Inférence statistique dans des modèles de moments conditionnels en présence de censure

Pascal Lavergne, Olivier Lopez, Valentin Patilea

Inférence statistique dans des modèles de moments conditionnels en présence de censure

Pascal Lavergne, Olivier Lopez & Valentin Patilea

Simon Fraser University et Toulouse School of Economics
Manufacture de Tabacs
21 allées de Brienne
31000 Toulouse
&
LSTA, Université Pierre et Marie Curie - Paris 6
175 rue du Chevaleret
75013 Paris
&
IRMAR-INSA
20, Avenue des Buttes de Coësmes
CS 70839, 35708 Rennes Cedex 7

Résumé: Une large littérature statistique, biostatistique, économétrique étudie l'inférence statistique dans des modèles semi paramétriques définis par des moments conditionnels lorsque les données sont censurées. La régression paramétrique classique ou quantile en présence d'une censure à droite sur la variable à expliquer sont des exemples de référence. Nous proposons une nouvelle classe d'estimateurs pour des modèles définis par l'espérance conditionnelle d'une fonction connue dépendante de vecteurs observés Y et X et d'un paramètre inconnu θ, sachant le vecteur aléatoire X, lorsque les observations Y sont soumises à un mécanisme de censure. L'estimateur proposé minimise une distance pondérée à l'aide d'un noyau et des poids qui prennent en compte la censure. Les résultats théoriques sont obtenus uniformément par rapport à la fenêtre du noyau et sous de conditions générales sur le mécanisme de censure. Plusieurs exemples de mécanismes de censure sont proposés.

Mots clés: Données de survie et données censurées, Modèles semi et non paramétriques.

Abstract: Much of the recent statistics, biostatistics and econometric literature addressed the problem of parameters estimation and testing in conditional moment restriction models in the presence of censored observations. Parametric censored regression and censored quantile regression models are benchmark examples. We propose a new class of estimators for models defined by conditional moment restrictions involving a variable Y and a vector of conditioning variables X when the variable Y is right-censored. Our generic estimator minimizes a weighted distance criterion based on kernel smoothing and suitable weights that account for the censoring mechanism. We develop a theory that

focuses on uniformity in bandwidth. We establish a root-n-asymptotic representation of our estimator as a process depending on the bandwidth within a wide range including fixed bandwidths. The results are obtained under a set of general conditions of the censoring mechanism. Several examples of right-censoring are discussed where the weights are based on Kaplan-Meier (respectively conditional Kaplan-Meier) estimator of the distribution function of the censoring variable.

Keywords: Survival analysis and censored data, Semi and nonparametric models.

1 Description du modèle

Soit $Z_1, ..., Z_n, ...$ des observations i.i.d. d'un vecteur $\widetilde{Z} = (Y, X')' \in \mathbb{R}^{s+q}$, $p, s \geq 1$. Soit $g(y, x, \theta)$ une fonction donnée à valeur en R^r, $r \geq 1$ qui dépend de $\widetilde{Z}$ et un paramètre inconnu $\theta \in \Theta \subset \mathbb{R}^d$, $d \geq 1$. Il est supposé que le paramètre est identifié par des restrictions de moment conditonnels définis pas g. Plus précisement, il existe un unique $\theta_0 \in \Theta$ tel que

$$\mathbb{E}\left[g(\widetilde{Z}, \theta_0)|X\right] = 0 \quad \text{a.s.} \tag{1}$$

Devant un tel modèle, les objectifs classiques sont d'estimer θ_0, tester des hypothèses sur les composantes de θ_0 et vérifier qu'un tel θ_0 existe. Il existe un vaste littérature proposant des nombreuses méthodes pour répondre à ces questions. Voir par exemple Lavergne et Patilea (2009) pour un survol. Pour simplifier, nous nous limiterons par la suite au cas $s = 1$ et $r = 1$.

Ici on s'intéresse au cas où le vecteur Y subit un mécanisme de censure et donc $\widetilde{Z}$ n'est pas complétement observable. Nous proposons une technique qui est adaptée à différents types de censure (censure à droite, censure bilatérale...) sous réserve qu'un certain nombre de conditions soient vérifiées. Nous fournissons quelques exemples de situations dans lesquelles notre technique s'applique.

Exemples :

1. *Censure à droite.* Dans ce modèle, la variable Y n'est pas observée directement, mais on observe des copies i.i.d. de
$$\begin{aligned} T_i &= \inf(Y_i, C_i), \\ \delta_i &= \mathbf{1}_{\{Y_i \leq C_i\}}. \end{aligned}$$

2. *Censure bilatérale.* Voir le modèle Patilea et Rolin (2006). On introduit une variable auxiliaire de Bernoulli Δ_i indépendante de la censure et de Y. On observe, en lieu et place de Y,
$$\begin{aligned} T_i &= \min(Y_i, C_i) + (1 - \Delta_i)\max(C_i - Y_i, 0), \\ \delta_i &= \mathbf{1}_{\{T_i = C_i, \Delta_i = 1\}} + \mathbf{1}_{\{T_i = Y_i, \Delta_i = 1\}} + \mathbf{1}_{\{\Delta_i = 0\}}. \end{aligned}$$

2 Procédure d'estimation et propriétés asymptotiques

Soit $Z_i = (T_i, X_i, \delta_i)$, et $Z_i' = (T_i, X_i)$. On suppose qu'il existe une fonction $\omega(Z)$ telle que, pour toute fonction ϕ,

$$E\left[\omega(Z)\phi(Z')|X\right] = E[\phi(\tilde{Z})|X].$$

L'existence d'une telle fonction ω dépend des conditions d'identifiabilité du modèle, et notamment des hypothèses qui sont faites sur les éventuelles relations de dépendance entre la variable de censure et les variables explicatives.

Par exemple, dans le cas de la censure à droite, si on suppose que C est indépendant de (Y, X), on peut considérer

$$\omega(Z_i) = \frac{\delta_i}{1 - G(T_i-)},$$

où $G(t) = \mathbb{P}(C \leq t)$ est la fonction de répartition de la censure. En revanche, si C est indépendant de Y conditionnellement à X (mais éventuellement peut dépendre de X), on pourra considérer

$$\omega(Z_i) = \frac{\delta_i}{1 - G(T_i - |X_i)}.$$

En général, sauf hypothèses supplémentaires, la distribution de la censure est inconnue, et la fonction ω n'est donc pas calculable. En conséquence, il est nécessaire de s'appuyer sur une version estimée $\hat{\omega}$. Dans les deux exemples précédents, cette estimation peut être réalisée en remplaçant G par son estimateur de Kaplan-Meier (premier cas) ou par son estimateur Kaplan-Meier conditionnel (deuxième cas).

Munis de ces estimateurs, on peut construire une forme quadratique

$$Q_n(\theta) = \sum_{i \neq j} \hat{\omega}(Z_i) g(Z_i', \theta) K\left(\frac{X_i - X_j}{h}\right) \hat{\omega}(Z_j) g(Z_j', \theta),$$

où l'on a introduit une fonction noyau K et un paramètre de lissage h.

Un estimateur du paramètre θ_0 est ensuite obtenu par minimisation de cette forme quadratique. En effet, cette forme quadratique peut-être vue comme un estimateur de la quantité

$$Q(\theta) = E[E[g(\tilde{Z}, \theta)|X]^2 f(X)],$$

où f est la densité du vecteur X. Cette quantité est strictement positive si $\theta \neq \theta_0$ (si le modèle est identifiée), et égale à zéro en θ_0. Ainsi, l'estimateur obtenu généralise celui introduit par Lavergne et Patilea (2009).

Sous des hypothèses générales de convergence de $\hat{\omega}$ vers ω, nous montrons que l'estimateur obtenu, comme celui de Lavergne et Patilea (2009) est asymptotiquement normal et converge à la vitesse $n^{-1/2}$. Ces conditions sur $\hat{\omega}$ sont reliées à des estimateurs de la

fonction de répartition multivariée de (X, Y), voir par exemple Lopez (2007). Par ailleurs, la forme quadratique Q_n à la base de notre procédure d'estimation peut être utilisée pour tester non paramétriquement l'adéquation au modèle (1).

Bibliographie

[1] Lavergne, P. et Patilea, V. (2009). *Smooth Minimum Distance Estimation and Testing with Conditional Estimating Equations: Uniform in Bandwidth Theory.* Working Paper Simon Fraser University.
[2] Lopez O. (2007) *On the estimation of the joint distribution in a censored regression model*, Document Crest 2007-11.
[3] Patilea, V., Rolin , J.-M.(2006) *Product-limit estimators of the survival function for two modified forms of current-status data,* Bernoulli, 12, p. 801–819.