Estimation non paramétrique de la fonction de répartition d'une variable censurée à droite sur petits domaines

Sandrine Casanova [1], Hélène Couprie [2] & Eve Leconte [3]

[1] *TSE, Université Toulouse 1 Capitole,*
21, allée de Brienne, 31042 TOULOUSE, France, sandrine.casanova@tse-fr.eu
[2] *CEREQ (DEEVA)*
10, place de la Joliette, 13002 Marseille, France, helene.couprie@gmail.com
[3] *TSE, Université Toulouse 1 Capitole,*
21, allée de Brienne, 31042 TOULOUSE, France, eve.leconte@tse-fr.eu

Résumé.
L'estimation de la fonction de répartition (fdr) en population finie est très utile pour déduire des estimateurs de paramètres complexes tels que les quantiles. Nous considérons le cas où la variable d'intérêt est censurée à droite. Dans ce contexte, Casanova et Leconte (2015) ont proposé un nouvel estimateur *model-based* non paramétrique de la fdr sur la population. Nous nous intéressons ici à l'estimation de la fdr sur petits domaines. Quand le domaine est de taille suffisante, les estimateurs basés sur les données du domaine sont de précision acceptable. Quand le domaine est de taille plus petite, de l'information doit alors être "empruntée" aux autres domaines pour améliorer la précision. Dans ce contexte, en utilisant de l'information auxiliaire fournie par une covariable, nous adaptons au cas censuré la technique de Casanova (2012) : la partie de la fdr associée aux individus hors échantillon est prédite à l'aide d'un quantile conditionnel dont l'ordre doit être préalablement estimé. Les performances du nouvel estimateur seront comparées par simulation à celles de l'estimateur de Kaplan-Meier et à celui de Casanova et Leconte (2015) calculés à partir des valeurs échantillonnés du domaine. La méthodologie sera illustrée sur des données du CEREQ concernant les temps d'accès au premier emploi de jeunes diplômées, les domaines correspondant ici aux différentes formations.

Mots-clés. Fonction de répartition, information auxiliaire, données censurées, quantile conditionnel.

Abstract. In survey analysis, the estimation of the cumulative distribution function (cdf) is of great interest in order to derive mean or median estimators for the population or for sub-populations (domains). We consider the case where the response variable is right censored. In this framework, nonparametric model-based estimators of the cdf in a finite population have been proposed by Casanova and Leconte (2015). We consider now the case of small domains. The new estimator uses auxiliary information brought by a continuous covariate and is based on nonparametric quantile regression adapted to the censored case. The obtained estimator has been compared by simulations with

the Kaplan-Meier estimator and the Casanova and Leconte (2015) estimator computed with the sampled individuals. Data of the CEREQ concerning qualified girls are used to illustrate the new methodology : the duration of interest is the time required to obtain the first job.

Keywords. Cumulative distribution function, auxiliary information, right censored data, conditional quantile.

1 Introduction

L'estimation de la fonction de répartition (fdr) en population finie est très utile pour déduire des estimateurs de paramètres complexes tels que les quantiles. Nous considérons le cas où la variable d'intérêt est censurée à droite. Cela se produit lorsqu'on étudie une variable de durée que l'on observe durant une période de temps limitée. Par exemple, si l'on considère des durées de chomâge, les individus qui n'auront pas retrouvé d'emploi à la fin de l'étude verront leurs durées de chômage censurées. A notre connaissance, dans le cadre des sondages, l'estimation de la fdr d'une variable censurée a seulement été étudié par Casanova et Leconte (2015) dans une population finie.

Nous nous intéressons ici à l'estimation de la fdr dans des sous-populations (domaines) qui peuvent être de petite taille en nous restreignant au cadre d'estimateurs *model-based*. Dans le cadre paramétrique, sans censure et sans domaines, Chambers et Dunstan (1986) proposent d'améliorer l'estimation de la fdr en prédisant les valeurs de la variable d'intérêt pour les individus non échantillonnés en utilisant l'information auxiliaire apportée par une covariable. Dorfman et Hall (1993) ont défini des versions non paramétriques des estimateurs de Chambers et Dunstan et en ont étudié les propriétés asymptotiques. Dans le cadre de l'estimation de la fdr sur un domaine, si ce domaine est de taille suffisante, l'estimation des paramètres d'intérêt peut se suffire des données relatives aux individus échantillonnés du domaine pour produire des estimateurs de précision acceptable. Cependant, dans la plupart des applications, les tailles d'échantillons correspondant à des petits domaines ne sont pas suffisantes. De l'information doit alors être "empruntée" aux autres domaines pour améliorer la précision. La technique de référence en petits domaines est le modèle linéaire mixte à effets aléatoires (Rao, 2003). Chambers et Tzavidis (2006) ont proposé une alternative en prédisant la variable d'intérêt des individus non échantillonnés à l'aide de M-quantiles conditionnels paramétriques. Casanova (2012) a étendu cette dernière technique au cadre non paramétrique.

Dans la section 2, nous proposons un nouvel estimateur de la fdr en généralisant la méthode de Casanova (2012) pour petits domaines au cas censuré. Cet estimateur est basé sur l'estimation de quantiles conditionnels. La section 3 compare par simulation les performances du nouvel estimateur à l'estimateur naïf de Kaplan-Meier et à celui de Casanova et Leconte (2015) calculés à partir des individus échantillonnés du domaine. La section 4 illustre ces méthodes sur des données du CEREQ concernant le temps d'accès

au premier emploi de jeunes diplômées.

2 Estimation non paramétrique de la fdr sur un petit domaine en présence de censure

2.1 Notations

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. 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 $\delta_{ij} = \mathbb{1}(t_{ij} < c_{ij})$. Nous disposons d'une information auxilaire mesurée par une covariable continue x_{ij} pour le j-ième individu du domaine U_i, connue sur tout U_i.

Dans le cadre des sondages, la fonction de répartition de la variable d'intérêt T sur le domaine U_i s'écrit $F_i(t) = \frac{1}{N_i} \sum_{j \in U_i} \mathbb{1}(t_j \leq t)$ que l'on peut décomposer en

$$F_i(t) = \frac{1}{N_i}(\sum_{j \in s_i} \mathbb{1}(t_j \leq t) + \sum_{j \in U_i \setminus s_i} \mathbb{1}(t_j \leq t)). \tag{1}$$

2.2 Un estimateur naïf de la fdr

La fonction de répartition empirique calculée sur les individus échantillonnés du domaine ne fournit pas un estimateur convergent en présence de censure. Par contre, un estimateur adapté qui généralise la fdr empirique au cas censuré est l'estimateur de Kaplan-Meier (Kaplan et Meier, 1958).

Dans sa version originale, l'estimateur de Kaplan-Meier est indéterminé après le dernier temps observé si celui-ci est censuré. Afin d'obtenir une fonction de répartition, nous préférons donc utiliser la version d'Efron (1967) qui vaut 1 après le dernier temps observé $y_{(n)}$ du domaine U_i :

$$\hat{F}^i_{\mathrm{KM}}(t) = \begin{cases} 1 - \prod_{j \in s_i} \left\{ 1 - \frac{1}{\sum_{r \in s_i} \mathbb{1}(y_r \geq y_j)} \right\}^{\mathbb{1}(y_j \leq t, \delta_j = 1)} & \text{if } \ t < y_{(n)} \\ 1 & \text{sinon.} \end{cases} \tag{2}$$

2.3 Le nouvel estimateur

Nous proposons un estimateur *model-based* de la fdr sur un domaine en estimant les deux termes de (1). 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 :

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

on reconnaît dans le terme entre parenthèses la fdr 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 (cf. section précédente).

Pour ce qui est du second terme, nous adaptons Casanova (2012) au cas censuré. En nous inspirant de Chambers et Tzavidis (2006), nous supposons le modèle de superpopulation ξ suivant :

$$t_{ij}=m(q_i,x_{ij})+\varepsilon_{ij},\quad i=1,\ldots,m,\quad j=1,\ldots,N_i,$$

où les ε_{ij} sont des variables i.i.d. de fdr G_i, q_i est un coefficient dans $(0,1)$ caractérisant la position du domaine U_i par rapport aux autres domaines et $m(q_i,x_{ij})$ est le quantile conditionnel d'ordre q_i de T sachant $X=x_{ij}$. D'autre part, chaque valeur t_{ij} peut être considérée comme le quantile conditionnel de T sachant $X=x_{ij}$ pour un ordre qu'on notera $q(t_{ij},x_{ij})$. Il paraît donc naturel d'estimer l'ordre q_i du domaine U_i par un résumé numérique des estimations des ordres quantiles conditionnels $q(t_{ij},x_{ij})$ des individus échantillonnés j du domaine U_i.

Etape 1 : estimation non paramétrique de l'ordre du domaine U_i

Les ordres quantiles conditionnels $q(t_{ij},x_{ij})$ des individus échantillonnés sont estimés à l'aide de l'estimateur de Kaplan-Meier généralisé (Beran, 1981) de la fdr conditionnelle de T sachant $X=x$, calculé sur tout l'échantillon s :

$$\hat{F}_{\mathrm{GKM}}\left(t\mid x\right)=\begin{cases}1-\displaystyle\prod_{j\in s}\left\{1-\frac{B_j\left(x\right)}{\displaystyle\sum_{r\in s}B_r\left(x\right)\mathbb{1}(y_r\geq y_j)}\right\}^{\mathbb{1}\left(y_j\leq t,\,\delta_j=1\right)} & \text{if } \ t<y_{(n)}\\ 1 & \text{sinon,}\end{cases}\qquad(3)$$

où les $B_j(x)$ sont les poids de Nadaraya-Watson définis par :

$$B_j\left(x\right)=\frac{K\left(\dfrac{x-X_j}{h_X}\right)}{\displaystyle\sum_{k\in s}K\left(\frac{x-X_k}{h_X}\right)}.$$

K est un noyau et h_X une fenêtre adéquate.

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

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

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. Nous avons donc :

$$\hat{q}(t_{ij}, x_{ij}) = \hat{F}_{\text{SGKM}}(y_{ij} \mid x_{ij}).$$

Du fait de la censure des t_{ij}, les ordres associés aux individus censurés sont également censurés à droite. Pour en tenir compte, le coefficient q_i du domaine U_i sera estimé par la médiane des ordres quantiles conditionnels estimés des individus échantillonnés du domaine, obtenue par inversion de l'estimateur de Kaplan-Meier calculé sur ces ordres. Nous la noterons $\hat{q}_i$.

Etape 2 : prédiction pour les points non échantillonnés

Comme $\mathbb{E}_{\xi}\left(\mathbb{1}(t_{ij} \leq t)\right) = P(t_{ij} \leq t) = G_i(t - m(q_i, x_{ij}))$, l'indicatrice $\mathbb{1}(t_{ij} \leq t)$ peut être prédite en estimant $G_i(t - m(q_i, x_{ij}))$. Un estimateur naturel $\hat{m}(\hat{q}_i, x_{ij})$ de $m(q_i, x_{ij})$ est le quantile conditionnel sachant x_{ij} d'ordre $\hat{q}_i$ solution en θ de $\hat{F}_{\text{SGKM}}\left(\theta \mid x_{ij}\right) = \hat{q}_i$. Il est obtenu par inversion de $\hat{F}_{\text{SGKM}}$. On peut remarquer que, comme pour l'estimation de l'ordre quantile q_i, tout l'échantillon s est utilisé pour le calcul de cet estimateur, ce qui permet "d'emprunter de la force" aux autres domaines.

Les résidus $\hat{\varepsilon}_{ij} = y_{ij} - \hat{m}(\hat{q}_i, x_{ij})$ étant censurés à droite comme le sont les y_{ij}, nous estimons la fdr G_i des erreurs par l'estimateur de Kaplan-Meier appliqué aux résidus $\hat{\varepsilon}_{ij}$ de s_i, que nous noterons $\hat{G}^i_{\text{KM}}$.

On en déduit l'estimateur suivant de la fdr du domaine U_i :

$$\hat{F}^i_{\text{Q}}(t) = \frac{1}{N_i}\left(n_i \hat{F}^i_{\text{KM}}(t) + \sum_{j \in U_i \setminus s_i} \hat{G}^i_{\text{KM}}(t - \hat{m}(\hat{q}_i, x_j))\right)$$

3 Simulations *model-based*

Pour comparer les performances du nouvel estimateur à celles de l'estimateur naïf, des simulations ont été réalisées. Nous avons également implémenté l'estimateur $\hat{F}^i_{\text{M}}$ de Casanova et Leconte (2015) qui n'emprunte pas de force aux autres domaines et prédit les valeurs t_{ij} des points non échantillonnés par la médiane conditionnelle $\hat{m}(x_{ij}) = \hat{F}^i_{\text{SGKM}}(y_{ij} \mid x_{ij})$, où $\hat{F}^i_{\text{SGKM}}$ désigne l'estimateur de Kaplan-Meier généralisé lissé calculé sur s_i.

Nous avons généré des populations de grande taille avec 30 domaines de tailles différentes (générées selon les valeurs entières d'une loi uniforme sur $[50, 150]$ et maintenues fixées pour toutes les itérations), selon le modèle de survie accéléré suivant : $ln(t_{ij}) = 4 - \nu x_{ij} + u_i + \varepsilon_{ij}$ où les x_{ij} suivent des lois uniformes sur $[1, 4]$. ν est un paramètre réel qui mesure l'effet de la covariable sur la durée ; u_i correspond à un effet aléatoire du domaine i et suit une loi normale centrée de variance σ^2. Plus σ^2 est élevé, plus les domaines sont différents. Les erreurs ε_{ij} suivent une loi de valeur extrême de façon à

obtenir une loi de Weibull pour les t_{ij}. Les délais de censure c_{ij} sont générés selon une loi uniforme sur $[0, c]$, le paramètre c permettant de régler le taux de censure (10 %, 25 % ou 50 %). Pour chaque itération, une nouvelle population est générée, dans laquelle nous tirons pour chaque domaine un échantillon avec un taux de sondage égal à 1/10 ou à 1/20. Les MASE (Mean Averaged Square Error) des trois estimateurs sont comparés pour les différents taux de censure et différentes valeurs de ν et σ^2.

4 Exemple d'application

Nous considérons comme illustration des données du CEREQ concernant les temps d'accès à l'emploi de jeunes diplômées, temps qui sont censurés pour les jeunes filles n'ayant pas trouvé d'emploi à la fin de l'étude. Les 39 domaines correspondent aux formations diplômantes et les tailles des échantillons des domaines varient de 1 à 39. La variable auxiliaire utilisée est le taux de chômage local de la commune où réside la jeune fille.

Bibliographie

[1] Casanova, S. (2012), Using M-quantiles to estimate a cumulative distribution function in a domain, *Annales d'Economie et de Statistique*, 107-108, 287–297.
[2] Casanova, S. et Leconte, E. (2015), A nonparametric model-based estimator for the cumulative distribution function of a right censored variable in a finite population, *Journal of Surveys: Statistics and Methodology*, 3, 317–338.
[3] Chambers, R.L. et Dunstan, R. (1986), Estimating distribution functions from survey data, *Biometrika*, 73, 597–804.
[4] Chambers, R.L. et Tzavidis, N. (2006), M-quantile models for small area estimation, *Biometrika*, 93, 255–268.
[5] Dabrowska, D.M. (1992), Nonparametric quantile regression with censored data, *Sankhya Journal*, 54, 252–259.
[6] Dorfman, A.H. et Hall, P. (1993), Estimators of the finite population distribution function using nonparametric regression, *Annals of Statistics*, 21, 1452–1475.
[7] Efron, B. (1967), The two sample problem with censored data, *Proc 5th Berkeley Symp*, 4, 831–853.
[8] Kaplan, E. et Meier, P. (1958), Nonparametric estimation for incomplete observation, *Journal of the American Statistical Association*, 53, 457–481.
[9] 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.
[10] Rao, J.N.K. (2003), *Small area estimation*, Wiley, New-York.