Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 337 (2003) 745–748
COMPTES RENDUS MATHEMATIQUE

Statistique/Probabilités

Estimation de la fonction de survie sous biais de longueur pour des risques concurrents et proportionnels

Jean-Yves Dauxois [a], Agathe Guilloux [a], Syed N.U.A. Kirmani [b]

[a] *CREST-ENSAI, campus de Ker Lann, 35170 Bruz, France*
[b] *University of Northern Iowa, Department of Mathematics, Cedar Falls, IA 50614-0506, États-Unis*

Reçu le 7 mai 2003 ; accepté après révision le 14 octobre 2003

Présenté par Paul Deheuvels

Résumé

Considérons une population d'individus sujets à deux causes de mort. Nous observons les individus vivants au temps t_0 et nous les suivons jusqu'à la mort. A partir de cet échantillon biaisé en longueur, nous proposons un estimateur de la fonction de survie pour les durées de vie initiales (i.e. pour toute la population) sous l'hypothèse des risques proportionnels pour les deux causes de mort. Le comportement asymptotique de notre estimateur est également étudié. ***Pour citer cet article : J.-Y. Dauxois et al., C. R. Acad. Sci. Paris, Ser. I 337 (2003).***

Abstract

Estimation of the survival function under length-biased sampling for competing risks and proportional hazards. Consider a population of individuals who experience two causes of death. We observe the ones alive at time t_0 and follow them until death. Given this length bias sample, we propose an estimator of the survival function of 'initial survival times' (i.e., for the entire population) under the assumption of proportional hazards for the two causes of death. The large sample behaviour of our estimator is also studied. ***To cite this article: J.-Y. Dauxois et al., C. R. Acad. Sci. Paris, Ser. I 337 (2003).***

1. Introduction

Le problème central en analyse des durées de vie est l'estimation de la distribution du temps Z entre deux événements donnés sous différentes méthodes d'échantillonnage. Fréquemment, cette distribution doit être estimée à partir d'un échantillon en coupe, c'est-à-dire en ne prenant en compte que les individus vivants à un temps t_0. On dit couramment que ces données sont biaisées en longueur car la densité f_{Z^b} de la durée de vie échantillonnée Z^b est proportionnelle à $\mathrm{Id}(\cdot)g_Z(\cdot)$ où g_Z est la densité de la durée de vie initiale Z et Id est la fonction identité.

Adresses e-mail : dauxois@ensai.fr (J.-Y. Dauxois), guilloux@ensai.fr (A. Guilloux), kirmani@math.uni.edu (S.N.U.A. Kirmani).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2003.10.019

Vardi [8] a été le premier à considérer l'estimation non-paramétrique en présence de biais de longueur. Plus récemment Asgharian et al. [3] ont obtenu un estimateur non-paramétrique du maximum de vraisemblance pour la fonction de survie de Z et son comportement asymptotique.

Dans cet article, nous étudions le cas d'un échantillon en biais de longueur dans lequel les individus sont soumis à deux causes de mort dont les risques sont supposés proportionnels. Cette situation a été étudiée pour un échantillon sans biais de longueur (formé des durées de vie initiales) par Cheng et Lin [5] et Csörgö [6]. A notre connaissance, la conjonction de ces deux phénomènes n'a pas été étudiée jusqu'alors.

Dans la Section 2, nous établissons une relation exprimant la fonction de survie d'une des deux durées de vie initiales en concurrence en fonction de celle du temps observé Z^b sous échantillonnage en biais de longueur. Nous utilisons cette relation dans la Section 3 pour proposer un estimateur de cette fonction de survie sous l'hypothèse de biais de longueur. Notre principal résultat est alors d'étudier le comportement asymptotique de notre estimateur (cf. Théorème 3.1).

2. Risques compétitifs et biais de longueur

Considérons une population d'individus $\{i \in I\}$ soumis à deux causes de mort. A chaque individu i on associe donc deux variables aléatoires positives X_i et Y_i, durées de vie associées à la première et seconde cause de mort. On suppose ici que les durées de vie sont indépendantes et en concurrence, c'est-à-dire que l'on ne peut observer que la v.a. $Z_i = \min(X_i, Y_i) = X_i \wedge Y_i$ et l'indicateur de la cause de mort $\delta_i = \mathbb{I}_{(X_i \leqslant Y_i)}$. Nous supposons de plus que les deux causes de mort sont à risques proportionnels, c'est-à-dire qu'il existe $\beta > 0$ tel que :

$$\overline{G}_Y(x) = \big(\overline{G}_X(x)\big)^\beta \text{ pour tout } x > 0, \tag{1}$$

où $\overline{G}_X$ et $\overline{G}_Y$ sont les fonctions de survie des v.a. X et Y. Armitage [2] a montré que cette hypothèse était équivalente à l'indépendance entre les variables aléatoires Z et δ.

Supposons maintenant que cette population, appelée dans la suite « population initiale », soit telle que les naissances soient distribuées suivant un processus de Poisson mélangé. Dans cet ensemble d'individus, nous supposons de plus n'observer que ceux vivants à un instant t_0 donné, c'est le phénomène du biais de longueur. Dans cet échantillon observé, les durées de vie X^b et Y^b associées aux deux causes de mort n'ont alors pas la même distribution que les durées de vie initiales (sur toute la population). Sous ces deux hypothèses, on peut établir les relations suivantes entre la densité $g_X g_Y$ du couple (X, Y) et celles f_{X^b}, f_{Y^b} et f_{X^b,Y^b} de X^b, Y^b et (X^b, Y^b) :

$$\begin{aligned} f_{X^b}(x) &= \frac{x}{\mathbb{E}(X)} g_X(x) \mathbb{I}_{\mathbb{R}_+}(x), \\ f_{Y^b}(x) &= \frac{y}{\mathbb{E}(Y)} g_Y(y) \mathbb{I}_{\mathbb{R}_+}(y), \\ f_{X^b,Y^b}(x,y) &= \frac{1}{\mathbb{E}Z} (x \wedge y) g_X(x) g_Y(y) \mathbb{I}_{\mathbb{R}_+^2}(x,y). \end{aligned} \tag{2}$$

En utilisant l'hypothèse des risques proportionnels, la densité de $Z^b = X^b \wedge Y^b$ peut alors s'écrire sous la forme :

$$f_{Z^b}(z) = \frac{1}{\mathbb{E}Z}(1+\beta) z g_X(z) \big(\overline{G}_X(z)\big)^\beta \mathbb{I}_{\mathbb{R}_+}(z). \tag{3}$$

La relation (3) est alors équivalente à :

$$\overline{G}_X(y) = \left(\frac{\int_y^\infty \frac{1}{z} \mathrm{d}\overline{F}_{Z^b}(z)}{\int_0^\infty \frac{1}{z} \mathrm{d}\overline{F}_{Z^b}(z)} \right)^{1/(1+\beta)}. \tag{4}$$

3. Estimation et comportement asymptotique

Avant d'utiliser l'Éq. (4) pour proposer un estimateur de la fonction de survie $\overline{G}_X$ à partir de l'observation en biais de longueur décrite précédemment, nous montrons de plus que les variables aléatoires Z^b et $\delta^b = \mathbb{I}_{(X^b \leqslant Y^b)}$ sont indépendantes. On peut en effet montrer, en utilisant (2), que l'on a pour tout $0 \leqslant a \leqslant b < \infty$:

$$\mathbb{P}\big(X^b \leqslant Y^b\big) = \mathbb{P}\big(X^b \leqslant Y^b / Z^b \in [a,b]\big) = \frac{1}{1+\beta}.$$

Dans la suite nous noterons α cette dernière probabilité.

On introduit alors $\widehat{\overline{F}}_{Z^b}$, l'estimateur empirique de $\overline{F}_{Z^b}$, et $\hat{\alpha}_n$, proportion de morts ayant pour cause 1. La relation (4) incite alors à estimer $\overline{G}_X$ par :

$$\widehat{\overline{G}}_X(y) = \left(\frac{\int_y^\infty \frac{1}{z}\,\mathrm{d}\widehat{\overline{F}}_{Z^b}(z)}{\int_0^\infty \frac{1}{z}\,\mathrm{d}\widehat{\overline{F}}_{Z^b}(z)}\right)^{\hat{\alpha}_n}. \tag{5}$$

Nous obtenons alors notre résultat principal qui décrit le comportement asymptotique de notre estimateur :

Théorème 3.1. *Dans l'espace des fonctions càdlàg sur* $[0,\infty[$ *muni de la norme de Skorohod, on a, quand n tend vers* ∞*, la convergence faible suivante* :

$$\sqrt{n}\big(\widehat{\overline{G}}_X(\cdot) - \overline{G}_X(\cdot)\big) \xrightarrow{\mathcal{D}} \xi(\cdot) = \alpha \overline{G}_Z(\cdot)^{\alpha-1} L(\cdot) + \ln\big[\overline{G}_Z(\cdot)\big]\overline{G}_Z(\cdot)^\alpha U,$$

où U *est une v.a. suivant une loi* $\mathcal{N}(0,\alpha(1-\alpha))$ *et indépendante du processus* $M(\cdot)$*,*

$$L(\cdot) = -\frac{\int_\cdot^\infty \frac{1}{x}\,\mathrm{d}\overline{F}_{Z^b}(x) M(0)}{(\int_0^\infty \frac{1}{x}\,\mathrm{d}\overline{F}_{Z^b}(x))^2} + \frac{M(\cdot)}{\int_0^\infty \frac{1}{x}\,\mathrm{d}\overline{F}_{Z^b}(x)},$$

$$M(\cdot) = \int_\cdot^\infty x\,\mathrm{d}\big[\overline{F}_{Z^b} B(W)\big](x),$$

$$W(\cdot) = \frac{F_{Z^b}(\cdot)}{\overline{F}_{Z^b}(\cdot)},$$

et B *est le mouvement brownien sur* $[0,\infty[$.

Nous donnons les grandes lignes de la preuve du théorème. En introduisant la fonction $F \mapsto \Phi_t(F) = \int_t^\infty \frac{1}{x}\,\mathrm{d}F(x) / \int_0^\infty \frac{1}{x}\,\mathrm{d}F(x)$ définie sur les fonctions càdlàg à variations bornées, on a alors

$$\overline{G}_X(t) = \big(\Phi_t\big(\overline{F}_{Z^b}\big)\big)^\alpha \quad \text{et} \quad \widehat{\overline{G}}_X(t) = \big(\Phi_t\big(\widehat{\overline{F}}_{Z^b}\big)\big)^{\hat{\alpha}_n}.$$

On fait alors un développement de Taylor à l'ordre 1 :

$$\begin{aligned}\widehat{\overline{G}}_X(t) - \overline{G}_X(t) = {} & \alpha \overline{G}_Z(t)^{\alpha-1}\big(\Phi_t\big(\widehat{\overline{F}}_{Z^b}\big) - \Phi_t\big(\overline{F}_{Z^b}\big)\big) + \ln\big[\overline{G}_Z(t)\big]\overline{G}_Z(t)^\alpha(\hat{\alpha}_n - \alpha) \\ & + \frac{1}{2}\big(\Phi_t\big(\widehat{\overline{F}}_{Z^b}\big) - \Phi_t\big(\overline{F}_{Z^b}\big)\big)^2 \alpha^*(\alpha^*-1)\overline{G}_Z(t)^{\alpha^*-2} + \frac{1}{2}(\hat{\alpha}_n - \alpha)^2\big(\ln\big[\overline{G}_Z(t)^*\big]\big)^2 \\ & + \big(\Phi_t\big(\widehat{\overline{F}}_{Z^b}\big) - \Phi_t\big(\overline{F}_{Z^b}\big)\big)(\hat{\alpha}_n - \alpha)\alpha^* \ln\big[\overline{G}_Z(t)^*\big]\big(\overline{G}_Z(t)^*\big)^{\alpha^*-1}.\end{aligned}$$

Ayant montré l'indépendance entre Z^b et δ^b et en appliquant le théorème de la limite centrale à $\hat{\alpha}_n$, on obtient le résultat du théorème en étudiant la convergence faible de $\sqrt{n}(\Phi_\cdot(\widehat{\overline{F}}_{Z^b}) - \Phi_\cdot(\overline{F}_{Z^b}))$.

Pour cela, on note $D\Phi_t$ la différentielle de Gâteaux de Φ_t et, finalement, on écrit :

$$\sqrt{n}\big(\Phi_t\big(\widehat{\overline{F}}_{Z^b}\big) - \Phi_t\big(\overline{F}_{Z^b}\big)\big) = D\Phi_t\big(\sqrt{n}\big(\widehat{\overline{F}}_{Z^b} - \overline{F}_{Z^b}\big)\big) + R_t\big(\widehat{\overline{F}}_{Z^b}, \overline{F}_{Z^b}\big),$$

où $R_t(\widehat{\overline{F}}_{Z^b}, \overline{F}_{Z^b}) = \sqrt{n}(\Phi_t(\widehat{\overline{F}}_{Z^b}) - \Phi_t(\overline{F}_{Z^b})) - D\Phi_t(\sqrt{n}(\widehat{\overline{F}}_{Z^b} - \overline{F}_{Z^b}))$. On montre que :

$$D\Phi_t\big(\sqrt{n}\big(\widehat{\overline{F}}_{Z^b} - \overline{F}_{Z^b}\big)\big) = -\frac{\int_t^\infty \frac{1}{x}\overline{F}_{Z^b}(x)\int_0^\infty \frac{1}{x}\,\mathrm{d}[\sqrt{n}(\widehat{\overline{F}}_{Z^b} - \overline{F}_{Z^b})](x)}{(\int_0^\infty \frac{1}{x}\overline{F}_{Z^b}(x))^2} + \frac{\int_t^\infty \frac{1}{x}\,\mathrm{d}[\sqrt{n}(\widehat{\overline{F}}_{Z^b} - \overline{F}_{Z^b})](x)}{\int_0^\infty \frac{1}{x}\overline{F}_{Z^b}(x)}.$$

Grâce au théorème d'application continue (cf., e.g., Billingsley [4]), pour obtenir la convergence faible de $D\Phi_t(\sqrt{n}(\widehat{\overline{F}}_{Z^b} - \overline{F}_{Z^b}))$, il suffit alors d'utiliser le lemme suivant, que l'on établit en utilisant des outils présents dans [1] et [7]

Lemme 3.2. *On a, quand $n \to \infty$, la convergence faible dans l'espace de Skorohod $\mathbb{D}[0, \infty[$:*

$$\widehat{M}(\cdot) = \sqrt{n}\int_\cdot^\infty \frac{1}{x}\,\mathrm{d}\big[\widehat{\overline{F}}_{Z^b} - \overline{F}_{Z^b}\big](x) \xrightarrow{D} M(\cdot) = \int_\cdot^\infty \frac{1}{x}\,\mathrm{d}\big[\overline{F}_{Z^b}B(W)\big](x),$$

où $W(\cdot) = F_{Z^b}(\cdot)/\overline{F}_{Z^b}(\cdot)$.

On achève la démonstration en prouvant la convergence faible de $R_\cdot(\widehat{\overline{F}}_{Z^b}, \overline{F}_{Z^b})$ vers 0.

Références

[1] P.K. Andersen, O. Borgan, R.D. Gill, N. Keiding, Statistical Models Based on Counting Processes, Springer-Verlag, 1992.
[2] P. Armitage, The comparison of survival curves, J. Roy. Soc. Ser. A 122 (1959) 279–300.
[3] M. Asgharian, C.E. M'Lan, D.B. Wolfson, Length-biased sampling with right-censoring: an unconditional approach, J. Amer. Statist. Assoc. 97 (2002) 201–209.
[4] P. Billingsley, Convergence of Probability Measures, Wiley, 1968.
[5] P.E. Cheng, G.D. Lin, Maximum likelihood estimation of a survival function under the Koziol–Green proportional hazards model, Statist. Probab. Lett. 5 (1987) 75–80.
[6] S. Csörgö, Estimation in the proportional hazards model of random-censorship, Statistics 19 (1988) 437–463.
[7] J.Y. Dauxois, A new method for proving weak convergence results applied to nonparametric estimators in survival analysis, Stochastic Process. Appl. 90 (2000) 327–334.
[8] Y. Vardi, Nonparametric estimation in presence of length-bias, Ann. Statist. 10 (1982) 616–620.