Estimation non paramétrique de la fonction de répartition d'une variable censurée à droite en population finie

Sandrine Casanova [1] & Eve Leconte [2]

[1] *TSE (GREMAQ), Université Toulouse 1 Capitole,*
21, allée de Brienne, 31042 TOULOUSE, France, sandrine.casanova@tse-fr.eu
[2] *TSE (GREMAQ), 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 ici le cas où la variable d'intérêt est censurée à droite. Dans ce contexte, nous proposons un nouvel estimateur *model-based* non paramétrique de la fdr sur la population. 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 de la médiane conditionnelle. Pour obtenir des estimations du biais et de la variance de l'erreur de prédiction du nouvel estimateur, nous généralisons au cas censuré la méthode de rééchantillonnage par bootstrap proposée par Lombardia *et al.* (2004). Le problème du traitement de la non-réponse est abordé. Des simulations comparent les performances du nouvel estimateur à celles de l'estimateur de Kaplan-Meier calculé à partir des points échantillonnés. Un exemple d'application à des données de durées de bien-être est présenté et des simulations *design-based* sont effectuées pour différents plans de sondage.

Mots-clés. Fonction de répartition, information auxiliaire, données censurées, estimateur de Kaplan-Meier généralisé, estimation bootstrap.

1 Introduction

En sondage, la littérature étudie principalement l'estimation de totaux ou de moyennes mais dans beaucoup d'applications, les paramètres d'intérêt sont plus complexes. Par exemple, l'estimation de la fonction de répartition (fdr) en population finie est très utile pour déduire des estimateurs tels que les quantiles. Nous considérons en plus ici 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 n'a jamais été étudiée.

Nous nous plaçons dans l'approche *model-based*. Dans le cadre paramétrique sans censure, 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 la section 2, nous proposons un nouvel estimateur de la fdr en généralisant l'estimateur de Dorfman et Hall (1993) au cas censuré. La section 3 propose une estimation bootstrap du biais et de la variance de l'erreur de prédiction des estimateurs. Dans la section 4, nous abordons le problème du traitement de la non-réponse. La section 5 compare par simulation les performances du nouvel estimateur (et de ses versions adaptées à la non-réponse) à l'estimateur naïf de Kaplan-Meier sur l'échantillon. Un exemple d'application à des données de bien-être est présenté en section 6 et les performances des estimateurs sont comparées pour différents plans de sondage par des simulations *design-based*.

2 Estimation non paramétrique de la fdr en présence de censure

2.1 Notations

Soit une population $\mathcal{P}$ de taille N et soit s un échantillon de $\mathcal{P}$ de taille n. Nous nous intéressons à la fdr d'une durée T (variable positive). t_j est la valeur de T pour l'individu j de la population. T est seulement connu pour les individus appartenant à s et éventuellement censuré à droite par C, une variable délai de censure indépendante de T. Avec les notations d'Efron, nous observons, sur l'échantillon s, $y_j = min(t_j, c_j)$ et $\delta_j = \mathbb{1}(t_j < c_j)$. Nous disposons d'une information auxiliaire mesurée par une covariable continue X qui vaut x_j pour l'individu j de la population et est connue sur toute la population.

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

$$F(t) = \frac{1}{N}(\sum_{j\in s} \mathbb{1}(t_j \leq t) + \sum_{j\in\mathcal{P}\setminus s} \mathbb{1}(t_j \leq t)). \quad (1)$$

2.2 Un estimateur naïf de la fdr

La fonction de répartition empirique calculée à partir des points 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 $y_{(n)}$:

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

2.3 Le nouvel estimateur

Nous proposons un estimateur *model-based* de la fdr 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 :

$$\frac{1}{N} \sum_{j \in s} \mathbb{1}(t_j \leq t) = \frac{n}{N} \left(\frac{1}{n} \sum_{j \in s} \mathbb{1}(t_j \leq t) \right),$$

on reconnaît dans le terme entre parenthèses la fdr sur l'échantillon s. Ce terme peut donc être estimé dans le cas censuré par l'estimateur de Kaplan-Meier sur l'échantillon s (cf. section précédente).

Pour ce qui est du second terme, nous adaptons Dorfman et Hall (1993) au cas censuré. Pour cela, nous supposons le modèle de superpopulation ξ suivant :

$$t_j = m(x_j) + \varepsilon_j \quad (j \in \mathcal{P})$$

où les ε_j sont des variables i.i.d. de fdr G et $m(x_j)$ est la médiane conditionnelle de T sachant $X = x_j$. Nous avons choisi de modéliser la relation entre T et X par la médiane conditionnelle qui est plus simple à estimer que la moyenne conditionnelle en présence de censure.

Comme $\mathbb{E}_\xi(\mathbb{1}(t_j \leq t)) = P(t_j \leq t) = G(t - m(x_j))$, l'indicatrice $\mathbb{1}(t_j \leq t)$ peut être prédite en estimant $G(t - m(x_j))$. Nous devons donc en premier lieu estimer la médiane conditionnelle $m(x_j)$ de T sachant $X = x_j$. Nous estimons donc la fdr conditionnelle de T

sachant $X = x$ à l'aide de l'estimateur de Kaplan-Meier généralisé (Beran, 1981) calculé sur s :

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

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

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

K est un noyau et h_X une fenêtre adéquate.

Comme $\hat{F}_{\text{GKM}}$ est une fonction en escalier, nous proposons d'utiliser plutôt la version lissée en t de Leconte *et al.* (2002), afin d'obtenir une meilleure estimation de la médiane :

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

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 $\hat{m}(x_j) = \hat{F}^{-1}_{\text{SGKM}}(0.5 \mid x_j)$.

Revenons à l'estimation de $G(t - m(x_j))$. Les résidus $\hat{\varepsilon}_j = y_j - \hat{m}(x_j)$ étant censurés à droite comme le sont les y_j, nous estimons la fdr G des erreurs par l'estimateur de Kaplan-Meier appliqué aux résidus $\hat{\varepsilon}_j$ de s, que nous noterons $\hat{G}_{\text{KM}}$. On en déduit l'estimateur suivant de la fdr :

$$\hat{F}_{\text{M}}(t) = \frac{1}{N}\left(n\hat{F}_{\text{KM}}(t) + \sum_{j \in \mathcal{P} \setminus s} \hat{G}_{\text{KM}}(t - \hat{m}(x_j)) \right)$$

3 Estimation boostrap du biais et de la variance de l'erreur de prédiction

Nous nous inspirons de Lombardia *et al.* (2004) qui ont proposé des estimateurs bootstrap du biais et de la variance de l'erreur de prédiction pour l'estimateur non paramétrique de la fdr sans censure de Dorfman et Hall (1993). Dans les problèmes de population finie, la

population $\mathcal{P}$ joue le même rôle que la fdr dans les approches de bootstrap avec *plug-in*. Pour estimer une fonctionnelle $\theta(\mathcal{P})$, on peut donc utiliser la fonctionnelle correspondante d'une population empirique $\mathcal{P}^*$: $\theta(\mathcal{P}^*)$.

Voici les étapes de la procédure bootstrap. Nous partons d'un échantillon (y_j, δ_j, x_j) où $j \in s$. La covariable x est connue sur tout $\mathcal{P}$. Nous supposons le modèle de superpopulation ξ défini à la section 2.

Etape 1 : Calcul des résidus : $\hat{\varepsilon}_j = y_j - \hat{m}(x_j)$ et calcul de l'estimateur $\tilde{G}_{\text{KM}}$ de la fdr des résidus par une version lisse de l'estimateur de Kaplan-Meier pour laquelle les paramètres de lissage seront calculés par validation croisée.

Etape 2 : Génération de B populations $\mathcal{P}^*$ de taille N :

$t_k^* = \hat{m}(x_k) + \varepsilon_k^*$ où les ε_k^* sont générés à partir de $\tilde{G}_{\text{KM}}$.

Les délais de censure c_k^* sont générés à partir de l'estimation par Kaplan-Meier inverse de la fdr de la variable délai de censure estimée à partir de l'échantillon s.

On en déduit la population $\mathcal{P}^*$: (y_k^*, δ_k^*, x_k) où $y_k^* = min(t_k^*, c_k^*)$ et δ_k^* est l'indicatrice de non censure correspondante.

Etape 3 : Tirage de R échantillons dans chaque population $\mathcal{P}^*$.

Le biais de l'erreur de prédiction $E(\widehat{F}(t) - F(t)|\mathcal{P})$ est estimé par $E_*(E(\hat{F}^*(t) - F^*(t)|\mathcal{P}^*))$, qui peut être approché par $\frac{1}{B}\frac{1}{R}\sum_{b=1}^{B}\sum_{r=1}^{R}[\widehat{F}^{*br}(t) - F^{*b}(t)]$, où $\widehat{F}^{*br}$ désigne l'estimateur de la fdr (estimateur naïf $\hat{F}_{\text{KM}}$ ou nouvel estimateur $\hat{F}_{\text{M}}$) calculé sur l'échantillon r de la population bootstrappée b. F^{*b} est l'estimateur de Kaplan-Meier de la fdr calculé sur la population b. De même, la variance de l'erreur de prédiction $Var(\widehat{F}(t) - F(t)|\mathcal{P})$ est estimée par $E_*(Var(\hat{F}^*(t) - F^*(t)|\mathcal{P}^*))$ et peut être appprochée par $\frac{1}{B}\frac{1}{R}\sum_{b=1}^{B}\sum_{r=1}^{R}[\widehat{F}^{*br}(t) - \widehat{F}^{*b}(t)]^2$. Nous pouvons de plus obtenir un intervalle de confiance de $F(t)$ au niveau de confiance $1-\alpha$ par la formule suivante : $[\hat{F}(t) - q^*_{1-\frac{\alpha}{2}}, \hat{F}(t) + q^*_{\frac{\alpha}{2}}]$ où les q^* sont les quantiles de l'estimation bootstrap de $H(u) = P(\hat{F}(t) - F(t) \leq u \mid \mathcal{P})$.

4 Prise en compte de la non-réponse

Dans les enquêtes longitudinales, le phénomène de censure mélange deux mécanismes très différents : les "exclus-vivants", qui correspondent aux individus pour lesquels l'événement d'intérêt ne s'est pas encore produit à la fin de l'étude, et les "perdus de vue", individus qui ont quitté l'étude avant la fin, ce qui correspond à des non-répondants. L'estimateur de la section précédente traite ces deux cas de la même façon, en considérant que la durée d'intérêt est supérieure à la durée observée pour ces individus, mais il serait peut-être plus judicieux de traiter le cas des perdus de vue avec des techniques adaptées à la non-réponse, comme l'imputation. On pourrait également imaginer une technique d'imputation sous contrainte, qui tienne compte du fait que la variable d'intérêt est supérieure au délai de

surveillance. On peut noter que si l'enquête est transversale, les non-répondants ne pourront pas être traités comme des individus censurés et seules des techniques d'imputation seront possibles.

5 Simulations *model-based*

Pour comparer les performances du nouvel estimateur à celles de l'estimateur naïf de Kaplan-Meier ainsi que les différentes approches de traitement de la non-réponse, des simulations ont été réalisées.

A chaque itération, une population de taille N a été générée selon un modèle de durée de vie accélérée : $ln(t_j) = -3 + 0.2 * x_j + \sigma * \varepsilon_j$ où la covariable x_j suit une loi uniforme sur (1,4). Les erreurs ε_j suivent une loi de valeur extrême de façon à obtenir une loi de Weibull pour les t_j. Ce modèle est un modèle des risques proportionnels avec un risque relatif (RR) égal à $exp(0.2/\sigma)$. Les délais de censure c_j sont générés selon une loi uniforme sur $[0, c]$, le paramètre c permettant de régler le taux de censure (10 %, 25 % ou 50 %). Deux valeurs de σ sont choisies, correspondant à une faible ou une forte liaison entre la variable d'intérêt et la variable auxiliaire. Les non-répondants sont tirés selon une loi de Bernoulli parmi les individus censurés. Pour chaque population, nous tirons un échantillon de taille $N/10$ selon un plan aléatoire simple sans remise.

Les MASE (Mean Averaged Square Error) des estimateurs sont comparés pour les différents taux de censure.

6 Exemple d'application et simulations *design-based*

Nous considérons des données du programme AFDC (Aid to Families with Dependent Children) (Hu et Ridder, 2012). Il s'agit d'une aide financière attribuée de 1935 à 1996 aux USA aux familles monoparentales, le montant de l'aide variant selon les ressources de la famille. Nous utilisons aussi les données du SIPP (Survey of Income and Program Participation) (SIPP) de 1992 et 1993. Le panel est divisé en 4 groupes de rotation. Chaque mois, un groupe est interrogé sur les 4 mois précédents et chaque membre du panel est suivi pendant 36 mois. La durée T d'intérêt est la durée de l'épisode de bien-être de la mère (nous nous limitons au premier épisode de bien-être). T est censurée pour les familles pour lesquelles l'épisode n'est pas terminé quand elles quittent le panel. Il a été montré que le montant de l'aide financière est négativement et significativement lié à la probabilité de quitter l'état de bien-être (par un modèle de Cox : $\beta = -0,00131$, $RR = e^{\beta} = 0,999$, $p = 0,0013$), ce qui nous incite à choisir le montant de l'aide comme variable auxiliaire. 520 épisodes de bien-être sont observés, dont 269 sont censurés (51,7 %).

Pour les simulations *design-based*, nous considérons les données échantillonnées ci-dessus comme notre population, dans laquelle nous tirons des échantillons selon différents plans de sondage (aléatoire simple, stratifié avec allocation proportionnelle, stratifié avec

allocation optimale). Les MASE des deux estimateurs seront comparés selon les différents plans de sondage.

Bibliographie

[1] Chambers, R.L. et Dunstan, R. (1986), Estimating distribution functions from survey data, *Biometrika*, 73, 597–804.
[2] Dabrowska, D.M. (1992), Nonparametric quantile regression with censored data, *Sankhya Journal*, 54, 252–259.
[3] Dorfman, A.H. et Hall, P. (1993), Estimators of the finite population distribution function using nonparametric regression, *Annals of Statistics*, 21, 1452–1475.
[4] Efron, B (1967), The two sample problem with censored data, *Proc 5th Berkeley Symp*, 4, 831–853.
[5] Kaplan, E et Meier, P (1958), Nonparametric estimation for incomplete observation, *Journal of the American Statistical Association*, 53, 457–481.
[6] Hu, Y. et Ridder, G. (2012), Estimation of nonlinear models with mismeasured regressors using marginal information, *Journal of Applied Econometrics*, 27(3), 347–385.
[7] 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.
[8] Lombardia, M.J., Gonzalez-Manteiga, W. et Prada-Sanchez, W. (2004), Bootstrapping the Dorfman-Hall-Chambers-Dunstan estimate of a finite population distribution function, *Journal of Nonparametric Statistics*, 16(1-2), 63–90.