Recueil du Symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche sociale et économique

MODÉLISATION ET ANALYSE DES DONNÉES SUR LA DURÉE PROVENANT D'ENQUÊTES LONGITUDINALES

J.F. Lawless et C. Boudreau[1]

RÉSUMÉ

Les données provenant d'enquêtes longitudinales sont utilisées pour faciliter la compréhension des processus survenant au cours de la vie, comme la scolarité, l'emploi, la fécondité et la santé. L'un des aspects de ce genre d'analyse a trait à la durée des périodes que les individus passent dans certains états. Le présent article porte sur les méthodes d'analyse des données sur la durée qui tiennent compte des caractéristiques des données d'enquêtes, comme l'utilisation de plans d'échantillonnage complexes, le suivi périodique des sujets, l'absence ou l'inexactitude des données et l'érosion de l'échantillon au fil du temps. À titre d'exemple, nous analysons les données sur les périodes sans emploi provenant de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée par Statistique Canada.

1. INTRODUCTION

L'utilisation de données provenant d'enquêtes longitudinales pour faciliter l'interprétation des processus qui surviennent au cours de la vie, comme la scolarité, l'emploi, la fécondité, le mariage et la santé, suscite un vif intérêt. L'un des aspects de cette approche a trait à la durée des périodes que passent les individus dans certains états, comme le chômage ou le mariage. L'analyse des données sur la durée et, de façon plus générale, l'analyse des bibliographies, sont des domaines bien développés offrant nombre de méthodes normalisées (p. ex., Andersen et coll., 1993; Kalbfleisch et Prentice, 2002; Lawless, 2002). Cependant, les données recueillies grâce aux enquêtes longitudinales posent divers problèmes que les méthodes standard ne permettent pas de résoudre complètement. Notre objectif est de discuter de ce genre de problèmes et de passer en revue les moyens permettant de les résoudre.

Les populations sur lesquelles sont basées les enquêtes longitudinales ont tendance à être fort hétérogènes et les individus sont habituellement échantillonnés selon un plan complexe comportant une stratification et une mise en grappes. Même si l'analyse vise directement une sous-population ou un processus particulier, l'utilisation d'un plan de sondage complexe pose des difficultés, comme nous le montrons plus loin. La façon dont les données longitudinales sont recueillies peut aussi causer des problèmes. Ainsi, le suivi périodique (p. ex. annuel) des individus peut donner lieu à des données manquantes ou incorrectes et rendre difficile la définition des temps (dates) de censure pour les personnes perdues de vue lors des suivis (érosion), le processus d'érosion ne peut pas toujours ne pas être pris en compte, le suivi après une courte période ne fournit que des renseignements limités sur les processus biographiques qui évoluent sur une longue période et, souvent, les facteurs variant avec le temps qui influent sur un processus ne sont pas mesurés.

Le présent article porte sur l'utilisation des données d'enquêtes longitudinales pour comprendre les événements qui surviennent au cours de la vie. Le cadre conceptuel utilisé pour la modélisation et l'analyse est le processus qui génère les individus et leur histoire. À la section 2, nous passons en revue les méthodes d'inférence concernant les données sur la durée dans ce cadre de travail, en insistant sur la nécessité de tenir compte du plan de sondage. À la section 3, nous considérons les problèmes que posent les données manquantes ou incorrectes ayant trait aux covariables. À la section 4, nous discutons de ces problèmes dans le contexte des données sur les périodes de chômage provenant de l'Enquête sur la dynamique du travail et du revenu (EDTR). À la section 5, nous tirons certaines conclusions.

[1] Université de Waterloo

2. ANALYSE DE LA DURÉE D'APRÈS DES DONNÉES D'ENQUÊTE

Supposons qu'une variable de durée *Y* (parfois appelée durée de vie ou de survie) et un vecteur de covariables *x* sont associés à des individus dans une population et que nous nous intéressons à la loi marginale de Y étant donné *x*. Puisque nombre de problèmes de durée portent sur le temps que passe un individu dans un état particulier, il sera commode ici de faire correspondre Y à la durée d'une période passée dans l'état représenté par *A*. En général, on s'attend à un certain degré d'association entre les *Y*, étant donné les *x* correspondants, pour les individus appartenant à certains groupes ou grappes de la population. Nous tiendrons compte de ce fait dans l'analyse qui suit, mais nous ne modéliserons pas l'association explicitement.

Supposons que la loi de Y sachant que *x* a une densité de probabilité $f(y|x;\theta)$. Si nous disposons d'un échantillon aléatoire simple *S* d'individus provenant de la population, nous pourrions fonder les inférences au sujet des paramètres du modèle θ sur la fonction de pseudo-vraisemblance

$$L(\theta)=\prod_{i\in s} f(y_i \mid x_i;\theta), \tag{2.1}$$

en supposant provisoirement que les durées y_i sont observées complètement. Cependant, habituellement, la population d'enquête, qui comprend *N* individus, est divisée en strate U_1, U_2, ..., U_H qui contiennent des grappes d'individus constituant les unités primaires d'échantillonnage (UPE) et l'échantillon d'enquête S est obtenu par la sélection aléatoire d'UPE dans chaque strate, puis par la sélection aléatoire d'individus dans chaque UPE. Utilisons la notation $R_i = I(i \in S)$ pour indiquer si l'individu i (i = 1, ..., *N*) est compris dans l'échantillon et représentons par le vecteur z_i les facteurs liés au plan de sondage, comme les renseignements sur les strates et les grappes, de sorte que les probabilités d'inclusion fondées sur le plan de sondage π_i satisfassent à

$$\pi_i = \Pr(R_i = 1 \mid y_i, x_i, z_i) = \Pr(R_i = 1 \mid z_i) \qquad i = 1,..., N\ . \tag{2.2}$$

Notons que x_i et z_i peuvent avoir certaines composantes en commun. Si Y_i et Z_i sont indépendants, sachant x_i, alors $\Pr(y_i \mid x_i, R_i = 1) = f(y_i \mid x_i)$ et l'estimation de θ peut être fondée sur (2.1) ou sur la fonction d'estimation du pseudo-score correspondante $U(\theta) = \partial \log L(\theta) / \partial\theta$, ou

$$U(\theta)=\sum_{i=1}^{N} R_i \frac{\partial \log f(y_i \mid x_i;\theta)}{\partial\theta} = \sum_{i=1}^{N} R_i U_i(\theta). \tag{2.3}$$

Nous appelons U(θ) une fonction de pseudo-score, car les Y_i ($i \in S$) ne sont généralement pas indépendants, de sorte que *(2.1)* n'est pas leur densité de probabilité conjointe. Cependant, la résolution de U(θ) = 0 donne un estimateur convergent $\hat{\theta}$ de θ sous des conditions faibles.

Si $Pr(y_i|x_i,z_i) \neq Pr(y_i|x_i)$, alors $Pr(y_i|x_i,R_i = 1) \neq f(y_i|x_i)$ et, en général, (2.3) produit des estimations non convergentes de θ. Dans ce cas, nous pouvons utiliser une fonction de pseudo-score pondérée

$$U_W(\theta)=\sum_{i=1}^{N} \frac{R_i}{\pi_i} \frac{\partial \log f(y_i \mid x_i;\theta)}{\partial\theta} = \sum_{i\in s} w_i U_i(\theta) \tag{2.4}$$

pour l'estimation (p.ex. Pfefferman, 1993; Thompson, 1997, sec. 6.2). La résolution de $U_W(\theta) = 0$ donne un estimateur $\hat{\theta}$ qui est convergent pour θ et dont la variance est plus importante que celle des estimateurs fondés sur (2.3) quand les deux sont convergents.

Pfefferman *(1993)* donne une synthèse intéressante de l'estimation pondérée contre non pondérée et des utilisations possibles du modèle $f(y|\ x;\theta)$ lorsque *Y* n'est pas indépendant des facteurs liés au plan de sondage Z, étant donné *x*. Nous supposons, pour la plupart de la discussion présentée ici, que les covariables x contiennent suffisamment de renseignements pour qu'il soit possible de ne pas tenir compte du plan d'échantillonnage et d'utiliser (2.3).

Cependant, il est facile d'étendre la méthode afin d'inclure des poids, comme dans (2.4).

2.1 Estimation pour les lois paramétriques de durée

Supposons que *Y* sachant *x* est caractérisé par une fonction de survie S(y | x; θ) = *Pr(Y* ≥ y | x; θ) et uen fonction de risque λ(y|x; θ) = *f(y|x; θ)/S(y|x; θ*). Supposons aussi que l'individu *i* est suivi durant la période de temps calendaire *(L*$_i$*, R*$_i$*)* et considérons l'observation du séjour dans un état A couvrant la période de temps *(t*$_i$*, t*$_i$ +$\tilde{y}_i$) qui chevauche *(L*$_i$*, R*$_i$*).* Si $t_i \le L_i$,alors la durée de l'épisode $\tilde{y}_i$ satisfait $\tilde{y}_i \ge L_i - t_i$ et est tronquée à gauche à $v_i = L_i - t_i \ge 0$. Si $t_i + \tilde{y}_i > R_i$ alors nous observons uniquement que $\tilde{y}_i > R_i - t_i$ et $\tilde{y}_i$ est censuré à droite à $c_i = Ri - t_i$. En utilisant la notation $y_i = \min(\tilde{y}_i, c_i)$ et $\delta_i = I(y_i = \tilde{y}_i)$, dans (2.3), la composante $U_i(\theta)$ du pseudo-score est alors remplacée par (p.ex., Lawless *2003*, sec. 5.2)

$$U_i(\theta) = \delta_i \frac{\partial \log \lambda(y_i \mid x_i; \theta)}{\partial \theta} - \frac{\partial}{\partial \theta} \int_{v_i}^{y_i} \lambda(t \mid x_i; \theta) dt \,. \tag{2.5}$$

En nous fondant sur le plan d'échantillonnage décrit plus haut, nous pouvons écrire la fonction d'estimation du pseudo-score (2.3) avec les composantes (2.5) sous la forme

$$U(\theta) = \sum_{h=1}^{H} \sum_{c=1}^{C_h} \sum_{i \in s_{hc}} U_i(\theta), \tag{2.6}$$

où C_h est le nombre de grappes ou d'UPE tirées à partir de la strate *h* et S_{hc} représente les individus échantillonnés à partir de la grappe *c* dans la strate *h.* Nous supposons que les durées associées aux individus compris dans des grappes distinctes sont indépendantes. Sous des conditions faibles, l'estimateur $\hat{\theta}$ obtenu en résolvant U(θ) = 0 est alors asymptotiquement normal et caractérisé par une moyenne θ et une matrice des covariances estimée de façon convergente par

$$V(\theta) = I\left(\hat{\theta}\right)^{-1} \hat{Var}(U(\theta)) I\left(\hat{\theta}\right)^{-1}, \tag{2.7}$$

où $I(\theta) = -\partial U(\theta)/\partial \theta'$ et

$$\hat{Var}(U(\theta)) = \sum_{h=1}^{H} \sum_{c=1}^{C_h} \left(\sum_{i \in s_{hc}} U_i\left(\hat{\theta}\right) \right) \left(\sum_{i \in S_{hc}} U_i\left(\hat{\theta}\right) \right)' . \tag{2.8}$$

Cela tient compte de l'association entre les durées liées aux individus appartenant à une même grappe. Les logiciels existants pour des modèles du temps accéléré de défaillance ou d'autres modèles paramétriques peuvent être utilisés pour obtenir $\hat{\theta}$ et $I\left(\hat{\theta}\right)$;une programmation supplémentaire est généralement nécessaire pour calculer (2.8).

2.2 Modèle semi-paramétrique de Cox

L'utilisation du modèle des risques proportionnels ou multiplicatifs de Cox (1972) est très répandue. La fonction de risque de *Y* sachant *x* est de la forme

$$\lambda(y \mid x) = \lambda_0(y) \exp(\beta' x) \tag{2.9}$$

où $\lambda_0(y)$ est une fonction de risque « de base » et β est un vecteur de coefficients de régression. Les méthodes bien connues de la vraisemblance partielle fournissent des estimations de β et des estimations non paramétriques de $\Lambda_0(y) = \int_0^y \lambda_0(t) dt$ sous échantillonnage aléatoire simple, ainsi que des scénario d'observation permettant la troncature à gauche ou la censure à droite (p. ex., Kalbfleisch et Prentice, 2002 ch. 4).

Les méthodes standard peuvent être étendues au traitement de l'association des durées à l'intérieur des grappes (p.ex. Lee et coll., 1992). Spiekerman et Lin (1998) et Boudreau et Lawless (2001) ont montré comment tenir compte à la fois des effets de grappe et de strate et des covariables variant avec le temps. Par exemple, dans certaines conditions, il est intéressant d'utiliser le modèle de Cox stratifié à fonctions de risque

$$\lambda\left(t \mid x_i, i \in U_h\right) = \lambda_{0h}(t)\exp\left(\beta' x_i(t)\right) \tag{2.10}$$

L'estimation de β et des fonctions de risque cumulé $\Lambda_{0h}(t)$ peut se fonder sur des fonctions d'estimation standard appliquées en cas d'échantillonnage aléatoire simple (Kalbfleisch et Prentice 2002, sec. 4.4). Boudreau et Lawless (2001) calculent des estimations de la variance analogues à (2.7) et démontrent qu'il est facile de les obtenir en utilisant la fonction coxph du logiciel S-Plus, en choisissant les options grappe et strate. Leur méthodologie permet également d'inclure une pondération. Nous donnons un exemple à la section 4.

Notons que Binder (1992) et Lin (2000) ont discuté de l'estimation pondérée sous le modèle de Cox. Ces auteurs ont considéré l'un et l'autre l'estimation de la variance fondée sur le plan de sondage et Lin a, en outre, discuté de l'estimation fondée sur un modèle comme nous le faisons ici. Lin permet l'utilisation de plans de sondage plus généraux que ceux examinés ici, mais ne tient pas compte de l'association entre les durées au niveau de la population. Les estimations de la variance produites par notre méthode (avec extension de (2.8) – (2.9) pour inclure la pondération) et par la sienne (en utilisant un logiciel tel que SUDAAN pour la partie fondée sur le plan de sondage de la variance) sont très proches dans les conditions que nous avons étudiées.

3. DONNÉES MANQUANTES OU INCORRECTES

3.1 Conditions initiales et troncature à gauche

Lors d'une enquête longitudinale, on suit habituellement un individu au cours d'une période déterminée (L_i, R_i) mais l'analyse peut nécessiter des renseignements sur l'individu antérieurs au temps L_i. Il en est, par exemple, ainsi quand un individu se trouve déjà dans l'état A et que l'on s'intéresse à la durée Y de la période qu'il passe dans cet état. Si l'on connaît le temps $t_i \leq L_i$ auquel un individu entre dans l'état en question, alors, à condition que t_i et Y_i soient indépendants, la loi pertinente de la durée de la durée de la période est la densité de probabilité tronquée à gauche $f(y)/S(vi)$, où $v_i = L_i - t_i$, ce qui donne la fonction de pseudo-score (pseudo-vraisemblance) (2.5). Dans certains cas, la distribution de Y_i peut dépendre du temps t_i d'entrée dans l'état A, auquel cas on remplace $f(y)$ et $S(y)$ par $f(y|t)$ et $S(y|t)$. Toutefois, si Y_i et t_i dépendent de facteurs non mesurés, alors (2.5) ne convient plus (p.ex. Keiding, 1992; Lawless et Fong, 1999). L'utilisation de poids de sondage ne permet pas de remédier à cette situation. Les problèmes de « conditions initiales » de ce genre ont fait l'objet de discussions considérables en économie et en sciences sociales (p.ex. Heckman et Singer, 1985; Hoem, 1985).

Il arrive aussi que l'on ne connaisse pas le temps d'entrée t_i dans l'état A ou qu'on le détermine incorrectement. Dans le premier cas, représentons par $g(v)$ la densité de probabilité de $V_i = L_i - t_i$, sachant qu'un individu est dans l'état A au temps L_i et en supposant que Y_i et t_i sont indépendants. Par souci de simplicité, nous supprimons la dépendance des covariables dans la notation. Si t_i et, donc, v_i ne sont pas observés, la densité de probabilité appropriée pour les données observées est

$$f^*(w_i) = \int_0^\infty g(v)\frac{f(v+w_i)}{S(v)}dv\,, \tag{3.1}$$

où $w_i = t_i + y_i - L_i$ est la durée de la période encore passé dans l'état A auprés le temps L_i. Nous supposons qu'on observe soit w_i soit un temps de censure pour cet état.

Un modèle de $g(v)$ est donc nécessaire lorsque t_i n'est pas observé. On peut parfois le spécifier d'après le taux de passage par l'état A au fil du temps si celle-ci est connue. Parfois, un argument est offert pour justifier l'hypothèse que $g(v) = S(v)/\mu$, où $\mu = E(Y)$ est la durée moyenne de la période passée dans l'état A, ce qui correspond à un taux de passage par l'état A constante au cours du temps et peut également être obtenu par des arguments d'équilibre

dans un processus de renouvellement. Si $g(u) = S(v)/\mu$, alors (3.1) se réduit à $S(wi)/\mu$, qui est la distribution bien connue de la durée de retour (*forward recurrence time distribution*) en théorie du renouvellement (jp.ex., Ross, 1983). Notons qu'il est possible de tester un modèle hypothétique pour g(v). En particulier, s'il existe des individus pour lesquels la période Y_i passée dans l'état A n'est pas tronquée, on peut obtenir pour ceux-ci une estimation $\hat{S}(y)$. La distribution empirique des w_i pour les périodes tronquées à gauche peut alors être comparée à $\hat{S}(y)/\hat{\mu}$,ou, de façon plus générale, à la distribution impliquée par (3.1).

S'il n'existe pas de modèle approprié pour *g(v),* une méthode acceptable consiste simplement à éliminer les y_i tronqués à gauche, et à utiliser uniquement les périodes passées dans l'état A qui débutent au moment où un individu commence à faire l'objet d'un suivi (p.ex., Aalen et Husebye, 1991). Boudreau et Lawless (2002) examinent la perte d'information que cause cette méthode lorsque (3.1) est effectivement disponible. Un problème pratique tient au fait que, dans certaines conditions, la plupart des observations peuvent être sujettes à un certain degré de troncature à gauche.

Si les valeurs de t_i enregistrées sont entachées d'une erreur de mesure due à un problème de remémoration ou à d'autres facteurs, une approche consiste à modéliser l'erreur de mesure ainsi que la distribution conjointe $g_1(t_i)\, f(y_i) / S(v_i)$ de t_i et Y_i pour obtenir une pseudo-vraisemblance fondée sur les données observées. Toutefois, l'exercice peut être très difficile et il faut tenir compte de la dépendance éventuelle entre l'erreur de mesure et v. Au lieu d'intégrer une modélisation formelle dans le processus d'estimation, on pourrait réaliser une analyse de sensibilité pour évaluer l'effet de l'erreur de mesure sur les estimés. Lawless (2003, sec. 8) donne un exemple portant sur la remémoration de la date à laquelle s'est terminé l'allaitement maternel des nourrissons.

3.2 Effets de l'observation intermittente

Si les membres du panel longitudinal sont interviewés à intervalle de temps assez grand, l'exactitude du moment où se sont produits les événements et celle d'autres variables relatives à la période écoulée depuis l'entrevue précédente doit être prise en considération. On observe parfois des effets de troncature, quand des individus ont tendance à situer les événements dans le voisinage du moment des entrevues (p. ex., Kalton et Miller, 1991). Les problèmes que posent les données manquantes ou incorrectes sur les covariables sont très difficiles à résoudre dans le cas des données sur la durée, particulièrement quand le processus de censure est relié aux covariables (p. ex., Kalbfleisch et Prentice, 2002, ch. 11). Dans nombre de cas, la modélisation des covariables et des processus de censure est inévitable. Les méthodes utilisées couramment en échantillonnage, comme l'imputation, n'ont pas été étudiées en détail, mais nécessiteraient aussi des hypothèses fortes concernant la censure. Ce domaine devrait être étudié plus en profondeur.

La question des données manquantes sur le moment des événements ou sur les durées parce que l'observation est intermittente peut être résolue plus facilement, mais une modélisation supplémentaire est parfois nécessaire. Par exemple, si l'entrée dans l'état A a eu lieu à un moment inconnu entre les dates de l'entrevue précédente et de l'entrevue courante, alors il faut spécifier un modèle du temps d'entrée, tel que discuté à la section 3.1.

L'observation intermittente peut aussi influer sur la validité des hypothèses standard de censure. Par exemple, si un individu est en train de vivre une période dans l'état A à l'année d'interview t, puis est perdu de vue à l'année de suivi t + 1, il est courant, en pratique, d'attribuer un temps de censure en se fondant sur la durée dans l'état A à l'année d'entrevue t. Ceci exige, toutefois, que le mécanisme de perte de vue de la personne soit indépendant du processus individuel de durée sur la période qui suit l'entrevue, étant donné les covariables mesurées jusqu'à l'année t. Cette hypothèse est vraisemblablement violée dans de nombreuses conditions. Le seul moyen de résoudre nettement ce problème consiste à dépister certains individus perdus de vue lors du suivi (p. ex. Farewell et coll., 2002), mais on peut recourir à l'analyse de sensibilité pour évaluer l'effet d'un mécanisme de perte de vue durant le suivi hypothétiquement dépendant.

4. EXEMPLE : PÉRIODES DE CHÔMAGE DANS L'EDTR

Nous allons maintenant illustrer brièvement les problèmes susmentionnés à l'aide des données de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée par Statistique Canada. Nous considérerons les données recueillies auprès du premier panel, dont le suivi a débuté en janvier 1993 et s'est poursuivi pendant six ans (sauf en cas de perte de vue durant le suivi). La sélection des membres du panel se fonde sur un plan d'échantillonnage stratifié à plusieurs degrés en vertu duquel les UPE sont des groupes de ménages formant des secteurs de dénombrement (SD). L'échantillonnage comporte la sélection aléatoire de ménages dans les UPE échantillonnées; dans chaque ménage échantillonné, tous les individus de plus de 16 ans font partie du panel longitudinal de l'enquête. Les membres du panel sont interviewés annuellement en janvier. Durant l'interview, on leur demande de fournir des renseignements sur les événements survenus au cours de l'année civile précédente. Des renseignements détaillés sur l'EDTR figurent sur le site Web de l'enquête qui peut être consulté à partir de http://www.statcan.ca.

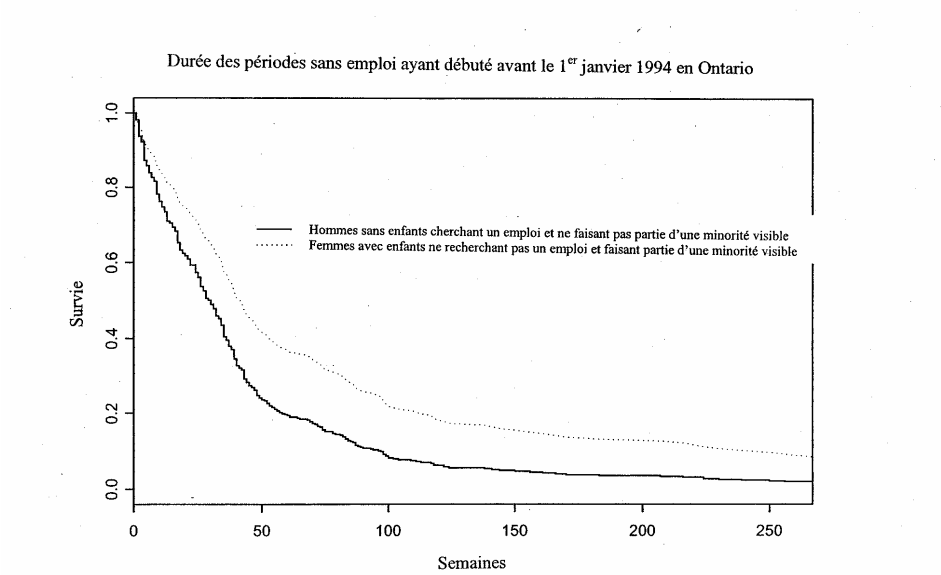
L'exemple donné ici porte sur la durée des périodes sans emploi des personnes faisant partie de la population active et ne souffrant pas d'une incapacité de longue durée. Nous discutons des problèmes que posent les données manquantes ou incorrectes plus bas, mais nous commençons par donner un exemple d'analyse semi-paramétrique des risques proportionnels (RP) de la durée des périodes de chômage qui ont débuté durant l'année civile 1993. Ces données portent sur 2 313 périodes observées, dont 104 étaient censurées à droite.

Le tableau 1 donne les résultats pour un modèle comprenant neuf covariables : SEX (= 1 si féminin, 0 si masculin); LOOKED (= 1 si la personne est à la recherche de travail, 0 autrement); CHILDREN (= 1 si la personne a au moins un enfant, 0 autrement); AGE (= âge en janvier 1993); SCHOOL (= nombre d'années de scolarité); HOURWAGE (= salaire horaire en dollars); EI (= 1 si la personne reçoit des prestations d'assurance-emploi, 0 autrement); MINORITY (= 1 si la personne est membre d'un groupe de minorités inclusivement visibles, 0 autrement); WINTER (= 1 si la personne a perdu son emploi antérieur entre novembre et avril). Nous tenons également compte des interactions bifactorielles entre ces covariables. Nous avons ajusté un modèle à risques proportionnels de Cox stratifié (Kalbfleisch et Prentice, 2002, section 4.4) où les provinces forment les strates. Les estimations et les erreurs-types calculées en s'appuyant sur les fonctions d'estimation pondérées et non pondérées (Boudreau et Lawless, 2001) sont présentées au tableau 1. Les estimations pondérées ont été calculées en se servant des poids d'échantillonnage individuels fournis dans les fichiers de données de l'EDTR qui tiennent compte à la fois du plan d'échantillonnage et d'une correction pour l'érosion du panel. Nous présentons pour ces estimations deux ensembles supplémentaires d'erreurs-types, c'est-à-dire les erreurs-types fondées sur le plan de sondage en échantillon fini de Binder (1992) et les erreurs-types analogues de Lin (2000), qui sont fondées sur les estimations de la variance de Binder auxquelles est ajouté un terme pour le processus de superpopulation. Tant pour les estimations pondérées que non pondérées, nos erreurs-types sont calculées en considérant les ménages comme des grappes. L'utilisation de secteurs de dénombrement (SD) comme grappes a produit des erreurs-types qui s'approchaient fortement de celles-ci. Dans le cas des estimations non pondérées, nous présentons les erreurs-types « sous hypothèse d'indépendance » (et techniquement incorrectes); elles sont calculées en analysant les données non pondérées comme si elles provenaient d'un échantillon de durées mutuellement indépendantes.

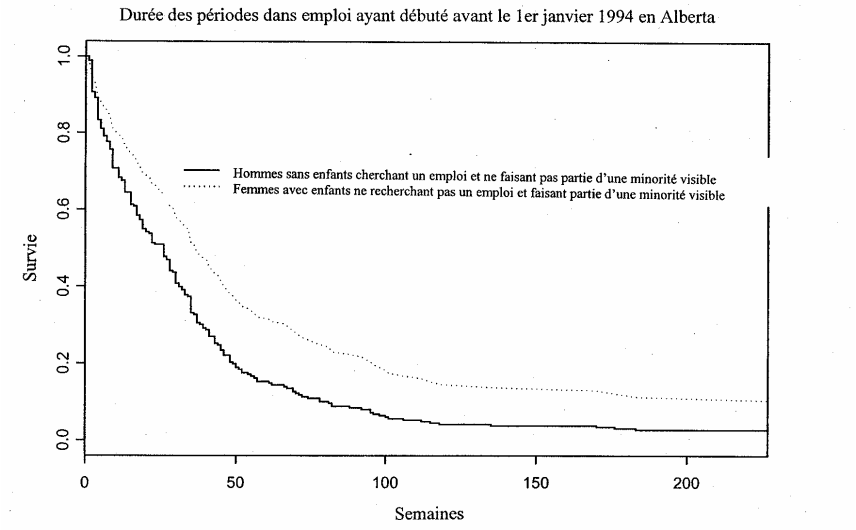
Dans le cas de l'analyse des données pondérées, les trois erreurs-types sont virtuellement identiques, indiquant que le terme de correction pour la superpopulation de Lin (2000) n'ajoute qu'une quantité très faible aux erreurs-types basées sur le plan de sondage de Binder (1992) et que nos méthodes d'estimation de la variance simples et robustes donnent de bons résultats ici. Les méthodes pondérées et non pondérées d'estimation produisent des estimations qui concordent bien dans l'ensemble; l'écart le plus important est celui observé pour la covariable LOOKED, où elles diffèrent d'un peu moins de deux erreurs-types (pondérées). Dans le cas de l'analyse des données non pondérées, les erreurs-types sous hypothèse d'indépendance diffèrent peu des erreurs-types robustes, témoignant uniquement d'un faible effet dû à l'association des durées des périodes sans emploi pour les personnes appartenant à un même ménage.

Les tests diagnostiques appliqués au modèle fondé sur l'analyse non pondérée (p. ex. Lawless 2002, section 7.2) n'indiquent qu'un seul écart significatif par rapport aux hypothèses : l'effet de la covariable WINTER diminue avec le temps. Ce résultat laisse entendre que la perte d'emploi durant l'hiver est associée à une chance croissante de retrouver un emploi au cours des 12 mois suivants, mais qu'elle influe peu sur la fonction de risque de retour à une période d'emploi après environ 12 mois sans emploi.

Les figures 1 et 2 donnent des exemples de tracés des fonctions de survie pour la durée d'une période sans emploi.



Âge=38,5 (moyenne) – Années de scolarité=11,7 (moyenne) – Salaire horaire=10,46 $ (moyenne) – Pas de prestations d'assurance-emploi



Âge=38,5 (moyenne) – Années de scolarité=11,7 (moyenne) – Salaire horaire=10,46 $ (moyenne) – Pas de prestations d'assurance-emploi

La méthode fondée inspirée de Boudreau et Lawless (2001) permet aussi de calculer des erreurs-types et des intervalles de confiance pour les probabilités de survie.

Tableau 1. Estimations et erreurs-types des effets des covariables du modèle à risques proportionnels pour les périodes sans emploi

Covariable	Estimation non pondérée	E.-T.	E.-T. (ind.)	Estimation pondérée	E.-T.	E.-T. (Lin)	E.-T. (Binder)
SEX	-0,561	0,224	0,212	-0,628	0,294	0,295	0,295
LOOKED	-0,119	0,051	0,056	0,023	0,075	0,075	0,075
CHILDREN	-0,714	0,132	0,128	-0,847	0,196	0,196	0,196
AGE	-0,025	0,006	0,005	-0,031	0,009	0,009	0,009
SCHOOL	0,001	0,009	0,010	-0,001	0,012	0,012	0,012
HOURWAGE	0,062	0,013	0,012	0,037	0,021	0,021	0,021
EI	-0,368	0,130	0,125	-0,216	0,178	0,178	0,178
MINORITY	-0,183	0,204	0,233	-0,098	0,244	0,244	0,244
WINTER	0,205	0,092	0,080	0,242	0,135	0,135	0,135
SEX*SCHOOL	0,033	0,015	0,015	0,036	0,020	0,020	0,020
SEX*CHILDREN	-0,007	0,088	0,089	-0,015	0,122	0,122	0,122
AGE*CHILDREN	0,023	0,004	0,004	0,025	0,006	0,006	0,006
SEX*HOURWAGE	-0,022	0,008	0,008	-0,008	0,012	0,012	0,012
WINTER*LOOKED	-0,100	0,105	0,097	-0,170	0,150	0,150	0,150
AGE*HOURWAGE	-0,001	0,0004	0,0003	-0,001	0,0005	0,0005	0,0005
AGE*EI	0,002	0,004	0,004	0,015	0,006	0,006	0,006

Les données de l'EDTR posent bon nombre des problèmes dont nous avons discutés à la section 3. Les durées des périodes sans emploi sont tronquées à gauche pour les membres du panel qui étaient chômeurs le 1er janvier 1993. Cependant, comme l'enquête ne fournit pas de renseignements sur le début de ces périodes, nous avons choisi ici de fonder l'estimation uniquement sur les périodes qui ont débuté après le 1er janvier 1993. Pour pouvoir utiliser également les périodes tronquées à gauche, nous devrions modéliser le taux d'entrées dans l'état de chômage pour les personnes qui sont devenues chômeuses avant 1993 et qui l'étaient encore le 1er janvier 1993.

La perte de vue durant le suivi, ou érosion du panel, est courante dans le cadre de l'EDTR. Par exemple, le panel initial de 1993 comprenait au départ environ 15 000 ménages et 31 000 personnes de plus de 16 ans. Les proportions de ces personnes qui ont été interviewées de nouveaux en janvier de 1994, 1995 et 1996 étaient de 89,6 %, 86,5 % et 85,2 %, respectivement. L'effet de l'érosion du panel sur l'estimation de la distribution des durées est plus important pour les longues durées que pour les autres. Par exemple, à la figure 1, nous voyons que certaines périodes sans emploi ont été très longues; cependant, l'estimation de cette partie de la distribution pourrait être influencée fortement par toute tendance caractérisant l'érosion du panel dont on ne peut omettre de tenir compte.

La mesure de la durée d'emploi pose divers problèmes, y compris les effets de l'erreur de remémoration et de la façon dont les débuts et fins de périodes d'emploi sont enregistrés. Dans le cadre de l'EDTR, ces derniers sont enregistrés en fonction de la semaine de l'occurrence, si bien que l'effet de groupement est faible pour les durées. D'autres facteurs, y compris les effets de troncature, entrent également en jeu (p. ex., voir Cotton et Giles, 1998).

Dans le cas d'enquêtes telles que l'EDTR, la question de la pondération est assez compliquée. Il est possible de calculer le poids « longitudinal » pour chaque membre du panel initial crée en 1993, mais, comme nous l'avons mentionné plus haut, les poids utilisés pour l'analyse des durées doivent être ajustés en fonction du temps afin de tenir compte de l'érosion du panel si l'on veut éviter les effets dus à l'échantillonnage et à l'érosion dont on ne peut omettre de tenir compte. Par exemple, une période qui a débuté en 1993 et s'est terminée en 1994 aurait une date de début enregistrée comme étant janvier 1994 et une date de fin enregistrée comme étant janvier 1995. Pour simplifier l'analyse, on utiliserait probablement soit les poids calculés pour janvier 1994 soit ceux calculés pour janvier 1995,

mais ni l'un ni l'autre choix ne serait entièrement satisfaisant dans toutes les circonstances. Il convient aussi de souligner que les personnes qui se joignent aux ménages faisant partie du panel original sont interviewées durant les années subséquentes. Bien que leur poids longitudinal initial soit nul, elles peuvent être incluses dans les analyses telles que celles décrites dans le présent article.

5. CONCLUSIONS

Les méthodes types d'analyse de la survie ou de la durée peuvent être adaptées aux données d'enquêtes longitudinales grâce à l'utilisation de méthodes d'estimation robustes de la variance et, au besoin, de poids pour contrecarrer les effets de l'échantillonnage ou de l'érosion du panel dont on ne peut omettre de tenir compte. Cependant, dans certains cas, les poids doivent varier avec le temps et les méthodes appliquées pour les choisir mériteraient d'être étudiées plus en détail. Ces travaux présenteraient des liens avec ceux de Robins et d'autres (p. ex., Robins et Rotnitzky, 1995).

D'autres sujets mériteraient aussi d'être étudiés. Premièrement, le désir d'élaborer des modèles décrivant les processus qui surviennent durant la vie des individus rend nécessaire l'élaboration de tests diagnostiques applicables aux modèles de durée. Des méthodes existent pour l'analyse non pondérée standard, mais les effets de l'échantillonnage par grappe ont été peu étudiés. On pourrait aussi élaborer des méthodes fondées sur des comparaisons d'estimations pondérées et non pondérées (p. ex., Pfefferman, 1993).

L'étude des périodes répétées ou des transitions entre états chez un individu, qui est fort intéressante, soulève la question des modèles de durée multivariés et des mesures d'association. La modélisation multiniveaux représente une approche (p. ex., Bandeen-Roche et Liang, 1996), mais il serait souhaitable d'élaborer des méthodes plus simples permettant de modéliser les processus au niveau individuel, en utilisant des techniques robustes d'estimation de la variance qui permettent de tenir compte des effets de grappe de niveau plus élevé.

Comme nous l'avons mentionné précédemment, l'étude empirique et les méthodes raisonnablement simples de traitement de l'erreur de mesure, des valeurs manquantes des covariables et de l'érosion éventuellement non négligeable du panel mériteraient qu'on leur accorde une certaine attention.

Enfin, les processus qui surviennent durant la vie des individus dépendent souvent considérablement de facteurs qui varient avec le temps. Il existe des méthodes d'analyse de la durée qui intègrent des covariables variant en fonction du temps, mais certaines questions concernant le meilleur moyen de quantifier ou de décrire les processus particuliers restent néanmoins à résoudre, si l'on veut arriver à mieux comprendre leur évolution.

REMERCIEMENTS

Cette étude a été financée en partie par une subvention accordée au premier auteur par le Conseil de recherches en sciences naturelles et en génie du Canada.

RÉFÉRENCES

Aalen, O. et Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine* ***10,*** 1227-40.

Andersen, P.K., Borgan, O., Gill, R.D. et Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Bandeen-Roche, K.J. et Liang, K.-Y. (1996). Modeling failure-time associations in data with multiple levels of clustering. *Biometrika* ***83,*** 29-39.

Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika 79,* 139-147.

Boudreau, C. et Lawless, J.F. (2001). Survival analysis based on Cox proportional hazards models and survey data. University of Waterloo Dept. of Statistics and Actuarial Science Working Paper 2001-10.

Boudreau, C. et Lawless, J.F. (2002). Efficiency and robustness issues involving missing or mismeasured left-truncation times in survival analysis. Manuscript.

Cotton, C. et Giles, P. (1998). The seam effect in the survey of labour and income dynamics. Statistics Canada Income and Labour Dynamics Working Paper Series, No. 98-18.

Farewell, V.T., Lawless, J.F., Gladman, D.D. et Urowitz, M.B. (2002). Analysis of the effect of lost-to-followup on the estimation of mortality from patient registry data. Submitted to *Applied Statistics*.

Heckman, J.J. et Singer, B. (1985). Social science duration analysis. In *Longitudinal Analysis of Labor Market Data*. Cambridge University Press, Ch. 2, eds. J.J. Heckman and B. Singer.

Hoem, J. (1985). Weighting, misclassification, and other issues in the analysis of survey samples of life histories, In *Longitudinal Analysis of Labor Market Data*. Cambridge University Press, Ch. 5, eds. J.J. Heckman and B. Singer.

Kalbfleisch, J.D. et Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data,* 2nd edition. John Wiley and Sons, New York.

Kalton, G. et Miller, M.E. (1991). The seam effect with social security income in the Survey of Income and Program Participation. J. *Official Statistics 7,* 235-245.

Keiding, N. (1992). Independent delayed entry. In *Survival Analysis: State of the Art,* 309-326. J.P. Klein and P.K. Goel, editors. Kluwer, Dordrecht.

Lawless, J.F. (2002). *Statistical Models and Methods for Lifetime Data,* 2nd edition. John Wiley and Sons, New York.

Lawless, J.F. (2003). Event history analysis and longitudinal surveys. To appear in *Analysis of Survey Data,* eds. R.L. Chambers and C.J. Skinner. John Wiley and Sons, Chichester.

Lawless, J.F. et Fong, D.Y.T. (1999). State duration models in clinical and observational studies. *Statist. Med.* ***18,*** 2365-2376.

Lin, D.Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika 87,* 37-47.

Pfefferman, D. (1993). The role of sampling weights when modeling survey data. *Int. Statist. Rev.* ***61,*** 317-337.

Robins, J.M. et Rotnitzky, A. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika,* ***82,*** 805-820.

Ross, S.M. (1983). *Stochastic Processes*. John Wiley and Sons, New York.

Spiekerman, C.F. et Lin, D.Y. (1998). Marginal regression models for multivariate failure time data. *J. Amer. Statist. Assoc.* ***93,*** 1164-1175.

Thompson, M.E. (1997). *Theory of Sample Surveys*. Chapman and Hall, London.