

Modèles de durée / Examen / Janvier 2018

Durée 2h – aucun document n'est autorisé

Corrigé

Vraisemblance d'un modèle paramétrique tronqué et censuré

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

On considère une variable de durée X dont la fonction de hasard sous-jacente est notée h_X et la fonction de survie S_X .

Les durées de survie (X_1, \dots, X_n) sont censurées à droite par des durées (C_1, \dots, C_n) indépendantes de l'échantillon d'intérêt et, au lieu d'observer directement (X_1, \dots, X_n) on observe $(T_1, D_1), \dots, (T_n, D_n)$ avec :

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

Question n°1 (4 points) : Déterminez l'expression de la vraisemblance en fonction des fonctions de survie et de hasard de X et de C .

On écrit (voir le [support de cours](#) pour les notations et le détail) :

$$\begin{aligned} P(T_i > t_i, D_i = 1) &= P(X_i \wedge C_i > t_i, X_i \leq C_i) = P(t_i < X_i \leq C_i) \\ &= \int_{t_i}^{+\infty} P(t_i < X_i \leq c) f_C(\theta, c) dc = \int_{t_i}^{+\infty} \left(\int_{t_i}^c f_X(\theta, x) dx \right) f_C(\theta, c) dc \end{aligned}$$

puis par Fubini on inverse les intégrales pour obtenir :

$$\begin{aligned} P(T_i > t_i, D_i = 1) &= \int_{t_i}^{+\infty} f_X(\theta, x) \left(\int_x^{+\infty} f_C(\theta, c) dc \right) dx \\ &= \int_{t_i}^{+\infty} f_X(\theta, x) S_C(\theta, x) dx \end{aligned}$$

et finalement

$$P(T_i \in [t_i, t_i + dt_i], D_i = 1) = -\frac{d}{dt_i} P(T_i > t_i, D_i = 1) = f_X(\theta, t_i) S_C(\theta, t_i) dt_i.$$

En utilisant ensuite un argument de symétrie pour calculer $P(T_i > t_i, D_i = 0)$ directement à partir des expressions ci-dessus, on trouve finalement :

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

Question n°2 (4 points) : Quelle hypothèse supplémentaire faut-il faire pour éliminer (au sens de « considérer comme une constante pour l'estimation de la loi de X ») la loi de la censure dans l'expression ci-dessus ? Donnez l'expression qu'on en déduit de la log-vraisemblance en fonction de $S_X = S_\theta$ et $h_X = h_\theta$. Donnez également un exemple dans lequel la loi de la censure ne peut pas être éliminée de la vraisemblance, alors que la censure est indépendante de la durée d'intérêt. Connaissez-vous une expérience dans laquelle la censure n'est pas indépendante de la durée modélisée ?

On doit supposer que les lois de X et C n'ont pas de paramètre en commun (« censure non informative ») et on trouve alors, qu'à une constante additive près :

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

Si par exemple on avait $S_\theta(x) = S_C(x)^\beta$, l'égalité ci-dessus ne serait plus vraie (voir [ce document](#) pour la présentation de ce modèle).

Dans le cas d'une censure « au $r^{\text{ème}}$ décès », la censure n'est pas indépendante de la variable modélisée, puisque $C = X_{(r)}$.

On suppose dans la suite que la condition permettant d'éliminer la loi de la censure de l'expression de la log-vraisemblance est vérifiée.

Question n°3 (2 points) : Qu'appelle-t-on troncature gauche ? Donnez un exemple de troncature gauche dans un contexte d'assurance. Comment faut-il adapter l'expression de la log-vraisemblance établie à la question précédente dans ce cas ?

Avec des notations évidentes (et vues en cours), la troncature gauche désigne le fait de ne pas observer X mais $X | X > E$ avec E une variable indépendante de X. L'adaptation de la vraisemblance est immédiate et conduit à

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

Dans le cas d'une franchise en arrêt de travail par exemple, E est égal à la durée de la franchise.

Question n°4 (4 points) : On fait l'hypothèse que la fonction de hasard est constante sur un intervalle $[x, x+1[$; à l'aide de ce qui précède, calculez la log-vraisemblance pour

L'estimation du paramètre θ , valeur de la fonction de hasard sur $[x, x+1[$ et en déduire l'estimateur du maximum de vraisemblance de la fonction de hasard sur cet intervalle. Quel lien avec la loi de Poisson peut-on faire ?

La fonction de hasard étant constante égale à $\theta = \theta_x$, la fonction de survie est $S(x) = \exp(-\theta x)$ et la log-vraisemblance du modèle est, à une constante près :

$$\ln L(\theta) = \sum_{i=1}^n [d_i \ln(\theta) + \theta \times (t_i - e_i)] = d_x \times \ln(\theta) + \theta \times E_x$$

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 $[x, x+1[$ était une loi de Poisson de paramètre $\theta \times E_x$;

En effet, dans ce cas $\ln(P(D_x = d)) = cste + d_x \times \ln(\theta) - \theta \times E_x$.

L'estimateur du maximum de vraisemblance de θ est donc $\hat{\theta} = \frac{D_x}{E_x}$.

Question n°5 (2 points) : Rappelez les définitions de la fonction de survie conditionnelle S_x et de la probabilité conditionnelle de sortie q_x ? Quel est le lien entre S_x et q_x ?

La fonction de survie conditionnelle est la fonction de survie de la variable aléatoire $(X - x) | X > x$, ce qui conduit à $S_x(u) = \frac{S(x+u)}{S(x)}$. Par définition, $q_x = P(X \leq x | X > x)$ et donc $q_x = 1 - S_x(1)$.

Question n°6 (2 points) : En utilisant les résultats des questions 4 et 5, indiquez quel est l'estimateur du maximum de vraisemblance de q_x lorsque la fonction de hasard est constante sur $[x, x+1[$; quelle hypothèse supplémentaire faut-il faire pour que l'on puisse utiliser l'estimateur de Hoem, $\hat{q}_x = \frac{D_x}{E_x}$?

On obtient facilement $q_x = 1 - \exp(-\theta_x)$ et donc $\hat{q}_x = 1 - \exp\left(-\frac{D_x}{E_x}\right)$, si la fonction de hasard est petite, un développement limité permet d'utiliser $\hat{q}_x \approx \frac{D_x}{E_x}$.

Question n°7 (2 points) : Si l'on souhaite utiliser l'approximation ci-dessus comme estimateur de q_x même lorsque l'hypothèse supplémentaire de la question 6 n'est pas satisfaite, quelle correction faut-il apporter au calcul de l'exposition au risque ? Donnez un exemple de l'intérêt de cette correction.

Les individus sortis non censurés dans l'intervalle $[x, x+1[$ doivent être comptés pour un dans l'exposition et non pour leur durée de présence dans l'intervalle ; pour se persuader de l'intérêt de cet ajustement, il suffit de considérer le cas d'un unique individu qui entre en x et sort au milieu de l'intervalle.