham, voir [2], définit la fonction de hasard d'une v.a  $T > 0$  par

$$h(t) = b \cdot c^t, \text{ pour } t \geq 0.$$

avec  $b, c \geq 0$ . il s'agit d'un modèle adapté à la modélisation de la durée de vie humaine, les paramètres  $b$  et  $c$  accommodent l'augmentation progressive du risque avec l'âge.

- (a) Donner l'expression de la fonction de hasard cumulé de  $T$ .

**Solution:** La fonction de hasard cumulé est donnée par

$$H(t) = \int_0^t h(s) ds = \frac{b}{\ln c} (c^t - 1).$$

(b) Soit  $\theta = (b, c)$  et

$$q_{\theta}(x) = \mathbb{P}(T \leq x + 1 | T > x)$$

la probabilité de décès à l'âge  $x$ . Montrer que les probabilités de décès s'écrivent sous la forme

$$q_{\theta}(x) = 1 - f^{c^x(c-1)},$$

où vous exprimerez  $f$  en fonction de  $b$  et  $c$ .

**Solution:** On a

$$\begin{aligned} q_{\theta}(x) &= 1 - \exp\left(-\int_x^{x+1} bc^t dt\right) \\ &= 1 - \exp(-H(x+1) + H(x)) \\ &= 1 - \exp\left(-\frac{b}{\log(c)}c^x(c-1)\right) \end{aligned}$$

On identifie  $f = e^{-b/\log(c)}$ .

(c) Montrer que

$$\log(q_{\theta}(x)) \approx \alpha + \beta x$$

Indication: On pourra utiliser un développement limité.

**Solution:** On fait l'approximation  $1 - \exp\left(-\frac{b}{\log(c)}c^x(c-1)\right) \approx \frac{b}{\log(c)}c^x(c-1)$  qui correspond à un développement limité à l'ordre 1 faisant l'hypothèse que  $\frac{b}{\log(c)}c^x(c-1)$  est proche de 0 puis on passe au log pour obtenir

$$\log(q_{\theta}(x)) = \log(f) + \log(c-1) + \log(c)x$$

(d) En utilisant le modèle de l'exercice 1 avec  $v_0 = 0, v_1 = 1, v_2 = 2, \dots$ , exprimer

$$q_{\text{hoem}}(x) = \mathbb{P}(T \leq x + 1 | T > x)$$

en fonction de  $\alpha_x$  pour  $x = 0, 1, \dots$

**Solution:** Il vient  $q_{\text{hoem}}(x) = 1 - e^{-\alpha_x}$ .

(e) Soit un échantillon de  $n$  observations i.i.d. et censurée à droite

$$\mathcal{D} = (x_i, \delta_i) = (t_i \wedge c_i, \mathbb{I}_{t_i \leq c_i}), \quad i = 1, \dots, n.$$

Déduire des deux questions précédentes une méthode d'estimation pour  $c$  et  $f$ .

**Solution:** On estime  $\alpha_x$  par  $\hat{\alpha}_x = d(x)/e(x)$ , où

$$d(x) = \sum_{i=1}^n \delta_i \mathbb{I}_{[x, x+1)}(x_i)$$

et

$$e(x) = \sum_{i=1}^n (x_i \wedge (x+1) - x)_+.$$

Puis on estime  $\hat{q}_{\text{hoem}}(x) = 1 - e^{-\hat{\alpha}_x}$ . On trouve  $\alpha$  et  $\beta$  dans

$$\log(\hat{q}_{\text{hoem}}(x)) \approx \hat{\alpha} + \hat{\beta}x$$

par les moindres carrés ordinaires puis on trouve  $c$ ,  $f$  et  $b$ .

- (f) Ecrire un code R permettant de faire l'inférence du modèle de Gompertz-Makeham. Le code doit comprendre

1. Ecrire une fonction pour générer des données depuis le modèle de GM. Simuler un échantillon de 1,000 observations  $t_1, \dots, t_{1000}$  avec

$$b = 0.0001 \text{ et } c = 1.1.$$

Puis tracer un histogramme.

2. Vérifier que l'échantillonneur fonctionne bien en comparant les probabilités de décès  $q(x)$  théorique et empirique. Les taux de hasards empiriques doivent être estimés de manière non paramétrique.
3. Ecrire un code qui retourne l'estimateur des paramètres du modèle de Gompertz-Makeham à partir d'un échantillon comprenant des données censurées à droite. Donner la valeur estimée des paramètres sur l'échantillon généré précédemment en ajoutant une censure à droite non informative tel que

$$c_1, \dots, c_{1000} \sim \text{Exp}(1/\bar{t}).$$

Vous pouvez utiliser la méthode de l'exercice ou bien utiliser le maximum de vraisemblance.

4. Comparer les fonctions de survies théoriques, empirique estimée via Kaplan-Meier et estimé paramétriquement suivant le modèle de Gompertz Makeham.

## References

- [1] Jan M. Hoem. Point estimation of forces of transition in demographic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(2):275–289, jul 1971.
- [2] William Matthew Makeham. On the law of mortality and the construction of annuity tables. *The Assurance Magazine and Journal of the Institute of Actuaries*, 8(6):301–310, jan 1860.