Inférence semi-paramétrique pour des évènements récurrents en présence de censure et d'un évènement terminal

Olivier Bouaziz, Olivier Lopez, Ségolen Geffray

Inférence semi-paramétrique pour des évènements récurrents en présence de censure et d'un évènement terminal

Olivier Bouaziz[1] & Olivier Lopez[1] & Ségolen Geffray[2]

[1] *Laboratoire de Statistique Théorique et Appliquée, Université Paris VI, 175 Rue du Chevaleret, 75013 Paris*

[2] *Université de Nantes*

1 Introduction

On s'intéresse à l'étude d'évènements récurrents en présence de décès et de censure aléatoire droite. Dans cet exposé on étudie le processus de comptage $N^*(t)$ qui compte le nombre d'évènements récurrents se produisant dans l'intervalle de temps $[0,t]$ et qui reflète l'état de santé du patient quand de nombreuses récurrences sont attendues (en cas de crises d'asthme ou de crises épileptiques, par exemple). On suppose qu'on observe des covariables $\mathbf{Z}$ indépendantes de loi absolument continue. On propose une nouvelle procédure semi-paramétrique : on travaille sur un modèle single index adapté à la moyenne cumulée, définie par $\mu(t|\mathbf{z}) = \mathbb{E}[N^*(t)|\mathbf{Z}=\mathbf{z}]$ et on suppose l'existence d'un paramètre θ_0 tel que $\mu(t|\mathbf{z}) = \mathbb{E}[N^*(t)|\theta_0'\cdot\mathbf{Z} = \theta_0'\cdot\mathbf{z}]$. Le but est d'obtenir un estimateur du paramètre θ_0 ainsi que de la fonction $\mu_{\theta_0}(\cdot) = \mathbb{E}[N^*(t)|\theta_0'\cdot\mathbf{Z}=\cdot]$. Nous obtenons alors des résultats de consistance et de normalité asymptotique pour ces estimateurs.

Mots clés : réduction de la dimension, évènements récurrents, censure aléatoire à droite, modèle à direction révélatrice unique, évènement terminal.

Interest is centered on a situation in which patients may experience recurrent events over time in the presence of dependent death and independent right-censoring. In this paper, we are concerned with the recurrent event process $N^*(t)$ which counts the number of recurrent events occurring over the time interval $[0,t]$ and which reflects the patient state of health when several recurrences are expected (in the case of asthma or epileptic seizures, for instance). We suppose that a time-independent vector covariate $\mathbf{Z}$ with an absolutely continuous distribution is observed. We propose a new semi-parametric inference procedure based on a single-index model adapted to the cumulative mean function defined as $\mu(t|\mathbf{z}) = \mathbb{E}[N^*(t)|\mathbf{Z}=\mathbf{z}]$. Specifically, we assume the existence of a parameter vector θ_0 such that $\mu(t|\mathbf{z}) = \mathbb{E}[N^*(t)|\theta_0'\cdot\mathbf{Z} = \theta_0'\cdot\mathbf{z}]$. Our aim consists of obtaining an estimator of the parameter vector θ_0 as well as of the function $\mu_{\theta_0}(\cdot) = \mathbb{E}[N^*(t)|\theta_0'\cdot\mathbf{Z}=\cdot]$ and of deriving the consistency and the weak convergence of the estimators.

Keywords: dimension reduction, recurrent event process, right-censoring, single-index model, terminal event.

Dans cet exposé on s'intéresse à la fréquence des évènements récurrents se produisant chez des patients comme dans le cas de crises d'asthme ou d'épilepsie par exemple. Une des principales difficultés dans l'analyse de ces évènements récurrents à partir d'étude clinique vient de la possibilité de décès du patient due à la maladie étudiée qui nous empêche d'observer les évènements récurrents postérieurs à son décès. De plus, dans un essai clinique, la survenue d'une censure aléatoire droite peut stopper définitivement l'observation des évènements récurrents. Les principales causes de censure sont : la perte de vue du patient, la fin de l'étude ou le décès du patient due à une cause indépendante de la maladie étudiée.

Ainsi, dans l'analyse des évènements récurrents les quantités d'intérêt peuvent être : les temps d'apparition des évènements récurrents, les intervalles de temps entre deux évènements récurrents successifs ou encore le processus $N^*(t)$ qui compte le nombre d'évènements récurrents survenus dans l'intervalle de temps $[0, t]$. Ici nous étudierons $N^*(t)$ qui est un bon indicateur de l'état de santé du patient.

Il existe trois approches principales pour étudier le processus $N^*(t)$: les modèles conditionnels, l'approche marginale et les modèles à frailty. Les modèles conditionnels s'intéressent à la quantité $\mathbb{E}[N^*(t)|\mathcal{F}_{t-}]$ qui représente l'intensité du processus des évènements récurrents conditionnellement aux évènements antérieurs à t. Dans les modèles à frailty, une variable latente est utilisée pour prendre en compte un effet aléatoire spécifique à chaque patient. Ici nous nous focaliserons sur l'approche marginale qui consiste à s'intéresser à la quantité $\mu(t) = \mathbb{E}[N^*(t)]$, la moyenne cumulée.

Dans les milieux médicaux il est essentiel de prendre en compte des variables explicatives $\mathbf{Z}$ (de dimension d) qui nous donnent des informations sur chaque patient par rapport à la maladie étudiée. Dans l'approche paramétrique on suppose que la fonction $\mu(t|z) = \mathbb{E}[N^*(t)|\mathbf{Z} = z]$ appartient à une certaine famille paramétrique, c'est à dire qu'on pose $\mu(t|z) = \mu_0(t, \theta_0, z)$ où μ_0 est une fonction connue et θ_0 un paramètre inconnu de dimension finie à estimer. Ces modèles requièrent donc de fortes hypothèses qui sont rarement vérifiées en pratique. Au contraire, les modèles purement non paramétriques nécessitent moins d'hypothèses puisqu'ils consistent à estimer directement la fonction $\mu(t)$ sans faire d'hypothèses sur sa forme. Cependant, il est connu que cette approche souffre du "fléau de la dimension" dès que la dimension des covariables devient trop grand ($d \geq 3$ en pratique). Une méthode intermédiaire entre ces deux modèles est alors l'approche semi paramétrique. Un des plus célèbres de ces modèles est le modèle de Cox qui consiste à faire l'hypothèse que $\mu(t|z) = \mu_0(t) \exp(\theta_0' \cdot \mathbf{Z})$ où μ_0 est une fonction et θ_0 un paramètre de dimension d. Il s'agit alors d'estimer μ_0 et θ_0, tous deux inconnus. Ces modèles ont été étudiés par de nombreux auteurs depuis un certain nombre d'années.

Par exemple, en l'absence de décès, Prentice *et al.* (1981) ont considéré des modèles de régression qui permettent à l'intensité du processus des évènements récurrents de dépendre du passé du patient en utilisant des modèles stratifiés. De même, Andersen et Gill (1982) ont étudié des modèles de Cox avec variables censurées où les covariables dépendent du temps. Ils ont aussi supposé que le processus $N^*(t)$ était un processus de Poisson. Andersen *et al.* (1993) quant à eux ont étudié un modèle de Markov non-homogène basé sur le processus $N^*(t)$ en présence de censures. Lin et al. (2000) ont considéré l'approche marginale dans un modèle de Cox sans faire d'hypothèse de Poisson sur $N^*(t)$. Enfin, Lawless et Nadeau (1995) ont proposé un modèle semi paramétrique où la moyenne cumulée est proportionnelle à une fonction de base inconnue avec un coefficient qui dépend des covariables.

Toutes ces méthodes cependant ne prennent pas en compte la présence de décès pendant l'étude. Récemment, des travaux ont été effectués sur les évènements récurrents en présence de décès. Ghosh et Lin (2002) ont étudié un modèle de régression où la dérivée de la moyenne cumulée est proportionnelle à une fonction de base inconnue avec un coefficient de regression. C'est donc un modèle semi paramétrique en présence de décès qu'ils ont étudiés en utilisant la méthode IPCW (Inverse Probability of Censoring Weighting) et la méthode IPSW (Inverse Probability of Survival Weighting). Milosvlasky *et al.* ont eux aussi construit un estimateur IPCW pour le paramètre de régression dans le modèle d'Andersen et Gill qui est consistent sous censure.

Toutes ces méthodes cependant s'appuient sur des hypothèses fortes qui peuvent ne pas être vérifiées en pratique. Mais il est quand même important d'introduire des techniques de réduction de la dimension en utilisant des modèles semi-paramétriques pour trouver un bon compromis entre la flexibilité du modèle et le besoin de pallier au fléau de la dimension.

2 Modèle et procédure d'estimation

On rappelle que $N^*(t)$ représente le nombre d'évènements récurrents survenus dans l'intervalle de temps $[0, t]$. On note D la date de décès, de fonction de répartition F. Puisque les patients qui décèdent ne peuvent plus avoir d'évènements récurrents, le processus $N^*(\cdot)$ ne saute pas après D. Soit C la date de censure, de fonction de répartition G. L'hypothèse single-index sur μ suppose l'existence d'un paramètre inconnu $\theta_0 \in \Theta \subset \mathbb{R}^d$ tel que :

$$\mu(t|\mathbf{z}) = \mu_{\theta_0}(t|\mathbf{Z} = \mathbf{z}) = \mathbb{E}[N^*(t)|\theta_0' \cdot \mathbf{Z} = \theta_0 \cdot \mathbf{z}], \tag{1}$$

où, pour tout $\theta \in \Theta$, $\mu_\theta(t|u) = \mathbb{E}(N^*(t)|\theta' \cdot \mathbf{Z} = u)$.

A cause des censures, $N^*(\cdot)$ et D ne sont pas toujours observées. A la place, nos observations consistent en :

$$\begin{cases} N(t) = N^*(t \wedge C) \\ T = D \wedge C \\ \delta = I_{D \leq C} \\ \mathbf{Z} \in \mathbb{R}^d. \end{cases}$$

Dans toute la suite on fera l'hypothèse que C est indépendant du couple $(D, \mathbf{Z})$. Cette hypothèse est satisfaite par exemple dans le cas où la censure est causée par la fin de l'étude ou parce que le patient est perdu de vue. Cependant il existe des cas où cette hypothèse n'est pas vérifiée et il est alors possible de la remplacer par les hypothèses suivantes :

$$\begin{cases} D \text{ et } C \text{ sont indépendants} \\ \mathbb{P}(D \leq C | \mathbf{Z}, D) = \mathbb{P}(D \leq C | D). \end{cases}$$

Ce type de conditions ont été très largement étudié dans la littérature sur la censure, entre autres par Stute (1993, 1996, 1999), Delecroix, Lopez et Patilea (2008), Lu et Burke (2005). Il est à noter que ces nouvelles hypothèses permettent une forme de dépendance entre la censure C et la covariable $\mathbf{Z}$. Même si nos résultats restent valides sous ces hypothèses, nous présenterons nos résultats seulement sur la première hypothèse (C indépendant de $(D, \mathbf{Z})$) par souci de clarté.

Un calcul rapide nous donne le lemme suivant :

Lemme 1. *Sous certaines hypothèses (dont (1)) on a :*

$$\mathbb{E}\left[\mu_\theta(t|\theta' \cdot \mathbf{Z})^2\right] \leq \mathbb{E}\left[\mu_{\theta_0}(t|\theta_0' \cdot \mathbf{Z})^2\right].$$

Comme conséquence directe de ce lemme,

$$\theta_0 = \arg\max_{\theta \in \Theta} \int_{\mathbb{R}} w(t) \mathbb{E}[\mu_\theta(t|\theta' \mathbf{Z})]^2 dt \tag{2}$$

où w représente une fonction positive à valeurs réelles. On définit maintenant

$$Y = \iint_0^t \frac{dN(s)}{1 - G(s-)} dw(t).$$

Tout d'abord, sous (1) on remarque qu'on a :

$$\mathbb{E}[Y|\mathbf{Z}] = \mathbb{E}[Y|\theta_0' \cdot \mathbf{Z}] = \iint_0^t \frac{\mathbb{E}[dN(s)|\theta_0' \cdot \mathbf{Z}]}{1 - G(s-)} dw(t) = \int \mu_{\theta_0}(t|\theta_0' \cdot \mathbf{Z}) dw(t).$$

A l'aide de cette dernière égalité et de (2), on peut montrer le lemme suivant :

Lemme 2. *Sous certaines hypothèses,*

$$\theta_0 = \arg\min_{\theta\in\Theta} \mathbb{E}\left[\left(Y - \mathbb{E}(Y|\theta'\cdot\mathbf{Z})\right)^2\right].$$

On propose donc un estimateur de θ basé sur ce lemme. Tout d'abord pour obtenir un estimateur de Y il faut estimer G, la fonction de répartition de la censure. En effet, sous présence de censure la fonction de répartition de C est indisponible puisque les censures ne sont pas directement observées. Un moyen classique pour estimer G consiste alors à utiliser l'estimateur de Kaplan Meier. On note $\hat{G}$ cet estimateur. Ensuite, on peut estimer $\mathbb{E}(Y|\theta'\mathbf{Z})$ en utilisant un estimateur à noyau classique. Finalement, on définit l'estimateur de θ_0 de la façon suivante :

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \left[\hat{Y}_i - \frac{\sum_{j\neq i} \hat{Y}_j K\left(\frac{\theta'\cdot\mathbf{Z}_j - \theta'\cdot\mathbf{Z}_i}{h}\right)}{nh\hat{f}_{\theta'\mathbf{Z}}(\theta'\cdot\mathbf{Z}_i)}\right]^2$$

où

$$\hat{Y}_i = \iint_0^t \frac{dN_i(s)}{1-\hat{G}(s-)} dw(t), \quad 1 \leq i \leq n,$$

$$\hat{f}_{\theta'\mathbf{Z}}(\theta'\cdot\mathbf{z}) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\theta'\cdot\mathbf{Z}_i - \theta'\cdot\mathbf{z}}{h}\right),$$

K est un noyau satisfaisant des hypothèses classiques et h est la fenêtre.

3 Résultats asymptotiques

Dans cette section nous donnons tout d'abord les résultats de convergence asymptotique obtenus pour notre estimateur $\hat{\theta}$ de θ_0 et nous donnons ensuite un résultat sur l'estimation de la fonction μ_θ elle même. Nous voulons préciser ici qu'à partir d'un estimateur consistant de θ_0, il est possible de choisir tout type d'estimateur non paramétrique de μ_θ qui vérifie un certain nombre d'hypothèses. On pourra utiliser par exemple un estimateur à noyau comme défini précédemment ou par exemple un estimateur par projection...

Théorème 1. *Sous certaines hypothèses, on a :*

1. *Consistence :*

$$\hat{\theta} \underset{\mathbb{P}}{\longrightarrow} \theta_0.$$

2. *Normalité asymptotique :*

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0, \Sigma),$$

où Σ est une matrice de variance qui peut être estimée de façon consistante.

Pour prouver la normalité asymptotique de notre estimateur, nous avons dû faire des hypothèses de régularité sur la fonction $\mu_\theta(t|\theta'\mathbf{Z})$: cela nous permet de dire que le gradient de μ_θ (par rapport à θ) appartient à une certaine classe de fonctions qui est une classe de Donsker. On a ainsi pu utiliser des outils de processus empiriques nécessaires pour prouver le théorème 1.

Soit $\hat{\mu}_{\theta,h}$ un estimateur de μ_θ appartenant à une certaine classe de fonctions où h est un paramètre de lissage.

Théorème 2. *Sous certaine hypothèses,*

$$\hat{\mu}_{\hat{\theta},h}(t|\hat{\theta}z) = \hat{\mu}_{\theta_0}(t|\theta_0 z) + R_n(z,t,h)$$

où $\sup_{z,t,h} |R_n(z,t,h)| = O_P(n^{-1/2})$.

Bibliographie

[1] Andersen, P., Borgan, O., Gill, R. et Keiding N. (1993) *Statistical models based on counting processes*, New-York: Spinger Verlag.

[2] Andersen, P. et Gill, R. (1982) Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, 10 , 1100–1120.

[3] Delecroix, M., Lopez, O. et Patilea, V. (2005) Nonlinear censored regression using synthetic data. *Scand. J. Statist.*, 35(2), 248–265.

[4] Lu, X. et Burke M.D. (2005) Censored multiple regression by the method of average derivatives. *J. Multivariate Anal.*, 95(1), 182–205.

[5] Ghosh, D. et Lin, D. (2000) Nonparametric analysis of recurrent events and death. *Biometrics*, 56, 554–562.

[6] Ghosh, D. et Lin, D. (2002) Marginal regression models for recurrent and terminal events. *Statist. Sinica*, 12, 663–688.

[7] Lawless, J. et Nadeau, C. (1995) Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37, 158–168.

[8] Lin, D., Wei, L., Yang, I. et Ying, Z. (2000) Semiparametric regression for the mean and rate functions of recurrent events. *J. Roy. Statist. Soc. B*, 62, 711–730.

[9] Miloslavsky, M., Keles, S., Van der Laan, M. et Butler, S. (2004) Recurrent events analysis in the presence of time-dependent covariates and dependant censoring. *J. Roy. Statist. Soc. B*, 66, 239–257.

[10] Prentice, L., Williams, B. et Peterson A. (1981) On the regression analysis of multivariate time data. *Biometrika*, 68, 373–379.

[11] Stute, W. (1993) Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal.*, 45(1), 89–103.

[12] Stute, W. (1996) Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.*, 23(4), 461–471.

[13] Stute, W. (1999) Nonlinear censored regression. *Statist. Sinica*, 9(4), 1089–1102.