Estimation sur petits domaines de la fonction de répartition d'une variable censurée à l'aide de quantiles conditionnels

Sandrine Casanova, Eve Leconte

Estimation sur petits domaines de la fonction de répartition d'une variable censurée à l'aide de quantiles conditionnels

Sandrine Casanova & Eve Leconte

TSE (GREMAQ), Université des Sciences Sociales,
21, allée de Brienne, 31000 TOULOUSE, France

E-mail : sandrine.casanova@univ-tlse1.fr, leconte@cict.fr

Résumé

L'estimation de la fonction de répartition (f.d.r.) en sondage est d'un grand intérêt pour déduire des estimateurs de paramètres classiques tels que la moyenne ou la médiane sur la population mais aussi sur une sous-population (domaine). Si un domaine est de taille suffisante, l'estimation des paramètres d'intérêt est basée sur les données relatives aux individus du domaine et les estimateurs produits sont de précision acceptable. Cependant, dans la plupart des applications, les tailles d'échantillons correspondant à des petits domaines ne sont pas suffisantes. L'estimation se fonde alors sur une information auxiliaire fournie par une covariable et de l'information est "empruntée" aux autres domaines. Dans ce contexte, nous proposons de nouveaux estimateurs non paramétriques de la f.d.r. sur un domaine dans le cas où la variable d'intérêt est censurée à droite. Pour cela, adaptant au cas censuré la technique de Casanova (2009), nous estimons les ordres quantiles des individus de l'ensemble des échantillons à l'aide de l'estimateur de Kaplan-Meier généralisé lissé (Leconte *al.*, 2002). Chaque domaine peut donc être décrit par les ordres quantiles des individus échantillonnés de ce domaine. Nous prédisons alors non paramétriquement la variable d'intérêt d'un individu hors échantillon par les quantiles conditionnels associés aux ordres qui décrivent son domaine. Ces prédictions nous permettent d'obtenir un nouvel estimateur de la f.d.r. du domaine. Des simulations comparent cette méthode avec l'estimateur de Kaplan-Meier calculé sur les points échantillonnés du domaine. Un exemple d'application à des données de durées de chômage illustre la méthode.

Abstract

In survey analysis, the estimation of the cumulative distribution function (c.d.f.) is of great interest in order to derive mean or median estimators of the population or for subpopulations (domains). When the size of the domain is big enough, the estimation is based on individuals of the domain and the resulting estimators are sufficiently precise. However, in most applications, sizes of the domain samples are not sufficient. In this case, the

estimation uses auxiliary information of a covariate and some information is "borrowed" from the others domains. In this framework, we propose nonparametrics estimators of the c.d.f. of a domain when the interest variable is right censored. The new estimators are adaptations to the censored case of techniques proposed in Casanova (2009) : quantiles orders of the individuals of the whole sample are estimated using a smoothed generalized Kaplan-Meier estimator (Leconte *al.*, 2002). Each domain can be described by quantiles orders of the sampled individuals of the domain. The interest variable of a non-sampled individual is then non parametrically predicted usind conditional quantiles associated to the orders describing its domain. This predictions allow us to derive new estimators of the c.d.f. of the domain. The obtained estimators are compared by simulations with the Kaplan-Meier estimator computed with sampled individuals of the domain. The method is illustrated with unemployment duration data.

Mots-clés : sondages, fonction de répartition, information auxiliaire, domaine, données censurées, Kaplan-Meier, quantiles conditionnels, estimation non paramétrique.

1 Introduction

Soit une population U partitionnée en m sous-populations ou domaines U_i de taille N_i, $i = 1, \cdots, m$. Soient s un échantillon de U de taille n et $s_i = s \cap U_i$ un échantillon du domaine U_i de taille n_i. Dans le cadre des sondages, la fonction de répartition (f.d.r.) d'une variable d'intérêt T sur le domaine U_i s'écrit $F_i(t) = \frac{1}{N_i} \sum_{j \in U_i} \mathbb{1}(t_{ij} \leq t)$ que l'on peut décomposer en

$$F_i(t) = \frac{1}{N_i}(\sum_{j \in s_i} \mathbb{1}(t_{ij} \leq t) + \sum_{j \in U_i \setminus s_i} \mathbb{1}(t_{ij} \leq t)) \quad (1)$$

où t_{ij} est la variable d'intérêt mesurée pour le j-ième individu du domaine U_i. On suppose que t_{ij} est seulement connu sur s_i et éventuellement censuré à droite par c_{ij}. Avec les notations d'Efron, nous observons, sur l'échantillon s_i, $y_{ij} = min(t_{ij}, c_{ij})$ et $\delta_{ij} = \mathbb{1}(t_{ij} < c_{ij})$.

L'estimateur classique en sondage pour estimer la f.d.r. sur le domaine U_i dans le cas non censuré est l'estimateur de Horvitz-Thompson qui correspond ici à la f.d.r. empirique calculée sur les points échantillonnés du domaine. Il s'agit du premier terme dans la décomposition de la f.d.r. ci-dessus. Cet estimateur est de précision acceptable si la taille de l'échantillon du domaine est suffisante. Dans le cas censuré, la généralisation de la f.d.r. empirique est l'estimateur de Kaplan-Meier (1958).

Si la taille d'échantillon est trop faible, ce qui est le cas pour de petits domaines, on peut améliorer l'estimation à l'aide d'une information auxiliaire apportée par une covariable et "emprunter de la force aux voisins". Ceci nous permet d'estimer le deuxième terme dans la décomposition de la f.d.r. en estimant les t_{ij} des individus non échantillonnés.

Dans le cas non censuré, un modèle classique utilisant l'information auxiliaire est le modèle mixte (voir Rao, 2003). Une alternative au modèle mixte est l'utilisation des quantiles et des M-quantiles conditionnels pour prédire les t_{ij} (voir Chambers et Tzavidis (2006) dans le cadre paramétrique et Casanova (2009) dans le cadre non-paramétrique).

Les estimateurs que nous proposons dans la section 3 sont des adaptations au cas censuré de Casanova (2009) dont les méthodes seront rappelées à la section 2. La section 4 présentera une étude de simulations et un exemple sera évoqué en section 5.

2 Estimation non paramétrique de la f.d.r. sur un petit domaine dans le cas non censuré

Nous allons décrire la méthode proposée par Casanova (2009). Elle se fait en deux étapes :
Etape 1 : estimation non paramétrique des ordres quantiles des points échantillonnés de l'ensemble des domaines

Pour une observation (y_k, x_k), il existe un ordre $q \in [0,1]$ tel que y_k est le quantile conditionnel à x_k d'ordre q. Cet ordre, appelé ordre-quantile conditionnel, classe l'observation dans l'échantillon. Les ordres conditionnels des observations d'un domaine le situent par rapport à l'ensemble de tous les domaines. Une estimation naturelle de l'ordre quantile conditionnel peut se faire à l'aide de l'estimateur de Nadaraya-Watson de la f.d.r. conditionnelle :

$q_k(y_k, x_k) = \dfrac{\sum_{l=1}^{n} \mathbb{1}(y_l \leq y_k) K\left(\dfrac{x_k - x_l}{h}\right)}{\sum_{l=1}^{n} K\left(\dfrac{x_k - x_l}{h}\right)}$, où K est un noyau de densité et h une fenêtre appropriée.

Chaque domaine U_i peut donc être décrit par l'ensemble des ordres quantiles estimés sur le sous-échantillon s_i. Nous noterons ces ordres $\{q_{ik},\ k = 1, \dots n_i\}$.

Etape 2 : estimations de la variable d'intérêt pour les points non échantillonnés du domaine

Pour chaque individu j de $U_i \setminus s_i$ de covariable x_{ij}, il y a donc n_i estimations possibles $m(q_{ik}, x_{ij})$ de t_{ij} à l'aide d'estimateurs des quantiles conditionnels à x_{ij} d'ordre q_{ik}. Casanova (2009) propose d'utiliser les polynômes locaux pour estimer ces quantiles conditionnels. Un estimateur de la f.d.r. sur le domaine U_i s'en déduit donc d'après la formule (1) :

$$F_C^i(t) = \frac{1}{N_i}\left(\sum_{j \in s_i} \mathbb{1}(t_{ij} \leq t) + \sum_{j \in U_i \setminus s_i}\left(\frac{1}{n_i}\sum_{k \in s_i} \mathbb{1}(m(q_{ik}, x_{ij}) \leq t)\right)\right)$$

Cette estimation utilise l'échantillon s de la population totale et de la "force" est donc empruntée à tous les domaines.

De plus, bien que les distributions des ordres quantile ne dépendent pas de x dans la population, la distribution de ces ordres peut dépendre de x dans un domaine. Par conséquent, il peut être judicieux pour améliorer l'estimation d'utiliser un noyau afin de donner plus de poids aux prédictions dont la covariable est proche de x_{ij}. Cela conduit à l'estimateur suivant

$$F_{CK}^{i}(t) = \frac{1}{N_i}\left(\sum_{j\in s_i} \mathbb{1}(t_{ij}\leq t) + \sum_{j\in U_i\setminus s_i} \frac{\sum\limits_{k\in s_i} K\left(\frac{x_{ik}-x_{ij}}{h}\right)\mathbb{1}(m(q_{ik},x_{ij})\leq t)}{\sum\limits_{k\in s_i} K\left(\frac{x_{ik}-x_{ij}}{h}\right)}\right)$$

où h est une fenêtre appropriée.

3 Estimation de la f.d.r. d'un petit domaine dans le cas censuré

Contrairement au cas non censuré, le premier terme de (1) n'est plus connu en raison de la censure à droite et doit être estimé. En remarquant qu'il peut s'écrire :

$$\frac{1}{N_i}\sum_{j\in s_i}\mathbb{1}(t_{ij}\leq t) = \frac{n_i}{N_i}\left(\frac{1}{n_i}\sum_{j\in s_i}\mathbb{1}(t_{ij}\leq t)\right),$$

on reconnait dans le terme entre parenthèses la f.d.r. sur l'échantillon s_i. Ce terme peut donc être estimé dans le cas censuré par l'estimateur de Kaplan-Meier sur l'échantillon s_i.

Pour ce qui est du second terme, nous adaptons Casanova (2009) au cas censuré.

Etape 1 : estimation des ordres quantiles q_k de s

La f.d.r. conditionnelle dans le cas censuré peut être estimée par l'estimateur de Kaplan-Meier généralisé (Dabrowska, 1992) :

$$F_{\text{GKM}}(t\mid x) = \begin{cases} 1-\prod\limits_{j=1}^{n}\left\{1-\dfrac{B_j(x)}{\sum\limits_{r=1}^{n}\mathbb{1}_{(y_r\geq y_j)}B_r(x)}\right\}^{\mathbb{1}(y_j\leq t,\,\delta_j=1)} & \text{si } \; t<y_{(n)} \\ 1 & \text{sinon,}\end{cases}$$

où les $B_j(x)$ sont des poids de type Nadaraya-Watson.

Nous proposons d'en utiliser la version lissée en t de Leconte *et al* (2002) :

$$F_{\text{SGKM}}(t\mid x) = \sum_{j=1}^{d}\left(F_{\text{GKM}}\left(y_{(j)}^{\dagger}\mid x\right) - F_{\text{GKM}}\left(y_{(j-1)}^{\dagger}\mid x\right)\right)H\left(\frac{t-y_{(j)}^{\dagger}}{h_T}\right)$$

où les $y^{\dagger}_{(j)}$ sont les observations non censurées ordonnées ($y^{\dagger}_{(d)} = y_{(n)}$), H est un noyau intégré et h_T est une fenêtre adéquate. On en déduit un estimateur de l'ordre quantile conditionnel : $q_k(y_k, x_k) = F_{\text{SGKM}}(y_k \mid x_k)$.

Comme l'information fournie par les observations censurées manque de précision, chaque domaine U_i sera décrit par l'ensemble des ordres quantiles estimés sur le sous-échantillon de s_i correspondant aux observations non censurées : $\{q_{ik},\ k = 1, \dots n_i^{\dagger}\}$.

Etape 2

Nous allons maintenant calculer les $n_i^{\dagger}$ estimations de la variable d'intérêt pour chaque individu non échantillonné j du domaine i : $m(q_{ik}, x_{ij})$. Ces estimations sont solutions de $F_{\text{SGKM}}(m \mid x_{ij}) = q_{ik}$. Ils s'obtiennent par inversion de F_{SGKM}.

On en déduit l'estimateur suivant de la f.d.r. du domaine U_i :

$$F^i_{CL}(t) = \frac{1}{N_i}\left(n_i F^i_{\text{KM}}(t) + \sum_{j \in U_i \setminus s_i}\left(\frac{1}{n_i^{\dagger}}\sum_{k \in s_i^{\dagger}} \mathbb{1}\!\mathrm{I}(m(q_{ik}, x_{ij}) \leq t)\right)\right)$$

Une version à noyau est également possible comme dans le cas non censuré (nous la noterons F_{CLK}).

4 Simulations

Des populations de taille 300 (150 individus par domaine) ont été générées selon 2 modèles différents :

- Modèle 1 :

  domaine U_1 : $t_{1j} = max(0, 1 + 2 * (1 + 4 * x_{1j}) + 5 * x_{1j} * u_j)$
  domaine U_2 : $t_{2j} = max(0, 1 + 5 * (1 + 4 * x_{2j}) + 1.5 * u_j)$
  où les covariables x suivent des lois uniformes sur $[0, 1]$ et le terme d'erreur u suit la loi normale centrée réduite.

- Modèle 2 :

  domaine U_1 : $t_{1j} = max(0, 1 + 2 * (1 + 4 * x_{1j}) + 1.5 * x_{1j} * u_j)$
  domaine U_2 : $t_{2j} = max(0, 1 + 30 * x_{2j} - 30 * x_{2j}^2 + 1.5 * u_j)$
  avec les mêmes lois pour x et u que dans le modèle 1.

t_{ij} est censuré par c_{ij} où $c \sim U(0, dp)$, dp étant choisi de façon à obtenir 0%, 10%, 20% ou 30% de censure.

1000 échantillons de taille 15 ont été tirés dans chaque domaine (1/10 du domaine).

Les résultats des MASE (Mean Averaged Square Error) sont dans le tableau 1. On constate que les nouveaux estimateurs ont des MASE plus faibles que l'estimateur de Kaplan-Meier, quels que soient le modèle et le taux de censure.

		Modèle 1			Modèle 2		
τ	Domaine	KM	CL	CLK	KM	CL	CLK
0 %	U1	0.0063	0.0057	0.0043	0.0084	0.0064	0.0023
	U2	0.0083	0.0040	0.0018	0.0079	0.0056	0.0042
10 %	U1	0.0071	0.0061	0.0047	0.0068	0.0053	0.0021
	U2	0.0100	0.0043	0.0022	0.0087	0.0055	0.0047
20 %	U1	0.0086	0.0080	0.0059	0.0081	0.0066	0.0027
	U2	0.0098	0.0053	0.0026	0.0106	0.0061	0.0056
30 %	U1	0.0087	0.0066	0.0065	0.0098	0.0060	0.0032
	U2	0.0134	0.0076	0.0034	0.0103	0.0065	0.0060

Table 1: MASE obtenus pour les trois estimateurs. KM est l'estimateur de Kaplan-Meier sur l'échantillon, CL est le nouvel estimateur et CLK est la version avec noyau ; τ désigne le taux de censure.

5 Exemple

Un exemple d'application à des durées de chomage censurées à droite sera présenté, l'information auxiliaire étant apportée par le niveau d'études. Les domaines correspondent à deux départements français : la Seine et la Haute-Garonne.

6 Bibliographie

[1] Aragon Y., Casanova S. Chambers R. et Leconte E. (2005). Conditional ordering using nonparametric expectiles. *Journal of official Statistics*, 21, 617–633.
[2] Casanova S. (2009). Using M-quantiles to estimate a cumulative distribution function in a domain. Soumis aux *Annales d'Economie et de Statistique*.
[3] Leconte, E, Poiraud-Casanova, S. et Thomas-Agnan, C. (2002). Smooth conditional distribution function and quantiles under random censorship. *Lifetime Data Analysis*, 8, 229–246.
[4] Chambers R.L. et Dunstan R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597–804.
[5] Chambers R.L. et Tzavidis N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255–268.
[6] Rao J.N.K. (2003). *Small area estimation.* Wiley, New-York.