

Modèles de durée / Examen du 8 janvier 2008

Durée 2h – aucun document n'est autorisé

Corrigé

Problème : un exemple de censure informative

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 vraisemblance de l'échantillon $(T_1, D_1), \dots, (T_n, D_n)$ s'écrit, avec des notations évidentes :

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

On fait l'hypothèse qu'il existe $\beta > 0$ tel que $S_C(x) = S_X(x)^\beta$, pour tout $x \geq 0$.

Question n°1 (1 point) : Calculer la densité de C en fonction de f_X et S_X ; en déduire la fonction de hasard de C en fonction de celle de X .

Par hypothèse $S_C(x) = S_X(x)^\beta$ et donc $f_C(x) = -\frac{d}{dx}S_C(x) = \beta S_X(x)^{\beta-1} f_X(x)$, d'où il découle que $f_C(x) = -\frac{d}{dx}S_C(x) = \beta S_X(x)^\beta h_X(x)$ puisque $h_X(x) = \frac{f_X(x)}{S_X(x)}$; on obtient donc $h_C(x) = \frac{f_C(x)}{S_C(x)} = \beta h_X(x)$.

Question n°2 (2 points) : Déterminer la loi de $D = 1_{\{X \leq C\}}$; en déduire un estimateur simple de β .

La loi de D est une loi de Bernoulli de paramètre $p = \Pr(D=1)$; mais $\Pr(D=1) = \mathbf{E}(\Pr(X \leq C | X)) = \mathbf{E}S_C(X)$, par indépendance de X et C . On en déduit donc :

$$\Pr(D=1) = \int_0^{+\infty} (1 - F_X(x))^\beta f_X(x) dx.$$

Comme $(1-F_X(x))^\beta f_X(x) = \frac{-1}{\beta+1} \frac{d}{dx} (1-F_X(x))^{\beta+1}$, on obtient finalement :

$$\Pr(D=1) = \frac{1}{\beta+1}.$$

Comme un estimateur naturel de p est $\hat{p} = \frac{1}{n} \sum_{i=1}^n D_i = \bar{D}$, on peut proposer comme estimateur de β $\hat{\beta} = \frac{1}{\bar{D}} - 1$.

Question n°3 (3 points) : Préciser la loi limite de l'estimateur défini ci-dessus ; donner un intervalle de confiance asymptotique au niveau α (on rappelle la méthode delta :

$$\mathbf{V}(\varphi(X)) \approx \left(\frac{d\varphi}{dx}(\mathbf{E}(X)) \right)^2 \mathbf{V}(X).$$

\bar{D} est évidemment asymptotiquement gaussien et sans biais et comme $\hat{\beta} = \varphi(\bar{D})$ avec $\varphi(x) = \frac{1}{x} - 1$, $\hat{\beta}$ est asymptotiquement gaussien et sans biais. Il faut donc calculer la variance de $\hat{\beta}$; on utilise la méthode delta avec $\varphi(x) = \frac{1}{x} - 1$ et $\frac{d}{dx} \varphi(x) = \frac{-1}{x^2}$, ce qui donne

$$\mathbf{V}(\hat{\beta}) \approx \left(\frac{d\varphi}{dx}(\mathbf{E}(X)) \right)^2 \mathbf{V}(X) = \frac{p(1-p)}{p^4} = \frac{(1-p)}{p^3}$$

et donc un estimateur de la variance est $\hat{\mathbf{V}}(\hat{\beta}) = \frac{(1-\bar{D})}{\bar{D}^3}$. L'intervalle de confiance découle immédiatement de $\hat{\sigma}^{-1}(\hat{\beta} - \beta) \rightarrow N(0,1)$.

Question n°4 (1 points) : Calculer la fonction de survie de T en fonction de S_X et β ; en déduire une alternative à l'estimateur de Kaplan-Meier de S_X .

On a $S_T(t) = \Pr(X \wedge C > t) = \Pr(X > t) \Pr(C > t) = S_X(t)^{\beta+1}$.

Question n°5 (1,5 points) : Commenter les cas particuliers du modèle $\beta=1$, $\beta \rightarrow 0$ et $\beta \rightarrow +\infty$.

Lorsque $\beta = 1$ C et X ont même loi ; lorsque $\beta \rightarrow 0$, la censure disparaît ($C = +\infty$) et lorsque $\beta \rightarrow +\infty$, la censure se concentre en 0, il n'est plus possible d'observer des durées positives.

A partir de maintenant on suppose que $S_X(t) = \exp(-\theta t)$, ie X suit une loi exponentielle de paramètre θ .

Question n°6 (4 points): Ecrire la log-vraisemblance de (θ, β) en fonction de n , $\bar{D} = n^{-1} \sum D_i$ et $\bar{T} = n^{-1} \sum T_i$. En déduire les estimateurs du maximum de vraisemblance de :

- (θ, β) lorsque les 2 paramètres sont inconnus ;
- θ lorsque β est connu ;
- β lorsque θ est connu.

On part de $L(\theta) = \prod_{i=1}^n [f_X(T_i, \theta) S_C(T_i, \theta)]^{D_i} [f_C(T_i, \theta) S_X(T_i, \theta)]^{1-D_i}$ et on utilise les relations établies à la question n°1 :

- $f_C(x) = \beta S_X(x)^{\beta-1} f_X(x)$,
- $S_C(x) = S_X(x)^\beta$.

On trouve ainsi que $L(\theta, \beta) = \prod_{i=1}^n \beta^{1-D_i} f_X(T_i) S_X(T_i)^\beta$.

Dans le cas de la loi exponentielle on obtient donc à partir de cette formule :

$$\ln L(\theta, \beta) = n(1 - \bar{D}) \ln(\beta) + n \ln(\theta) - (\beta + 1) \theta n \bar{T}$$

Les équations de vraisemblance en découlent simplement :

$$\begin{aligned} - \frac{\partial}{\partial \theta} \ln L(\theta, \beta) &= \frac{n}{\theta} - (\beta + 1) n \bar{T} \\ - \frac{\partial}{\partial \beta} \ln L(\theta, \beta) &= \frac{n(1 - \bar{D})}{\beta} - \theta n \bar{T} \end{aligned}$$

En annulant ces 2 dérivées on obtient les équations :

$$- \hat{\theta} = \frac{1}{(\hat{\beta} + 1) \bar{T}}$$

$$\text{- } \hat{\beta} = \frac{(1 - \bar{D})}{\hat{\theta} \bar{T}}$$

De la première égalité on tire $\hat{\theta}(\hat{\beta} + 1) = \frac{1}{\bar{T}}$ et de la seconde $\hat{\theta}\hat{\beta} = \frac{(1 - \bar{D})}{\bar{T}}$; en faisant le quotient de ces 2 égalités puis en reportant la valeur de $\hat{\beta}$ dans l'une des équations de vraisemblance on trouve donc que :

$$\text{- } \hat{\beta} = \frac{1}{\bar{D}} - 1 ;$$

$$\text{- } \hat{\theta} = \frac{\bar{D}}{\bar{T}} .$$

On peut noter que $\hat{\theta}$ est le classique estimateur d'un modèle exponentiel censuré, rapport entre le nombre de sorties non censurées et l'exposition au risque. $\hat{\beta}$ est l'estimateur trouvé à la question n°2.

Lorsque β est connu on trouve simplement avec la première équation de vraisemblance $\hat{\theta} = \frac{1}{(\beta+1)\bar{T}}$. De manière symétrique, lorsque θ est connu on a $\hat{\beta} = \frac{(1 - \bar{D})}{\theta \bar{T}}$.

Question n°7 (2,5 points) : Calculer l'espérance et la variance de $\hat{\theta}$ lorsque β est connu ; on rappelle que lorsque Z suit une loi $\gamma(r, \lambda)$, $\mathbf{E}(Z^p) = \lambda^{-p} \frac{\Gamma(r+p)}{\Gamma(r)}$, pour $p > -r$.

On a vu à la question précédente que dans ce cas $\hat{\theta} = \frac{1}{(\beta+1)\bar{T}}$; on observe alors que comme il résulte de la question n°4 que T_i suit une loi exponentielle de paramètre $\lambda = (\beta+1)\theta$, $n\bar{T}$ suit une loi gamma $\gamma(n, \lambda)$. On calcule donc :

$$\text{- } \mathbf{E}(\bar{T}^{-1}) = n\lambda^1 \frac{\Gamma(n-1)}{\Gamma(n)} = \frac{n}{n-1} \lambda^1 ;$$

$$\text{- } \mathbf{E}(\bar{T}^{-2}) = n^2 \lambda^2 \frac{\Gamma(n-2)}{\Gamma(n)} = \frac{n^2}{(n-1)(n-2)} \lambda^2$$

et l'espérance et la variance de $\hat{\theta}$ s'en déduisent directement.

Question n°8 (2,5 points) : Calculer l'information de Fisher sur (θ, β) ; on rappelle que dans un modèle paramétrique, l'information de Fisher est définie par $I(\omega) = -\mathbf{E}\left(\frac{\partial^2 \ln L(\omega)}{\partial \omega \partial \omega'}\right)$.

En repartant des expressions :

$$\begin{aligned} -\frac{\partial}{\partial \theta} \ln L(\theta, \beta) &= \frac{n}{\theta} - (\beta + 1)n\bar{T} \\ -\frac{\partial}{\partial \beta} \ln L(\theta, \beta) &= \frac{n(1 - \bar{D})}{\beta} - \theta n\bar{T} \end{aligned}$$

on trouve :

$$\begin{aligned} -\frac{\partial^2}{\partial \theta^2} \ln L(\theta, \beta) &= -\frac{n}{\theta^2} \\ -\frac{\partial}{\partial \beta \partial \theta} \ln L(\theta, \beta) &= -n\bar{T} \\ -\frac{\partial^2}{\partial \beta^2} \ln L(\theta, \beta) &= -\frac{n(1 - \bar{D})}{\beta^2} \end{aligned}$$

Comme $\mathbf{E}(\bar{T}) = \frac{1}{(\beta+1)\theta}$ et $\mathbf{E}(\bar{D}) = \frac{1}{(\beta+1)}$ on trouve finalement que :

$$I(\theta, \beta) = \frac{n}{\beta+1} \begin{bmatrix} (\beta+1)\theta^{-2} & \theta^{-1} \\ \theta^{-1} & \beta^{-1} \end{bmatrix}$$

Question n°9 (2,5 points) : En déduire la loi limite de la statistique

$$Q_n(\theta, \beta) = n \left(\left(\frac{\hat{\theta}}{\theta} - 1 \right)^2 + \frac{\beta}{\beta+1} \left(\frac{\hat{\beta}}{\beta} - 1 \right)^2 + \frac{2\beta}{\beta+1} \left(\frac{\hat{\theta}}{\theta} - 1 \right) \left(\frac{\hat{\beta}}{\beta} - 1 \right) \right)$$

En remarquant que $Q_n(\theta, \beta) = (\hat{\theta} - \theta, \hat{\beta} - \beta) I(\theta, \beta) \begin{pmatrix} \hat{\theta} - \theta \\ \hat{\beta} - \beta \end{pmatrix}$ on conclut que $Q_n(\theta, \beta)$ suit asymptotiquement un Khi-2 à 2 degrés de liberté.