Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 346 (2008) 225–228
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique/Probabilités

Loi du logarithme uniforme pour un estimateur non paramétrique de la régression en données censurées

Vivian Viallon

L.S.T.A., Université Paris 6, 175, rue du Chevaleret, 75013 Paris, France

Reçu le 20 juin 2006 ; accepté après révision le 22 novembre 2007

Disponible sur Internet le 16 janvier 2008

Présenté par Paul Deheuvels

Résumé

Dans cette Note nous présentons une loi du logarithme uniforme pour un estimateur non paramétrique de la régression en présence de données censurées. Cette loi est analogue à celle obtenue, notamment, par Einmahl et Mason [U. Einmahl, D.M. Mason, J. Theor. Probab. 13 (2000) 1–3] dans le cas non censuré. ***Pour citer cet article : V. Viallon, C. R. Acad. Sci. Paris, Ser. I 346 (2008).***

Abstract

A uniform law of the logarithm for a nonparametric estimate of the regression function under random censorship. In this Note, a uniform law of the logarithm is established for a nonparametric estimate of the regression function under random censorship. This law is analogous to that obtained by Einmahl and Mason [U. Einmahl, D.M. Mason, J. Theor. Probab. 13 (2000) 1–3] in the uncensored case. ***To cite this article: V. Viallon, C. R. Acad. Sci. Paris, Ser. I 346 (2008).***

1. Introduction

A travers l'étude du processus empirique multivarié indexé par des classes de fonctions, Einmahl et Mason [6] ou encore Deheuvels et Mason [5] ont récemment obtenu des lois du logarithme uniformes, notamment pour des estimateurs non paramétriques de la régression dans le cas non censuré. L'objectif de cette Note est d'étendre ce type de résultat au cas des données censurées. Après avoir rappelé la définition de l'estimateur à noyau de type Inverse Probability of Censoring Weighted (I.P.C.W.) développé par Kohler et al. [12] (voir également Carbonez et al. [2]), nous établissons une loi du logarithme uniforme pour cet estimateur de la fonction de régression. Nous obtenons également en corollaire de ce résultat une loi analogue pour un estimateur de la fonction de répartition conditionnelle en données censurées.

2. Notations et hypothèses

Considérons le triplet $(Y, C, \mathbf{X})$ à valeurs dans $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$, où Y est la variable d'intérêt, C la variable de censure et $\mathbf{X}$ un vecteur de $\mathbb{R}^d$ de variables concomitantes, dont le lien avec Y est à étudier. Nous travaillons par la suite

Adresse e-mail : viallon@ccr.jussieu.fr.

1631-073X/$ – see front matter
doi:10.1016/j.crma.2007.11.030

sur un échantillon $\{(Y_i, C_i, \mathbf{X}_i)_{1\leqslant i\leqslant n}\}$ de triplets indépendants et de même loi que $(Y, C, \mathbf{X})$. En pratique, on observe $Z_i := \min\{Y_i, C_i\}$ et $\delta_i := \mathbb{1}_{\{Y_i\leqslant C_i\}}$, où $\mathbb{1}_E$ désigne la fonction indicatrice de E. Notons, pour $t \in \mathbb{R}$, $F(t) := \mathbb{P}(Y \leqslant t)$, $G(t) := \mathbb{P}(C \leqslant t)$ et $H(t) := \mathbb{P}(Z \leqslant t)$, les versions continues à droite des fonctions de répartition de Y, C et Z.

Nous travaillerons par la suite sur une boule I de $\mathbb{R}^d$ d'intérieur non vide. Nous supposerons qu'il existe un réel $\alpha > 0$ tel que $(\mathbf{X}, Y)$ ait une densité jointe $f_{\mathbf{X},Y}$ par rapport à la mesure de Lebesgue sur $I^\alpha \times \mathbb{R}$, avec $I^\alpha := \{\mathbf{x} \in \mathbb{R}^d: \inf_{\mathbf{y}\in I} \|\mathbf{x}-\mathbf{y}\|_{\mathbb{R}^d} \leqslant \alpha\}$, $\|\cdot\|_{\mathbb{R}^d}$ désignant la norme euclidienne usuelle sur $\mathbb{R}^d$. Notons de plus $f_{\mathbf{X}}$ la densité marginale de $\mathbf{X}$ sur I^α. Les hypothèses suivantes seront faites pour établir nos résultats :

(F.1) C et $(\mathbf{X}, Y)$ sont indépendants.
(F.2) C admet une densité f_C par rapport à la mesure de Lebesgue sur $\mathbb{R}$.
(F.3) $f_{\mathbf{X}}$ est continue et strictement positive sur I^α.
(F.4) Pour tout $\mathbf{x} \in I^\alpha$, $\lim_{\mathbf{x}'\to\mathbf{x};\mathbf{x}'\in I^\alpha} f_{\mathbf{X},Y}(\mathbf{x}', y) = f_{\mathbf{X},Y}(\mathbf{x}, y)$ pour presque tout $y \in \mathbb{R}$.

Étant donnée une fonction ψ borélienne et à valeurs dans $\mathbb{R}$, nous nous intéressons à l'espérance conditionnelle de $\psi(Y)$ sachant que $\mathbf{X} = \mathbf{x}$, pour tout $\mathbf{x} \in I$,

$$m_\psi(\mathbf{x}) := \mathbb{E}(\psi(Y)|\mathbf{X} = \mathbf{x}).$$

Soit $(h_n)_{n\geqslant 1}$ une suite de réels positifs et K un noyau, i.e. une fonction mesurable à valeurs réelles et d'intégrale 1, défini sur $\mathbb{R}^d$. Pour estimer m_ψ, nous nous inspirons des idées de Kohler et al. [12] en introduisant dans un premier temps l'estimateur défini sur I,

$$\widehat{m}^{(1)}_{\psi;n}(\mathbf{x}) := \sum_{i=1}^n W_{n,i}(\mathbf{x}) \frac{\delta_i \psi(Z_i)}{1 - G(Z_i)}, \quad \text{avec } W_{n,i}(\mathbf{x}) := \frac{K((\mathbf{x}-\mathbf{X}_i)/h_n)}{\sum_{j=1}^n K((\mathbf{x}-\mathbf{X}_j)/h_n)}.$$

Remarquons que d'après l'hypothèse (F.3), la fonction de poids $W_{n,i}$ est définie à partir d'un certain rang pour tout $\mathbf{x} \in I$. En pratique la fonction G est inconnue, et $\widehat{m}^{(1)}_{\psi;n}(\mathbf{x})$ est alors remplacé par

$$\widehat{m}_{\psi;n}(\mathbf{x}) := \sum_{i=1}^n W_{n,i}(\mathbf{x}) \frac{\delta_i \psi(Z_i)}{1 - G_n(Z_i)}, \tag{1}$$

où G_n est l'estimateur de Kaplan–Meier [11] de G, et avec la convention $0/0 = 0$. Lorsque la variable Y est censurée à droite, les fonctionnelles de la loi conditionnelle de Y ne peuvent généralement pas être estimées sur l'ensemble du support de Y [9]. Par la suite, nous travaillerons donc sous la condition (**A**), qui sera dite vérifiée si l'une au moins des conditions (**A**)(i) ou (**A**)(ii) introduites ci-dessous l'est. Notons $T_L := \sup\{t \in \mathbb{R}:\ L(t) < 1\}$ pour toute fonction de répartition L continue à droite sur $\mathbb{R}$.

(**A**)(i) Il existe un réel $\tau < T_H$ tel que $\psi = 0$ sur (τ, ∞).
(**A**)(ii) (a) Pour un réel $0 < p \leqslant 1/2$ fixé, $\int_0^{T_H} (1-F)^{-p/(1-p)}\, dG < \infty$ et $n^{2p-1} h_n^{-d} |\log(h_n)| \to \infty$ lorsque $n \to \infty$;
(b) $T_F < T_G$ et $(Y, C) \in \mathbb{R}^+ \times \mathbb{R}^+$.

Le choix $\psi = \mathbb{1}_{[0,t]}$, pour $t < T_H$, en (1) conduit à estimer la fonction de répartition conditionnelle $F(t|\mathbf{X} = \mathbf{x}) := \mathbb{P}(Y \leqslant t|\mathbf{X} = \mathbf{x})$ par

$$\widehat{F}_n(t|\mathbf{X} = \mathbf{x}) := \sum_{i=1}^n W_{n,i}(\mathbf{x}) \frac{\delta_i \mathbb{1}_{\{Z_i\leqslant t\}}}{1 - G_n(Z_i)}. \tag{2}$$

A l'instar d'Einmahl et Mason [6] et Deheuvels et Mason [5], nous n'étudierons ici que le comportement asymptotique de nos estimateurs convenablement centrés. Il peut être montré que les termes résiduels de type biais sont négligeables devant les termes de type variance traités dans nos résultats, sous des conditions de régularité générales (cf. [5]). Soient alors $m_{\psi;n}(\mathbf{x})$ et $F_n(t|\mathbf{X} = \mathbf{x})$ les termes de centrage définis sur I comme suit

$$m_{\psi;n}(\mathbf{x}) := \frac{\mathbb{E}\{\psi(Y) K((\mathbf{x}-\mathbf{X})/h_n)\}}{\mathbb{E}\{K((\mathbf{x}-\mathbf{X})/h_n)\}} \quad \text{et} \quad F_n(t|\mathbf{X} = \mathbf{x}) := \frac{\mathbb{E}\{\mathbb{1}_{\{Y\leqslant t\}} K((\mathbf{x}-\mathbf{X})/h_n)\}}{\mathbb{E}\{K((\mathbf{x}-\mathbf{X})/h_n)\}}.$$

Introduisons enfin, pour tout $\mathbf{x} \in I$, la quantité $\sigma_\psi^2(\mathbf{x}) = \mathbb{E}\{\psi^2(Y)/[1-G(Y)] \mid \mathbf{X} = \mathbf{x}\} - m_\psi^2(\mathbf{x})$.

Pour établir nos résultats, nous supposerons que la fonction ψ est bornée, i.e.

(F.5) Il existe $M > 0$ tel que $\sup_{t\in\mathbb{R}} |\psi(t)| \leqslant M < \infty$.

Concernant le noyau K, nous travaillerons sous les conditions (K.1-2) ci-dessous.

(K.1) K est positif, à support compact et uniformément borné sur son support.
(K.2) Il existe un polynôme P de d variables réelles et une fonction réelle ϕ à variation bornée tels que $K(\mathbf{x}) = \phi(P(\mathbf{x}))$.

Enfin, nous supposerons que la suite de réels $(h_n)_{n\geqslant 1}$ vérifie les conditions suivantes, lorsque $n \to \infty$,

(H.1) $h_n \to 0$; (H.2) $nh_n^d/\log n \to \infty$; (H.3) $(h_n^d \log\log n)/\big|\log(h_n)\big| \to 0$.

3. Résultats

Dans ce qui suit, « $\xrightarrow{\mathbb{P}}$ » dénote la convergence en probabilité.

Théorème 3.1. *Soit $(h_n)_{n\geqslant 1}$ une suite de réels vérifiant les conditions* (H.1-2-3). *Sous les hypothèses* (**A**), (F.1-2-3-4-5) *et* (K.1-2)*, on a,*

$$\sup_{\mathbf{x}\in I} \frac{\sqrt{nh_n^d} \pm \{\widehat{m}_{\psi;n}(\mathbf{x}) - m_{\psi;n}(\mathbf{x})\}}{\sqrt{2\log(1/h_n^d)}} \xrightarrow{\mathbb{P}} \left\{ \int_{\mathbb{R}^d} K^2(\mathbf{t})\,\mathrm{d}\mathbf{t} \sup_{\mathbf{x}\in I} \frac{\sigma_\psi^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right\}^{1/2}. \tag{3}$$

Corollaire 3.2. *Soient un réel $\tau < T_H$ et une suite $(h_n)_{n\geqslant 1}$ de réels vérifiant les conditions* (H.1-2-3). *Sous les hypothèses* (F.1-2-3-4) *et* (K.1-2) *on a, pour tout $-\infty < t \leqslant \tau$,*

$$\sup_{\mathbf{x}\in I} \frac{\sqrt{nh_n^d} \pm \{\widehat{F}_n(t|\mathbf{x}) - F_n(t|\mathbf{x})\}}{\sqrt{2\log(1/h_n^d)}} \xrightarrow{\mathbb{P}} \left\{ \int_{\mathbb{R}^d} K^2(\mathbf{t})\,\mathrm{d}\mathbf{t} \sup_{\mathbf{x}\in I} \frac{\sigma_t^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right\}^{1/2}, \tag{4}$$

avec $\sigma_t^2(\mathbf{x}) = \mathbb{E}\{\mathbb{1}_{\{Y\leqslant t\}}/[1-G(Y)] \mid \mathbf{X} = \mathbf{x}\} - F^2(t|\mathbf{X} = \mathbf{x})$.

Remarque 1. Dans [4], la convergence presque sûre d'un estimateur de la fonction de répartition conditionnelle en données censurées est obtenue uniformément en $t \in [0, \tau]$, et en $h_n \in [a'_n\, a''_n]$ (où $(a'_n)_{n\geqslant 1}$ et $(a''_n)_{n\geqslant 1}$ vérifient certaines conditions de croissance). Cependant, la vitesse correspondante n'est pas fournie. Remarquons premièrement que l'estimateur $\widehat{F}_n(t|\mathbf{x})$ introduit ici est différent de l'estimateur classique de la fonction de répartition conditionnelle en données censurées étudié notamment par Deheuvels et Derzko [4] (pour plus de détails, voir la section Discussion de Carbonez et al. [2]). D'autre part, l'uniformité en t et h_n pourrait être obtenue dans notre cadre d'étude en utilisant des arguments analogues à ceux présentés dans [6] et [7]. Ce problème sera considéré ultérieurement.

3.1. Éléments de preuve

Supposons dans un premier temps que la fonction G est connue. Considérons la fonction Ψ définie de $\mathbb{R}^2$ dans $\mathbb{R}$ par $\Psi(y,c) = [\mathbb{1}_{\{y\leqslant c\}}\psi(y\wedge c)]/[1-G(y\wedge c)]$. L'estimateur $\widehat{m}_{\psi;n}^{(1)}$ peut se réécrire sous la forme d'un estimateur non paramétrique linéaire de la régression généralisée, analogue au cas non censuré [5],

$$\widehat{m}_{\psi;n}^{(1)}(\mathbf{x}) = \sum_{i=1}^n W_{n,i}(\mathbf{x})\Psi(Y_i, C_i).$$

En utilisant des arguments de conditionnement, il vient sous l'hypothèse (F.1),

$$\mathbb{E}\left\{ \left(\frac{\delta\psi(Z)}{1-G(Z)} \right)^j \middle| \mathbf{X} \right\} = \mathbb{E}\left\{ \frac{\psi^j(Y)}{(1-G(Y))^{j-1}} \middle| \mathbf{X} \right\}, \quad \text{pour } j = 1, 2. \tag{5}$$

Le Théorème 3.1 est alors une conséquence directe du Théorème 2.1 de Blondin [1], (5), et un résultat similaire, permettant d'obtenir les expressions des termes de variance σ_{ψ}^2 et de centrage $m_{\psi;n}$.

Pour étendre le résultat au cas où G est inconnue, supposons tout d'abord que l'hypothèse (**A**)(i) est vérifiée. D'après (F.2), (F.5) et (K.1), on a,

$$\sup_{\mathbf{x}\in I}\left|\widehat{m}_{\psi;n}(\mathbf{x}) - \widehat{m}^{(1)}_{\psi;n}(\mathbf{x})\right| \leqslant M \sup_{-\infty<t\leqslant\tau}\left|G_n(t) - G(t)\right|, \text{ pour un réel } 0 \leqslant M < \infty.$$

La loi du logarithme itéré de Földes et Rejtő [8] combinée aux hypothèses (H.1-2-3) permet alors d'achever la démonstration du Théorème 3.1. Sous l'hypothèse (**A**)(ii), la preuve s'obtient de façon analogue en remplaçant la loi du logarithme itéré de Földes et Rejtő par celle de Gu et Lai [10] (si $p = 1/2$) ou par le Théorème 2.1 de Chen et Lo [3] (si $0 < p < 1/2$).

Références

[1] D. Blondin, Nonparametric, multidimensional estimation of regression derivatives (Estimation nonparamétrique multidimensionnelle des dérivées de la régression), C. R. Acad. Sci. Paris Ser. I 339 (10) (2004) 713–716.
[2] A. Carbonez, L. Györfi, E.C. van der Meulen, Partitioning-estimates of a regression function under random censoring, Statist. Decisions 13 (1) (1995) 21–37.
[3] K. Chen, S.H. Lo, On the rate of uniform convergence of the Product-Limit estimator: strong and weak laws, Ann. Statist. 25 (3) (1997) 1050–1087.
[4] P. Deheuvels, G. Derzko, Uniform consistency for conditional lifetime distribution estimators under random right-censorship, in: J.-L. Auget, N. Balakrishnan, M. Mesbah, G. Molenberghs (Eds.), Advances in Statistical Methods in the Health Sciences, Applications to Cancer and AIDS Studies, Genome Sequence Analysis and Survival Analysis, Birkhäuser, Boston, ISBN-10: 0-8176-4368-0, 2007, pp. 195–209.
[5] P. Deheuvels, D.M. Mason, General confidence bounds for nonparametric functional estimators, Stat. Inference Stoch. Process. 7 (2004) 225–277.
[6] U. Einmahl, D.M. Mason, An empirical process approach to the uniform consistency of kernel type estimators, J. Theor. Probab. 13 (2000) 1–13.
[7] U. Einmahl, D.M. Mason, Uniform in bandwidth consistency of kernel-type function estimators, Ann. Stat. 33 (3) (2005) 1380–1403.
[8] A. Földes, L. Rejtő, A LIL type result for the product-limit estimator, Z. Wahrsch. Verw. Gebiete 56 (1981) 75–86.
[9] S. Gross, T.L. Lai, Nonparametric estimation and regression analysis with left-truncated and right-censored data, J. Am. Stat. Assoc. 91 (426) (1996) 1166–1180.
[10] M. Gu, T.L. Lai, Functional laws of the iterated logarithm for the product-limit estimator of a distribution function under random censorship or truncation, Ann. Probab. 18 (1990) 160–189.
[11] E.L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, J. Am. Stat. Assoc. 53 (1958) 457–481.
[12] M. Kohler, K. Máthé, M. Pintér, Prediction from randomly right censored data, J. Multivar. Anal. 80 (1) (2002) 73–100.