Modèles de durée / Examen / Janvier 2022

Durée 2h – aucun document n'est autorisé

Corrigé

La qualité de la rédaction, des justifications apportées et de la présentation de la copie seront prises en compte dans la notation. Chaque réponse doit être correctement justifiée : une réponse juste sans justification sera considérée comme fausse.

Prise en compte d'un effet « température » dans un modèle prospectif de mortalité

On se place dans le cadre d'un modèle tronqué à gauche et censuré à droite avec une censure aléatoire droite non informative, avec les notations suivantes :

$$T_i = X_i \wedge C_i \text{ et } D_i = \begin{cases} 1 \ si \ X_i \leq C_i \\ 0 \ si \ X_i > C_i \end{cases}$$

On note $E_i,\ i = 1,\dots,n$ les instants de troncature gauche.

Question n°1 (2 points) : Rappelez les définitions des termes ci-dessus. : censure aléatoire droite, censure non informative, troncature gauche.

> Cf. la section 1.2.1 du support de cours.

Question n°2 (4 points) : Démontrez que la vraisemblance peut s'écrire

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

> Cf. la section 1.2.1 du support de cours.

Question n°3 (4 points) : Montrez à l'aide de l'expression ci-dessus, qu'en supposant la fonction de hasard constante μ_x sur un intervalle $[x, x+1[$, on peut, pour estimer μ_x, supposer que le nombre D_x de sorties observées non censurées sur $[x, x+1[$ suit une loi de Poisson dont on précisera le paramètre en fonction de μ_x et de l'exposition au risque E_x. Vous préciserez la définition des instants de troncature gauche et de censure droite pour l'observation de l'intervalle $[x, x+1[$ en fonction des instants e_i et t_i.

> à l'aide de ce qui précède on trouve que la log-vraisemblance du modèle est, à une constante près :
>
> $$\ln L(\theta) = \sum_{i=1}^{n} \left[d_i \ln(\theta) - \theta \times (t_i - e_i) \right] = d_x \times \ln(\theta) - \theta \times E_x$$
>
> avec $d_x = \sum_{i=1}^{d} d_i$ et $E_x = \sum_{i=1}^{d} (t_i - e_i)$. On remarque alors que tout se passe comme si la variable D_x qui compte le nombre de sorties sur l'intervalle

$E_x = \sum_{i=1}^{d} (t_i - e_i)$

$[x, x+1[$ était une loi de Poisson de paramètre $\theta \times E_x$; en effet, dans ce cas $\ln\left(P\left(D_x = d\right)\right) = cste + d_x \times \ln(\theta) - \theta \times E_x$.

Dans ce qui précède, $e_i = e_i(x) = e_i \vee x$ et $t_i = t_i(x) = t_i \wedge (x+1)$.

Pour la prise en compte de l'effet du temps sur les taux de mortalité, il est donc possible de proposer la spécification suivante :[1]

$$D_{x,t} \approx P\left(E_{x,t} \times \mu_{x,t}\right) \text{ avec } \ln\left(\mu_{x,t}\right) = \alpha_x + \beta_x \times k_t$$

Question n°4 (2 points) : précisez les contraintes à ajouter sur les paramètres pour que la relation ci-dessus soit suffisante pour définir de manière univoque le lien entre la force de mortalité et les variables du modèle .

Pour toute constante c non nulle le modèle est invariant par les transformations suivantes :

$$\left(\alpha_x, \beta_x, k_t\right) \rightarrow \left(\alpha_x, \frac{\beta_x}{c}, c \times k_t\right)$$

$$\left(\alpha_x, \beta_x, k_t\right) \rightarrow \left(\alpha_x - c\beta_x, \beta_x, k_t + c\right)$$

Le modèle n'est donc pas identifiable et il convient alors d'imposer deux contraintes sur les paramètres. On retient en général les contraintes suivantes :

$$\sum_{x=x_m}^{x_M} \beta_x = 1 \text{ et } \sum_{t=t_m}^{t_M} k_t = 0 .$$

Question n°5 (2 points) : Indiquez deux manières d'estimer les paramètres du modèle prospectif de mortalité

L'estimation de ce modèle peut être effectuée par minimisation d'un critère de type « moindres carrés » (Lee-Carter) ou par maximum de vraisemblance (« log-Poisson »).

On souhaite intégrer une variable explicative supplémentaire dépendant du temps, pour tenir compte de l'effet des températures observées sur la mortalité et on propose donc la modèle suivant :

$$\ln\left(\mu_{x,t}\right) = \alpha_x + \beta_x \times k_t + w_t$$

[1] Voir la section 4.4 de ce support.

Question n°6 (4 points) : Montrez que $D_t = \sum_x D_{x,t}$ suit une loi de Poisson dont le paramètre λ_t vérifie une équation $\mathbf{ln}(\lambda_t) = u_t + w_t$ avec un u_t que l'on précisera.

> En sommant sur tous les âges on trouve que $D_t = \sum_x D_{x,t}$ suit une loi de Poisson
>
> $$D_x \approx P\left(\lambda_t = \sum_x E_{x,t} \times \mu_{x,t}\right)$$
>
> Mais
>
> $$\sum_x E_{x,t} \times \mu_{x,t} = \sum_x E_{x,t} \times \mathbf{exp}(\alpha_x + \beta_x \times k_t + w_t) = \sum_x E_{x,t} \times \mathbf{exp}(\alpha_x + \beta_x \times k_t) \times e^{w_t}$$
>
> En notant $u_t = \mathbf{ln}\left(\sum_x E_{x,t} \times \mathbf{exp}(\alpha_x + \beta_x \times k_t)\right)$ on obtient $\mathbf{ln}(\lambda_t) = u_t + w_t$

Pour modéliser l'influence des variations de température sur la mortalité, le modèle suivant, appelé CSDL (*Constrained Segmented Distributed Lag Model*), a été proposé[2] :

$$\mathbf{ln}(\lambda_t) = \eta(t) + \sum_{i=0}^{L_1} \beta_{1,i} \times (z_{t-i} - \Psi_1) + \sum_{i=0}^{L_2} \beta_{2,i} \times (z_{t-i} - \Psi_2)$$

avec z_t la température moyenne du jour *t* et $\eta(t)$ une fonction intégrant les variables explicatives supplémentaires telles que l'année, le jour de la semaine ou le mois.

Question n°7 (2 points) : montrez que l'expression ci-dessus peut se mettre sous la forme $\mathbf{ln}(\lambda_t) = u_t + w_t$ avec des coefficients *u* et *w* que l'on précisera. En déduire que le modèle CSDL est un cas particulier du modèle $\mathbf{ln}(\mu_{x,t}) = \alpha_x + \beta_x \times k_t + w_t$ avec un choix adapté de $\eta(t)$.

> Cette expression est bien analogue à la précédente avec
>
> $$u_t = \eta(t) \text{ et } w_t = \sum_{i=0}^{L_1} \beta_{1,i} \times (z_{t-i} - \Psi_1) + \sum_{i=0}^{L_2} \beta_{2,i} \times (z_{t-i} - \Psi_2)$$
>
> La spécification CSDL est donc bien cohérente avec un modèle Poisson classique, il faut juste choisir le terme de tendance pour rendre compte des effets de structure par âge et de dérive tendancielle de mortalité, avec
>
> $$\eta(t) = \mathbf{ln}\left(\sum_x E_{x,t} \times \mathbf{exp}(\alpha_x + \beta_x \times k_t)\right).$$

[2] Cf. MUGGEO, VMR [2010] Analyzing Temperature Effects on Mortality Within the R Environment : The Constrained Segmented Distributed Lag Parameterization. *Journal of Statistical Software* ainsi que la presentation détaillée qui en est faite dans
PINCEMIN G. [2021] Risques climatiques et mortalité, impact du risque canicule à l'horizon 2070, Mémoire d'actuaire, EURIA.

Dans la décomposition ci-dessus de $\mathbf{ln}\left(\mu_{x,t}\right)$, on a donc deux composantes : le terme $\eta\left(t\right)$ pour rendre compte des effets de structure par âge et de dérive tendancielle de mortalité et le terme w_t pour le effets d'un scénario de température. Pour le scénario « central » on a par construction du modèle $w_t = 0$ (en d'autres termes, la dérive de mortalité de référence incluse dans $\eta\left(t\right)$ correspond à $w_t = 0$.