© Théo Jalabert

Modèles de durée / Examen du 12 janvier 2015

Durée 2h – aucun document n'est autorisé

Corrigé

Sur le modèle de Thatcher / Bongaarts

La qualité de la rédaction, des justifications apportées et de la présentation de la copie seront prises en compte dans la notation.

On s'intéresse, dans le cadre de la modélisation de la mortalité, au modèle

$$h(x) = \gamma + \frac{\alpha \, \mathbf{exp}(\beta x)}{1 + \alpha \, \mathbf{exp}(\beta x)}$$

avec h la fonction de hasard du modèle (utilisé par Thatcher et Bongaarts).

Question n°1 (2 points) : Rappelez la définition et les principales propriétés de la fonction logistique. Pourquoi l'utilise-t-on souvent dans les modèles de mortalité ?

> *Cf.* le cours (*cf.* ce support p. 41)

Question n°2 (2 points) : Montrez que la transformation logistique appliquée à la fonction $h(x) - \gamma$ conduit à une fonction affine de l'âge que vous expliciterez.

> En notant $\mathbf{lg}(x) = \mathbf{ln}\left(\frac{x}{1-x}\right)$ la fonction logistique et en utilisant le fait que l'inverse de cette fonction s'écrit $g(x) = \frac{e^x}{1+e^x}$, il est immédiat que
>
> $$\mathbf{lg}\left(h(x) - \gamma\right) = \mathbf{ln}(\alpha) + \beta x .$$

Question n°3 (2 points) : déterminez la fonction de survie de ce modèle.

> On sait que $S(x) = \mathbf{exp}\left(-\int_0^x h(u)\,du\right)$ et après quelques calculs on en déduit que
>
> $$S(x) = e^{-\gamma x}\left(\frac{v_{\alpha,\beta}(x)}{v_{\alpha,\beta}(0)}\right)^{-\frac{1}{\beta}}$$
>
> avec $v_{\alpha,\beta}(u) = 1 + \alpha\,\mathbf{exp}(\beta u)$.

À partir de maintenant, on suppose que $\gamma = 0$ et on note $a = \mathbf{ln}(\alpha)$ et on s'intéresse à l'estimation des paramètres du modèle avec des données censurées. On notera $(t_i, d_i)_{1\le i\le n}$ la réalisation d'un échantillon censuré à droite avec t la durée observée et d l'indicatrice de non censure.

Question n°4 (3 points) : Rappelez, dans un cadre général, l'expression de la log-vraisemblance en fonction de S et h. Sans donner de démonstration de cette expression, vous fournirez une justification intuitive des termes associés à chaque observation. Comment cette expression doit-elle être modifiée en présence de troncature gauche ? Vous rappellerez ce qu'est la troncature gauche.

> Le détail de la réponse se trouve dans ce support. On rappelle que
> $$\ln L(\theta) = \sum_{i=1}^{n} \left(d_i \ln h(t_i) + \ln S(t_i) \right).$$
> Dans cette expression le terme avec la fonction de survie indique que l'individu i a été observé vivant juste avant t_i et celui avec la fonction de hasard, présent uniquement pour les individus dont la sortie non censurée a été observée à cet instant, est la contribution des observations complètes. En présence de troncature gauche, la loi sous-jacente est remplacée par la loi conditionnelle (avec des notations évidentes) $X \mid X > E$ et donc la fonction de hasard est inchangée et la fonction de survie remplacée par la fonction de survie conditionnelle :
> $$\ln L(\theta) = \sum_{i=1}^{n} \left(d_i \ln h(t_i) + \ln S(t_i) - \ln S(e_i) \right).$$

On revient maintenant au cas particulier du modèle $h(x) = \dfrac{\alpha \exp(\beta x)}{1 + \alpha \exp(\beta x)}$.

Question n°5 (4 points) : Calculez les expressions de $\frac{\partial}{\partial a} \ln h$, $\frac{\partial}{\partial \beta} \ln h$, $\frac{\partial}{\partial a} \ln S$ et $\frac{\partial}{\partial \beta} \ln S$ en fonction de h et S.

> Comme $\ln h = \ln e^{a+\beta x} - \ln\left(1 + e^{a+\beta x}\right)$, $\frac{\partial}{\partial a} \ln h = 1 - h$ et $\frac{\partial}{\partial \beta} \ln h = x \times (1 - h)$; pour la fonction de survie, on a montré à la question 3 que $\ln S(x) = \frac{-1}{\beta} \ln\left(\frac{1 + e^{a+\beta x}}{1 + e^{a}} \right)$ et on en déduit $\frac{\partial}{\partial a} \ln S(x) = -\frac{h(x) - h(0)}{\beta}$ et $\frac{\partial}{\partial \beta} \ln S = -\frac{1}{\beta}(xh + \ln S)$.

Question n°6 (4 points) : En déduire les équations normales à résoudre pour l'estimation de (α, β).

> On observe d'abord qu'estimer (α, β) par maximum de vraisemblance est équivalent à estimer (a, β). En utilisant la question 4 on écrit la log-vraisemblance
> $$\ln L(a, \beta) = \sum_{i=1}^{n} \left(d_i \ln h(t_i) + \ln S(t_i) \right)$$

© Théo Jalabert

et on remarque que $\frac{\partial}{\partial v}\mathbf{ln}\,L(a,\beta)=\sum_{i=1}^{n}\left(d_i\frac{\partial}{\partial v}\mathbf{ln}\,h(t_i)+\frac{\partial}{\partial v}\mathbf{ln}\,S(t_i)\right)$ pour $v=a,\beta$

et l'expression des équations à résoudre $\frac{\partial}{\partial a}\mathbf{ln}\,L(a,\beta)=0$ et $\frac{\partial}{\partial \beta}\mathbf{ln}\,L(a,\beta)=0$

s'en déduit facilement.

Question n°7 (3 points) : Comment résoudre ces équations ? Vous décrirez l'algorithme standard en détaillant chacun des termes mis en jeu. Quelle méthode d'estimation alternative pourriez-vous utiliser ?

La méthode de NR doit être détaillée. La méthode alternative est celle des moindres carrés.