Techniques d'enquête, juin 2007
Vol. 33, N° 1, pp. 15-25
Statistique Canada, N° 12-001 au catalogue

Modélisation des durées de périodes multiples à partir de données d'enquête longitudinale

Milorad S. Kovačević et Georgia Roberts [1]

Résumé

Nous étudions certaines modifications du modèle de Cox à période unique classique afin de traiter les périodes multiples chez une même personne lorsque les données sont recueillies dans le cadre d'une enquête longitudinale à plan d'échantillonnage complexe. L'une des modifications est l'utilisation d'une approche fondée sur le plan de sondage pour l'estimation des coefficients du modèle et de leurs variances; dans l'estimation de la variance, chaque individu est traité comme une grappe de périodes, ce qui ajoute un degré supplémentaire de mise en grappes dans le plan de sondage. D'autres modifications du modèle ont pour but de rendre souple la spécification du risque de base afin de tenir compte de la dépendance différentielle éventuelle du risque à l'égard de l'ordre et de la durée des périodes successives, et de tenir compte aussi des effets différentiels des covariables sur les périodes de différents ordres. Ces approches sont illustrées en utilisant des données provenant de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée au Canada.

Mots clés : Régression de Cox; inférence fondée sur le plan de sondage; inférence fondée sur un modèle; ordre de la période; EDTR.

1. Introduction

Le problème de modélisation abordé dans le présent article est connu sous divers noms, tels que modélisation des temps de défaillance corrélés, modélisation multivariée de la survie, modélisation de périodes multiples ou problème d'événements récurrents. Il a été étudié dans la littérature biomédicale (par exemple, Lin 1994, Hougaard 1999), sociale (Blossfeld et Hamerle 1989, Hamerle 1989) et économique (Lancaster 1979, Heckman et Singer 1982). Généralement, ce type de modélisation est requis pour résoudre les problèmes qui se posent dans les études du temps écoulé jusqu'à l'événement, lorsque deux événements ou plus surviennent chez le même sujet et que le but de la recherche est d'évaluer l'effet de diverses covariables sur la durée d'une période dans un état particulier. Les temps de défaillance sont corrélés chez un sujet donné et, donc, l'hypothèse d'indépendance des temps de défaillance conditionnellement à des covariables mesurées que requièrent les modèles de survie standard est vraisemblablement violée. Dans les études de durée des périodes (de pauvreté, de chômage, *etc.*), la « défaillance » équivaut à la sortie de l'état d'intérêt. Une propriété supplémentaire d'un grand nombre de périodes multiples, souvent ignorée, est que les périodes sont des « événements » ordonnés; autrement dit, la deuxième période ne peut pas survenir avant la première, et ainsi de suite. Le présent article a été motivé par une étude des périodes de chômage dont il est discuté plus en détail à la section 5.

L'interdépendance des périodes survenant chez un même individu est due au fait qu'elles ont en commun certaines caractéristiques inobservées de l'individu. L'effet de ces caractéristiques inobservées peut être modélisé explicitement sous forme d'un effet aléatoire (par exemple, Clayton et Cuzick 1985). Le cas échéant, il est supposé que l'effet aléatoire suit une loi statistique connue. La loi gamma de moyenne 1 et de variance inconnue est la loi privilégiée dans de nombreuses applications. Ensuite, des estimations des effets aléatoires et fixes peuvent être obtenues par une méthode appropriée (par exemple, vraisemblance en deux étapes (Lancaster 1979), en utilisant un algorithme EM (Klein 1992), *etc.*). Cette approche n'est pas explorée dans le présent article.

Une autre approche, qui est celle que nous utiliserons, consiste à adopter une méthode semi-paramétrique dans laquelle nous ne modélisons pas explicitement l'interdépendance des périodes multiples. Nous modélisons les lois marginales des périodes individuelles, en utilisant éventuellement l'ordre des périodes dans la spécification du modèle. Dans un contexte autre que les sondages, Lin (1994) décrit comment il est suffisant de modifier simplement la matrice de covariance « naïve » des coefficients estimés du modèle obtenue sous l'hypothèse d'indépendance, puisque les durées corrélées doivent être prises en compte dans les estimations de la variance, mais non dans les estimations des coefficients proprement dit.

Dans les études socioéconomiques des durées des périodes, les sources de données sont fréquemment des enquêtes longitudinales à plan d'échantillonnage complexe comportant une stratification, un échantillonnage à plusieurs degrés, la sélection avec probabilités inégales, des corrections stochastiques de l'érosion et de la non-réponse, le

1. Milorad S. Kovačević, conseiller en recherche méthodologique, Statistique Canada, Ottawa, Canada, K1A 0T6. Courriel : kovamil@statcan.ca; Georgia Roberts, chef du Centre de ressources en analyse de données à la Division des méthodes d'enquête sociales, Statistique Canada, Ottawa, Canada, K1A 0T6. Courriel : robertg@statcan.ca.

calage sur des paramètres connus, *etc*. Par conséquent, il est nécessaire de tenir compte de l'effet du plan d'échantillonnage sur la distribution des données d'échantillon lors de l'estimation des paramètres du modèle et des variances de ces estimations. Notre approche, lors de l'analyse de données d'enquête complexes, consiste à modéliser les lois marginales des périodes multiples selon des méthodes pour période unique, en traitant la dépendance entre les périodes comme une perturbation - aussi bien la dépendance due à la corrélation des périodes chez la même personne que la dépendance entre individus due au plan de sondage - mais à tenir compte des probabilités de sélection inégales à l'aide des poids de sondage. En fonction du modèle choisi, les paramètres de population finie sont définis et estimés comme dans Binder (1992). Les erreurs-types sont estimées par une méthode appropriée de linéarisation convergente sous le plan en posant l'hypothèse que les unités primaires d'échantillonnage sont échantillonnées avec remise dans les strates. Cette hypothèse est valide lorsque les fractions d'échantillonnage de premier degré sont faibles, comme cela est généralement le cas dans les enquêtes socioéconomiques. En outre, pour ce genre d'échantillon, la différence entre les inférences en population finie et en superpopulation (c'est-à-dire les erreurs-types et les statistiques de test) s'avère assez négligeable (Lin 2000). Par conséquent, les résultats de l'inférence basée sur notre approche peuvent s'étendre au-delà de la population finie étudiée.

À la section suivante, nous passons en revue la modélisation de périodes uniques et certaines méthodes d'estimation robuste des variances lorsque le modèle est spécifié incorrectement, d'abord dans un cadre fondé sur un modèle, puis dans un cadre fondé sur le plan de sondage. À la section 3, nous discutons plus en détail de l'estimation robuste de la variance pour des périodes multiples. À la section 4, nous introduisons trois modèles pour périodes multiples et décrivons comment les ajuster en utilisant des méthodes d'estimation robuste sous le plan de sondage. À la section 5, nous ajustons ces modèles aux données de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée au Canada et discutons des résultats. Enfin, à la section 6, nous présentons certaines remarques générales.

2. Inférence pour le modèle de taux de risque à période unique

La durée d'une période (ou simplement, une période) vécue par un individu est une variable aléatoire dénotée par T. Nous nous intéressons particulièrement à la fonction de risque $h(t)$ de T au temps t, définie comme étant le risque instantané d'achèvement de la période au temps t sachant qu'elle n'a pas été achevée avant le temps t, exprimée formellement par

$$h(t) = \lim_{dt \to 0} \frac{\text{Prob}\{t \le T < t + dt \mid T \ge t\}}{dt}.$$

La valeur de la fonction de risque au temps t est appelée taux de sortie afin de mettre l'accent sur le fait que l'achèvement de la période équivaut à la sortie de l'état d'intérêt. Les modèles de durée et l'analyse de la durée en général sont formulés et discutés en termes de la fonction de risque et de ses propriétés.

Du point de vue d'un spécialiste du domaine, l'intérêt principal est souvent d'étudier l'effet de certaines covariables clés sur la distribution de T. Un modèle à risques proportionnels est fréquemment choisi pour ce genre d'étude. Sous le modèle à risques proportionnels, la fonction de risque de la période T, étant donné un vecteur de covariables variant éventuellement en fonction du temps $\mathbf{x}(t) = (x_1(t), \ldots, x_p(t))'$ est

$$h(t \mid \mathbf{x}(t)) = \lambda_0(t)\, e^{\mathbf{x}'(t)\boldsymbol{\beta}}. \tag{1}$$

La fonction $\lambda_0(t)$ est une fonction de risque de base non spécifiée qui donne la forme de $h(t \mid \mathbf{x}(t))$. Le risque de base décrit la dépendance de la durée, comme préciser si le taux de risque dépend du temps déjà écoulé dans la période. Par exemple, une dépendance négative décrit la situation où la probabilité de sortie d'un état est d'autant plus faible que la période est longue. Si la valeur de toutes les variables $\mathbf{x}(t)$ d'un individu est fixée à 0, la valeur (niveau) de la fonction de risque est égale au risque de base.

2.1 Inférence fondée sur le modèle

Le vecteur $\boldsymbol{\beta}$ contient les paramètres de régression inconnus montrant la dépendance du risque par rapport au vecteur $\mathbf{x}(t)$ et peut être estimé en maximisant la fonction de vraisemblance partielle (Cox 1975) :

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{e^{\mathbf{x}_i'(T_i)\boldsymbol{\beta}}}{\sum_{j=1}^{n} Y_j(T_i)\, e^{\mathbf{x}_j'(T_i)\boldsymbol{\beta}}} \right]^{\delta_i}. \tag{2}$$

Ici, $T_1, \ldots, T_n$ représentent n durées éventuellement censurées à droite; $\delta_i = 1$ si T_i est une durée observée et $\delta_i = 0$ autrement; et $\mathbf{x}_i(t)$ est le vecteur de covariables correspondant observé sur $[0, T_i]$. Au dénominateur, la somme est calculée sur les périodes qui risquent d'être achevées au temps T_i, c'est-à-dire $Y_j = 1$ si $t \le T_j$, et égale à 0 autrement. L'estimation $\hat{\boldsymbol{\beta}}$ du paramètre $\boldsymbol{\beta}$ du modèle est obtenue en résolvant l'équation de score de vraisemblance partielle

$$U_0(\boldsymbol{\beta}) = \sum_{i=1}^{n} u_{i0}(T_i, \boldsymbol{\beta}) = 0, \tag{3}$$

où

$$u_{i0}(T_i,\ \boldsymbol{\beta}) = \delta_i \left\{ \mathbf{x}_i(T_i) - \frac{S^{(1)}(T_i, \boldsymbol{\beta})}{S^{(0)}(T_i, \boldsymbol{\beta})} \right\}, \qquad (4)$$

$$S^{(0)}(t,\ \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}, \qquad (5)$$

et

$$S^{(1)}(t, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\, \mathbf{x}_i(t) e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}. \qquad (6)$$

Si le modèle (1) est vérifié et que les durées sont indépendantes, la matrice des variances fondées sur le modèle de la fonction de score $U_0(\boldsymbol{\beta})$ est

$$J(\boldsymbol{\beta}) = -\partial U_0(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$$
$$= \sum_{i=1}^{n} \delta_i \left\{ \frac{S^{(2)}(T_i,\ \boldsymbol{\beta})}{S^{(0)}(T_i,\ \boldsymbol{\beta})} - \frac{S^{(1)}(T_i,\ \boldsymbol{\beta})[S^{(1)}(T_i,\ \boldsymbol{\beta})]'}{[S^{(0)}(T_i,\ \boldsymbol{\beta})]^2} \right\},$$

où

$$S^{(2)}(t, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\ \mathbf{x}_i(t)\ \mathbf{x}_i'(t)\ e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}.$$

La variance approximative de $\hat{\boldsymbol{\beta}}$, obtenue par linéarisation, est $J^{-1}(\hat{\boldsymbol{\beta}})$.

Si la forme de (1) est incorrecte, mais que les observations sont indépendantes, Lin et Wei (1989) donnent l'estimateur robuste de la variance pour $\hat{\boldsymbol{\beta}}$ sous la forme

$$J^{-1}(\hat{\boldsymbol{\beta}})\, G(\hat{\boldsymbol{\beta}})\, J^{-1}(\hat{\boldsymbol{\beta}}), \qquad (7)$$

où

$$G(\boldsymbol{\beta}) = \sum_{i=1}^{n} g_i(\boldsymbol{\beta})\, g_i'(\boldsymbol{\beta})$$

et

$$g_i(\boldsymbol{\beta}) = u_{i0}(T_i, \boldsymbol{\beta})$$
$$-\sum_{j=1}^{n} \delta_j \frac{Y_i(T_j) e^{\mathbf{x}_i'(T_j)\boldsymbol{\beta}}}{n S^{(0)}(T_j, \boldsymbol{\beta})} \left\{ \mathbf{x}_i(T_j) - \frac{S^{(1)}(T_j, \boldsymbol{\beta})}{S^{(0)}(T_j, \boldsymbol{\beta})} \right\}. \qquad (8)$$

2.2 Inférence fondée sur le plan de sondage

Pour les observations provenant d'une enquête à plan d'échantillonnage complexe, Binder (1992) a utilisé une méthode de pseudovraisemblance pour estimer les paramètres d'un modèle à risques proportionnels et leurs variances dans le cas d'une seule période par individu. En particulier, il a commencé par définir le paramètre d'intérêt en population finie comme étant la solution de l'équation de score de vraisemblance partielle (3) calculée d'après les périodes de la population finie visée par l'enquête :

$$U_0(\mathbf{B}) = \sum_{i=1}^{N} u_{i0}(T_i, \mathbf{B}) = 0,$$

où $u_{i0}(T_i, \mathbf{B})$ est le résidu de score défini de la même manière que $u_{i0}(T_i, \boldsymbol{\beta})$, excepté que les moyennes dans les définitions de $S^{(0)}(t, \mathbf{B})$ et de $S^{(1)}(t, \mathbf{B})$ portent sur N plutôt que n observations. Notons que, si les membres de la population finie visée par l'enquête ne connaissent pas tous des périodes de l'état étudié, N représente la taille de la sous-population qui vit de telles périodes et la sommation est faite sur ces N individus.

Une estimation $\hat{\mathbf{B}}$ du paramètre $\mathbf{B}$ est obtenue sous forme d'une solution de l'équation d'estimation du pseudo-score partiel

$$\hat{U}_0(\hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i(s)\, \hat{u}_{i0}(T_i, \hat{\mathbf{B}}) = 0,$$

où $w_i(s) = w_i$, le poids de sondage, si $i \in s$, et 0 autrement. La fonction $\hat{u}_{i0}(T_i, \hat{\mathbf{B}})$ prend la forme

$$\hat{u}_{i0}(T_i, \hat{\mathbf{B}}) = \delta_i \left\{ \mathbf{x}_i(T_i) - \frac{\hat{S}^{(1)}(T_i, \hat{\mathbf{B}})}{\hat{S}^{(0)}(T_i, \hat{\mathbf{B}})} \right\},$$

où

$$\hat{S}^{(0)}(t, \hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i(s)\, Y_i(t) e^{\mathbf{x}_i'(t)\hat{\mathbf{B}}},$$

et

$$\hat{S}^{(1)}(t, \hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i(s)\, Y_i(t)\, \mathbf{x}_i(t)\ e^{\mathbf{x}_i'(t)\hat{\mathbf{B}}}.$$

En général, la variance sous le plan de sondage d'une estimation $\hat{\boldsymbol{\theta}}$ qui satisfait une équation d'estimation de la forme $\hat{U}(\hat{\boldsymbol{\theta}}) = \sum w_i\, u_i(\hat{\boldsymbol{\theta}}) = 0$ peut être estimée, par linéarisation, en tant que

$$\hat{J}^{-1}\, \hat{V}(\hat{U}(\hat{\boldsymbol{\theta}}))\, \hat{J}^{-1}, \qquad (9)$$

où $\hat{J} = \partial \hat{U}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ est évaluée à $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, et $\hat{V}(\hat{U}(\hat{\boldsymbol{\theta}}))$ est la variance estimée du total estimé $\hat{U}(\boldsymbol{\theta})$ obtenue par une méthode standard d'estimation de la variance fondée sur le plan (voir, par exemple, Cochran (1977)) et évaluée à $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Binder (1983) énonce qu'afin d'utiliser cette approche pour calculer une estimation convergente de la variance, $\hat{U}(\hat{\boldsymbol{\theta}})$ doit être exprimée sous la forme d'une somme de vecteurs aléatoires indépendants. Dans le cas du modèle à risques proportionnels susmentionné, $\hat{U}_0(\hat{\mathbf{B}})$ ne satisfait pas cette condition, puisque chaque $\hat{u}_{i0}$ est une fonction de $\hat{S}^{(0)}(T_i, \hat{\mathbf{B}})$ et de $\hat{S}^{(1)}(T_i, \hat{\mathbf{B}})$, qui englobent tous deux de nombreux individus outre le i^{e}. Donc, Binder (1992) a trouvé pour $\hat{U}_0(\hat{\mathbf{B}})$ une expression de rechange conforme à

ces conditions, ce qui permet d'obtenir une estimation convergente sous le plan $\hat{V}(\hat{U}_0(\hat{\mathbf{B}}))$ par application d'une méthode d'estimation de la variance fondée sur le plan à l'expression de rechange, puis l'évaluation de cette estimation de la variance à $\mathbf{B} = \hat{\mathbf{B}}$. Si la méthode d'estimation de la variance sous le plan choisie est la méthode de linéarisation, alors la première étape consiste à calculer le résidu suivant pour chacun des individus échantillonnés :

$$\hat{u}_i(T_i, \hat{\mathbf{B}}) = \hat{u}_{i0}(T_i, \hat{\mathbf{B}}) - \sum_{j=1}^{N} w_j(s)\, \delta_j \frac{Y_i(T_j)\, e^{\mathbf{x}_i'(T_j)\hat{\mathbf{B}}}}{\hat{S}^{(0)}(T_j, \hat{\mathbf{B}})} \left\{ \mathbf{x}_i(T_j) - \frac{\hat{S}^{(1)}(T_j, \hat{\mathbf{B}})}{\hat{S}^{(0)}(T_j, \hat{\mathbf{B}})} \right\}. \quad (10)$$

Chaque individu compris dans l'échantillon appartient à une UPE particulière dans une strate donnée. Donc, au lieu d'identifier un individu à l'aide d'un indice inférieur unique i, nous utiliserons un indique inférieur triple hci où $h = 1, 2, ..., H$ identifie la strate, $c = 1, 2, ..., c_h$ identifie l'UPE dans la strate et $i = 1, 2, ..., n_{hc}$ identifie l'individu dans l'UPE. Alors,

$$\hat{V}(\hat{U}_0(\hat{\mathbf{B}})) = \sum_{h=1}^{H} \frac{1}{c_h(c_h - 1)} \sum_{c=1}^{c_h} (t_{hc} - \bar{t}_h)\, (t_{hc} - \bar{t}_h)',$$

où

$$t_{hc} = c_h \sum_{i=1}^{n_{hc}} w_{hci}\, \hat{u}_{hci} \quad \text{et} \quad \bar{t}_h = \sum_{c=1}^{c_h} t_{hc} \Big/ c_h.$$

3. Inférence pour les modèles du taux de risque à périodes multiples

3.1 Inférence fondée sur le modèle

Si plus d'une période est observée pour un individu, il est raisonnable de supposer que ces périodes ne sont pas indépendantes. Donc, la fonction de vraisemblance partielle (2) est spécifiée incorrectement pour des périodes multiples, puisqu'elle ne tient pas compte de la corrélation intra-individu des périodes observées chez le même individu. En s'inspirant de Lin et Wei (1989), il suffit de modifier seulement la matrice de covariance des paramètres du modèle estimé, puisque les durées corrélées affectent la variance, tandis que les paramètres du modèle peuvent être estimés de manière convergente sans tenir compte de cette corrélation. Lin (1994) démontre comment la matrice de covariance des paramètres du modèle estimé peut être estimée en cas de corrélation intra-individu des périodes, à condition que les périodes provenant d'individus différents soient indépendantes.

3.2 Inférence fondée sur le plan de sondage

Dans une enquête longitudinale à plan de sondage à plusieurs degrés, les événements multiples peuvent être corrélés à divers niveaux : les périodes sont regroupées dans un individu et les individus sont groupés dans les unités de degré d'échantillonnage élevé. La corrélation intra-grappe positive à tout niveau ajoute aux estimations calculées à partir de ce genre de données une variation supplémentaire, outre celle attendue sous des conditions d'indépendance. L'hypothèse d'indépendance des observations lorsque celles-ci sont corrélées dans les grappes donne lieu à une sous-estimation des erreurs-types réelles, ce qui exagère les valeurs des statistiques de test et, en dernière analyse, aboutit au rejet trop fréquent des hypothèses nulles. Donc, pour les périodes multiples chez un individu, où les données proviennent d'une enquête longitudinale, il ne suffit pas de tenir compte simplement de la corrélation intra-individu.

L'estimation de la variance sous le plan de sondage dans le cas de données hiérarchiques corrélées dans les grappes peut être grandement simplifiée quand il est raisonnable de supposer que les individus provenant d'unités primaires d'échantillonnage (UPE) différentes ne sont pas corrélés. Cela équivaut à supposer que les UPE sont échantillonnées avec remise. Cette hypothèse est aussi vérifiée approximativement quand les unités de premier degré sont obtenues par échantillonnage sans remise, à condition que la fraction d'échantillonnage au premier degré soit très faible. Dans ces conditions, une estimation de la variabilité inter-UPE reflète la variabilité entre les unités à tous les degrés d'échantillonnage subséquents, quelle que soit la structure de dépendance entre les observations dans chaque UPE. Pour un résumé récent de l'estimation robuste de la variance dans le cas de données corrélées dans les grappes, voir Williams (2000). Cela implique que l'approche d'estimation robuste de la variance du modèle à période unique dans le cas d'un plan de sondage comportant un échantillonnage avec remise au premier degré décrite par Binder (1992) peut être appliquée directement au cas des périodes multiples, puisqu'elle tient compte de l'effet de la corrélation intra-grappe à tous les niveaux dans chaque UPE.

4. Trois modèles pour les périodes multiples

Afin de tenir compte de covariables ayant des effets différents pour des périodes d'ordres différents, ainsi que de dépendances de durée différentes (risques de base), nous explorons trois modèles pour les périodes multiples. Ces modèles se distinguent par la définition du risque et les hypothèses au sujet du risque de base. Deux de ces modèles tiennent compte de l'ordre des périodes.

Il convient toutefois de souligner que, dans nos travaux, l'ordre des périodes fait uniquement référence à celles survenant durant la période d'observation pour laquelle les données sont recueillies et non pas à la biographie complète d'un individu (à moins que les deux périodes ne coïncident). Par exemple, par première période, nous entendons la première période de l'état étudié survenant durant la période d'observation, alors qu'il pourrait s'agit d'une période d'ordre absolu plus élevé au cours de la vie de la personne. Cette limite implique qu'il convient d'interpréter avec précaution toute incidence que l'ordre de la période peut avoir sur les effets des covariables ou sur la dépendance par rapport au temps.

Modèle 1 : Dans le premier modèle, l'ensemble de risques est défini avec précaution afin de tenir compte de l'ordre des périodes en ce sens qu'un individu ne peut pas être exposé au risque d'achèvement de la deuxième période avant que la première soit achevée, *etc*. Ce modèle, connu sous le nom de modèle d'ensemble de risques conditionnels, a été proposé par Prentice, Williams et Peterson (1981) et revu par Lin (1994). Il a également été discuté par Hamerle (1989) et par Blossfeld et Hamerle (1989) dans le contexte de la modélisation de processus à épisodes multiples. En général, l'ensemble de risques conditionnels au temps t pour l'achèvement d'une période d'ordre j comprend tous les individus qui sont dans leur j^{e} période. Ce modèle permet que l'ordre de la période influence à la fois l'effet des covariables et la forme de la fonction de risque de base.

La fonction de risque pour le i^{e} individu pour la période de j^{e} ordre est

$$h_j(t \mid \mathbf{x}_{ij}(t)) = \lambda_{0j}(t)\, e^{\mathbf{x}'_{ij}(t)\boldsymbol{\beta}_j},$$

où, pour chaque ordre de période, une fonction de risque de base différente et un vecteur de coefficients différent sont permis. Pour ce modèle et pour d'autres que nous examinerons à la présente section, le temps t est mesuré à partir du début de la j^{e} période. Bien que les périodes chez un même individu ne soient pas nécessairement indépendantes, la vraisemblance partielle qui suit reste valide pour l'estimation des $\boldsymbol{\beta}_j$:

$$L(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K) = \prod_{j=1}^{K} \prod_{i=1}^{N_j} \left[\frac{e^{\mathbf{x}'_{ij}(T_{ij})\boldsymbol{\beta}_j}}{\sum_{r=1}^{N_j} Y_{rj}(T_{ij})\, e^{\mathbf{x}'_{ij}(T_{ij})\boldsymbol{\beta}_j}} \right]^{\delta_{ij}}. \quad (11)$$

Ici, $T_{1j}, ..., T_{N_j j}$ sont N_j durées de période de j^{e} ordre éventuellement censurées à droite, $\delta_{ij} = 1$ si T_{ij} est une durée observée et $\delta_{ij} = 0$ autrement, et K est l'ordre le plus élevé des périodes qui doivent être incluses dans le modèle de Cox. Au dénominateur, la somme est calculée sur les j^{e} périodes risquant d'être achevées au temps T_{ij}, c'est-à-dire $Y_{rj}(t) = 1$ si $t \le T_{rj}$ et est égale à 0 autrement. Le vecteur de covariables correspondant observé sur $[0, T_{ij}]$ est $\mathbf{x}_{ij}(t)$. La vraisemblance partielle (11) peut être maximisée séparément pour chaque j si aucune contrainte supplémentaire n'est appliquée aux $\boldsymbol{\beta}_j$.

Les équations de score correspondantes qui définissent le paramètre de population finie $\mathbf{B} = (\mathbf{B}'_1, \mathbf{B}'_2, ..., \mathbf{B}'_K)'$ sont :

$$U_0(\mathbf{B}) = \sum_{j=1}^{K} \sum_{i=1}^{N_j} u_{ij0}(T_{ij}, \mathbf{B}_j) = 0, \quad (12)$$

avec

$$u_{ij0}(T_{ij}, \mathbf{B}_j) = \delta_{ij} \left\{ \mathbf{x}_{ij}(T_{ij}) - \frac{S^{(1)}(T_{ij}, \mathbf{B}_j)}{S^{(0)}(T_{ij}, \mathbf{B}_j)} \right\},$$

et avec $S^{(0)}(t, \mathbf{B}_j)$ et $S^{(1)}(t, \mathbf{B}_j)$ prenant la forme (5) et (6) respectivement, mais avec N_j remplaçant n et $\mathbf{B}_j$ remplaçant $\boldsymbol{\beta}$.

Les estimations sous le plan des paramètres $\mathbf{B}_j$ sont obtenues en résolvant les équations $\sum_{i=1}^{N_j} w_i(s)\hat{u}_{ij0}(T_{ij}, \hat{\mathbf{B}}_j) = 0$ séparément pour chaque j, où $\hat{u}_{ij0}$ a la forme de u_{ij0}, mais avec $S^{(0)}$ et $S^{(1)}$ remplacés par $\hat{S}^{(0)}$ et $\hat{S}^{(1)}$ respectivement. Notons que les poids d'échantillonnage correspondent aux individus et non aux périodes. De même, l'estimation de la matrice de covariance de chaque $\hat{\mathbf{B}}_j$ se fera séparément, en utilisant la méthode d'estimation robuste sous le plan décrite à la section 2.2. Techniquement, il s'agit d'un ensemble d'analyses distinctes selon l'ordre de la période.

Modèle 2 : Le deuxième modèle étudié est le modèle marginal (Wei, Lin et Weissfeld 1989) :

$$h_j(t \mid \mathbf{x}_{ij}(\mathrm{t})) = \lambda_{0j}(t)\, e^{\mathbf{x}'_{ij}(t)\boldsymbol{\beta}},$$

où, pour chaque ordre de période, nous permettons une fonction de risque de base différente tandis que les effets des covariables sont maintenus les mêmes pour différents ordres de période. La fonction de vraisemblance partielle correspondante, ainsi que l'ensemble de risques, sous l'hypothèse que les périodes chez le même individu sont indépendantes, sont les mêmes que pour le modèle 1, avec $\boldsymbol{\beta}$ remplaçant les $\boldsymbol{\beta}_j$. L'équation de score correspondante qui définit le paramètre de population finie est

$$U_0^*(\mathbf{B}) = \sum_{j=1}^{K} \sum_{i=1}^{N_j} u^*_{ij0}(T_{ij}, \mathbf{B}) = 0,$$

avec

$$u^*_{ij0}(T_{ij}, \mathbf{B}_j) = \delta_{ij} \left\{ \mathbf{x}_{ij}(T_{ij}) - \frac{S^{(1)}(T_{ij}, \mathbf{B})}{S^{(0)}(T_{ij}, \mathbf{B})} \right\},$$

où $S^{(0)}(t, \mathbf{B})$ et $S^{(1)}(t, \mathbf{B})$ sont définis par (5) et (6) respectivement, mais avec N_j remplaçant n et $\mathbf{B}$ remplaçant $\boldsymbol{\beta}$.

L'estimation sous le plan du paramètre $\mathbf{B}$ s'obtient en résolvant les équations de score pondérées

$$\sum_{i=1}^{K} \sum_{j=1}^{N_i} w_i(s) \hat{u}^*_{ij0}(T_{ij}, \hat{\mathbf{B}}) = 0,$$

où $\hat{u}^*_{ij0}$ a la forme de u^*_{ij0}, mais avec $S^{(0)}(t, \mathbf{B})$ et $S^{(1)}(t, \mathbf{B})$ remplacés par $\hat{S}^{(0)}(t, \hat{\mathbf{B}})$ et $\hat{S}^{(1)}(t, \hat{\mathbf{B}})$ respectivement.

L'estimation de la matrice de covariance de $\hat{\mathbf{B}}$ sera faite en utilisant la méthode d'estimation robuste sous le plan expliquée à la section 3.2.

Modèle 3 : Le dernier modèle étudié est le suivant :

$$h_j(t \mid \mathbf{x}_{ij}) = \lambda_0(t)\, e^{\mathbf{x}'_{ij}(t)\boldsymbol{\beta}}.$$

Dans ce modèle, nous supposons que les fonctions de risque de base et les effets des covariables sont communs aux divers ordres de période. L'ensemble de risques au temps T_{ij} est défini autrement que dans les modèles 1 et 2, et contient toutes les périodes pour lesquelles $t \le T_{ij}$, en supposant effectivement que toutes les périodes proviennent d'individus différents. Techniquement, ce modèle est un modèle à période unique, de sorte que l'estimation des coefficients et des variances par une méthode robuste sous le plan est simple.

5. Exemple de modélisation de périodes de chômage multiples

5.1 Les données

L'ensemble de données que nous utilisons pour l'illustration provient du premier panel de six ans (1993 à 1998) de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée au Canada. Dans ce panel, environ 31 000 personnes sélectionnées dans environ 15 000 ménages ont été suivies pendant six ans par la voie d'interviews annuelles. Certaines personnes sont sorties de l'échantillon au cours du temps pour diverses raisons, tandis que quelques autres, après avoir manqué une ou plusieurs interviews, ont recommencé à participer au panel. Une pondération compliquée des personnes répondant à l'EDTR chaque année tient compte des divers types d'érosion du panel, de sorte que chaque répondant durant une année particulière soit pondéré en fonction de la population de référence pertinente de 1993. Cette approche produit une pondération longitudinale distincte pour chaque cycle (c'est-à-dire année) de collecte de données. Pour la présente analyse, nous avons utilisé les poids longitudinaux provenant de la dernière année du panel, c'est-à-dire 1998, ce qui signifie que seules les données provenant des personnes qui ont répondu au dernier cycle du panel ont été incluses dans les analyses. Un bon résumé des questions relatives au plan d'échantillonnage de l'EDTR est donné dans Lavigne et Michaud (1998). Un examen des questions en rapport avec les études des périodes de chômage basées sur l'EDTR est donné dans Roberts et Kovačević (2001).

L'état d'intérêt est « être en chômage », défini ici comme étant l'état entre une mise à pied permanente dans le contexte d'un emploi à temps plein et le commencement d'un autre emploi à temps plein. Un emploi est « à temps plein » s'il exige au moins 30 heures de travail par semaine. L'évènement d'intérêt est « la sortie du chômage ». Seules les périodes débutant après le 1er janvier 1993 ont été incluses, puisque le 31 janvier 1993 est la date de début des observations auprès du panel. Les périodes de chômage qui n'étaient pas achevées à la fin de la période d'observation (31 décembre 1998) ont été considérées comme étant censurées. Les dénombrements d'échantillon des individus connaissant des périodes admissibles et des périodes en fonction de leur ordre sont présentés au tableau 1. Brièvement, 17 880 périodes ont été dénombrées auprès de 8 401 membres du panel longitudinal. Environ la moitié des individus échantillonnés (4 260) devenus chômeurs durant cette période ont connu deux périodes de chômage ou plus. En tout, 3 809 périodes sont demeurées inachevées à cause de la cessation du panel.

Sur une longue liste de covariables disponibles, nous n'en n'avons choisies que dix. La variable de sexe [SEX] de l'individu longitudinal est la seule qui demeure constante au cours des diverses périodes. Quatre variables ont des valeurs enregistrées à la fin de l'année durant laquelle la période a commencé, à savoir le niveau de scolarité [EDUCLEV] avec quatre catégories (faible, moyen-faible, moyen, élevé), l'état matrimonial [MARST] avec trois catégories (célibataire, marié(e)/union de fait, autre), le revenu familial par personne (en dollars canadiens) avec quatre catégories (<10 000, de 10 000 à 20 000, de 20 000 à 30 000, 30 000 et plus) et l'âge [AGE] (en années). Trois variables ont les valeurs correspondant à l'emploi avec mise à pied qui a précédé la période, à savoir le type de fin d'emploi [TYPJBEND] avec deux catégories (congédié(e) et départ volontaire), la profession [OCCUPATION] avec six catégories (professionnel, administration, secteur primaire, fabrication, construction et autre) et la taille de l'entreprise [FIRMSIZE] avec cinq catégories (<20, de 20 à 99, de 100 à 499, de 500 à 999, et 1 000 et plus employés). Deux variables binaires représentent la situation durant la période, à savoir travailler à temps partiel [PARTTJB] et être aux études [ATSCH].

L'ensemble de données a été préparé selon le mode de « dénombrement » où chaque individu présentant des

périodes admissibles est représenté par un ensemble de lignes dont chacune correspond à une période. Bien qu'une ligne contienne le temps d'entrée dans la période t_1 et le temps de sortie de la période t_2 ou le temps de censure t_c, pour l'analyse, la durée est toujours considérée sous la forme $(0,\ t_2 - t_1)$ ou $(0,\ t_c - t_1)$. Les covariables examinées sont reliées à chaque ligne. Sont également reliés à chaque ligne le poids longitudinal de 1998 et les identificateurs de la strate et de l'UPE de la personne dont la période est décrite par l'enregistrement en question.

5.2 Analyse

Pour les besoins de l'illustration, nous avons limité l'analyse aux quatre premières périodes, si bien que tous les individus échantillonnés présentant des périodes admissibles sont inclus dans l'analyse, mais que les enregistrements de période survenue après la quatrième ne sont pas pris en considération à cause de leur faible nombre dans l'échantillon.

Nous avons estimé les coefficients et leurs variances pour les trois modèles par les méthodes fondées sur le plan de sondage décrites à la section 4 en utilisant la procédures « SURVIVAL » du logiciel SUDAAN version 8. Pour chacun des trois modèles, nous avons spécifié un plan de sondage stratifié avec tirage des UPE avec remise (c'est-à-dire DESIGN = WR). Les trois modèles ont été ajustés au même nombre de périodes (16 307). Pour chaque modèle, nous avons ensuite calculé les fonctions de risque de base cumulatif empiriques en utilisant une approche produit-limite (voir Kalbfleisch et Prentice (2002), pages 114-116) telle qu'elle est implémentée dans la procédure SURVIVAL de SUDAAN.

L'approche robuste sous le modèle pour les périodes multiples décrites à la section 3.1 comporte un ajustement des estimations de la variance en vue de tenir compte de l'interdépendance éventuelle des périodes chez un même individu, en supposant que les périodes provenant d'individus différents sont indépendantes; cependant, dans cette approche, aucune mesure n'est prise pour tenir compte des probabilités inégales de sélection des individus échantillonnés, que ce soit dans les estimations des coefficients ou les estimations de la variance. Afin de le faire, pour les modèles 1 et 2, nous avons également utilisé la procédure SURVIVAL de SUDAAN version 8 pour estimer les variances des estimations pondérées des coefficients, où nous avons émis l'hypothèse d'indépendance inter-individus des périodes, mais avons tenu compte d'une corrélation intra-individu éventuelle des périodes. Pour cela, nous avons spécifié un plan d'échantillonnage non stratifié avec sélection des grappes avec remise en précisant que chaque individu formait sa propre grappe. Les hypothèses de dépendance sont les mêmes que celles utilisées par Lin (1994), mais nous avons tenu compte de l'utilisation de pondérations dans l'estimation des coefficients et des variances. Nous appellerons ces estimations des variances « estimations robustes modifiées de la variance sous le modèle des estimations pondérées des coefficients ».

5.3 Certaines statistiques descriptives

La durée moyenne estimée d'une période achevée est de 33,3 semaines, tandis que la durée moyenne estimée de la partie observée d'une période censurée (inachevée) est de 48,5 semaines.

Tableau 1 Dénombrement des individus du panel de six ans de l'EDTR ayant des périodes de chômage débutant entre janvier 1993 et décembre 1998, selon le nombre total de périodes et l'ordre de la période (A-achevée, I-inachevée)

Individus selon le nombre de périodes		Périodes par ordre									
		Première		Deuxième		Troisième		Quatrième		5e +	
		A	I	A	I	A	I	A	I	A	I
1 période	4 141	2 221	1 920	-	-	-	-	-	-	-	-
2 périodes	1 915	1 915	-	1 154	761	-	-	-	-	-	-
3 périodes	1 044	1 044	-	1 044	-	612	432	-	-	-	-
4 périodes	629	629	-	629	-	629	-	348	281	-	-
5 périodes et plus	672	672	-	672	-	672	-	672	-	1 158	415
Total	8 401	6 481	1 920	3 499	761	1 913	432	1 020	281	1 158	415

L'examen visuel des fonctions de survie estimées de Kaplan-Meier (non présentées) pour les périodes de chaque ordre révélait que, à mesure que croissait l'ordre, la valeur de la fonction de survie à n'importe quel temps fixe t diminuait, indiquant que les premières périodes sont les plus longues parmi les périodes achevées et que la durée d'une période multiple est d'autant plus courte que son ordre est élevé. Cette constatation est vraisemblablement une conséquence de la durée de vie limitée du panel, en ce sens qu'un individu comptant un plus grand nombre de périodes durant l'intervalle de temps de six ans donné est susceptible de présenter des périodes plus courtes.

5.4 Ajustement des modèles selon une approche fondée sur le plan

Comme nous l'avons mentionné plus haut, notre exemple est simplement une illustration de l'approche fondée sur le plan de sondage d'ajustement des modèles à risques proportionnels à des données sur des événements multiples provenant d'une enquête à plan de sondage complexe. Donc, nous consacrons peu de temps ici à discuter de la façon d'évaluer l'adéquation de ce ces modèles, dont l'adéquation des hypothèses de proportionnalité dans chacun des modèles ou le fait de savoir si un type de modèle est aussi bien ajusté qu'un autre.

Les coefficients estimés d'après l'ajustement des trois modèles aux données de l'EDTR sont présentés au tableau 2. Les valeurs des coefficients significatives au seuil de 5 % sur la base de tests t individuels sont en caractères gras.

Le modèle 1 est conditionnel à l'ordre de la période et comprend l'ajustement de quatre modèles distincts aux données provenant des quatre ordres de période différents. Comme le montre le tableau 2, les coefficients des variables SEXE, AGE et au moins une catégorie de la variable de revenu familial sont significatifs pour les périodes de tous ordres, quoique leur grandeur estimée varie en fonction de l'ordre de la période. Les coefficients estimés pour AGE sont négatifs, mais diminuent de grandeur à mesure que l'ordre de la période augmente, tandis qu'aucune tendance n'est discernable pour les coefficients estimés pour les deux autres variables. Les variables EDUCLEV, PARTJB et ATSCH ont des coefficients significatifs pour les périodes d'ordre 1, 2 et 3, mais non pour les périodes d'ordre 4. Ce résultat peut être attribué, en partie, à la petite taille d'échantillon pour les quatrièmes périodes. Pour chacune des trois autres variables du modèle (MARST, OCCUPATION et FIRMSIZE), un coefficient n'était significatif que pour un seul ordre de période.

Pour le modèle 2, les coefficients sont contraints d'être les mêmes pour tous les ordres de période. Comme le montre le tableau 2, numériquement, un grand nombre de valeurs des coefficients estimés, mais pas toutes, sont comprises entre les estimations calculées pour la première et pour la deuxième période à l'aide du modèle 1, ce qui pourrait être dû au fait qu'une proportion élevée de l'échantillon correspond aux événements de ces ordres. Toutes les variables, sauf OCCUPATION, ont un coefficient significatif. Les erreurs-types des coefficients sont plus faibles dans le cas du modèle 2 que dans celui du modèle 1.

Le modèle 3 est un modèle à période unique avec un seul ensemble de coefficients et une seule fonction de risque de base. Les coefficients estimés du modèle sont semblables aux estimations obtenues à l'aide du modèle 2.

Les fonctions de risque de base cumulatif estimées pour les modèles 1 à 3 sont illustrées aux figures 1 à 3 respectivement. Dans tous les cas, pour les durées allant jusqu'à environ 50 semaines, les fonctions ont une forme concave, ce qui sous-entend qu'il existe une dépendance positive par rapport au temps du taux de sortie (autrement dit, la probabilité de sortie est d'autant plus élevée que la période est longue). Pour les durées de plus de 50 semaines, la forme devient convexe, ce qui suggère une dépendance négative par rapport au temps dans le cas des périodes plus longues. À la figure 1, la position de la fonction de risque de base cumulatif estimée varie selon l'ordre de la période, la courbe pour les périodes d'ordre 1 étant la plus élevée et celle pour les périodes d'ordre 4, la plus basse. À la figure 2, pour le modèle 2, les positions des diverses courbes ne suivent pas l'ordre des périodes. Cette différence observée entre les figures 1 et 2 pourrait servir de diagnostic visuel indiquant qu'une étude plus approfondie est nécessaire afin d'évaluer lequel des modèles 1 ou 2 est un meilleur descripteur des données, puisque les coefficients estimés ont une incidence sur les risques de base estimés.

5.5 Comparaison aux estimations robustes modifiées de la variance sous le modèle

Comme il est décrit à la section 5.2, les estimations robustes modifiées de la variance sous le modèle tiennent compte de la corrélation éventuelle entre les périodes chez un même individu, sous l'hypothèse d'indépendance entre les individus. La comparaison, pour les modèles 1 et 2, des estimations des erreurs-types obtenues par cette approche aux estimations des erreurs-types sous le plan de sondage n'a révélé que des écarts faibles. Il semble donc que les estimations sous le plan reflètent toute corrélation entre les périodes chez un même individu et qu'il n'y ait pas de dépendance supplémentaire au-delà du niveau individuel dans notre exemple.

Tableau 2 Coefficients β estimés pour les trois modèles

	Modèle 1				Modèle 2	Modèle 3
	Ordre 1	**Ordre 2**	**Ordre 3**	**Ordre 4**		
SEX (F)						
M	**0,4417**	**0,3781**	**0,3299**	**0,4435**	**0,4049**	**0,4090**
EDUCLEV (É)						
F	**-0,4561**	**-0,5234**	**-0,3748**	-0,1065	**-0,4128**	**-0,4331**
MF	**-0,2330**	**-0,2700**	**-0,3310**	-0,1653	**-0,2436**	**-0,2474**
M	-0,0744	-0,1060	-0,1156	0,0668	-0,0684	-0,0671
MARST (M)						
Célibataire	**-0,1142**	-0,1290	-0,0622	-0,1375	**-0,1357**	**-0,1330**
Autre	0,0985	-0,0894	0,1124	-0,1072	0,0328	0,0401
TYPJBEND (Congédié(e))						
Départ volontaire	0,0704	**0,2752**	**0,4207**	**0,3413**	**0,1579**	**0,1284**
OCCUPATION (Autre)						
Professionnels	0,1592	-0,1364	-0,1388	0,0903	0,0490	0,0485
Administration	-0,0265	**-0,2930**	-0,1769	0,0579	-0,0971	-0,0938
Secteur primaire	-0,0211	-0,2175	-0,1187	0,2032	-0,0410	-0,0201
Fabrication	-0,0003	-0,0994	-0,1295	0,2862	-0,0093	-0,0088
Construction	0,1290	-0,1862	-0,0879	0,2339	0,0490	0,0813
FIRMSIZE (1 000+)						
<20	-0,0027	-0,0097	0,1005	**0,4403**	0,0441	0,0408
20 à 99	0,0358	0,0881	0,0815	0,3999	**0,0928**	**0,0951**
100 à 499	0,0436	-0,0905	0,0328	0,0257	0,0214	0,0278
500 à 999	-0,0006	0,0153	-0,0623	-0,0067	-0,0005	0,0020
PARTTJB (Non)						
Oui	**-0,2903**	**-0,5414**	**-0,5109**	-0,1407	**-0,3693**	**-0,3743**
ATSCH (Non)						
Oui	**-1,0832**	**-1,1516**	**-1,2956**	-1,3541	**-1,1205**	**-1,1266**
Revenu familial par personne (10 000 $ et moins)						
10 000 $ à 20 000 $	**0,1294**	**0,1802**	0,0692	**0,1117**	**0,1345**	**0,1330**
20 000 $ à 30 000 $	**0,1644**	**0,3611**	0,1572	**0,4900**	**0,2241**	**0,2141**
30 000 $ et plus	**0,1712**	**0,3916**	**0,3005**	**0,4241**	**0,2280**	**0,2115**
AGE	**-0,0491**	**-0,0311**	**-0,0269**	**-0,0207**	**-0,0424**	**-0,0435**
Périodes dans l'ensemble de risques	8 386	4 255	2 345	1 300	16 286	16 286
Censurées	1 913	759	432	281	3 385	3 385
Achevées	6 473	3 496	1 913	1 019	12 901	12 901

Les valeurs significatives au seuil de signification de 5 % sont en caractères gras.

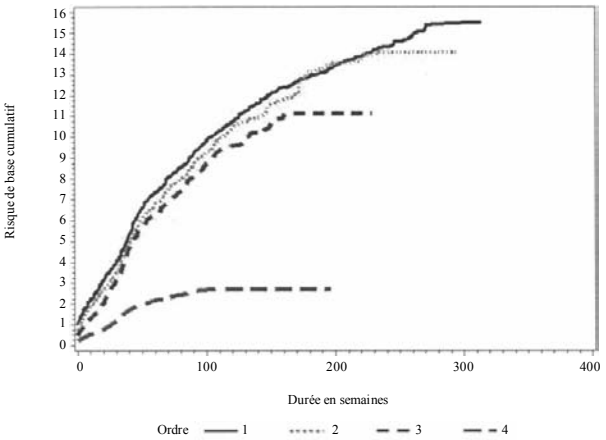

Figure 1 Risque de base cumulatif – Modèle 1

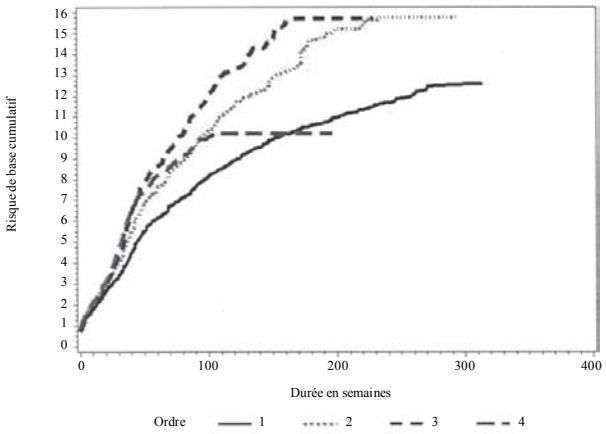

Figure 2 Risque de base cumulatif – Modèle 2

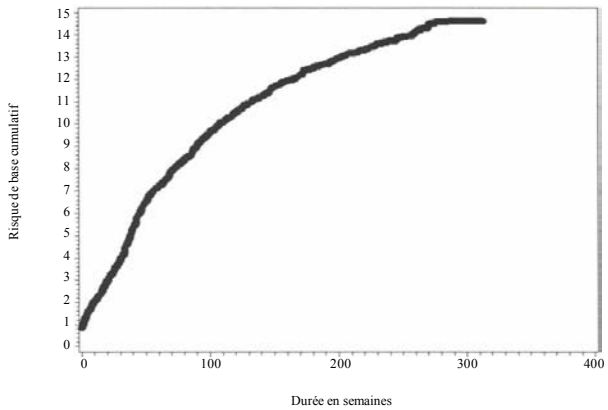

Figure 3 Risque de base cumulatif – Modèle 3

6. Conclusion

Nous avons étudié le problème de l'analyse de périodes multiples en considérant deux approches générales pour traiter le manque d'indépendance entre les temps de sortie, à savoir une approche robuste basée sur un modèle et une approche basée sur le plan de sondage. La première consiste à estimer les paramètres du modèle en supposant que les périodes sont indépendantes, puis à corriger la matrice de covariance naïve de façon à tenir compte des dépendances intra-individu postulées par le chercheur. Cette approche ne tient pas compte de la mise en grappes éventuelle entre individus (ou, en fait, de toute mise en grappes qui pourrait avoir lieu à un niveau d'agrégation plus élevé que l'individu) ni des probabilités inégales de sélection des individus (quoique, dans notre exemple, nous avons montré comment la méthode pourrait être étendue afin d'inclure les poids de sondage). La deuxième approche définit les coefficients du modèle comme des paramètres de population finie. Ces paramètres sont ensuite estimés en tenant compte des probabilités de sélection éventuellement inégales des individus. Une méthode d'estimation de la variance sous le plan de sondage qui tient compte des corrélations éventuelles entre individus dans la même UPE rend compte automatiquement des dépendances non spécifiées des périodes à des niveaux inférieurs à l'UPE, comme les dépendances intra-individu. Dans le cas des échantillons de grande taille, cette inférence sous le plan de sondage s'étend directement à la super-population à partir de laquelle la population finie a été hypothétiquement générée. Le défaut de la première approche est qu'elle ignore totalement la possibilité d'une mise en grappes entre individus. Un inconvénient éventuel de la deuxième approche, comme nous l'avons appliquée, est qu'elle repose sur l'hypothèse de l'échantillonnage avec remise des UPE comprenant les individus. Les deux approches coïncident dans le cas de l'échantillonnage aléatoire simple d'individus, où, dans le cas de l'approche robuste fondée sur un modèle, la dépendance entre les périodes chez un même individu est postulée explicitement et prise en compte dans la formule d'estimation de la variance et où, dans l'approche fondée sur le plan de sondage, les périodes chez un même individu sont traitées comme une grappe dans l'estimation de la variance sous le plan de sondage.

Nous avons appliqué l'approche fondée sur le plan de sondage à trois modèles de type à risques proportionnels. L'un tenait compte de risques de base non spécifiés différentiels et de coefficients différents pour chaque ordre de période. Le deuxième tenait encore compte de risques de base non spécifiés différentiels pour divers ordres de période, mais exigeait que les coefficients soient les mêmes pour tous les ordres. Le troisième était un simple modèle à période unique. Nous avons constaté que la façon dont l'information sur l'ordre de la période était utilisée avait une incidence sur les résultats de l'ajustement de nos modèles. Une comparaison visuelle des estimations des coefficients et des estimations des risques de base cumulatifs pour les modèles 1 et 2 indiquait des résultats différents. Comme l'a suggéré l'un des examinateurs, il serait bon d'élaborer un test formel en vue de déterminer si les coefficients diffèrent effectivement en fonction de l'ordre de la période (comme dans le modèle 1), étant donné des risques de base pouvant varier selon l'ordre de la période. Il est, en fait, facile de produire un tel test, de la façon suivante. Soit $\boldsymbol{\gamma} = (\mathbf{B}_1', \mathbf{B}_2', ..., \mathbf{B}_K')'$ le vecteur des K vecteurs de coefficients du modèle 1, où chacun est de longueur p, et soit $\mathbf{z}_{ij}(t) = (0', 0', ..., \mathbf{x}_{ij}(t)', 0', ..., 0')'$ le vecteur de longueur pK pour la j^{e} période du i^{e} individu où la j^{e} composante de ce vecteur contient le vecteur des covariables $\mathbf{x}_{ij}(t)$. Alors, le modèle 1 peut être exprimé par

$$h_j(t \mid \mathbf{z}_{ij}(\mathrm{t})) = \lambda_{0j}(t)\, e^{\mathbf{z}_{ij}'(t)\boldsymbol{\gamma}},$$

qui a la forme générale des risques de base variant avec l'ordre de la période, mais ayant un vecteur de coefficients fixe. Un test de la convergence des coefficients se rapportant à chaque ordre de période, c'est-à-dire $H_0 : \mathbf{B}_1 = \mathbf{B}_2 = ... \mathbf{B}_K$ équivaut à tester $H_0 : \mathbf{C}\boldsymbol{\gamma} = 0$ où C est la matrice $\mathbf{C} = I_p \otimes [1_{K-1} \; -I_{K-1}]$ de dimensions $(K-1)p \times Kp$. Étant donné une estimation $\hat{\boldsymbol{\gamma}}$ de $\boldsymbol{\gamma}$ et une estimation $\hat{V}(\hat{\boldsymbol{\gamma}})$ de la matrice de covariance de $\hat{\boldsymbol{\gamma}}$, obtenue comme il est décrit à la section 4 pour le modèle 2, une statistique de Wald peut être calculée pour tester l'hypothèse. Si l'hypothèse n'est pas rejetée, on peut conclure qu'un modèle à coefficients constants sur l'ordre de la période (mais à risques de base variables en fonction de l'ordre de la période) semble être aussi bien ajusté aux données qu'un modèle où les risques de base et les coefficients varient en fonction de l'ordre de la période. D'autres mesures de l'adéquation du modèle devraient également être faciles à élaborer dans le cadre d'estimations sous le plan de sondage.

Nous avons aussi comparé visuellement, dans le cas de notre exemple, les estimations des erreurs-types des coefficients obtenues sous le plan de sondage (en tenant compte de la mise en grappes au niveau de l'UPE et à des niveaux inférieurs), ainsi que sous une modification de l'approche robuste fondée sur un modèle (tenant compte de la mise en grappes au niveau de l'individu ou à un niveau inférieur) pour les modèles 1 et 2. Nous n'avons observé que des différences faibles, qui indiquaient l'absence d'effets de grappe à un niveau d'agrégation supérieur au niveau de l'individu pour les données en question. Nous avons également calculé les estimations des erreurrs-types en supposant que les périodes chez une même personne étaient indépendantes et de nouveau constaté des différences faibles seulement par rapport à celles obtenues suivant l'approche fondée sur le plan. Il semble donc que, pour l'exemple choisi, la dépendance inter-périodes soit faible. Cependant, en général, nous estimons qu'une approche fondée sur le plan de sondage protège contre l'omission de toute dépendance non postulée au niveau de l'UPE ou à un niveau inférieur dans les estimations de la variance.

Remerciements

Nous remercions Normand Laniel et Xuelin Zhang de leurs commentaires constructifs concernant une version antérieure du manuscrit. Nous remercions également le rédacteur adjoint et les examinateurs de leurs commentaires et suggestions qui ont accru considérablement la lisibilité du manuscrit.

Bibliographie

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-291.

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

Blossfeld, H.-P., et Hamerle, A. (1989). Using Cox models to study multiepisode processes. *Sociological Methods and Research*, 17, 4, 432-448.

Clayton, D., et Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (avec discussion). *Journal of the Royal Statistical Society*, Séries A, 1985, 148, 82-117.

Cochran, W.G. (1977). *Sampling Techniques*. Deuxième édition. New York : John Wiley & Sons, Inc.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, 62, 269-276.

Hamerle, A. (1989). Multiple-spell regression models for duration data. *Applied Statistics*, 38, 1, 127-138.

Heckman, J., et Singer, B. (1982). Population heterogeneity in demographic models. Dans *Multidimensional Mathematical Demography*, (Éds., K. Land et A. Rogers), New York : Academic Press, 567-599.

Hougaard, P. (1999). Fundamentals of survival data. *Biometrics*, 55, 1, 13-22.

Kalbfleisch, J.D., et Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*. 2ième Édition, New York : John Wiley & Sons, Inc.

Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48, 795-806.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*, 47, 939-956.

Lavigne, M., et Michaud, S. (1998). General Aspects of the Survey of Labour and Income Dynamics. Document de travail, Statistique Canada, 75F0002M No. 98-05.

Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: a marginal approach. *Statistics in Medicine*, 13, 2233-2247.

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

Lin, D.Y., et Wei, L.J. (1989). The robust Inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84, 1074-1078.

Prentice, R.L., Williams, B.J. et Peterson, A.V. (1981). On the regression analysis of multivariate failure data. *Biometrika*, 68, 373-379.

Roberts, G., et Kovačević, M. (2001). New research problems in analysis of duration data arising from complexities of longitudinal surveys. *Proceedings of the Survey Methods Section of the Statistical Society of Canada,* 111-116.

Wei, L.J., Lin, D.Y. et Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84, 1065-1073.

Williams, R.L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56, 645-646.