

Modèles de durée / Examen du 14 janvier 2011

Durée 2h – aucun document n'est autorisé

Corrigé

Problème : prise en compte de la sélection médicale

On considère une variable aléatoire T décrivant une durée de vie et on note S et h respectivement la fonction de survie et la fonction de hasard associées. On cherche à prendre en compte l'effet de la sélection médicale, dont on se dit *a priori* que cela conduit à minorer temporairement les taux de décès conditionnels et que l'effet s'atténue rapidement avec le temps. On rappelle que dans un modèle paramétrique avec une censure aléatoire droite non informative la log-vraisemblance a la forme suivante :

$$\ln L(\theta) = \text{cste} + \sum_{i=1}^n \left\{ d_i \ln(h_\theta(t_i)) + \ln S_\theta(t_i) \right\}.$$

On suppose qu'en plus de la censure, les observations sont tronquées à gauche : en pratique l'individu i est observé à partir de l'instant $e_i \leq t_i$.

Question n°1 (2 points) : Donner en la justifiant la forme de la log-vraisemblance du modèle tronqué et censuré.

On écrit simplement $\ln L(\theta) = \text{cste} + \sum_{i=1}^n \left\{ d_i \ln(h_\theta(t_i | e_i)) + \ln S_\theta(t_i | e_i) \right\}$ et on observe que $h_\theta(t_i | e_i) = h_\theta(t_i)$ et $S_\theta(t_i | e_i) = \frac{S_\theta(t_i)}{S_\theta(e_i)}$ d'où l'on tire :

$$\ln L(\theta) = \text{cste} + \sum_{i=1}^n \left\{ d_i \ln(h_\theta(t_i)) + \ln S_\theta(t_i) - \ln S_\theta(e_i) \right\}.$$

On veut introduire dans le modèle le fait que lorsqu'un individu entre dans la population à risque, sa probabilité conditionnelle de sortie se trouve diminuée temporairement du fait de la sélection médicale à l'entrée. Pour cela on propose d'utiliser le modèle suivant :

$$h_{\tilde{\theta}}(t|e) = \left(1 - \pi \times e^{-\delta(t-e)} \right) \times h_\theta(t), \quad \tilde{\theta} = (\pi, \delta, \theta).$$

Question n°2 (2 point) : Donnez une interprétation des paramètres π et δ . Que dire du cas particulier $\delta = 0$? Quelles conditions faut-il imposer aux paramètres π et δ ?

Lorsque $t = e$, $h_{\tilde{\theta}}(t|e) = (1 - \pi) \times h_{\theta}(t)$ et donc π s'interprète comme l'abattement associé à la présence de la sélection médicale. δ est la vitesse à laquelle cet effet s'atténue. Lorsque $\delta = 0$ on a un abattement constant en fonction du temps. On doit imposer $\pi \in [0,1]$ et $\delta \geq 0$.

On suppose que la fonction de hasard de base est donnée par $h_{\theta}(t) = \gamma + \alpha e^{\beta t}$ (Makeham, 1860).

Question n°3 (1,5 point) : Donner une interprétation du modèle $h_{\theta}(t) = \gamma + \alpha e^{\beta t}$.

Le coefficient γ indépendant de l'âge s'interprète comme un taux de décès accidentel, la composante $\alpha e^{\beta t}$ (modèle de Gompertz, 1825) comme l'effet du vieillissement. Ce modèle peut être vu comme un modèle à risques concurrents $T = T_a \wedge T_v$ avec T_a un modèle exponentiel et T_v un modèle de Gompertz.

Question n°4 (4 points) : Déterminer la log-vraisemblance du modèle avec sélection médicale.

On a

$$\begin{aligned} h_{\tilde{\theta}}(t|e) &= \left(1 - \pi \times e^{-\delta(t-e)}\right) \times \left(\gamma + \alpha e^{\beta t}\right) \\ &= \gamma - \gamma \times \pi \times e^{-\delta(t-e)} + \alpha e^{\beta t} - \alpha \times \pi \times e^{-\delta(t-e)+\beta t} \end{aligned}$$

et on en déduit après quelques calculs :

$$\begin{aligned} \int_e^t h_{\tilde{\theta}}(u|e) du &= \int_e^t \left(\gamma - \gamma \times \pi \times e^{-\delta(u-e)} + \alpha e^{\beta u} - \alpha \times \pi \times e^{-\delta(u-e)+\beta u} \right) du \\ &= \gamma(t-e) - \frac{\gamma \times \pi}{\delta} \times \left(1 - e^{-\delta(t-e)}\right) + \frac{\alpha}{\beta} \left(e^{\beta t} - e^{\beta e}\right) - \frac{\alpha \times \pi}{\beta - \delta} e^{\delta e} \left(e^{(\beta-\delta)t} - e^{(\beta-\delta)e}\right) \\ &= \gamma(t-e) - \gamma \times \pi \times g_{\delta}(t-e) + \alpha \times e^{\beta t} \times g_{\beta}(t-e) - \alpha \times \pi \times e^{\delta e + (\beta-\delta)t} \times g_{\beta-\delta}(t-e) \end{aligned}$$

$$\text{avec } g_a(x) = \frac{1 - e^{-ax}}{a}.$$

On trouve alors la log-vraisemblance en utilisant la formule

$$\ln L(\theta) = cste + \sum_{i=1}^n \left\{ d_i \ln(h_{\theta}(t_i|e_i)) + \ln S_{\theta}(t_i|e_i) \right\}$$

$$\text{avec } \ln S_{\theta}(t_i|e_i) = - \int_{e_i}^{t_i} h_{\tilde{\theta}}(u|e_i) du.$$

Question n°5 (2,5 points) : Décrire la procédure à suivre pour estimer les paramètres $(\pi, \delta, \alpha, \beta, \gamma)$.

On calcule les dérivées de la log-vraisemblance par rapport à chacun des paramètres pour obtenir le vecteur des scores dont on cherche une racine. En pratique le calcul peut être effectué par la méthode de Newton-Raphson.

On rappelle que plutôt que de chercher directement le maximum de la vraisemblance dans le cas général, il est souvent considéré la solution approchée obtenue en minimisant le critère de moindres carrés pondérés suivant :

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

Question n°6 (3 points) : Que désignent les quantités E_x et \hat{q}_x ? Rappeler le raisonnement permettant d'obtenir $\varphi(\theta)$ et préciser comment est calculé $q_x(\theta)$ à partir de la fonction de hasard sous-jacente.

cf. le cours.

Question n°7 (3 points) : On se propose d'adapter l'approche simplifiée rappelée ci-dessus au modèle dont on a calculé la log-vraisemblance à la question 4. Comment faire ?

On discrétise dans les dimensions âge et ancienneté et on fait comme dans le cours.

Question n°8 (2 points) : Montrer que dans le cas où la fonction de hasard de base est constante (*i.e.* $\alpha = 0$) on a l'expression simple :

$$q_{x,e} = 1 - \exp(-\gamma \times (1 - \pi \times e^{-\delta(x-e)} \times g_\delta(1)))$$

On utilise $q_{x,e} = 1 - \frac{S_{\tilde{\theta}}(x+1)}{S_{\tilde{\theta}}(x)}$ avec (d'après la question 4) :

$$S_{\tilde{\theta}}(x) = \exp\{-\gamma(x-e) + \gamma \times \pi \times g_\delta(x-e)\}$$

et en observant que $g_\delta(x+1-e) - g_\delta(x-e) = e^{-\delta(x-e)} g_\delta(1)$