Journal de la Société Française de Statistique
Vol. 155 No. 3 (2014)

Estimation dans le modèle de transformation linéaire avec données manquantes

Title: Estimation in the linear transformation model with missing data

Amel Mezaouer[1], Kamal Boukhetala[2] et Jean-François Dupuy[3]

Résumé : La classe des modèles de transformation linéaire est une classe de modèles de régression semi-paramétriques de durées. Elle comprend comme cas particuliers les modèles à risques proportionnels et à risques convergents, très utilisés en fiabilité. Cheng *et al.* (Biometrika, 1995) ont proposé des équations d'estimation simples pour en estimer le paramètre de régression. Dans cet article, nous considérons la situation où l'observation de la durée jusqu'à défaillance (éventuellement censurée) n'est possible que pour un sous-échantillon aléatoire de l'échantillon initial des items. Cette situation de données manquantes se rencontre en particulier en fiabilité lorsque des contraintes inattendues viennent interrompre un essai en cours. Tout d'abord, nous adaptons les équations d'estimation de Cheng *et al.* (Biometrika, 1995) à ce problème. Puis nous montrons la consistance de l'estimateur ainsi construit. Enfin, nous évaluons les propriétés de cet estimateur par simulations et nous illustrons la méthode sur un jeu de données réelles.

Abstract: The class of linear transformation models is a class of semi-parametric regression models for lifetime data. This class includes the proportional hazards and proportional odds models as special cases. Cheng *et al.* (Biometrika, 1995) proposed simple estimating equations for the regression parameter in this class of models. In the present paper, we consider the situation where the lifetime data is only observed in a random subset of the initial sample. This may happen, for example, in reliability testing where unexpected issues arising during the experiment may prevent engineers from observing the duration for all the tested items. We adapt Cheng *et al.*'s estimating equations to this setting and we prove the consistency of the resulting estimator. We evaluate its finite-sample properties via simulations and we illustrate our methodology on a real-data set.

Mots-clés : durées censurées, consistance, équation d'estimation, pondération par probabilité inverse, simulations
Keywords: censored data, consistency, estimating equation, inverse weighted probability, simulations
Classification AMS 2000 : 62N01, 62N02, 62E20

1. Introduction

Depuis l'introduction en 1972 par D. Cox (Cox, 1972) du modèle à risques proportionnels, la littérature consacrée aux modèles de régression de durées a connu un essor remarquable. De nombreux modèles, toujours plus sophistiqués, ont été proposés et leurs différents aspects : inférence, validation, sélection, application (en fiabilité notamment) ont été étudiés en détail, nourrissant une littérature foisonnante. Plusieurs ouvrages dressent un panorama de ce domaine de recherche en perpétuelle évolution. Citons, en toute subjectivité et sans prétendre à l'exhaustivité,

[1] Faculté des Sciences, Université Saad Dahlab, BP 270 Blida, Algérie
E-mail : a_mezaouer@univ-blida.dz
[2] Université des Sciences et Technologie Houari Boumédienne, Bab-Ezzouar, Algérie
E-mail : kboukhetala@usthb.dz
[3] IRMAR-Institut National des Sciences Appliquées de Rennes, France
E-mail : Jean-Francois.Dupuy@insa-rennes.fr

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

les ouvrages Andersen et al. (1993), Bagdonavičius and Nikulin (2002), Fleming and Harrington (1991), Klein and Moeschberger (1997), Lawless (2003), Martinussen and Scheike (2006) et Meeker and Escobar (1998) qui s'adressent aussi bien au lecteur intéressé par les aspects théoriques de l'inférence statistique dans les modèles de régression de durées qu'au praticien de la statistique à la recherche d'outils de modélisation et d'exemples concrets. Parmi la très grande variété des modèles qui ont été développés au cours des trente dernières années, la classe des modèles semi-paramétriques de transformation linéaire offre l'avantage d'inclure plusieurs des modèles les plus utilisés dans les applications. Cette classe de modèles a suscité une littérature très abondante depuis vingt ans. Le lecteur pourra consulter avec profit les articles Chen et al. (2002), Cheng et al. (1995), Dupuy (2008, article de synthèse), Fine et al. (1998), Fleming and Lin (2000), Kong et al. (2004), Kong et al. (2006), Slud and Vonta (2004) ainsi que les ouvrages Bagdonavičius and Nikulin (2002) et Martinussen and Scheike (2006) qui lui consacrent chacun un chapitre.

Notons T la durée aléatoire jusqu'à un instant de défaillance ou de panne et $Z = (Z^1, \ldots, Z^p)^\top$ un vecteur de dimension p de variables explicatives (dans la suite, $^\top$ désigne la transposée). La classe des modèles semi-paramétriques de transformation linéaire exprime la relation entre T et Z sous la forme

$$e(T) = -\beta_0^\top Z + \varepsilon, \tag{1}$$

où e est une fonction strictement croissante inconnue (dite fonction de transformation), $\beta_0 = (\beta_0^1, \ldots, \beta_0^p)^\top$ est un vecteur de paramètres de régression inconnus (paramètres d'intérêt du modèle) et ε désigne un terme d'erreur aléatoire (indépendent de Z) dont la loi de probabilité est supposée connue (on notera F_ε sa fonction de répartition). Si $H(u) = \exp(e(u))$ et h désigne la dérivée de H, on montre aisément que la fonction de risque instantané conditionnelle $\lambda(t) = \lim_{\delta \downarrow 0} \delta^{-1} \mathbb{P}(t \leq T < t + \delta | T \geq t, Z)$ de T sachant Z peut s'écrire sous la forme

$$\lambda(t) = \lambda_{e^\varepsilon}(e^{\beta_0^\top Z} H(t)) e^{\beta_0^\top Z} h(t) \tag{2}$$

où λ_{e^ε} désigne la fonction de risque instantané de $\exp(\varepsilon)$. On déduit facilement de (2) des cas particuliers remarquables du modèle (1). Ainsi, si ε suit la loi des valeurs extrêmes (c'est-à-dire : $F_\varepsilon(t) = 1 - \exp(-\exp(t))$), on montre facilement que $\exp(\varepsilon)$ suit une loi exponentielle de paramètre 1, d'où $\lambda_{e^\varepsilon} \equiv 1$ et (2) se réduit à $\lambda(t) = e^{\beta_0^\top Z} h(t)$ qui est le modèle à risques proportionnels de Cox de fonction de risque de base instantané h. Si ε suit la loi logistique (c'est-à-dire : $F_\varepsilon(t) = \exp(t)/(1+\exp(t))$), alors $\lambda_{e^\varepsilon}(t) = (1+t)^{-1}$ et (2) se ramène à $\lambda(t) = h(t)/\{H(t) + e^{-\beta_0^\top Z}\}$ qui est le modèle à risques convergents. On trouvera dans Kosorok and Song (2007) d'autres cas particuliers intéressants du modèle (1).

Plusieurs méthodes ont été proposées pour estimer le paramètre β_0 dans la classe de modèles (1) (le lecteur pourra notamment consulter Cheng et al., 1995, Fine et al., 1998, Kong et al., 2004, Kong et al., 2006, Martinussen and Scheike, 2006). Cheng et al. (1995) ont en particulier proposé des équations d'estimation simples à partir d'un échantillon d'observations indépendantes

$$(X_i, \Delta_i, Z_i), \quad i = 1, \ldots, n \tag{3}$$

du triplet (X, Δ, Z), où $X = \min(T, C)$ désigne la durée observée, C une censure aléatoire, Z un vecteur de variables explicatives, $\Delta = 1(T \leq C)$ et $1(\cdot)$ désigne la fonction indicatrice. L'estimateur proposé est consistant et asymptotiquement gaussien mais ne saurait être plus efficace que les

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

estimateurs développés spécifiquement pour des cas particuliers du modèle (1) (voir par exemple Andersen et al., 1993 pour le modèle à risques proportionnels ou Murphy et al., 1997 pour le modèle à risques convergents). Sa simplicité, en revanche, en fait un point de départ intéressant pour construire de nouveaux estimateurs dans des situations moins standards que celle décrite par les données (3). Ainsi, Kong et al. (2004, 2006) ont récemment adapté les équations d'estimation de Cheng *et al.* aux études cas-cohorte. Fine et al. (1998) les ont adaptées au cas où le support de la censure est inclus dans celui de la durée d'intérêt T.

Dans cet article, nous adaptons ces équations d'estimation à une situation de données manquantes. Supposons que l'on dispose d'un échantillon de n items. Pour chacun d'entre eux, on observe un vecteur de variables explicatives Z à l'instant $t = 0$ (début de l'étude) et l'on souhaite observer la durée jusqu'à la défaillance. On considère la situation où l'observation de la durée (éventuellement censurée) et de l'indicatrice de censure n'est possible que pour un sous-échantillon aléatoire de l'échantillon initial. Cette situation se rencontre en particulier en fiabilité lorsque des contraintes techniques inattendues viennent limiter les possibilités de recueil des données ou interrompre une partie d'un essai en cours. On dispose alors d'observations du triplet (X,Δ,Z) sur un sous-ensemble des n items tandis que pour les autres items, on ne dispose que des observations de Z. Dans ce contexte de données manquantes, une solution simple pour estimer β_0 consiste à mener une analyse en "cas complets" ("CC" par la suite) c'est-à-dire à : i) retirer de l'échantillon les items i pour lesquels l'information (X_i,Δ_i) est manquante, ii) calculer l'estimateur de Cheng et al. (1995) sur les items restants. Cette solution entraîne néanmoins une perte d'information et n'est donc pas satisfaisante. Nous proposons donc dans cet article une alternative basée sur le principe de la "pondération par probabilité inverse" (sans doute mieux connu sous sa traduction anglophone de : "inverse weighted probability (IWP)", dont l'article Seaman and White, 2013 dresse un panorama récent).

Plus précisément, nous construisons des équations d'estimation adaptées au problème de données manquantes décrit ci-dessus. Puis nous montrons la consistance de l'estimateur obtenu (section 2). Nous évaluons ensuite les propriétés de cet estimateur par simulations et nous le comparons à l'estimateur CC (section 3). Enfin, nous illustrons la méthode proposée sur un jeu de données réelles. Des perspectives et problèmes ouverts concluent l'article.

2. Estimation dans le modèle de transformation linéaire avec données manquantes

2.1. Préliminaires

On considère un essai de durée $\tau < \infty$. Durant cet essai, on observe n items indépendants qui fournissent chacun une copie (X_i,Δ_i,Z_i) du vecteur aléatoire (X,Δ,Z) où $X = \min(T,C)$ et $\Delta = 1(T \leq C)$ (un item atteignant la fin de l'essai sans avoir connu de défaillance est censuré à τ). On suppose que T et C sont indépendants conditionnellement à Z et que la censure est non-informative pour les paramètres du modèle (1). On suppose également que C et Z sont indépendants. Soit $t \geq 0$. On désigne par $G(t) = \mathbb{P}(C > t)$ la fonction de survie de C et par $Y(t) = 1(X \geq t)$ l'indicatrice à risque au temps t. Cheng et al. (1995) ont proposé d'estimer β_0 dans le modèle (1) par la solution de l'équation d'estimation

$$\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} Z_{ij} \left\{ \frac{\Delta_j Y_i(X_j)}{\widehat{G}^2(X_j)} - \xi(Z_{ij}^{\top}\beta) \right\} = 0 \tag{4}$$

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

où $\xi(s) = \int_{-\infty}^{+\infty}\{1-F_\varepsilon(t+s)\}dF_\varepsilon(t)$, $Z_{ij} = Z_i - Z_j$ et $\widehat{G}(\cdot)$ désigne l'estimateur de Kaplan-Meier de $G(\cdot)$. Sous des hypothèses de régularité appropriées, Cheng et al. (1995) montrent la consistance et la normalité asymptotique de l'estimateur ainsi obtenu.

Comme mentionné dans l'introduction, nous cherchons à estimer β_0 lorsque les variables explicatives Z sont observées sur tous les items de l'échantillon mais le couple (X,Δ) n'est disponible que pour un sous-échantillon. Notons R l'indicatrice qui vaut 1 si (X,Δ) est observé et 0 sinon. Les données disponibles sont donc :

$$\mathscr{O}_i := (X_iR_i, \Delta_iR_i, R_i, Z_i), i = 1,\ldots,n,$$

soit (X_i,Δ_i,Z_i) si $R_i = 1$ et Z_i si $R_i = 0$. Nous supposons dans la suite que les données sont "manquantes au hasard" (ou "missing at random" ; une typologie des données manquantes est donnée dans l'ouvrage récent Tsiatis, 2006) :

A. La probabilité d'observer le couple (X,Δ) ne dépend pas de (X,Δ) mais peut dépendre de Z. Autrement dit, $\mathbb{P}(R = r|X,\Delta,Z) = \mathbb{P}(R = r|Z)$, $r = 0,1$.

Dans la suite, nous notons $\eta_i = \mathbb{P}(R_i = 1|Z_i)$ la probabilité d'observer le vecteur complet (X,Δ,Z) pour l'item i et nous l'appelons probabilité de sélection.

2.2. *Equations d'estimation et asymptotique*

L'intuition à la base du principe de la pondération par probabilité inverse est la suivante. Soit un item i, dont la probabilité de fournir une observation complète (X_i,Δ_i,Z_i) est égale à η_i et tel que $R_i = 1$. Alors cet item peut être considéré comme le représentant observé d'un groupe de taille $1/\eta_i$ d'items similaires mais non observés. La pondération par probabilité inverse consiste donc à pondérer chaque item i tel que $R_i = 1$ par $1/\eta_i$ afin de tenir compte de la contribution d'items similaires mais incomplètement observés. Les probabilités de sélection η_i étant généralement inconnues, il est nécessaire de les estimer. Sous l'hypothèse A, elles peuvent être estimées (paramétriquement, non/semi-paramétriquement) à partir des observations (R_i,Z_i), $i = 1,\ldots,n$.

Par exemple, si l'on suppose que les η_i suivent un modèle de régression logistique (*ie* $\text{logit}(\eta_i) := \text{logit}(\eta_i(\gamma_0)) = \gamma_0^\top Z_i$), un estimateur $\widehat{\gamma}_n$ de $\gamma_0 \in \mathbb{R}^p$ est obtenu en maximisant la vraisemblance $\prod_{i=1}^n \eta_i^{R_i}\{1-\eta_i\}^{1-R_i}$. Si ce modèle est correct, η_i peut être estimée de manière consistante par $\widehat{\eta}_i := \eta_i(\widehat{\gamma}_n)$ (voir Fahrmeir and Kaufmann, 1985). Nous proposons alors d'estimer β_0 dans le modèle (1) par la solution $\widehat{\beta}_n$ de l'équation d'estimation

$$\Psi_n(\beta) := \sum_{i=1}^n \sum_{j=1,j\neq i}^n Z_{ij}\left\{\frac{U_{ij}}{\widehat{G}^2(X_j)\widehat{\eta}_i\widehat{\eta}_j} - \xi(Z_{ij}^\top\beta)\right\} = 0 \tag{5}$$

où $U_{ij} := \Delta_jY_i(X_j)R_iR_j$. Cette solution n'admettant pas d'expression explicite, il sera nécessaire d'utiliser un algorithme numérique pour l'approcher.

Remarque 1. Nous avons supposé l'indépendance de C et Z. Néanmoins, comme le mentionnent Cheng et al. (1995), cette hypothèse peut aisément être relâchée en remplaçant $\widehat{G}$ dans (4) (et dans notre problème, dans (5)) par un estimateur de la fonction de survie conditionnelle de C sachant Z (par exemple, un estimateur de type noyau si Z est continue, voir Dabrowska, 1992).

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

Remarque 2. La démarche décrite ci-dessus peut être généralisée à une modélisation semi-paramétrique (*eg*, $\text{logit}(\eta_i) = \gamma_0^\top Z_i + h(\theta_0^\top Z_i)$) ou non-paramétrique (*eg*, $\text{logit}(\eta_i) = h(\theta_0^\top Z_i)$) des η_i, où h désigne une fonction inconnue. Des méthodes de type vraisemblance locale peuvent alors être utilisées pour estimer les η_i (voir Carroll et al., 1997 par exemple). Dans la suite, nous nous plaçons dans un cadre paramétrique et supposons que les η_i suivent le modèle logistique $\text{logit}(\eta_i) = \gamma_0^\top Z_i$, où $\gamma_0 \in \mathbb{R}^p$, $\eta_i := \eta_i(\gamma_0)$ et $\eta_i(\gamma) := \mathbb{P}(R_i = 1|Z_i;\gamma)$.

Nous allons maintenant établir la consistance de la suite d'estimateurs $(\widehat{\beta}_n)$. Les conditions de régularité suivantes nous seront utiles :

B. Il existe des compacts $\mathscr{B}$ et $\mathscr{G}$ de $\mathbb{R}^p$ tels que $\beta_0 \in \mathscr{B}$ et $\gamma_0 \in \mathscr{G}$.

C. Les composantes $Z^1, \ldots, Z^p$ du vecteur Z sont bornées : il existe $c_1 < \infty$ tel que $Z^k \in [-c_1, c_1]$, $k = 1, \ldots, p$.

D. Il existe des constantes strictement positives c_2 et c_3 telles que $G(\tau) > c_2$ et $\inf_{\gamma \in \mathscr{G}} \mathbb{P}(R = 1|Z = z;\gamma) > c_3$ pour tout z.

E. La fonction de répartition $F_\varepsilon(\cdot)$ de ε est de classe $\mathscr{C}^1$ (dans la suite, nous noterons $f_\varepsilon(\cdot)$ sa dérivée et $\dot{\xi}(s) := \partial\xi(s)/\partial s = -\int_{-\infty}^{\infty} f_\varepsilon(t+s) f_\varepsilon(t) dt$).

F. la matrice $\mathbb{E}\left[Z_{12} Z_{12}^\top \dot{\xi}(Z_{12}^\top \beta_0)\right]$ est définie positive.

Nous pouvons maintenant énoncer notre résultat.

Théorème. *Supposons que les conditions A-F soient vérifiées. Alors la suite d'estimateurs* $(\widehat{\beta}_n)$ *converge en probabilité vers* β_0 *lorsque n tend vers l'infini. De plus,* $\sqrt{n}(\widehat{\beta}_n - \beta_0) = O_p(1)$.

Démonstration. Pour prouver ce résultat, nous vérifions les conditions du théorème de Foutz (voir Foutz, 1977). Ce résultat, basé sur le théorème d'inversion locale bien connu en analyse (voir Rudin, 1964 par exemple), énonce des conditions suffisantes pour l'existence et la consistance de la solution d'une équation d'estimation du type $S_n(\theta) = 0$, où θ est le paramètre à estimer (le lecteur intéressé pourra trouver dans Strawderman and Tsiatis, 1996 une extension du résultat de Foutz au cas où la dimension de θ croît avec la taille n de l'échantillon). En utilisant les notations de notre article, ces conditions sont :

1. $n^{-2}\Psi_n(\beta_0)$ converge en probabilité vers 0 lorsque n tend vers l'infini
2. $\partial\Psi_n(\beta)/\partial\beta^\top$ existe et est continue sur un voisinage de β_0
3. $-n^{-2}\partial\Psi_n(\beta)/\partial\beta^\top$ converge uniformémement en probabilité vers une fonction $A(\beta)$ sur un voisinage de β_0, et $A(\beta_0)$ est définie positive

Nous montrons tout d'abord que la condition 1 est vérifiée. Décomposons $n^{-2}\Psi_n(\beta)$ de la façon suivante :

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

$$
\begin{aligned}
n^{-2}\Psi_n(\beta) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} Z_{ij}\left\{\frac{U_{ij}}{\widehat{G}^2(X_j)\widehat{\eta}_i\widehat{\eta}_j}-\frac{U_{ij}}{\widehat{G}^2(X_j)\eta_i\eta_j}+\frac{U_{ij}}{\widehat{G}^2(X_j)\eta_i\eta_j}\right.\\
&\qquad\left.-\frac{U_{ij}}{G^2(X_j)\eta_i\eta_j}+\frac{U_{ij}}{G^2(X_j)\eta_i\eta_j}-\xi(Z_{ij}^\top\beta)\right\}\\
&= \frac{(n-1)}{n}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} Z_{ij}\left\{\frac{U_{ij}}{G^2(X_j)\eta_i\eta_j}-\xi(Z_{ij}^\top\beta)\right\}\\
&\qquad+\frac{(n-1)}{n}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{Z_{ij}U_{ij}}{\eta_i\eta_j}\left[\frac{1}{\widehat{G}^2(X_j)}-\frac{1}{G^2(X_j)}\right]\\
&\qquad+\frac{(n-1)}{n}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{Z_{ij}U_{ij}}{\widehat{G}^2(X_j)}\left[\frac{1}{\widehat{\eta}_i\widehat{\eta}_j}-\frac{1}{\eta_i\eta_j}\right]\\
&= \frac{(n-1)}{n}\left[\Psi_{n,1}(\beta)+\Psi_{n,2}+\Psi_{n,3}\right]. \qquad (6)
\end{aligned}
$$

Considérons tout d'abord le terme $\Psi_{n,1}(\beta)$. Sous les conditions A-F, on vérifie que pour tout $k=1,\ldots,p$ et $\beta\in\mathscr{B}$,

$$
\mathbb{E}\left[\left|Z_{12}^k\left\{\frac{U_{12}}{G^2(X_2)\eta_1\eta_2}-\xi(Z_{12}^\top\beta)\right\}\right|\right]\leq 2c_1\left(\frac{1}{c_2^2c_3^2}+1\right)<\infty.
$$

D'après la loi des grands nombres pour les U-statistiques (voir Hoeffding, 1961), lorsque n tend vers l'infini :

$$
\Psi_{n,1}(\beta)\xrightarrow{p}\mathbb{E}\left[Z_{12}\left\{\frac{U_{12}}{G^2(X_2)\eta_1\eta_2}-\xi(Z_{12}^\top\beta)\right\}\right]
$$

(où $\xrightarrow{p}$ désigne la convergence en probabilité). Considérons maintenant la k-ème composante de $\Psi_{n,2}$ (pour $k=1,\ldots,p$). On a :

$$
\begin{aligned}
\Psi_{n,2}^k &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{Z_{ij}^kU_{ij}}{\eta_i\eta_j}\left[\frac{1}{\widehat{G}^2(X_j)}-\frac{1}{G^2(X_j)}\right]\\
&= \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{Z_{ij}^kU_{ij}}{\eta_i\eta_j}\left[\frac{G(X_j)-\widehat{G}(X_j)}{G^3(X_j)}\right]\left[\frac{G(X_j)\{\widehat{G}(X_j)+G(X_j)\}}{2\widehat{G}^2(X_j)}\right]. \qquad (7)
\end{aligned}
$$

Or $\frac{G(X_j)\{\widehat{G}(X_j)+G(X_j)\}}{2\widehat{G}^2(X_j)}=1+o_p(1)$ et sous les hypothèses C et D,

$$
\begin{aligned}
\left|\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{Z_{ij}^kU_{ij}}{\eta_i\eta_j}\left[\frac{G(X_j)-\widehat{G}(X_j)}{G^3(X_j)}\right]\right| &\leq \frac{2c_1}{c_2^3c_3^2}\sup_{u\in[0,\tau]}|\widehat{G}(u)-G(u)| \qquad (8)\\
&= O_p(1)
\end{aligned}
$$

d'où

$$
\Psi_{n,2}^k=\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{Z_{ij}^kU_{ij}}{\eta_i\eta_j}\left[\frac{G(X_j)-\widehat{G}(X_j)}{G^3(X_j)}\right]+o_p(1).
$$

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

Enfin, l'inégalité (8) et la convergence uniforme en probabilité de l'estimateur de Kaplan-Meier (voir Fleming and Harrington, 1991, par exemple) entraînent : $\Psi^k_{n,2} \xrightarrow{p} 0$ lorsque n tend vers l'infini. Considérons enfin la k-ème composante de $\Psi_{n,3}$ (pour $k = 1,\ldots,p$). Sous les hypothèses B, C et D,

$$|\Psi^k_{n,3}| \leq \frac{1}{n(n-1)} \frac{c_1}{c_3^4} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \frac{1}{\widehat{G}^2(\tau)} \left|\eta_i\eta_j - \widehat{\eta}_i\widehat{\eta}_j\right|.$$

Par un développement de $\eta_i(\gamma) := \mathbb{P}(R_i = 1|Z_i;\gamma)$ au voisinage de γ_0, nous avons

$$\widehat{\eta}_i\widehat{\eta}_j - \eta_i\eta_j = (\widehat{\gamma}_n - \gamma_0)^\top(\eta_i\dot{\eta}_j(\widetilde{\gamma}_n) + \eta_j\dot{\eta}_i(\widetilde{\gamma}_n)) + (\widehat{\gamma}_n - \gamma_0)^\top\dot{\eta}_i(\widetilde{\gamma}_n)(\widehat{\gamma}_n - \gamma_0)^\top\dot{\eta}_j(\widetilde{\gamma}_n)$$

où $\dot{\eta}_i(\gamma) = \partial\eta_i(\gamma)/\partial\gamma$ et $\widetilde{\gamma}_n \xrightarrow{p} \gamma_0$ lorsque n tend vers l'infini. En utilisant successivement l'inégalité triangulaire et l'inégalité de Cauchy-Schwarz, nous obtenons :

$$\left|\widehat{\eta}_i\widehat{\eta}_j - \eta_i\eta_j\right| \leq \|\widehat{\gamma}_n - \gamma_0\|\|\eta_i\dot{\eta}_j(\widetilde{\gamma}_n) + \eta_j\dot{\eta}_i(\widetilde{\gamma}_n)\| + \|\widehat{\gamma}_n - \gamma_0\|^2\|\dot{\eta}_i(\widetilde{\gamma}_n)\|\|\dot{\eta}_j(\widetilde{\gamma}_n)\|$$

où $\|\cdot\|$ désigne la norme euclidienne sur $\mathbb{R}^p$. Sous les conditions B et C, il existe une constante $c_4 < \infty$ telle que $\|\dot{\eta}_i(\widetilde{\gamma}_n)\| < c_4$ pour tout $i = 1,\ldots,n$ et donc $\left|\widehat{\eta}_i\widehat{\eta}_j - \eta_i\eta_j\right| \leq 2c_4\|\widehat{\gamma}_n - \gamma_0\| + c_4^2\|\widehat{\gamma}_n - \gamma_0\|^2$. Finalement,

$$|\Psi^k_{n,3}| \leq \frac{c_1}{c_3^4} \frac{1}{G^2(\tau) + o_p(1)} \left(2c_4\|\widehat{\gamma}_n - \gamma_0\| + c_4^2\|\widehat{\gamma}_n - \gamma_0\|^2\right) \tag{9}$$

et donc $\Psi^k_{n,3} \xrightarrow{p} 0$ lorsque n tend vers l'infini. Ainsi, pour tout $\beta \in \mathscr{B}$, $n^{-2}\Psi_n(\beta)$ converge en probabilité vers

$$\Psi(\beta) := \mathbb{E}\left[Z_{12}\left\{\frac{U_{12}}{G^2(X_2)\eta_1\eta_2} - \xi(Z_{12}^\top\beta)\right\}\right]$$

lorsque n tend vers l'infini. Nous montrons maintenant que $\Psi(\beta_0) = 0$. Pour cela, notons que sous la condition A et par indépendance de C et Z,

$$\begin{aligned}
\mathbb{E}\left[Z_{12}\frac{U_{12}}{G^2(X_2)\eta_1\eta_2}\right] &= \mathbb{E}\left[\mathbb{E}\left[Z_{12}\frac{1(T_2 \leq C_2)Y_1(X_2)R_1R_2}{G^2(X_2)\eta_1\eta_2}\middle| Z_1,Z_2,T_2\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Z_{12}\frac{1(T_2 \leq C_2)1(T_1 \geq T_2)1(C_1 \geq T_2)R_1R_2}{G^2(T_2)\eta_1\eta_2}\middle| Z_1,Z_2,T_2\right]\right] \\
&= \mathbb{E}\left[Z_{12}\frac{G(T_2)G(T_2)}{G^2(T_2)}\mathbb{E}[1(T_1 \geq T_2)|Z_1,Z_2,T_2]\right] \\
&= \mathbb{E}[Z_{12}1(T_1 \geq T_2)] \\
&= \mathbb{E}[Z_{12}\mathbb{E}[1(T_1 \geq T_2)|Z_1,Z_2]].
\end{aligned}$$

Or $\mathbb{E}[1(T_1 \geq T_2)|Z_1,Z_2] = \xi(Z_{12}^\top\beta_0)$ (voir Cheng et al., 1995) donc

$$\mathbb{E}\left[Z_{12}\frac{U_{12}}{G^2(X_2)\eta_1\eta_2}\right] = \mathbb{E}\left[Z_{12}\xi(Z_{12}^\top\beta_0)\right]$$

et finalement,$\Psi(\beta_0) = 0$. Nous avons donc montré que $n^{-2}\Psi_n(\beta_0) \xrightarrow{p} 0$.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

Nous montrons maintenant que la condition 2 est vérifiée. Notons tout d'abord que

$$\partial\Psi_n(\beta)/\partial\beta^\top = -\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} Z_{ij}Z_{ij}^\top \dot{\xi}(Z_{ij}^\top\beta).$$

Sous la condition E, $\partial\Psi_n(\beta)/\partial\beta^\top$ est continue sur $\mathscr{B}$.

Nous montrons enfin que la condition 3 est vérifiée. Pour tout $\beta \in \mathscr{B}$,

$$-n^{-2}\partial\Psi_n(\beta)/\partial\beta^\top = \frac{(n-1)}{n}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} Z_{ij}Z_{ij}^\top \dot{\xi}(Z_{ij}^\top\beta)$$

et d'après la loi des grands nombres pour les U-statistiques (Hoeffding, 1961), lorsque n tend vers l'infini : $-n^{-2}\partial\Psi_n(\beta)/\partial\beta^\top \xrightarrow{p} A(\beta) = \int_{z_1,z_2} z_{12}z_{12}^\top \dot{\xi}(z_{12}^\top\beta)dF_Z(z_1)dF_Z(z_2)$ (où F_Z désigne la fonction de répartition de Z). La convergence uniforme en probabilité sur $\mathscr{B}$ de $-n^{-2}\partial\Psi_n(\beta)/\partial\beta^\top$ vers $A(\beta)$ s'ensuit par application de la loi des grands nombres uniforme pour les U-statistiques (Newey, 1991). Enfin, $A(\beta_0)$ est définie positive d'après la condition F.

Les conditions 1, 2 et 3 du théorème de Foutz (Foutz, 1977) sont vérifiées donc $\widehat{\beta}_n \xrightarrow{p} \beta_0$ lorsque n tend vers l'infini.

Maintenant, par un développement de $\Psi_n(\beta)$ au voisinage de β_0, nous avons :

$$\Psi_n(\widehat{\beta}_n) = \Psi_n(\beta_0) + (\partial\Psi_n(\widetilde{\beta}_n)/\partial\beta^\top)(\widehat{\beta}_n - \beta_0)$$

où $\widetilde{\beta}_n \xrightarrow{p} \beta_0$ lorsque n tend vers l'infini.

Nous pouvons alors écrire : $\sqrt{n}(\widehat{\beta}_n - \beta_0) = (-n^{-2}\partial\Psi_n(\widetilde{\beta}_n)/\partial\beta^\top)^{-1}n^{-3/2}\Psi_n(\beta_0)$. Or

$$\begin{aligned}\left|-n^{-2}\partial\Psi_n(\widetilde{\beta}_n)/\partial\beta^\top - A(\beta_0)\right| &\leq \left|-n^{-2}\partial\Psi_n(\widetilde{\beta}_n)/\partial\beta^\top - A(\widetilde{\beta}_n)\right| + \left|A(\widetilde{\beta}_n) - A(\beta_0)\right| \\ &\leq \sup_{\beta\in\mathscr{B}}\left|-n^{-2}\partial\Psi_n(\beta)/\partial\beta^\top - A(\beta)\right| + \left|A(\widetilde{\beta}_n) - A(\beta_0)\right| \\ &\leq o_p(1) + o_p(1).\end{aligned}$$

Comme de plus $A(\beta_0)$ est inversible (par la condition F), on obtient : $(-n^{-2}\partial\Psi_n(\widetilde{\beta}_n)/\partial\beta^\top)^{-1} \xrightarrow{p} A(\beta_0)^{-1}$. Ainsi, $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ et $A(\beta_0)^{-1}n^{-3/2}\Psi_n(\beta_0)$ convergent en loi vers la même limite. D'après (6), on a : $n^{-3/2}\Psi_n(\beta_0) = \frac{(n-1)}{n}[\sqrt{n}\Psi_{n,1}(\beta_0) + \sqrt{n}\Psi_{n,2} + \sqrt{n}\Psi_{n,3}]$. D'après le théorème central limite pour les U-statistiques (théorème 12.3 de van der Vaart, 1998), $\sqrt{n}\Psi_{n,1}(\beta_0)$ converge en loi vers un vecteur gaussien centré de matrice de variance-covariance

$$\mathbb{E}[\ell(\mathscr{O}_1,\mathscr{O}_2)\ell(\mathscr{O}_1,\mathscr{O}_2')^\top]$$

, où $\ell(\mathscr{O}_1,\mathscr{O}_2) = Z_{12}\left\{\frac{U_{12}}{G^2(X_2)\eta_1\eta_2} - \xi(Z_{12}^\top\beta_0)\right\}$ et $\mathscr{O}_1,\mathscr{O}_2,\mathscr{O}_2'$ sont des répliques indépendantes de $\mathscr{O}_i$. On montre à partir de (7) et (9) que $\sqrt{n}\Psi_{n,2} = O_p(1)$ et $\sqrt{n}\Psi_{n,3} = O_p(1)$. Finalement, $\sqrt{n}(\widehat{\beta}_n - \beta_0) = O_p(1)$. □

Remarque 3. Les termes $\sqrt{n}\Psi_{n,2}$ et $\sqrt{n}\Psi_{n,3}$ apparaissent lorsque G et η_i sont inconnus et doivent être estimés. D'après la démonstration ci-dessus, si G et η_i étaient connus, $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ converge-rait en loi vers un vecteur gaussien centré de variance $A(\beta_0)^{-1}\mathbb{E}[\ell(\mathscr{O}_1,\mathscr{O}_2)\ell(\mathscr{O}_1,\mathscr{O}_2')^\top]A(\beta_0)^{-1}$. Si G et η_i sont correctement estimés, il est raisonnable de penser que $\widehat{\beta}_n$ sera également asympto-tiquement gaussien (ce point sera examiné dans l'étude de simulation de la section 3).

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

2.3. *Estimation de la distribution conditionnelle de T sachant Z*

Notons $S_Z(t) := \mathbb{P}(T > t|Z)$ la fonction de survie conditionnelle de T sachant Z. On montre facilement à partir de (1) que $S_Z(t) = 1 - F_\varepsilon(\beta_0^\top Z + e(t))$. Il est donc nécessaire d'estimer la fonction inconnue $e(\cdot)$ pour pouvoir estimer $S_Z(\cdot)$. Nous proposons une équation d'estimation pour $e(t)$ et nous en déduisons un estimateur de $S_Z(t)$. Le lemme suivant sert de base à notre proposition :

Lemme. *Sous l'hypothèse A et les hypothèses d'indépendance conditionnelle de T et C et d'indépendance de C et Z, on a :*

$$\mathbb{E}\left[\frac{R1(X>t)}{\eta G(t)} - \left(1 - F_\varepsilon(\beta_0^\top Z + e(t))\right)\right] = 0.$$

Démonstration.

$$\begin{aligned}
\mathbb{E}\left[\frac{R1(X>t)}{\eta G(t)}\middle| Z\right] &= \frac{1}{\eta G(t)}\mathbb{E}[R1(X>t)|Z] \\
&= \frac{1}{\eta G(t)}\mathbb{E}[R|Z]\,\mathbb{E}[1(X>t)|Z] \\
&= \frac{1}{G(t)}\mathbb{E}[1(C>t)|Z]\,\mathbb{E}[1(T>t)|Z] \\
&= \mathbb{P}(T>t|Z) \\
&= 1 - F_\varepsilon(\beta_0^\top Z + e(t))
\end{aligned}$$

d'où le résultat, en prenant l'espérance de chaque côté de l'égalité ci-dessus. □

Au vu de ce lemme, un estimateur naturel de $e(t)$ peut être construit comme la solution $\widehat{e}(t)$ de l'équation d'estimation

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i 1(X_i>t)}{\widehat{\eta}_i \widehat{G}(t)} - \left(1 - F_\varepsilon(\widehat{\beta}_n^\top Z_i + e(t))\right)\right] = 0.$$

En pratique, $\widehat{e}(t)$ doit être approché numériquement. On déduit alors un estimateur de type "plug-in" de $S_{z^*}(t) = \mathbb{P}(T>t|Z=z^*)$:

$$\widehat{S}_{z^*}(t) = 1 - F_\varepsilon(\widehat{\beta}_n^\top z^* + \widehat{e}(t)).$$

3. Simulations et exemple

3.1. *Etude de simulation*

Dans cette étude réalisée avec le logiciel R (voir R Core Team, 2013), nous évaluons numériquement les propriétés de l'estimateur $\widehat{\beta}_n$ et le comparons à l'estimateur CC obtenu en ignorant les items pour lesquels le couple (X,Δ) est manquant puis en appliquant l'équation d'estimation (4) aux items restants.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

Le modèle considéré pour les simulations est le suivant : $e(T) = -\beta_0 Z + \varepsilon$, où e est la fonction ln, ε suit une distribution des valeurs extrêmes et Z est une variable explicative uni-dimensionnelle (par simplicité) de loi normale centrée réduite. Les valeurs suivantes : $\ln(1.5) \approx 0.405$ et 0 sont considérées pour β_0. La variable C est simulée suivant une loi exponentielle dont le paramètre $\lambda > 0$ est choisi pour produire 15% et 30% de durées censurées. Nous évaluons les propriétés de $\widehat{\beta}_n$ pour des tailles d'échantillon faible ($n = 75$) et modérée ($n = 150$). Les indicatrices R_i ($i = 1, \ldots, n$) de données manquantes sont simulées suivant une loi de Bernoulli de paramètre $\mathbb{P}(R = 1|Z = z) = \exp(\theta_0 + \theta_1 z)/(1 + \exp(\theta_0 + \theta_1 z))$ où θ_0 et θ_1 sont choisis pour produire 15% et 30% de données manquantes. Le cas où $\mathbb{P}(R = 1|Z = z) = 1$ pour tout z (pas de données manquantes) est également considéré. Ce cas (idéal dans les applications) fournira en effet la valeur de référence à laquelle nous comparerons les estimateurs $\widehat{\beta}_n$ et CC (noté $\widehat{\beta}_{n,CC}$). L'estimateur de Kaplan-Meier est utilisé pour estimer la fonction de survie $G(\cdot)$ de la censure. Les η_i sont estimées à l'aide d'un modèle logistique paramétrique, selon la procédure décrite dans la section 2.2. Pour chaque combinaison des différents paramètres de l'étude (taille d'échantillon, pourcentages de durées censurées et de données manquantes), nous simulons $N = 500$ échantillons. Pour chaque échantillon simulé j ($j = 1, \ldots, N$), nous calculons les estimations $\widehat{\beta}_n^j$ (donnée par l'équation d'estimation (5)) et CC (notée $\widehat{\beta}_{n,CC}^j$) de β_0. Les moyennes $N^{-1}\sum_{j=1}^N \widehat{\beta}_n^j$ et $N^{-1}\sum_{j=1}^N \widehat{\beta}_{n,CC}^j$ et variances empiriques de ces estimations sont données dans la Table 1 (lignes intitulées "moyenne" et "variance"). Au vu de ces résultats, l'estimateur $\widehat{\beta}_n$ semble nettement supérieur à l'estimateur CC en terme de biais comme de précision. Si les performances des deux estimateurs se dégradent lorsque le pourcentage de données manquantes augmente, $\widehat{\beta}_n$ semble plus robuste à cette augmentation que $\widehat{\beta}_{n,CC}$.

Nous avons montré dans la section 2 que $\sqrt{n}(\widehat{\beta}_n - \beta_0) = O_p(1)$. Comme mentionné dans la remarque 3, on peut de plus s'attendre à ce que $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ suive asymptotiquement une loi normale. La démonstration théorique de cette intuition reste un problème ouvert. Néanmoins, notre étude de simulation peut fournir des indications utiles sur ce point. Nous représentons les histogrammes et diagrammes quantile-quantile (QQ-plots) des $\widehat{\beta}_n^j$ et $\widehat{\beta}_{n,CC}^j$ ($j = 1, \ldots, N$). Les figures 1 et 2 donnent ces graphiques pour $n = 75$, $\beta_0 = 0.405$ et un pourcentage de censure égal à 15% et 30% respectivement. Il ressort de ces graphiques qu'une approximation gaussienne de la loi de $\widehat{\beta}_n$ semble raisonnable, y compris lorsque la taille n de l'échantillon est relativement faible. Cette approximation semble en revanche plus discutable pour la loi de l'estimateur CC (les graphiques obtenus pour les autres valeurs de n et β_0 conduisent aux mêmes observations et sont omis). Pour chaque échantillon j, nous obtenons enfin des estimations $\widehat{\text{e.t.}}(\widehat{\beta}_n^j)$ et $\widehat{\text{e.t.}}(\widehat{\beta}_{n,CC}^j)$ de l'écart-type asymptotique de $\sqrt{n}(\widehat{\beta}_n^j - \beta_0)$ (par la méthode du bootstrap) et $\sqrt{n}(\widehat{\beta}_{n,CC}^j - \beta_0)$ (en utilisant l'estimateur proposé dans Cheng et al., 1995). Les moyennes $\overline{\text{e.t.}}(\widehat{\beta}_n) := N^{-1}\sum_{j=1}^N \widehat{\text{e.t.}}(\widehat{\beta}_n^j)$ et $\overline{\text{e.t.}}(\widehat{\beta}_{n,CC}) := N^{-1}\sum_{j=1}^N \widehat{\text{e.t.}}(\widehat{\beta}_{n,CC}^j)$ de ces estimations sont fournies dans la Table 1 (lignes intitulées "moyenne ($\widehat{\text{e.t.}}$)"). Notons qu'asymptotiquement, la variance empirique de $\widehat{\beta}_n$ (respectivement $\widehat{\beta}_{n,CC}$) devrait être proche de $(\overline{\text{e.t.}}(\widehat{\beta}_n))^2/n$ (respectivement $(\overline{\text{e.t.}}(\widehat{\beta}_{n,CC}))^2/n$). Ceci est presque toujours vérifié pour $\widehat{\beta}_n$. Ce n'est en revanche pas le cas pour $\widehat{\beta}_{n,CC}$ ce qui indique que l'estimateur de la variance asymptotique proposé dans Cheng et al. (1995) et adapté à la méthode CC n'est ici pas satisfaisant (en raison sans doute de la faible taille d'échantillon résultant de l'approche CC).

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

TABLE 1. *Etude de simulation : résultats.*

β_0	n	% de censure	% de données manquantes	0	15		30	
			estimateur	DC	$\widehat{\beta}_n$	$\widehat{\beta}_{n,CC}$	$\widehat{\beta}_n$	$\widehat{\beta}_{n,CC}$
0.405	75	15	moyenne	0.386	0.393	0.705	0.382	0.825
			moyenne ($\widehat{\text{e.t.}}$)	1.193	1.355	2.712	1.481	4.719
			variance	0.019	0.027	0.152	0.031	0.259
			puissance	0.803	0.672	0.582	0.631	0.141
		30	moyenne	0.363	0.366	0.636	0.354	0.865
			moyenne ($\widehat{\text{e.t.}}$)	1.168	1.312	2.372	1.494	5.174
			variance	0.021	0.032	0.135	0.041	0.358
			puissance	0.756	0.689	0.653	0.545	0.116
	150	15	moyenne	0.382	0.379	0.639	0.383	0.834
			moyenne ($\widehat{\text{e.t.}}$)	1.218	1.382	2.411	1.591	4.564
			variance	0.011	0.016	0.062	0.026	0.129
			puissance	0.968	0.901	0.909	0.794	0.473
		30	moyenne	0.367	0.367	0.634	0.369	0.841
			moyenne ($\widehat{\text{e.t.}}$)	1.199	1.348	2.386	1.551	4.633
			variance	0.011	0.018	0.079	0.028	0.150
			puissance	0.956	0.874	0.885	0.764	0.501
0	75	15	moyenne	-0.003	0.002	0.003	0.010	0.015
			moyenne ($\widehat{\text{e.t.}}$)	1.125	1.351	1.697	1.583	2.476
			variance	0.019	0.026	0.064	0.031	0.154
			niveau	0.028	0.038	0.038	0.037	0.015
		30	moyenne	0.002	0.003	0.002	-0.002	0.004
			moyenne ($\widehat{\text{e.t.}}$)	1.137	1.367	1.726	1.636	2.962
			variance	0.022	0.030	0.077	0.047	0.293
			niveau	0.048	0.056	0.047	0.072	0.016
	150	15	moyenne	-0.002	0.000	0.000	0.000	-0.007
			moyenne ($\widehat{\text{e.t.}}$)	1.143	1.369	1.673	1.609	2.644
			variance	0.009	0.012	0.030	0.017	0.105
			niveau	0.032	0.028	0.032	0.052	0.041
		30	moyenne	-0.002	-0.004	-0.007	-0.008	-0.028
			moyenne ($\widehat{\text{e.t.}}$)	1.138	1.382	1.704	1.649	2.737
			variance	0.009	0.015	0.038	0.019	0.129
			niveau	0.024	0.040	0.044	0.052	0.028

Note. DC : données complètes.

Pour chaque échantillon j ($j = 1, \ldots, N$), la normalité asymptotique de l'estimateur de (Cheng et al., 1995) permet de calculer une statistique de test de type Wald pour tester $H_0 : \beta_0 = 0$ contre $H_1 : \beta_0 \neq 0$ à partir des cas complets. Une règle de décision de niveau asymptotique 5% consiste alors à rejeter H_0 si $|\widehat{\beta}^j_{n,CC}/\widehat{\text{e.t.}}(\widehat{\beta}^j_{n,CC})| > 1.96$. Lorsque $\beta_0 = 0.405$, on calcule à partir des N échantillons simulés la puissance empirique du test de Wald basé sur $\widehat{\beta}_{n,CC}$. Lorsque $\beta_0 = 0$, on peut calculer le niveau empirique de ce test. Nous n'avons pas démontré la normalité asymptotique de $\widehat{\beta}_n$ mais encouragés par les résultats des figures 1 et 2, nous calculons les puissance et niveau empiriques du test de Wald de H_0 basé sur $\widehat{\beta}_n$. L'ensemble de ces résultats est donné dans la Table 1 (lignes "puissance" et "niveau"). Le niveau empirique du test basé sur

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

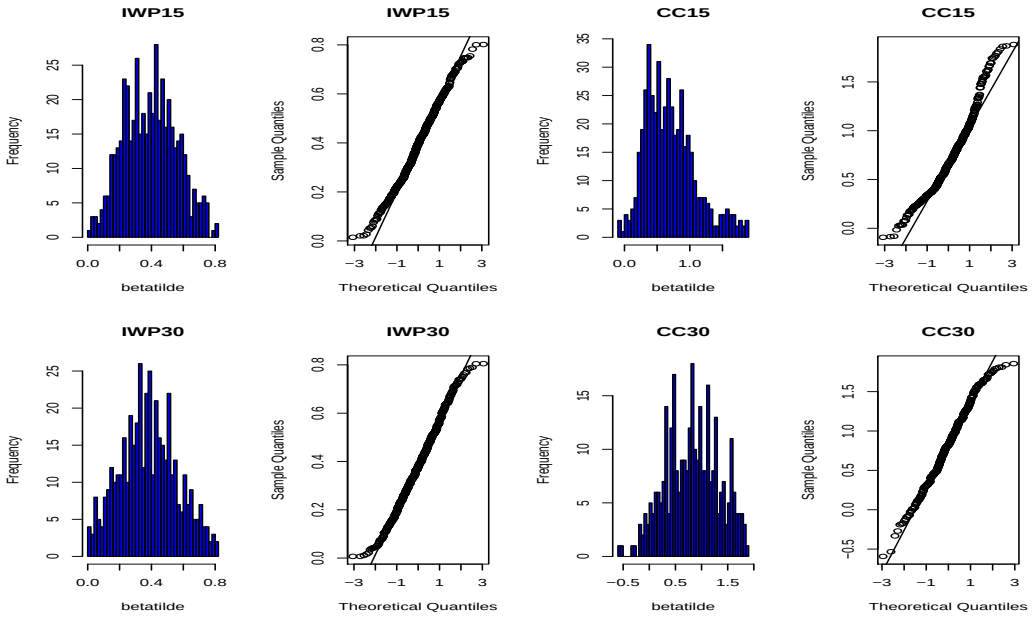

FIGURE 1. *Histogrammes et QQ-plots des* $\widetilde{\beta}_n^j$ *(IWP15 : 15% de données manquantes et IWP30 : 30% de données manquantes) et* $\widehat{\beta}^j$ *(CC15 : 15% de données manquantes et CC30 : 30% de données manquantes) (*$j = 1,\ldots,N$*) pour* $n = 75$, $\beta_0 = 0.405$, *et 15% de censure.*

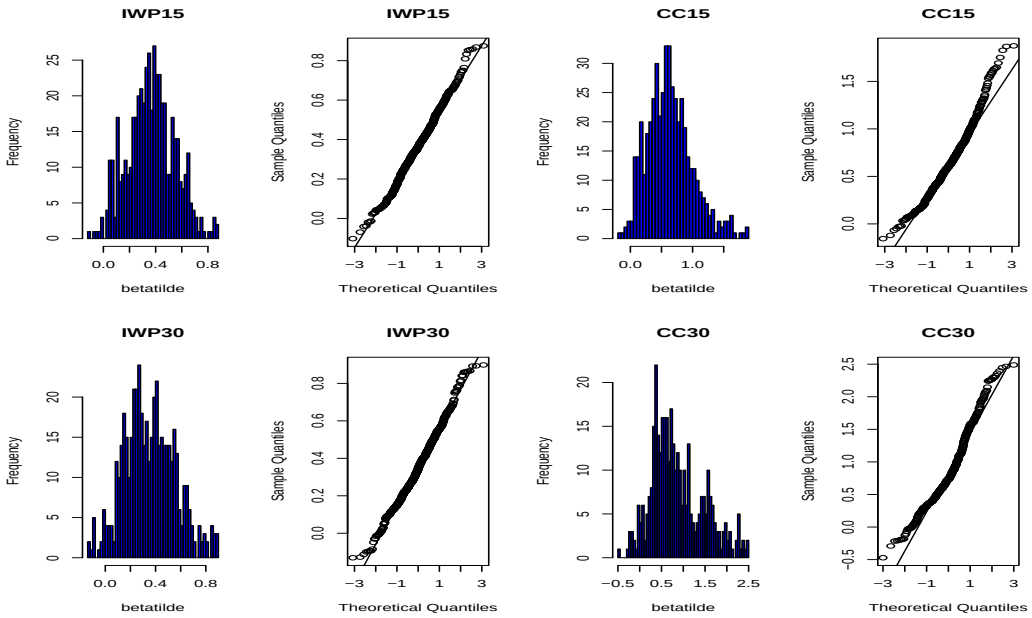

FIGURE 2. *Histogrammes et QQ-plots des* $\widetilde{\beta}_n^j$ *et* $\widehat{\beta}^j$ *(*$j = 1,\ldots,N$*) pour* $n = 75$, $\beta_0 = 0.405$, *et 30% de censure.*

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

$\widehat{\beta}_n$ (respectivement $\widehat{\beta}_{n,CC}$) varie entre 2.8% et 5.6% (respectivement 1.5% et 7.2%). On constate donc de meilleures performances pour le test basé sur $\widehat{\beta}_n$. Les puissances des deux tests diminuent lorsque la proportion de données manquantes augmente. Mais cette diminution est beaucoup plus marquée pour le test basé sur $\widehat{\beta}_{n,CC}$, le test construit sur $\widehat{\beta}_n$ conservant une puissance relativement élevée (en comparaison du cas sans données manquantes).

L'ensemble de ces résultats indiquent que l'estimateur proposé améliore de façon notable l'estimateur CC, qui constitue pour l'instant la seule solution disponible pour estimer les paramètres du modèle de transformation linéaire en présence de données manquantes. L'estimateur $\widehat{\beta}_n$ fournit des estimations pertinentes y compris lorsque la taille n de l'échantillon est faible (de l'ordre de quelques dizaines) et/ou la proportion de données manquantes (du couple (X,Δ)) est relativement élevée.

3.2. Exemple

Nous illustrons la méthode proposée sur un jeu de données réelles [1] constitué de $n = 142$ brins de kevlar soumis chacun à un stress donné (ici la suspension d'une charge accrochée au brin). On dispose pour 108 brins seulement de la charge Z qui leur est appliquée, de la durée X jusqu'à la rupture du brin ou censure et de l'indicatrice de censure Δ. Pour les 34 autres brins, on dispose seulement de la charge qui leur est appliquée. La censure intervient soit à la fin de l'essai (au bout de 11h et 24 minutes environ) soit en raison d'un détachement de la charge sans rupture du brin (du à un défaut d'attache de la charge). Dans ce dernier cas, la censure n'est pas liée à la valeur de la charge et peut être considérée comme indépendante de Z. Nous estimons le paramètre β_0 (paramètre de régression associé à la charge Z) dans le modèle (1) où ε suit une loi des valeurs extrêmes. Nous obtenons $\widehat{\beta}_n = 0.760(0.085)$ et $\widehat{\beta}_{n,CC} = 0.677(0.090)$ (les nombres entre parenthèses sont les écarts-type estimés). Si les deux estimateurs s'accordent sur l'existence d'un effet de la charge sur le risque de rupture des brins, on note néanmoins un écart relatif de 11% dans l'appréciation quantitative de cet effet. Au vu des résultats de la section 3.1, un crédit plus important pourra être donné à l'estimation $\widehat{\beta}_n$, dont on pourra penser qu'elle reflète de manière moins biaisée que $\widehat{\beta}_{n,CC}$ l'influence de la charge sur le risque de rupture.

4. Conclusion et perspectives

Dans cet article, nous avons adapté à un problème de données manquantes les équations d'estimation proposées par Cheng et al. (1995) pour estimer simplement le paramètre de régression d'un modèle de transformation linéaire. L'estimateur proposé repose sur le principe de la pondération par probabilité inverse, simple à mettre en oeuvre. Cet estimateur est consistant et l'étude de simulation que nous avons menée suggère que sa distribution peut être approchée par une loi normale lorsque la taille de l'échantillon est suffisamment grande. De plus, cet estimateur améliore de façon significative la seule solution disponible à ce jour pour implémenter les équations d'estimation de Cheng *et al.* dans un contexte de données manquantes, à savoir la méthode "cas complets". Au-delà de ces résultats encourageants, plusieurs problèmes restent ouverts. Citons en particulier l'étude de la robustesse de l'estimateur proposé à une mauvaise définition du modèle

[1] disponible à l'adresse suivante : `http://dupuy.perso.math.cnrs.fr/datakevlar.txt`

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

de $\mathbb{P}(R = 1|Z)$ et la construction d'estimateurs robustes de β_0. Les techniques exposées dans Tsiatis (2006) devraient pouvoir être utilisées avec profit pour résoudre cette question.

Remerciements

Nous souhaitons remercier deux rapporteurs pour leur nombreuses suggestions. L'ensemble de leurs commentaires, questions et propositions nous ont permis de proposer un article plus abouti. Nous remercions également Olivier Gaudoin pour son invitation à contribuer à ce numéro spécial, ainsi que pour sa patience.

Références

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer, New York.

Bagdonavičius, V. and Nikulin, M. S. (2002). *Accelerated Life Models. Modeling and Statistical Analysis*. Chapman & Hall.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92 :477–489.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, 89 :659–668.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 82 :835–845.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B*, 34 :187–220.

Dabrowska, D. M. (1992). Variable bandwidth conditional Kaplan-Meier estimate. *Scandinavian Journal of Statistics*, 19 :351–361.

Dupuy, J.-F. (2008). Transformation models for failure time data : an overview of some recent developments. *Proceedings of the Second International Conference on Accelerated Life Testing in Reliability and Quality Control (Bordeaux)*, pages 43–47.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13 :342–368.

Fine, J. P., Ying, Z., and Wei, L. J. (1998). On the linear transformation model for censored data. *Biometrika*, 85 :980–986.

Fleming, T. R. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Fleming, T. R. and Lin, D. Y. (2000). Survival analysis in clinical trials : past developments and future directions. *Biometrics*, 56 :971–983.

Foutz, R. V. (1977). On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association*, 72 :147–148.

Hoeffding, W. (1961). The strong law of large numbers for u-statistics. *Institute of Statistics Mimeo Series No. 302, University of North Carolina, Chapel Hill, N. C.*

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis : Techniques for Censored and Truncated Data*. Springer, New York.

Kong, L., Cai, J., and Sen, P. K. (2004). Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika*, 91 :305–319.

Kong, L., Cai, J., and Sen, P. K. (2006). Asymptotic results for fitting semiparametric transformation models to failure time data from case-cohort studies. *Statistica Sinica*, 16 :135–151.

Kosorok, M. R. and Song, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Annals of Statistics*, 35 :957–989.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. Wiley Series in Probability and Statistics. Wiley, Hoboken, second edition.

Martinussen, T. and Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data*. Springer, New York.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. Wiley, New York.

Murphy, S., Rossini, A., and van der Vaart, A. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92 :968–976.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59 :1161–1167.

R Core Team (2013). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http ://www.R-project.org, ISBN 3-900051-07-0.

Rudin, W. (1964). *Principles of Mathematical Analysis*. McGraw-Hill, New York.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22 :278–295.

Slud, E. V. and Vonta, F. (2004). Consistency of the NPML estimator in the right-censored transformation model. *Scandinavian Journal of Statistics*, 31 :21–41.

Strawderman, R. L. and Tsiatis, A. A. (1996). On consistency in parameter spaces of expanding dimension : an application of the inverse function theorem. *Statistica Sinica*, 6 :917–923.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238