HAL
archives-ouvertes.fr

Propriétés asymptotiques d'un estimateur du quantile conditionnel pour des données aléatoirement tronquées à gauche

Mohamed Lemdani, Ould-Saïd Elias, Nicolas Poulin

▶ **To cite this version:**

Mohamed Lemdani, Ould-Saïd Elias, Nicolas Poulin. Propriétés asymptotiques d'un estimateur du quantile conditionnel pour des données aléatoirement tronquées à gauche. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. inria-00386648

HAL Id: inria-00386648
https://hal.inria.fr/inria-00386648

Submitted on 22 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Propriétés asymptotiques d'un estimateur du quantile conditionnel pour des données aléatoirement tronquées à gauche

Mohamed Lemdani & Elias Ould-Saïd & Nicolas Poulin

Mohamed Lemdani
Lab. de Biomathématiques,
Univ. de Lille 2. Fac. de Pharmacie
3, rue du Pr. Laguesse, 59006 Lille France
e-mail: Mohamed.Lemdani@univ-lille2.fr

Elias Ould-Saïd
L.M.P.A. J. Liouville
Univ. du Littoral Côte d'Opale
BP 699, 62228 Calais, France
e-mail: ouldsaid@lmpa.univ-littoral.fr

Nicolas Poulin
Centre National de la Recherche Scientifique
Institut Pluridisciplinaire Hubert Curien (UMR 7178)
Département Ecologie, Physiologie et Ethologie
23, rue Becquerel
67087 Strasbourg Cedex 2, France
e-mail: nicolas.poulin@c-strasbourg.fr

Mots clefs : convergence, données de survie, données tronquées, fonction des quantiles, estimateur à noyau, quantile conditionnel, normalité asymptotique, VC-classe.

Key words: Asymptotic normality, Consistency, Kernel estimator, Quantile function, Survival data, Truncated data, VC-class.

Résumé

Considérons une suite de variables aléatoires (v.a) indépendantes $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$ de fonction de répartition (f.d.r) commune continue F inconnue. Ces v.a sont regardées comme les durées de vie des sujets étudiés. La troncature aléatoire à gauche peut notamment avoir lieu si le temps d'origine de la durée de vie étudiée précède le temps d'origine de l'étude. Seuls les sujets dont l'événement se produit après le début de l'étude peuvent être suivis, sinon ces observations sont tronquées. Ce modèle peut survenir dans différents

champs d'application comme l'astronomie et les études médicales.
Soit $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_N$ une suite de v.a indépendantes de f.d.r commune continue G. On suppose aussi que ces variables sont indépendantes des $\mathbf{Y}_i$. Dans le modèle de troncature à gauche, $(\mathbf{Y}_i, \mathbf{T}_i)$ est observé seulement lorsque $\mathbf{Y}_i \geq \mathbf{T}_i$, sinon rien n'est observé.
Soit $(Y_1, T_1), \ldots, (Y_n, T_n)$ l'échantillon observé (i.e $Y_i \geq T_i$) dont la taille n est aléatoire (avec $n \leq N$) et où $(Y_i, T_i)_{1 \leq i \leq n}$ est une sous-suite de $(\mathbf{Y}_i, \mathbf{T}_i)_{1 \leq i \leq N}$. Des estimateurs empiriques basés sur l'échantillon observé ne permettront pas d'estimer les f.d.r F et G mais les f.d.r $F^\star$ et $G^\star$ de $\mathbf{Y}|\mathbf{Y} \geq \mathbf{T}$ et $\mathbf{T}|\mathbf{Y} \geq \mathbf{T}$ (i.e de Y et de T).

Dans le cas i.i.d, Lynden-Bell (1971) introduit les estimateurs produits-limites de F et G suivants

$$F_n(y) = 1 - \prod_{Y_i \leq y} \left(\frac{nC_n(Y_i) - 1}{nC_n(Y_i)} \right) \qquad \text{et} \qquad G_n(y) = \prod_{T_i > y} \left(\frac{nC_n(T_i) - 1}{nC_n(T_i)} \right) \tag{1}$$

où

$$C_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{T_i \,\leq\, y \,\leq\, Y_i\}}.$$

Woodroofe (1985) établit les conditions d'identifiabilité du modèle ainsi que la convergence presque sûre des estimateurs de Lynden-Bell.

La fonction des quantiles de $\mathbf{Y}$ est la fonction inverse de F et est définie, pour tout p fixé dans $(0, 1)$, par

$$F^{-1}(p) = \inf \{y \ : \ F(y) \geq p\}. \tag{2}$$

La fonction des quantiles de l'estimateur de Lynden-Bell de F (resp. G) est un estimateur de la fonction des quantiles de F (resp. G). La convergence uniforme avec vitesse de convergence de l'estimateur de la fonction des quantiles de F est établie par Gűrler et al. (1993). Cet article présente aussi une représentation de type Bahadur pour la fonction des quantiles ainsi que la normalité asymptotique. L'extension pour des séries temporelles a été obtenue par Lemdani et al. (2005).

Nous considérons le problème, lorsque $\mathbf{Y}$ est tronquée par $\mathbf{T}$, de l'estimation de la fonction du quantile conditionnel de $\mathbf{Y}$ étant donné un vecteur de covariables $\mathbf{X}$ de dimension d de f.d.r V et de densité continue v. Soit $F(\cdot, \cdot)$ la f.d.r jointe de $(\mathbf{X}, \mathbf{Y})$, supposée de classe $\mathcal{C}^1\left(\mathbb{R}^{d+1}\right)$. La f.d.r conditionnelle de $\mathbf{Y}$ sachant $\mathbf{X} = \mathbf{x} = (x_1, \ldots, x_d)$ peut être écrite

$$F(\cdot|\mathbf{x}) = \frac{F_1(\mathbf{x}, \cdot)}{v(\mathbf{x})} \tag{3}$$

avec

$$F_1(\mathbf{x}, \cdot) = \frac{\partial F(\mathbf{x}, \cdot)}{\partial \mathbf{x}} := \frac{\partial^d F(\mathbf{x}, \cdot)}{\partial x_1 \cdots \partial x_d}.$$

Un estimateur v_n de la densité v ainsi que la convergence uniforme presque sûre avec vitesse de convergence de v_n sont établis par Ould-Saïd et Lemdani (2006).

Pour tout p fixé dans $(0,1)$, $\xi_p(\mathbf{x})$, le quantile conditionnel de $\mathbf{Y}$ sachant $\mathbf{X}=\mathbf{x}$, est défini par (2).
Il est bien connu que la fonction des quantiles conditionnels donne une bonne description des données (voir, par exemple, Chaudhury et al. (1997))

En utilisant des poids adaptés de type Nadaraya-Watson, nous définissons un estimateur à noyau de la f.d.r conditionnelle :

$$F_n\left(y|x\right)=\frac{\sum_{i=1}^{n}G_n^{-1}\left(Y_i\right)k_d\left(\frac{x-X_i}{h_n}\right)K_0\left(\frac{y-Y_i}{h_n}\right)}{\sum_{i=1}^{n}G_n^{-1}\left(Y_i\right)k_d\left(\frac{x-X_i}{h_n}\right)} \tag{4}$$

où K_0 est une f.d.r lisse, h_n une suite de fenêtres tendant vers 0 et k_d une fonction non-négative sur $\mathbb{R}^d$ qui intervient dans v_n.

La fonction du quantile de l'estimateur $F_n\left(\cdot|\mathbf{x}\right)$ est un estimateur de la fonction du quantile conditionnel de $\mathbf{Y}$ sachant $\mathbf{X}=\mathbf{x}$, noté $\xi_{p,n}(\mathbf{x})$.
Nous obtenons dans Lemdani et al. (2009), pour l'estimateur de la f.d.r conditionnelle, la convergence presque sûre avec taux de convergence. Nous montrons, dans le Théorème 1, que l'estimateur du quantile conditionnel hérite de ces vitesses de convergence. La preuve de ce résultat est donnée dans Lemdani et al. (2009).

Théorème 1 *Considérant des hypothèses classiques de l'estimation à noyau ainsi que sur la densité jointe de* $(\mathbf{X},\mathbf{Y})$*, et pour tout* p *fixé dans* $(0,1)$*, si le quantile conditonnel vérifie :*

$$\forall\varepsilon>0,\ \exists\beta>0,\ \forall\,\eta_p:\Omega\to\mathbb{R},\sup_{\mathbf{x}\in\Omega}|\xi_p(\mathbf{x})-\eta_p(\mathbf{x})|\geq\varepsilon\Rightarrow\sup_{\mathbf{x}\in\Omega}|F(\xi_p(\mathbf{x})|\mathbf{x})-F(\eta_p(\mathbf{x})|\mathbf{x})|\geq\beta.$$

alors

$$\sup_{x\in\Omega}|\xi_{p,n}(\mathbf{x})-\xi_p(\mathbf{x})|=O\left(\max\left\{\sqrt{\frac{\log n}{nh_n^d}},h_n^2\right\}\right)\ \mathbb{P}-p.s.\qquad quand\ n\to\infty. \tag{5}$$

Nous obtenons, dans le Théorème 2 de Lemdani et al. (2009), la normalité asymptotique pour $F_n\left(\cdot|\mathbf{x}\right)$ et celle pour $\xi_{p,n}(\mathbf{x})$:

Théorème 2 *Considérant des hypothèses classiques de l'estimation à noyau ainsi que sur la densité jointe de* $(\mathbf{X}, \mathbf{Y})$*, et pour tout* p *fixé dans* $(0,1)$ *et tout* $\mathbf{x} \in \Omega_0$ *tel que* $f^2\left(\xi_p(\mathbf{x})|\mathbf{x}\right) \neq 0$*,*

$$\sqrt{nh_n}\left(\xi_{p,n}(\mathbf{x}) - \xi_p(\mathbf{x})\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2(\mathbf{x}, \xi_p(\mathbf{x}))}{f^2\left(\xi_p(\mathbf{x})|\mathbf{x}\right)}\right).$$

Dans Lemdani et al. (2009), nous explicitons σ dont un estimateur est facilement calculable. Ceci nous permet notamment de définir des intervalles de confiance pour l'estimateur du quantile conditionnel.
Des simulations illustrent ces résultats pour des échantillons de tailles finies.

Abstract

Consider a sequence of random variables (r.v's) $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$ with common unknown continuous distribution function (d.f) F. The $\mathbf{Y}_i$'s are regarded as the lifetimes of the items under study and are supposed to be subject to left-truncation which may occur if the time origin of the lifetime precedes the time origin of the study. Only subjects that fail after the start of the study are being followed, the others being truncated. This model arises in various fields such as astronomy and medical studies.
Let $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_N$ be a sequence of independent r.v's with common unknown continuous d.f G. The $\mathbf{T}_i$'s are assumed to be independent of the $\mathbf{Y}_i$'s. In the left-truncation model, $(\mathbf{Y}_i, \mathbf{T}_i)$ is observed only when $\mathbf{Y}_i \geq \mathbf{T}_i$ otherwise nothing is observed.
Let $(Y_1, T_1), \ldots, (Y_n, T_n)$ be the actually observed sample (i.e $Y_i \geq T_i$). The size n of the observed sample is random (with $n \leq N$) and $(Y_i, T_i)_{1 \leq i \leq n}$ is a subsequence of $(\mathbf{Y}_i, \mathbf{T}_i)_{1 \leq i \leq N}$. Empirical estimators based on the observed sample do not estimate F and G but rather the d.f $F^\star$ and $G^\star$ related to $\mathbf{Y}|\mathbf{Y} \geq \mathbf{T}$ and $\mathbf{T}|\mathbf{Y} \geq \mathbf{T}$ (i.e to Y and T).

In the i.i.d framework, the product-limit estimators of F and G were obtained by Lynden-Bell (1971) :

$$F_n(y) = 1 - \prod_{Y_i \leq y}\left(\frac{nC_n\left(Y_i\right) - 1}{nC_n\left(Y_i\right)}\right) \qquad \text{and} \qquad G_n(y) = \prod_{T_i > y}\left(\frac{nC_n\left(T_i\right) - 1}{nC_n\left(T_i\right)}\right). \tag{2}$$

where

$$C_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i \,\leq\, y \,\leq\, Y_i\}}.$$

The almost sure consistency of these estimators and the identifiability's conditions were given by Woodroofe (1985).

For a p fixed in $(0,1)$, the p^{th} quantile of $\mathbf{Y}$ is the inverse function of F and is defined by

$$F^{-1}(p) = \inf\{y \ : \ F(y) \geq p\}. \tag{3}$$

The quantile function associated to the Lynden-Bell estimator of F (resp. G) is an estimator of the quantile function of F (resp. G). Uniform consistancy with rates of the quantile function estimator was established by Gűrler et al. (1993). In this paper, they also gave a Bahadur-type representation for the quantile function and asymptotic models. The extension to a time seies case was obtained by Lemdani et al. (2005).

We consider the problem of the classical conditional quantile estimation, where $\mathbf{Y}$ is truncated by a r.v $\mathbf{T}$. Let $\mathbf{X}$ be a random covariate vector taking its values in $\mathbb{R}^d$ with V and continuous density v.
Now consider the joint df $F(\cdot,\cdot)$ of $(\mathbf{X},\mathbf{Y})$ and suppose it is of class $\mathcal{C}^1\left(\mathbb{R}^{d+1}\right)$. Then the conditional df of $\mathbf{Y}$ given $\mathbf{X}=\mathbf{x}=(x_1,\ldots,x_d)$ can be written

$$F(\cdot|\mathbf{x})=\frac{F_1(\mathbf{x},\cdot)}{v\left(\mathbf{x}\right)} \tag{4}$$

with

$$F_1\left(\mathbf{x},\cdot\right)=\frac{\partial F(x,\cdot)}{\partial \mathbf{x}}:=\frac{\partial^d F(\mathbf{x},\cdot)}{\partial x_1\cdots\partial x_d}.$$

Ould-Saïd et Lemdani (2006) gave an estimator v_n of the density v and established the almost-sure uniform consistency of v_n.

For a fixed $p\in(0,1)$ the conditional quantile of $\mathbf{Y}$ given $\mathbf{X}=\mathbf{x}$ is defined by (3) and noted $\xi_p(\mathbf{x})$.
It is well known that the conditional quantile function gives a good description of the data (see, e.g. Chaudhuri et al. (1997)))

By using Nadaraya-Watson-type adapted weights, we set a kernel estimator of the conditional d.f :

$$F_n\left(y|x\right)=\frac{\sum_{i=1}^{n}G_n^{-1}\left(Y_i\right)k_d\left(\frac{x-X_i}{h_n}\right)K_0\left(\frac{y-Y_i}{h_n}\right)}{\sum_{i=1}^{n}G_n^{-1}\left(Y_i\right)k_d\left(\frac{x-X_i}{h_n}\right)} \tag{5}$$

where K_0 is a smooth d.f, h_n a sequence of bandwith that goes to 0 and k_d a non-negative function on $\mathbb{R}^d$ that is used in v_n.

The associated quantile to $F_n\left(\cdot|\mathbf{x}\right)$ is an estimator of the conditional quantile function of $\mathbf{Y}$ given $\mathbf{X}=\mathbf{x}$ noted $\xi_{p,n}(\mathbf{x})$.

We derive in Lemdani et al. (2009) the almost-sure consistency with rates as regards the estimator of the conditional d.f. Then we establish in Theorem 1 that the conditional quantile estimator inherits of these rates :

Theorem 1 *Under classical assumptions of the kernel quantile estimation, and for each fixed* $p \in (0,1)$*, if the conditional quantile satisfies:*

$$\forall \varepsilon > 0,\ \exists \beta > 0,\ \forall\, \eta_p : \Omega \to \mathbb{R}, \sup_{\mathbf{x}\in\Omega} |\xi_p(\mathbf{x}) - \eta_p(\mathbf{x})| \geq \varepsilon \Rightarrow \sup_{\mathbf{x}\in\Omega} |F(\xi_p(\mathbf{x})|\mathbf{x}) - F(\eta_p(\mathbf{x})|\mathbf{x})| \geq \beta.$$

then

$$\sup_{x\in\Omega} |\xi_{p,n}(\mathbf{x}) - \xi_p(\mathbf{x})| = O\left(\max\left\{\sqrt{\frac{\log n}{n h_n^d}}, h_n^2\right\}\right)\ \mathbb{P}-a.s. \qquad as\ n \to \infty. \tag{6}$$

The asymptotic normality for $F_n(\cdot|\mathbf{x})$ and for $\xi_{p,n}(\mathbf{x})$ is obtained in the Theorem 2 of Lemdani et al. (2009).

Theorem 2 *Under classical assumptions of the kernel quantile estimation and on the joint density* $f(\cdot,\cdot)$ *of* $(\mathbf{X},\mathbf{Y})$*. For any* $\mathbf{x} \in \Omega_0$ *such that* $f^2(\xi_p(\mathbf{x})|\mathbf{x}) \neq 0$*,*

$$\sqrt{n h_n}\left(\xi_{p,n}(\mathbf{x}) - \xi_p(\mathbf{x})\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\sigma^2(\mathbf{x}, \xi_p(\mathbf{x}))}{f^2(\xi_p(\mathbf{x})|\mathbf{x})}\right).$$

In Lemdani et al. (2009), the proof of this result is provided as well as an explicit expression of σ that could then be easily estimated. Using this results, we give confidence intervals for the estimator of the conditional quantile function of $\mathbf{Y}$ given $\mathbf{X}$.
Simulations are drawn to illustrate the results for finite samples.

Bibliographie

[1] Chaudhury, P., Doksum, K. et Samarov, A. (1997) On average derivative quantile regression, *Ann. Statist.*, 25, 715–744.
[2] Gürler, U., Stute, W. et Wang, J.L. (1993) Weak and strong quantile representations for randomly truncated data with applications. *Statist. Probab. Lett.*, 17, 139–148.
[3] Lemdani, M., Ould-Saïd, E. et Poulin, N. (2005) Strong representation of the quantile function for left truncated and dependent data. *Math. Meth. Statist.*,14, 332–345.
[4] Lemdani, M., Ould-Saïd, E. et Poulin, N. (2009) Asymptotic properties of a conditional quantile estimator with randomly truncated. *J. Multivariate Anal.*,100, 546–559.
[5] Lynden-Bell, D. (1971) A method of allowing for known observationnal selection in small samples applied to 3CR quasars. *Monthly Notices Roy. Astron. Soc.*, 155, 95–118.
[6] Ould-Saïd, E. and Lemdani, M. (2006) Asymptotic properties of a nonparametric regression function estimator with randomly truncated data. *Ann. Inst. Statist. Math.*, 58, 357–378.
[7] Woodroofe, M. (1985) Estimating a distribution function with truncated data. *Ann. Statist.*, 13, 163-177.