TD 2: APPROCHE NON-PARAMÉTRIQUE DE LA SURVIE.

Modèle de duréee M1 DUAS– Semestre 2
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 $0 = v_1 < v_2 \ldots < v_{k+1} = \infty$, et $\alpha_1, \ldots, \alpha_k \geq 0$.

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

   Solution: Il s'agit de la loi exponentielle de paramètre θ_1

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

   Solution: On a

$$\begin{aligned} H(t) &= \int_0^t h(s)\mathrm{d}s \\ &= \sum_{l=1}^{k} \alpha_l (v_{l+1} - v_l)\mathbb{I}_{t > v_l} + \sum_{l=1}^{k} \alpha_l (t - v_l)\mathbb{I}_{t \in [v_l, v_{l+1})} \\ &= \sum_{l=1}^{k} \alpha_l (t \wedge v_{l+1} - v_l)_+ = \sum_{l=1}^{k} \max(x_i \wedge v_{l+1} - v_l, 0) \end{aligned}$$

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

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

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

   Donner l'estimateur du maximum de vraisemblance des paramètres α_l pour $l = 1, \ldots, k$. L'estimateur obtenu porte le nom d'estimateur de Hoem [1], il est très populaire en science actuarielle.
   Indication: L'estimateur doit faire apparaitre 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'interprête ici comme la somme des temps passés par les individus sur le segment $[v_l, v_{l+1})$.

Solution: La vraisemblance s'écrit

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

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

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

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

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

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

On en déduit que

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

et

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

(e) Peut on proposer un test pour l'hypothèse $\alpha_1 = \ldots = \alpha_k = \alpha_O$.

Solution: On peut proposer un test de Wald avec comme statistique de test

$${}^t(\alpha - \alpha_0)\widehat{I}_n^{-1}(\alpha - \alpha_0) \sim \mathsf{Normal}(0, \mathrm{Id})$$

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

$$h(t) = a + b \cdot c^t,$$

avec $a, b, c \geq 0$. il s'agit d'un modèle adaptté à la modélisation de la durée de vie humaine, les paramètre b et c accomode l'augmentation progressive du risque avec l'âge, tandis que le paramètre a permet de prendre en compte les décès accidentels.

(a) Donner l'expression de la fonction de hasard cumulé, de la survie et de la densité de T.

Solution: La fonction de hasard cumulé est donnée par

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

La fonction de survie est donnée par

$$S(t) = \exp(-H(t)),$$

et la fonction de densité par

$$f(t) = h(t)S(t).$$

(b) En présence de censure à droite, écrire la log vraisemblance du modèle pour un échantillon

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

Solution: La log vraisemblance en présence de données censurées s'écrit

$$\begin{aligned} l(\mathcal{D};\theta) &= \sum_{k=1}^{n} \delta_k \log h(x_k;\theta) + \sum_{k=1}^{n} \log S(x_k;\theta) \\ &= \delta_k \log(a + bc^t) + \sum_{k=1}^{n}\left[-ax_k - \frac{b}{\ln c}(c^{x_k} - 1)\right]. \end{aligned}$$

(c) Calculer la dérivée première par rapport à chacun des paramètres du modèle (soit le gradient).

Solution: On a

$$\frac{\partial}{\partial a} l(\mathcal{D};\theta) = \sum_{k=1}^{n} \frac{\delta_k}{a + bc^{x_k}} - \sum_{k=1}^{n} x_k,$$

$$\frac{\partial}{\partial b} l(\mathcal{D};\theta) = \sum_{k=1}^{n} \frac{\delta_k c^{x_k}}{a + bc^{x_k}} - \sum_{k=1}^{n} \frac{c^{x_k} - 1}{\ln c}$$

et

$$\frac{\partial}{\partial c} l(\mathcal{D};\theta) = \sum_{k=1}^{n} \frac{\delta_k b x_k c^{x_k - 1}}{a + bc^{x_k}} + \sum_{k=1}^{n}\left[\frac{b(c^{x_k} - 1)}{c(\ln c)^2} - \frac{b x_k c^{x_k}}{\ln c}\right].$$

(d) 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 $5,000$ observations $t_1, \ldots, t_{5000}$ avec

   $$a = 0.001, b = 0.0001 \text{ et } c = 1.1.$$

   Puis tracer un histogramme.

2. Vérifier que l'échantilloneur fonctionne bien en comparant les taux de hasard théorique et empirique. Les taux de hasard empirique doivent être estimé de manière non paramétrique. On définit une grille de points equidistants

$$v_0 < v_1 < \ldots < v_k$$

et on compare $d(v_k)/n(v_k)$ à $h(v_k)(v_{k+1} - v_k)$ via un graphique qui comprend les deux courbes de taux de hasard. (Prenez un écart de 1 entre les v_k).

3. Ecrire un code qui retourne l'estimateur du maximum de vraisemblance des paramètres du modèle de Gompertz-Makeham à partir d'un échantillon comprenant des données censurés à droite. On utilisera la fonction `optim` en renseignant le gradient de la log vraisemblance. Donner la valeur estimée des paramètre sur l'échantillon généré précédemment en ajoutant une censure à droite non informative tel que

$$c_1, \ldots, c_{5000} \sim \text{Exp}(1/\bar{t}),$$

où $\bar{t} = \dfrac{1}{n}\displaystyle\sum_k t_k$. Prenez comme valeurs initiales dans l'algorithme d'optimisation les valeurs suivantes

$$a = 0.01, b = 0.01, \text{ et } c = 1.2$$

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.