

Modèles de durée / Examen du 10 février 2009

Durée 2h – aucun document n'est autorisé

Corrigé

Problème : modèle Pareto censuré

On considère une situation de censure aléatoire droite :

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

dans laquelle on suppose les censures C_i indépendantes des durées X_i et on rappelle que la log-vraisemblance de l'échantillon $(T_1, D_1), \dots, (T_n, D_n)$ s'écrit, à une constante additive près et avec des notations évidentes :

$$\ln L(\theta) = \sum_{i=1}^n \ln S_\theta(T_i) + \sum_{i=1}^n D_i \ln h_\theta(T_i).$$

On fait l'hypothèse qu'il existe $\theta > 0$ tel que $S_\theta(x) = \frac{1}{(1+x)^\theta}$, pour tout $x \geq 0$.

Question n°1 (2,5 points) : Calculer la fonction de hasard h_θ et la densité f_θ du modèle. Que dire du comportement de la fonction de hasard en fonction de $x \geq 0$? Calculer également l'information de Fisher pour une observation.

On utilise le fait que $h_\theta(x) = -\frac{d}{dx} \ln S_\theta(x) = \frac{\theta}{1+x}$, puis $f_\theta(x) = h_\theta(x)S_\theta(x)$ et donc $f_\theta(x) = \frac{\theta}{(1+x)^{\theta+1}}$. La fonction de hasard est croissante : il y a donc rodage.

L'information de Fisher est $I_\theta = -\mathbf{E}_\theta \left(\frac{\partial^2 \ln f_\theta}{\partial \theta^2}(X) \right) = \frac{1}{\theta^2}$.

Question n°2 (3 points) : Calculer l'espérance de vie résiduelle $e_\theta(x) = \mathbf{E}_\theta(X - x | X > x)$. Donner la condition sur $\theta > 0$ pour que cette espérance existe.

L'espérance de vie résiduelle est égale à $e_\theta(x) = \int S_{\theta,x}(u) du$ avec $S_{\theta,x}(u)$ la fonction de survie conditionnelle définie par $S_{\theta,x}(u) = \frac{S_\theta(x+u)}{S_\theta(x)}$. On a ici

$$S_{\theta,x}(u) = \left(\frac{1+x}{1+x+u} \right)^\theta \text{ et donc :}$$

$$e_\theta(x) = \int_0^{+\infty} \left(\frac{1+x}{1+x+u} \right)^\theta du = \left[\frac{-(1+x)^\theta}{\theta-1} \left(\frac{1}{1+x+u} \right)^{\theta-1} \right]_0^{+\infty}$$

ce qui conduit facilement à $e_\theta(x) = \frac{x+1}{\theta-1}$ dès lors que $\theta > 1$.

Question n°3 (3 points) : On fait l'hypothèse qu'on observe un échantillon issu de la loi f_θ sur un intervalle de temps $[0, c]$, $c > 0$ fixé et connu. Déterminer l'estimateur du maximum de vraisemblance de θ ainsi que l'information de Fisher dans le modèle censuré.

On écrit en posant $r = \sum_{i=1}^n 1_{\{x_i \leq c\}}$:

$$\ln L(\theta, t) = r \ln \theta - (\theta + 1) \sum_{i=1}^r \ln(1 + t_{(i)}) - \theta \sum_{i=r+1}^n \ln(1 + t_{(i)})$$

ce qui conduit à :

$$\frac{\partial}{\partial \theta} \ln L(\theta, t) = \frac{r}{\theta} - \sum_{i=1}^r \ln(1 + t_{(i)}) \text{ et donc } \frac{\partial}{\partial \theta} \ln L(\theta, t) = 0 \text{ implique :}$$

$$\hat{\theta} = \frac{r}{\sum_{i=1}^r \ln(1 + t_{(i)}) + \sum_{i=r+1}^n \ln(1 + c)} = \frac{r}{\sum_{i=1}^n \ln(1 + t_i)}$$

On note que $\frac{\partial^2}{\partial \theta^2} \ln L(\theta, t) = -\frac{r}{\theta^2} < 0$ et donc il s'agit bien d'un maximum.

L'information de Fisher est $nI_c(\theta) = -\frac{E_\theta(r)}{\theta^2} = \frac{n \Pr(X \leq c)}{\theta^2}$ avec

$$\Pr(X \leq c) = 1 - \frac{1}{(1+c)^\theta} \text{ et donc } I_c(\theta) = \frac{1}{\theta^2} \left(1 - \frac{1}{(1+c)^\theta} \right).$$

Question n°4 (2,5 points) : Donner un intervalle de confiance asymptotique au niveau $1-\alpha$ pour $\hat{\theta}$.

On sait que $\sqrt{n}(\hat{\theta} - \theta) \rightarrow N\left(0, \frac{1}{I_c(\theta)}\right)$, d'où il résulte que :

$$\Pr\left\{\theta \in \left[\frac{\hat{\theta}}{1+u_{\frac{1-\alpha}{2}}\frac{(1+c)^{-\hat{\theta}}}{\sqrt{n}}}, \frac{\hat{\theta}}{1-u_{\frac{1-\alpha}{2}}\frac{(1+c)^{-\hat{\theta}}}{\sqrt{n}}}\right]\right\} \rightarrow 1-\alpha$$

Question n°5 (2 points) : Calculer le quantile d'ordre q de la loi f_θ et en déduire une méthode graphique pour vérifier l'adéquation du modèle aux données.

Par définition, on a $S_\theta(x) = 1 - q = (1+x)^{-\theta}$ d'où l'on déduit la fonction quantile :

$$F_\theta^{-1}(q) = (1-q)^{1/\theta} - 1.$$

Si \hat{S} est un estimateur non paramétrique de la fonction de survie (par exemple celui de Kaplan-Meier), les points $(\hat{S}(t_i)^{-1/\hat{\theta}} - 1, t_i)$ doivent donc être approximativement alignés.

Question n°6 (4 points) : On suppose maintenant que la censure a pour loi $S_{\beta\theta}(x)$ pour un paramètre $\beta > 0$ inconnu. En se souvenant que la vraisemblance du modèle est de la forme

$$L(\pi) = \prod_{i=1}^n [f_X(T_i, \pi) S_C(T_i, \pi)]^{D_i} [f_C(T_i, \pi) S_X(T_i, \pi)]^{1-D_i},$$

calculer l'estimateur du maximum de vraisemblance du paramètre (θ, β) ; comment s'appelle ce type de censure ? Que devient l'estimateur de θ si β est connu ?

En utilisant la formule générale rappelée ci-dessus on trouve que :

$$\ln L(\theta, \beta, t) = n \ln \theta + (n-r) \ln \beta - ((\beta+1)\theta + 1) \sum_{i=1}^n \ln(1+t_i)$$

Les équations de vraisemblance en découlent simplement :

$$\begin{aligned} -\frac{\partial}{\partial \theta} \ln L(\theta, \beta) &= \frac{n}{\theta} - (\beta+1) \sum_{i=1}^n \ln(1+t_i) \\ -\frac{\partial}{\partial \beta} \ln L(\theta, \beta) &= \frac{(n-r)}{\beta} - \theta \sum_{i=1}^n \ln(1+t_i) \end{aligned}$$

En annulant ces 2 dérivées on obtient les équations :

$$\begin{aligned} - \hat{\theta} &= \frac{n}{(\hat{\beta}+1) \sum_{i=1}^n \ln(1+t_i)} \\ - \hat{\beta} &= \frac{(n-r)}{\hat{\theta} \sum_{i=1}^n \ln(1+t_i)} \end{aligned}$$

$$\text{On trouve donc que } \hat{\beta} = \frac{n}{r} - 1 \text{ et } \hat{\theta} = \frac{r}{\sum_{i=1}^n \ln(1+t_i)}.$$

Lorsque β est connu on trouve simplement avec la première équation de vraisemblance et on trouve immédiatement $\hat{\theta} = \frac{1}{(\beta+1) \sum_{i=1}^n \ln(1+t_i)}$.

Question n°7 (3 points) : Calculer la probabilité qu'une observation soit censurée.

La probabilité cherchée est $\Pr(X > C) = \mathbf{E}(\Pr[X > C | C])$ ce qui, compte tenu de $f_\theta(x) = \frac{\theta}{(1+x)^{\theta+1}}$, conduit à :

$$\Pr(X > C) = \mathbf{E}(\Pr[X > C | C]) = \int_0^{+\infty} \frac{\beta\theta}{(1+c)^{\beta\theta+1}} \left(\int_c^{+\infty} \frac{\theta}{(1+x)^{\theta+1}} dx \right) dc$$

Comme $\int_c^{+\infty} \frac{\theta}{(1+x)^{\theta+1}} dx = \frac{1}{(1+c)^\theta}$, on trouve finalement après quelques calculs :

$$\Pr(X > C) = \frac{\beta}{\beta+1}.$$