Numéro 3 **Février 2013**

Cahiers
Recherche et Méthodes

Analyse des événements de l'histoire de vie : estimation de modèles logistiques à temps discret avec SPSS

Jean-Marie Le Goff & Yannic Forney

Jean-Philippe Antonietti & André Berchtold Eds.

Université de Lausanne
Faculté des SSP
CH-1015 LAUSANNE

Les Cahiers Recherche et Méthodes (CREM) sont disponibles sur le site web suivant :

http://www.unil.ch/consultation-statistique

Anciens numéros :

1. *Multiple imputation in a longitudinal context: A simulation study using the TREE data*. André Berchtold & Joan-Carles Surís. Janvier 2012.
2. *Méthodes non-paramétriques de l'analyse des événements du parcours de vie (Event history Analysis). Estimations avec SPSS. Méthode de Kaplan-Meier et méthode actuarielle.* Jean-Marie Le Goff & Yannic Forney. Février 2013.
3. *Analyse des événements de l'histoire de vie : estimation de modèles logistiques à temps discret avec SPSS.* Jean-Marie Le Goff & Yannic Forney. Février 2013.

Analyse des événements de l'histoire de vie : estimation de modèles logistiques à temps discret avec SPSS

Jean-Marie Le Goff

Yannic Forney

(Version 2)

Lines
Pôle national de recherché Lives
Université de Lausanne

Analyse des événements de l'histoire de vie : Estimation de modèles logistiques à temps discret avec SPSS

Préambule

Le présent article a été publié une première fois fin 2003 sous le titre « Mise en œuvre des modèles logistiques à temps discret avec SPSS» sur une page du site internet du centre Pavie, cette dernière étant consacrée aux méthodes de l'analyse des biographies (*Event history Analysis*). Le site ayant été fermé en 2010, cet article trouve une place naturelle dans la collection des Cahiers de Recherche et Méthodes.

Par rapport à sa version originale, l'article a peu changé. Nous avons effectué plusieurs corrections stylistiques. Néanmoins, nous avons ajouté dans l'introduction quelques mots concernant les modèles *cloglog* à temps discret qui, théoriquement, devraient plutôt être estimés en présence de données dans lesquelles le temps est découpé en intervalles assez longs, alors que, concrètement, c'est rarement le cas. Lors de la rédaction de ce cahier, nous avons utilisé la version 11.5 de SPSS. Si cette version de SPSS est plutôt ancienne, nous avons décidé de maintenir les sorties output que nous avions obtenu avec cette version ainsi que les syntaxes. Les unes et les autres ont, en effet, peu changé. Nous espérons que la deuxième édition donnera lieu à autant d'interactions entre nous et ses lecteurs qu'il y en a eu lors de sa première publication.

Le 30 janvier 2013
JMLG et YF

Jean-Marie Le Goff est actuellement chercheur au centre de recherche sur les parcours de vie et les inégalités de l'Université de Lausanne et collabore au pole national de recherche LIVES « Surmonter la vulnérabilité : perspective du parcours de vie ». Contacts : jean-marie.Legoff@unil.ch

Yannic Forney est actuellement chef de projet à la Fédération romande des entreprises (Genève)

Sommaire

1. Introduction

Ce présent article a pour objet de présenter la mise en œuvre des modèles logistiques à temps discret avec SPSS. L'usage de ces techniques dans le cadre de l'analyse des biographies correspond à une utilisation particulière des modèles de régression logistique. Nous n'aurons pas pour objectif de décrire les techniques de régression logistique d'un point de vue général, mais uniquement dans le cadre particulier de l'analyse des biographies (*Event history analysis*). Nous ferons toutefois en sorte que le lecteur qui ne connaîtrait pas ces techniques puisse les mettre en œuvre facilement avec la commande « LOGISTIC REGRESSION » de SPSS[1].

La première question qui se pose concernant l'usage des modèles logit à temps discret est de savoir quand ils doivent être appliqués. La réponse à cette question dépend des données dont on dispose. La modélisation de l'occurrence d'un événement au cours du temps repose sur une description de chaque individu par un couple de deux variables. La première, l'indice de censure, indique si l'individu a connu ou non l'événement analysé au cours de la période d'observation. La deuxième indique la durée qui s'est écoulée entre le moment où l'individu est entré dans la population à risque et le moment d'occurrence de l'événement s'il le connaît, le moment de sortie d'observation s'il ne le connaît pas. L'estimation d'un modèle de l'analyse des biographies consiste, entre autres, à estimer le rôle joué par des caractéristiques individuelles sur le risque d'occurrence de l'événement au cours du temps. Toutefois, l'estimation de modèles de Cox ou de modèles paramétriques repose sur l'idée que le temps est considéré de manière continue. Ceci signifie notamment que les individus connaissent les uns après les autres l'événement étudié (s'ils le connaissent). En d'autres termes, il ne peut y avoir deux personnes (ou plus) qui connaissent l'événement au même moment.

Dans la réalité, les événements sont toujours mesurés sur un temps qui est discrétisé. L'unité de temps à partir de laquelle sont mesurées les durées d'occurrence des événements peut être le mois, l'année, voire la décade (Allison, 1982). Ceci explique que des individus peuvent connaître l'événement à la même durée observée. La littérature anglo-saxonne utilise le terme de *tie* (nœud) pour désigner les intervalles de temps dans lesquels il y a occurrence de deux ou plusieurs événements[2]. La probabilité qu'il y ait de nombreux *ties* au cours de la période d'observation est d'autant plus forte que l'unité de temps est grande. Or, pour Cox (1972), qui est l'inventeur des modèles semi-paramétriques, ces derniers sont particulièrement sensibles à des problèmes de *ties* en raison du fait que les estimations s'appuient sur le rang d'occurrence des événements (Yamaguchi 1991, Vermunt 1997). Les résultats d'un modèle de Cox peuvent être fortement biaisés si un *nombre important* d'individus connaît l'événement étudié à un même moment. Encore s'agit-il de préciser ce que signifie l'expression *nombre important* : selon Yamaguchi (1991), le biais peut devenir conséquent si plus de 5% de la population soumise au risque connaît l'événement à un moment donné.

En présence de *ties*, ce sont alors des modèles dans lesquels le temps est considéré discret qui doivent être estimés. Dans les modèles à temps discret, l'intérêt ne porte plus sur le risque de connaître l'événement durant un court laps de temps, mais sur la probabilité conditionnelle de connaître cet événement durant un intervalle de temps (le mois ou l'année). La mise en œuvre de ces modèles en sciences sociales a été initiée au début des années quatre-vingt par Allison

[1] Indiquons seulement qu'il existe quelques ouvrages d'initiation à ces modèles que les lecteurs pourront consulter : Hosner & Lemeshow (2000); Menard (2001) ; Kleinbaum & Klein (2002).

[2] Pour une discussion détaillée sur cette question des *ties*, cf. Therneau & Grambsch (2000), pp 31-37 et 48-53, particulièrement le paragraphe « recommandations », p 52.

(1982, 1984, 1995). Parmi les modèles à temps discret, le plus utilisé est sans doute le modèle *logit à temps discret* (*discrete time logit model*) auquel cet article est entièrement consacré. En toute rigueur, ce modèle devrait être estimé dans le cas des processus discrets dans lesquels les échéances ne peuvent avoir lieu qu'à des moments particuliers, par exemple, le mois de septembre si l'on s'intéresse à la probabilité de passage d'une classe à une autre. Dans le cas de processus continus mais qui est discrétisé en raison de l'unité de mesure, typiquement, si les durées sont mesurées en année, il faudrait plutôt estimer des modèles binomiaux avec un lien *complémentaire log log* (*cloglog*) plutôt qu'un lien *logit* (Allison, 1982). Néanmoins, dans les faits, le modèle de régression logistique à temps discret est le modèle le plus souvent utilisé, quel soit le type de temps discret. La popularité de ce modèle est vraisemblablement due à sa facilité de compréhension et d'application (Vermunt, 1997), d'autant qu'une comparaison de l'estimation d'un modèle *logit* et d'un modèle *cloglog*[3] à temps discret sur les mêmes données donne des résultats extrêmement comparables. Dès lors, les modèles de régression logistique à temps discret sont particulièrement adéquats à appliquer sur des données telles que celles des enquêtes par panel dans lesquels on ne dispose pas toujours de la date de changement de situation entre deux vagues d'enquête[4].

Dans la deuxième partie de ce cahier sont développés les aspects théoriques du modèle logistique à temps discret. La mise en œuvre de ce modèle présente néanmoins une difficulté se rapportant à la préparation des données. Dans la plupart des logiciels statistiques, dont SPSS, les modèles non-paramétriques ou semi-paramétriques sont estimés à partir de données dans lesquelles une ligne correspond à un épisode (*fichier épisode*), chaque épisode étant caractérisé par un indice de censure, une durée, ainsi que des caractéristiques individuelles. Les modèles logit à temps discret exigent une autre organisation des données, le fichier *personne-période*[5]. Dans ce deuxième type de fichier, une ligne correspond à un intervalle de temps, le nombre de lignes décrivant un épisode correspondant à la durée qui sépare le début de cet épisode et sa fin ou sa censure. La troisième partie sera consacrée à la construction d'un tel fichier à partir d'un fichier épisode. Dans une quatrième partie, nous développerons un exemple d'estimation d'un modèle logistique à temps discret, celui-ci portant sur la durée du premier emploi d'hommes nés entre 1945 et 1964.

2. Présentation des modèles logit à temps discret et principe d'estimation

2.1 Notions probabilistes de l'analyse des biographies en temps discret

Soit T, une variable aléatoire discrète qui indique la durée écoulée avant l'occurrence d'un événement pour chaque individu. Si $T= t_l$, cela signifie que l'événement se produit au temps t_l où l indique le $l^{ème}$ moment du temps et satisfait la condition $t_1<t_2<\ldots$. Par extension, t_l peut désigner un intervalle de temps. La probabilité $f(t_l)$ de connaître l'événement à l'instant t_l dans la population est[6] (Yamaguchi, 1991, Vermunt 1997) :

$$f(t_l) = P(T = t_l) \qquad (2.1)$$

[3] Un tel modèle peut être estimé avec SPSS avec la commande PLUM qui permet, plus généralement, d'estimer des modèles de régression pour des variables ordinales.
[4] Cf., par exemple, les travaux de Hank 2002 sur la fécondité en Allemagne à partir des données du GSOEP.
[5] Sur l'importance de la personne période en démographie, cf. Preston et al. (2001).
[6] Cette probabilité est donc l'équivalent en temps discret de la densité de probabilité.

La fonction de survie $S(t_l)$ est la probabilité de ne pas avoir connu l'événement avant t_l[7] :

$$S(t_l) = P(T > t_l) = \sum_{k=l+1}^{L^*} f(t_k) \qquad (2.2)$$

Où L^* indique le nombre total d'intervalle de temps pris en compte. Le risque au temps t_l est la probabilité conditionnelle de connaître l'événement au temps t_l, *sachant que* les individus n'ont pas encore connu l'événement :

$$P(t_l) = P(T = t_l \,/\, T \geq t_l) = f(t_l) / S(t_{l-1}) \qquad (2.3)$$

$P(t_l)$ correspond à la notion de *quotient* tel que celui-ci est utilisé dans l'analyse démographique classique (Pressat, 1983). Il est l'équivalent en temps discret du risque (*hazard rate*). Dans cet article, nous le noterons par $P(t_l)$ plutôt que par l'écriture conventionnelle $h(t_l)$ afin de souligner qu'il s'agit d'une probabilité. Il est important de souligner qu'il s'agit d'une *probabilité conditionnelle*. Le complémentaire à 1 de cette probabilité, $1 - P(t_l)$, représente la probabilité de ne pas connaître l'événement au temps t_l sachant que les personnes n'avaient pas connu l'événement auparavant :

$$S(t_l) = S(t_{l-1})(1 - P(t_l)) \qquad (2.4)$$

$S(t_l)$ correspond donc au produit de toutes les probabilités conditionnelles de ne pas avoir connu l'événement depuis le début de l'observation :

$$S(t_l) = \prod_{k=1}^{l} (1 - P(t_k)) \qquad (2.5)$$

Et dans ce cas, la probabilité $f(t_l)$ de connaître l'événement en t_l peut alors aussi s'écrire de la manière suivante :

$$f(t_l) = P(t_l) \prod_{k=1}^{l-1} (1 - P(t_k)) \qquad (2.6)$$

2.2 Formulation du modèle logit à temps discret

La modélisation de la distribution de l'occurrence d'un événement du parcours de vie repose sur l'idée que le risque de connaître l'événement au temps *t* est une fonction du temps et des caractéristiques des individus. En temps discret, ceci peut être formulé par (pour simplifier, nous abandonnons ici les indices *l*) :

$$P(t) = f^{o}(t; x_t) \qquad (2.7)$$

[7] Quelques auteurs définissent cette probabilité de survie par $S(t_l) = P(T \geq t_l)$. Cf. Vermunt (1997, p 83, n 3).

x_t représente un vecteur de caractéristiques $x_t=(x_{t1},x_{t2},x_{t3},...,x_{tk})$. Ces caractéristiques sont soit des variables quantitatives, soit des caractéristiques qualitatives. Dans ce dernier cas, il s'agit alors de variables binaires 0 ou 1 qui indiquent respectivement l'absence ou la présence de la caractéristique chez l'individu. Il est à noter que ces variables peuvent varier au cours du temps (par exemple, la situation maritale si l'intérêt porte sur la première naissance), raison pour laquelle, on ajoute un indice *t* à chacune des caractéristiques. Dans le cas où le temps est considéré continu, la modélisation du rôle joué par les caractéristiques individuelles s'appuie le plus souvent sur l'hypothèse de proportionnalité des risques. Cette hypothèse revient à supposer que le risque des individus qui possèdent une caractéristique donnée est multiplié par une constante, en comparaison avec les individus qui ne possèdent pas cette caractéristique. Ainsi, si $h(t,x_t)$ représente le risque en temps continu des individus possédant les caractéristiques x_t, alors selon l'hypothèse de risques proportionnels :

$$h(t) = h_0(t)\exp\left(\sum_{j=1}^{k} b_j x_j\right) \qquad (2.8)$$

Où $h_0(t)$ représente le risque des individus pour qui l'ensemble des caractéristiques x_j sont égales à 0 et les paramètres b_j sont les logarithmes des coefficients de proportionnalité, à estimer. En passant par le logarithme, le modèle s'écrit :

$$\log h(t) = \log h_0(t) + \sum_{j=1}^{k} b_j x_j \qquad (2.9)$$

L'hypothèse de risque proportionnel pose problème en temps discret, puisque la probabilité conditionnelle de connaître l'événement au temps *t* ne peut être supérieure à 1, contrairement au risque tel que celui-ci est analysé en temps continu. Un modèle à risque proportionnel pourrait donc aboutir à l'estimation de coefficients tels que la probabilité conditionnelle pourrait devenir supérieure à 1. En revanche, la transformation *logit de P(t),* c'est-à-dire, *le logarithme du rapport entre P(t, x_t) et (1-P(t, x_t),* varie entre moins l'infini et plus l'infini. Il devient alors intéressant de développer une modélisation du rôle joué par les caractéristiques des individus selon un modèle logit. Rappelons que, si *P(x)* représente une probabilité ou une proportion que l'on souhaite analyser, et si *x* représente un vecteur de caractéristiques possédées par les individus, la formulation générale du modèle logit s'écrit : (Kleinbaum and Klein, 2002) :

$$\log\left(\frac{P(x)}{1-P(x)}\right) = \log\left(\frac{P_o}{1-P_o}\right) + \sum_{j=1}^{k} b_j x_j \qquad (2.10)$$

Où P_0 représente la probabilité analysée des individus de référence, c'est-à-dire, des individus dont les caractéristiques sont égales à 0. L'exponentiel des coefficients *b* à estimer est appelé le *odds ratio* ou plus simplement le *odds*. Le modèle logit s'écrit plus simplement :

$$\log\left(\frac{P(x)}{1-P(x)}\right) = a + bx \qquad (2.11)$$

où *a* représente le logarithme de $P_0/(1-P_0)$ et $bx = \sum_{j=1}^{k} b_j x_j$. En d'autres termes, le logarithme du rapport entre *P(x)* et son complémentaire à 1 est une fonction linéaire des caractéristiques *x* plus une constante. Dans le cas présent du modèle logit à temps discret, la probabilité à estimer n'est pas seulement fonction des différentes caractéristiques des individus, mais aussi une fonction du temps, raison pour laquelle on peut symboliser cette probabilité par $P(t, x_t)$. Par similarité avec le modèle précédent, une première écriture du modèle logit à temps discret est (Allison, 1982) :

$$\log\left(\frac{P(t, x_t)}{1 - P(t, x_t)}\right) = a + bx_t \qquad (2.12)$$

Ce modèle revient à considérer que le logit de la probabilité conditionnelle de connaître l'événement est, à chaque instant, une fonction linéaire des caractéristiques en présence, ces caractéristiques pouvant varier avec le temps, plus une constante. Tel que formulé, ce modèle est restrictif puisqu'il revient à considérer qu'il n'y a pas de dépendance du risque (ou du *odds*) au temps. En d'autres termes, la probabilité conditionnelle de connaître l'événement est la même quel que soit le moment auquel on se situe. La fonction de séjour *S(t)* correspondante suit une distribution géométrique et l'on parle alors dans ce cas d'un modèle géométrique[8]. Cette hypothèse est sans doute forte ; on peut dès lors introduire l'idée d'une dépendance de l'odds ratio au temps. Le modèle se formalise alors ainsi :

$$\log\left(\frac{P(t, x_t)}{1 - P(t, x_t)}\right) = a(t) + bx_t \qquad (2.13)$$

Où *a(t)* est une fonction du temps. Il est à remarquer que si $P(t,x_t)$ est petit, ce qui est souvent le cas lorsque l'unité de temps est petite (le mois, dans le cas d'événements socio-démographiques), alors $1- P(t,x_t)$ est quasiment égal à 1 et le rapport $P(t,x_t) / [1- P(t,x_t)]$ est quasiment égal à $P(t,x_t)$. On se trouve alors quasiment dans les conditions d'un modèle à risque proportionnel, et les différents coefficients *b* estimés seront semblables à ceux qui seraient estimé à partir d'un modèle à risque proportionnel en temps continu (Yamaguchi, 1991). En outre, de manière similaire à un *modèle de Gompertz* en temps continu, on peut considérer que *a(t)* est une fonction linéaire du temps. Dans ce cas, le modèle s'écrit :

$$\log\left(\frac{P(t, x_t)}{1 - P(t, x_t)}\right) = c + at + bx_t \qquad (2.14)$$

Où *c* est une constante alors que *a* correspondra à la pente de la fonction de Gompertz. De même, une modélisation similaire au *modèle de Weibull* en temps continu peut être proposée, en considérant que *a(t)* est une fonction linéaire du logarithme du temps :

$$\log\left(\frac{P(t,x_t)}{1-P(t,x_t)}\right) = c + a\log(t) + bx_t \qquad (2.15)$$

[8] De la même manière que, en temps continu, la distribution de la fonction de séjour est une distribution exponentielle et que l'on parle d'un modèle exponentiel lorsque le risque est constant.

Aussi bien dans le modèle de type Gompertz que dans celui de type Weibull, le rapport entre la probabilité de connaître l'événement et celle de ne pas le connaître est une fonction monotone du temps (toujours croissante ou toujours décroissante). Une telle modélisation dans laquelle la dépendance au temps est toujours croissante ou décroissante peut poser problème, en ce sens qu'il a été observé pour plusieurs événements du parcours de vie que la distribution du risque de leur occurrence était d'abord croissante, puis décroissante (Dieckmann, 1990). On peut alors être plutôt intéressé à estimer un modèle similaire au modèle *Piecewise constant* en temps continu (Blossfeld and Rohwer, 2002). On considère ainsi que le logit de la probabilité conditionnelle est constant sur un intervalle de temps, puis prend une autre valeur sur l'intervalle de temps suivant. On estime alors autant de coefficients que l'on définit d'intervalles de temps. Un tel modèle s'écrit :

$$\log\left(\frac{P(t,x_t)}{1-P(t,x_t)}\right)=c+\sum_{k=1}^{l} a_k+bx_t \qquad (2.16)$$

c est une constante à estimer, alors que a_k sera constant uniquement sur un intervalle de temps $k=(t,\ t+t_k)$. Une telle modélisation présente une grande souplesse d'utilisation, puisqu'il ne préjuge en rien de la distribution de la probabilité conditionnelle. Nous verrons dans la quatrième partie de cet article que ce modèle est aisé à estimer avec SPSS.

2.3 Equation de vraisemblance

Comme dans la régression logistique classique, les différents paramètres sont estimés par la méthode du maximum de vraisemblance (Kleinbaum et Klein 2002, Allison, 1982 et 1984). Dans le cadre d'un modèle de l'analyse des biographies, un individu i contribue à la vraisemblance par $f(t_i)$ s'il connaît l'événement et $S(t_i)$ s'il ne le connaît pas. L'équation de vraisemblance pour l'ensemble de la population d'effectif n soumise au risque de connaître l'événement correspond donc au produit de la contribution à la vraisemblance de chaque individu i :

$$L=\prod_{i=1}^{n}\left[f(t_i)\right]^{\delta_i}\left[S(t_i)\right]^{1-\delta_i} \qquad (2.17)$$

Où δ_i est égal à 1 si l'individu a connu l'événement, 0 sinon. En temps discret, l'équation de vraisemblance se formalise ainsi (Allison, 1982 :74) :

$$L=\prod_{i=1}^{n}\left[\Pr(T_i=t_i)\right]^{\delta_i}\left[\Pr(T_i>t_i)\right]^{1-\delta_i} \qquad (2.18)$$

Or, d'après la relation 2.6 :

$$\Pr(T_i=t_i)=P_{it}\prod_{k=1}^{t-1}(1-P_{ik}) \qquad (2.19)$$

et l'équation 2.5 :

$$\Pr(T_i>t_i)=\prod_{k=1}^{t}(1-P_{ik}) \qquad (2.20)$$

En conséquence, en substituant (2.19) et (2.20) dans (2.18) et en passant par le logarithme de la vraisemblance :

$$\log L = \sum_{i=1}^{n} \delta_i \log \left\{ P_{it_i} /(1 - P_{it_i}) \right\} + \sum_{i=1}^{n} \sum_{k=1}^{t_i} \log(1 - P_{ik}) \qquad (2.21)$$

Toujours en suivant Allison (1982 : 75), en définissant une variable aléatoire y_{it} égale à 1 si la personne connaît l'événement au temps t et 0 sinon, la formule devient :

$$\log L = \sum_{i=1}^{n} \sum_{k=1}^{t_i} y_{i\,t} \log \left\{ P_{ik} /(1 - P_{ik}) \right\} + \sum_{i=1}^{n} \sum_{k=1}^{t_i} \log(1 - P_{ik}) \qquad (2.22)$$

Cette dernière équation indique une propriété importante des modèles logit en temps discret : *l'estimation d'un modèle logistique à temps discret revient à estimer un modèle logistique de la probabilité de connaître l'événement sur un fichier de données dans lequel chaque individu est décomposé, et cela de manière indépendante, en autant d'intervalles de temps que cet individu est soumis au risque.* Par exemple, si un individu connaît l'événement au temps t=5, cinq observations différentes seront créées. Pour chacune de ces observations est associé un indice, équivalent à un indice de censure, égal à 1 si l'individu connaît l'événement et à 0 sinon. Dans le cas des quatre premières observations, cet indice est donc égal à 0. Dans le cas de la cinquième observation, la variable dépendante est codée 1. Au cas où l'individu serait sorti d'observation sans avoir connu l'événement, la dernière observation devrait être codée 0. Une telle décomposition revient alors à construire une base de données dans laquelle l'individu est une *personne-période*. On parle alors d'un *fichier personne-période*. Si l'unité de temps à partir de laquelle sont mesurées les différentes durées avant l'occurrence de l'événement est l'année, on parlera d'un fichier *personne-année*. Si l'unité de temps est le mois, on parlera d'un fichier *personne-mois,* etc.

Cette propriété des modèles logit à temps discret que nous venons d'énoncer a pour conséquence pratique que ceux-ci peuvent être estimés à partir des procédures usuelles d'estimation des régressions logistiques qui sont implantées dans la plupart des logiciels de statistiques, à l'exemple de la procédure « LOGISTIC REGRESSION » dans SPSS (Kleinbaum & Klein, 2002). Il est à noter que le fait que les individus soient décomposés en plusieurs observations n'a, non seulement aucune conséquence sur les valeurs estimées des différents coefficients, mais n'a pas non plus d'effets sur l'estimation des variances de ces coefficients. En d'autres termes, la significativité des coefficients restent semblable lorsque le modèle est estimé sur des données de personne-période (Allison, 1982). L'estimation d'un modèle logit à temps discret apparaît finalement assez simple, surtout pour des utilisateurs habitués à estimer des modèles logistiques classiques. Il reste toutefois qu'à cette facilité d'estimation correspond une difficulté en amont de la procédure d'estimation qui est celle de la préparation des données sous la forme d'un fichier *personne-période*. La partie suivante de ce chapitre est consacrée à la création de ce type de fichier avec SPSS.

3. La préparation d'un fichier de données

Dans ce point, nous développons un exemple de *construction d'un fichier personne-période à partir d'un fichier épisode*. Il s'agit ici de décomposer un fichier portant sur la durée du premier emploi de plus de trois mois d'hommes nés en Suisse. Ces données sont issues de l'enquête suisse sur la famille qui a été réalisée par l'OFS en 1994 (Gabadinho 1998,

Gabadinho & Wanner 1999)[9]. Le fichier épisode est composé de 1 100 premiers emplois qui ont été occupés par des hommes nés entre 1945 et 1964. Ces emplois sont détaillés par quelques caractéristiques relatives à la fonction occupée, etc. Notre objectif consiste donc à analyser l'effet de ces caractéristiques sur la probabilité conditionnelle de départ d'emploi[10]. Ces caractéristiques, qui sont fixes au cours du temps, sont : la fonction exercée (variable *csp*), que nous subdiviserons de manière simple en emploi de type primaire (agriculture, sylviculture), secondaire (industriel, c'est-à-dire, ouvrier) et tertiaire (services) ; le statut de l'emploi (*statut*), c'est-à-dire, principalement si les personnes étaient indépendantes, cadres, collaborateurs ou dans une autre situation ; le temps de travail, à plein-temps, à temps partiel, ou temps indéterminé (*temps*).

Nous avons, en outre, pour intérêt d'analyser l'effet d'un mariage sur la probabilité de quitter son activité professionnelle. Il s'agit là donc d'une caractéristique dépendante du temps. Chaque individu est caractérisé par une date de mariage (*datmar2*). S'il y a eu mariage avant la fin de l'activité professionnelle, cette date est codée par le nombre de mois écoulé depuis janvier 1945[11]. Dans le cas où les individus observés ne se sont pas mariés durant leur premier emploi, cette date est codée 999. Ce codage est arbitraire mais le choix d'un nombre nettement plus élevé que la date du moment de l'enquête (octobre 1994 à mai 1995, soit les mois 597 à 605), nous sera utile lors de la construction du fichier personne-période (Blossfeld and Rohwer, 2002). Une autre variable dépendante du temps prise en compte sera l'âge (en mois écoulé depuis la date de naissance). Cette variable nécessite que l'on prenne en compte la date de naissance des personnes enquêtées, qui dans le fichier épisode a pour nom « *f011t* ».

Les emplois sont caractérisés, en outre, par leur date d'entrée (*debut*) et de fin (*datfind*) calculés en nombre de mois écoulés depuis janvier 1945 ainsi que par un indice de censure, *censurd.* Il y a censure (*censurd=0*) si les individus n'ont pas quitté leur emploi au moment de l'enquête : *datfind* correspond alors à la date de l'enquête. Toutefois, en raison de l'intérêt que nous portons à l'effet du mariage, nous avons décidé de censurer les quelques individus qui non-seulement se sont mariés durant leur premier emploi, mais qui ont aussi divorcé. Ces derniers, qui sont seulement au nombre de six, sont ainsi censurés à la date de leur divorce. A l'exception de ces six hommes, tous les autres individus ayant quitté leur premier emploi sont suivis jusque la date de départ. Soulignons pour terminer, que chaque épisode enregistré dans le fichier épisode est caractérisé aussi par un numéro d'identification (*intnr*).

Du fait que notre unité de mesure des durées est le mois, notre fichier *personne-période* sera un fichier *personne-mois*. Durant la phase de construction de ce fichier, chaque épisode sera décomposé en *T* lignes. A toutes ces lignes, à l'exception de la dernière, sera associé un indice de censure qui sera égal à 0. A la dernière ligne sera associé un indice de censure égal à 0 ou 1 indiquant si l'individu est sorti d'observation ou s'il a connu l'événement. En outre, le fichier personne période que l'on souhaite construire doit contenir les variables explicatives relatives à la description de l'emploi exercé. En ce qui concerne la variable dépendante du temps (le mariage), l'idée est de construire une variable dichotomique : l'indice sera égal à 1 si les individus sont mariés et à 0 si ce n'est pas le cas. Nous retranscrivons ici de manière fidèle, les différentes commandes de syntaxe qui ont été nécessaires pour la construction du fichier personne-période à partir du fichier épisode (tableau 1)[12] :

[9] Cf. dans cette même série des cahiers recherche et méthodes, notre article sur la mise en œuvre des méthodes non-paramétriques avec SPSS pour une description de ce fichier (Le Goff & Forney, 2012).
[10] Cf. la quatrième section de cet article.
[11] Janvier 1945 est le mois 1, février 1945, le mois 2, et ainsi de suite. Décembre 1994 sera ainsi le mois 600.
[12] L'écriture de cette syntaxe nous a largement été inspirée par l'exemple détaillé décrit dans Mills (1999).

- 1) les trois premières lignes d'instruction « *write outfile* » servent à transférer le fichier épisode initial en un fichier texte ASCII qui prend le nom de « *rawdata* ». Chaque variable a une largeur de 8 colonnes (F8).

- 2) l'instruction « *set mxloops*» en quatrième ligne définit un nombre maximum d'itérations qui seront effectuées dans une boucle[13]. Dans le cas présent, cette instruction constitue un préalable à l'ensemble des instructions suivantes (lignes 5 à 17) qui seront consacrées à décomposer les épisodes en personnes-périodes. Ici, le nombre de boucles va correspondre à la durée qui sépare le début de la fin de l'emploi. En se donnant pour nombre maximum de boucles la valeur de 600, nous dépassons ici largement la durée maximale qui a été observée, qui est de 420 ;

- 3) l'instruction « *input program* » (ligne 5) est une commande utilisée lorsque l'on veut créer des « *cas* » ou des « individus statistiques » (*case*), ici les personnes-périodes. Entre cette instruction et celle qui clôt le processus de création de cas (« *end input program* », ligne 17), sont définies trois étapes de création des données :

  a) Lecture du fichier « *rawdata* » qui vient d'être créé avec la commande « *data list file* » (lignes 6 et 7). Les formats des variables doivent être rigoureusement définis de la même manière qu'ils avaient été transcris dans la commande « *write* » en ligne 2 ;

  b) Création de la variable « *tmonth* », qui indiquera pour chaque cas (personne-période) le mois auquel on se situe, en nombre de mois écoulés depuis janvier 1945, ainsi que la variable « *censurdp* », qui sera égale à 0 pour toute les lignes sauf la dernière et à 0 ou 1 dans le cas de la dernière ligne, selon que les individus on vu leur épisode d'emploi censuré ou non (ligne 8) ;

  c) Création des personnes-périodes et calcul des différentes variables (ligne 9 à 17). L'instruction « *leave all* » (ligne 9) permet d'éviter que les variables ne soient réinitialisées à chaque fois qu'il y aura création d'une personne-période. Les variables prendront ainsi pour valeur celle qu'elles avaient dans le cas précédent (SPSS, 1994). Cette instruction est valable pour toutes les variables, c'est-à-dire, aussi pour les caractéristiques de l'emploi, la date du mariage, etc. Toutefois, chaque fois qu'il y aura lecture d'un nouvel épisode dans le fichier « *rawdata* », la variable *tmonth* sera initialisée à la valeur de début (ligne 10). Les commandes suivantes se situent à l'intérieur d'une boucle (instruction *loop*, ligne 11 à 16). Cet ensemble d'instruction va créer un nombre de cas égal à la durée. Dans chaque cas, la variable *censurdp* sera définie comme étant égal à 0 (ligne 12). Toutefois, s'il s'agit du dernier cas créé (lorsque *tmonth=datfin*), *censurdp* prend la valeur de l'indice de *censurd* (ligne 13). A ce point précis, toutes les variables du nouveau cas créé (une personne-période) ont été définies soit par la valeur du cas précédent, soit par 0 ou 1 dans le cas de la variable *censurdp*. L'instruction « *end case* » (ligne 14) signale ainsi que l'ensemble des instructions se rapportant à la génération de ce cas est terminé. Il reste toutefois qu'ainsi créée, la variable *tmonth* n'est pas correcte. Elle a, en effet, pris pour valeur celle qu'elle avait dans le cas précédent. Il suffit donc de lui

[13] Le nombre de boucles par défaut est défini dans SPSS à 40.

ajouter 1 (ligne 15). Ceci signifie que le premier *tmonth* sera égal à la date de début plus un mois. La boucle s'arrête lorsque *tmonth* est égal à la date de fin d'emploi plus 1 ;

- 4) Les deux lignes suivantes de syntaxe (lignes 19 et 20) permettent de sauvegarder le fichier nouvellement créé sous le nom de « *periode.sav* ». Les commandes suivantes sont plus classiques et visent à créer dans ce nouveau fichier, la variable de durée d'emploi « *dureep* », qui est égale, dans chaque cas, à la différence entre *tmonth* et *debut*, mois de début d'emploi (ligne 21). De même, la variable *agep* va correspondre à l'âge de l'individu en faisant la différence entre *tmonth* et *f011t* (ligne 22). Il est ensuite créé une variable « *mar* » qui prend la valeur 0 si les individus ne sont pas (encore) mariés et 1 dès qu'ils se marient (lignes 23 à 28). Comme on le voie, la construction de variables dépendantes du temps est extrêmement aisée dans le cas d'un fichier personne-période, dès lors que l'on dispose de variables adéquates importées du fichier épisode (date de naissance, date du mariage). Pour terminer, le fichier personne-mois venant d'être créé est sauvegardé sous le nom de *periodep*.

Tableau 1 : syntaxe de création d'un fichier personne-période

```
write outfile="c:\base\rawdata"                                                                    1
     /intnr(f8) debut(f8) censurd(f8) datfind(f8) f011t(f8) csp(f8) statut(f8) temps(f8) datmar2(f8).   2
   exe.                                                                                            3

set mxloops=600.                                                                                   4
   input program.                                                                                  5
     data list file="c:\base\rawdata"                                                              6
      /intnr(f8) debut(f8) censurd(f8) datfind(f8) f011t(f8) csp(f8) statut(f8) temps(f8) datmar2(f8).   7

    numeric tmonth(f4) censurdp (f4).                                                              8
      leave all.                                                                                   9
       compute tmonth=debut.                                                                      10
         loop.                                                                                    11
            compute censurdp=0.                                                                   12
            if tmonth=datfind censurdp=censurd.                                                   13
            end case.                                                                             14
           compute tmonth=tmonth+1.                                                               15
        end loop if tmonth=datfind+1.                                                              16
  end input program.                                                                               17
 exe.                                                                                              18
SAVE OUTFILE='c:\base\periode.sav'                                                                 19
 /COMPRESSED.                                                                                      20

compute dureep=tmonth-debut.                                                                       21
compute agep=tmonth-f011t.                                                                         22
do if tmonth<datmar2.                                                                              23
   compute mar=0.                                                                                  24
else.                                                                                              25
   compute mar=1.                                                                                  26
end if.                                                                                            27
exe.                                                                                               28

SAVE OUTFILE='c:\base\periodep.sav'                                                                29
 /COMPRESSED.                                                                                      30
```

Note : la numérotation des lignes a ici été ajoutée

D'un fichier initialement composé de 1 100 épisodes, nous arrivons ainsi à la création d'un fichier composé de 56 486 *personnes-mois*. Lors du développement des modèles logit à temps discret, d'autres variables ont été construites. Il s'agit essentiellement de variables dichotomiques qui ont été créées à partir des variables existantes. La construction de ces variables ne posant pas de difficultés particulières, nous ne développerons pas les commandes de syntaxe qui ont été nécessaires à leur définition. Il nous paraît plus intéressant de développer l'estimation de différents modèles de régression logistique à temps discret.

4. Un exemple d'analyse : la durée du premier emploi d'hommes nés entre 1945 et 1964

Le fichier que nous venons de construire constitue la base de données sur laquelle sont estimés des modèles de régression logistique à temps discret en vue d'analyser la durée des emplois. Dans ce point, nous considérons que le logit de la probabilité conditionnelle de départ d'un premier emploi des hommes nés entre 1945 et 1964 est une fonction du temps et des différentes caractéristiques que nous avons décrites dans le point précédent. Nous aurons pour démarche de partir de modèles simples, dans lesquels nous étudierons la dépendance au temps. Puis, nous introduirons les différentes caractéristiques individuelles et de l'emploi. Avant de commencer, signalons qu'avant même la construction du fichier « *personne-mois* », nous avons réalisé une estimation non-paramétrique de la probabilité de survie, c'est-à-dire, la probabilité de ne pas avoir quitté son emploi au cours du temps (figure 1[14]). Cette probabilité diminue fortement durant les tous premiers mois d'emploi. Moins de la moitié des emplois dure plus de trois ans. Il semble, en revanche, qu'un groupe d'individus reste très longtemps dans la situation de premier emploi.

4.1 La dépendance au temps

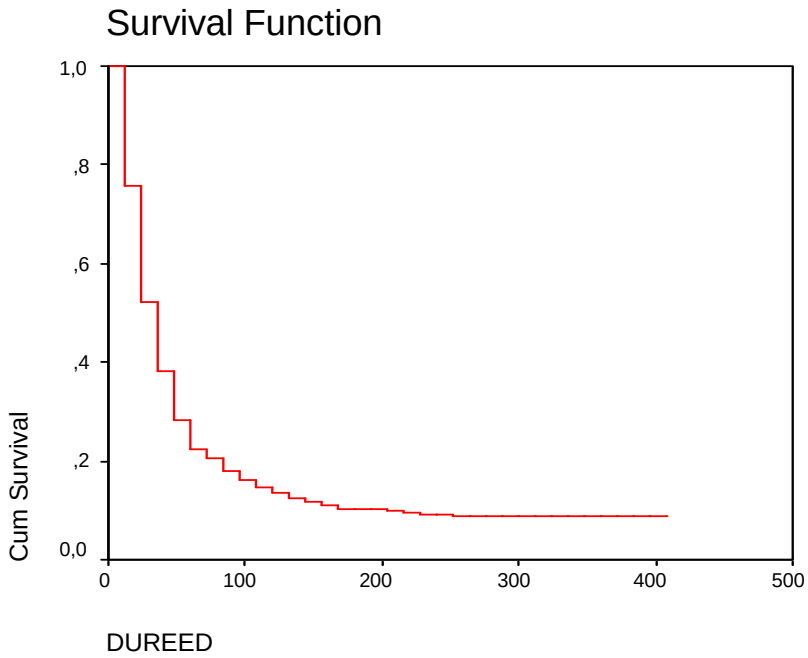
4.1.1 Hypothèses et formalisation de différents modèles prenant en compte la dépendance au temps

Une hypothèse couramment admise en socio-économie repose sur l'idée d'une acquisition d'expérience professionnelle spécifique au cours de l'exercice d'un emploi (Blossfeld & Rohwer, 2002). Plus la durée de l'activité professionnelle s'allonge, et plus il y a accumulation d'expérience professionnelle spécifique. En d'autres termes, plus la durée de l'emploi s'allonge et plus un départ devient coûteux pour l'employé, comme pour l'employeur. La probabilité conditionnelle de départ d'emploi devrait alors diminuer au cours du temps[15].

[14] Nous avons utilisé ici la méthode actuarielle d'estimation : cf. cahier sur les méthodes non-paramétriques de l'analyse des biographies (Le Goff & Forney, 2012). Les résultats que nous avons ici sont « presque » les mêmes que ceux qui avaient été présentés dans cet article : presque, car nous avons censuré les quelques individus qui ont divorcé alors qu'ils étaient encore dans leur premier emploi.

[15] Nous reprenons là une hypothèse très similaire à celle que nous avions formulée dans notre article sur les méthodes non-paramétriques (Le Goff & Forney, 2012). Nous avions vu dans cet article aussi une critique à ce type d'hypothèse en relation avec la notion d'hétérogénéité non-observée.

Figure 1 : Probabilité de ne pas avoir quitté sa première activité professionnelle (Estimation actuarielle, avec SPSS)



Cette hypothèse signifie aussi que le rapport de cette probabilité sur son complémentaire devrait aussi diminuer au cours du temps. Par conséquent, le logarithme de ce rapport devrait décroître. Il reste toutefois à se demander plus précisément de quel type est cette dépendance négative au temps. S'agit-il d'une diminution qui est linéaire ? En d'autres termes, y aurait-il une relation linéaire entre accumulation d'expérience professionnelle et le logit de la probabilité de départ d'emploi ? Ou bien, cette dépendance s'amenuiserait-elle au cours du temps ? En d'autres termes, l'expérience professionnelle (ou plutôt son manque) jouerait-elle seulement un rôle important sur la probabilité de départ d'emploi durant les premiers mois qui suivent l'embauche, mais perdrait par la suite de son importance ?

En termes plus techniques, notre questionnement pourrait être ainsi posé : La dépendance au temps du logit de la probabilité conditionnelle de départ d'emploi est-elle mieux modélisée par une fonction de Gompertz ou par une fonction de Weibull ? Pour répondre à cette question, nous allons estimer trois modèles, le premier dans lequel la dépendance au temps est paramétrée par une fonction de Gompertz (cf. équation 2.13), le deuxième, par une fonction de Weibull (cf. équation 2.14) et le troisième par une fonction de type Piecewise (cf. équation 2.15). L'idée est alors de comparer les résultats de l'ensemble des trois modèles, afin de déterminer lequel, parmi les deux premiers, présente des résultats qui se rapprochent le plus du troisième.

Avant de comparer les résultats des trois modèles, nous allons décrire de manière plus spécifique le modèle que l'on souhaite estimer dans une syntaxe SPSS ainsi que les informations obtenues dans « l'output » relatif aux résultats du modèle venant d'être spécifié. Dans le point suivant, nous détaillons l'estimation du modèle Piecewise.

4.1.2 Lecture d'un résultat de modèle logit à temps discret

Les intervalles de temps sélectionnés pour l'estimation du modèle de type Piecewise sont, 1) les mois 0 à 11 (première année), 12 à 23 (deuxième année), 24 à 35 (troisième année), 36 à 59 (quatrième et cinquième année), 60 à 119 (sixième à dixième année) et, 120^{e} mois et plus. Le modèle peut donc s'écrire ainsi :

$$\log\left(\frac{P(t)}{1-P(t)}\right) = constante + b_1(duree\,12\,à\,23) + b_2(duree\,24\,à\,35) + b_3(duree\,36\,à\,59) + b_4(duree\ 60\,à\,119) + b_5(duree\,120\,et\ plus) \quad (4.1)$$

La constante correspondra à la valeur du logit durant la première année. Tous les coefficients *b* mesurés pour les autres intervalles vont tenir compte de la diminution ou de l'augmentation de cette constante. En code-syntaxe SPSS, ce modèle est ainsi spécifié :

```
LOGISTIC REGRESSION VAR=censurdp
  /METHOD=ENTER dur1224 dur2436 dur3660 dur60120 dur120p.
```

La première ligne permet d'appeler la régression logistique et de déclarer la variable binaire que l'on cherche à expliquer, ici « *censurdp* ». La ligne suivante, commençant par la commande « /METHOD=ENTER » permet de déclarer les variables se rapportant aux intervalles de temps, dont on va estimer les coefficients *b*[16] (par défaut, SPSS estime une constante au modèle, qui correspond ici au logit de la probabilité de l'événement durant la première année). Une fois la procédure d'estimation du modèle lancée, les résultats sont affichés dans le fichier de sortie (output). Nous ne détaillerons pas ici l'ensemble du listing des résultats[17]. Nous nous concentrons plutôt sur les résultats qui permettent d'évaluer la qualité du modèle que l'on vient d'estimer. Le fichier de sortie présente d'abord un tableau de classification (*Case Processing Summary*) qui indique le nombre d'individus pris en compte, et le cas échéant, le nombre de données manquantes. Le tableau suivant (*Dependant Variable Encoding*) indique la manière dont SPSS recode la variable que l'on cherche à expliquer (ici, *censurdp*). Il est utile à ce stade de vérifier que la valeur recodée 1 correspond bien à la valeur 1 de la variable à expliquer (ici, *censurd*) car en dépend ici la valeur des coefficients. S'ensuit un petit tableau qui indique la valeur de la constante lorsque aucune variable n'est introduite dans le modèle (tableau 2). Dans la suite du texte, nous appellerons ce modèle le modèle 0. Il est important de souligner qu'il correspond à un modèle dans lequel la probabilité conditionnelle est considérée constante au cours du temps. Il s'agit donc du modèle géométrique (dans lequel la fonction de séjour suit une distribution géométrique).

Tableau 2 : Valeur de la constante lorsque aucune variable n'est introduite dans le modèle (modèle 0) ; détail du listing SPSS

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-4,032	,032	15730,167	1	,000	,018

Ce tableau en est suivi par d'autres qui présentent les résultats lorsque sont prises en compte les caractéristiques, dans notre cas les caractéristiques se rapportant aux intervalles de temps.

[16] Une autre possibilité consiste à passer par la boîte de dialogue. Dans ce cas, sélectionner Analyse, Régression et Binary Logistic. Choisir ensuite la variable « censurdp » comme variable dépendante du temps et les différentes covariables, puis coller l'ensemble sur une feuille de syntaxe.

[17] Pour l'ensemble de ces détails, nous invitons les lecteurs à consulter l'ouvrage de Kleinbaum & Klein (2002).

Un premier tableau (*Omnibus Tests of Model Coefficients*) contient la significativité du modèle et les caractéristiques introduites (tableau 3) :

Tableau 3 : Significativité du modèle (modèle 1) ; détail du listing SPSS

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	411,998	5	,000
	Block	411,998	5	,000
	Model	411,998	5	,000

Ce tableau donne en surtout les résultats d'un *test du logarithme de vraisemblance* dont le résultat est important à prendre en compte car ce test permet de voir en quoi l'introduction de nouvelles caractéristiques améliore le modèle de base, ici le modèle 0. Ce test[18] repose sur le calcul de la différence entre l'opposé du double du logarithme du maximum de vraisemblance (-2LMV) du modèle 1 et celui du modèle 0. Le nombre obtenu est alors comparé à un $\chi 2$ dont les degrés de liberté correspondent au nombre de caractéristiques qui sont introduites dans le modèle. Dans le cas présent, la statistique obtenue est de 411,998 que l'on compare avec un $\chi 2$ à 5 degrés de liberté (cf. dernière ligne du tableau 3). Le test est significatif, ce qui permet de conclure que l'introduction des caractéristiques améliore le modèle. Les deux premières lignes du tableau donnent les mêmes résultats. Toutefois, il deviendra utile de les consulter lorsque l'on introduira de nouvelles caractéristiques dans le modèle (cf. points 4.2 et 4.3 de cet article). Remarquons que le tableau suivant donne les résultats de la valeur de -2LMV pour le modèle 1[19] (tableau 4).

Tableau 4 : Valeur du maximum de vraisemblance (extrait de listing SPSS)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	9517,465	,007	,045

Enfin, un dernier tableau (*Variables in the Equation*) indique les résultats des coefficients *b* (tableau 5).

Tableau 5 : Valeur des coefficients b du modèle de type Piecewise (extrait du listing SPSS)

[18] Ce test est valable dans l'ensemble des modèles de régression faisant appel à une résolution par la technique du maximum de vraisemblance

[19] A noter que certains logiciels statistiques n'indiquent que la valeur du logarithme du maximum de vraisemblance.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	DUR1224	.085	.087	.965	1	.326	1.089
	DUR2436	-.091	.102	.794	1	.373	.913
	DUR3660	-.303	.100	9.160	1	.002	.739
	DUR60120	-1.146	.117	96.331	1	.000	.318
	DUR120P	-2.225	.176	160.483	1	.000	.108
	Constant	-3.590	.055	4276.037	1	.000	.028

a. Variable(s) entered on step 1: DUR1224, DUR2436, DUR3660, DUR60120, DUR120P.

Il est important d'indiquer ici, que les coefficients *b* (dans le tableau, les coefficients B) sont exprimés par rapport à la constante qui est en fait la valeur du logit de la probabilité de connaître l'événement durant la première année. Le logit de quitter son emploi entre, par exemple le $36^{ième}$ et le $59^{ième}$ mois est égal à (-3,590) + (-0,303)=-3,893. De même, le logit de sortir d'observation entre le $60^{ième}$ et $119^{ième}$ mois sera égal à (-3,590) + (-1,146)=-4,736. Le tableau donne la valeur de l'écart-type de chacun des coefficients estimés (2^{e} colonne). Il présente ensuite les résultats d'un test de Wald pour chacun des coefficients (3^{e} colonne). Le test de Wald permet de tester l'hypothèse nulle selon laquelle B n'est pas différent de 0. La statistique de Wald se distribue selon une loi du χ^2 à 1 degré de liberté. Dans le cas présent, le test est significatif à partir de la troisième année qui suit l'embauche dans l'emploi. Ceci signifie que la probabilité de quitter son emploi devient plus faible après la troisième année d'exercice de son activité. La dernière colonne du tableau donne les valeurs de l'exponentiel des coefficients estimés, c'est-à-dire, les *odds ratio.*

Le modèle aurait pu être spécifié autrement, c'est-à-dire, sans estimer une constante, mais directement avec un coefficient pour l'intervalle de temps correspondant à la première année :

$$\log\left(\frac{P(t)}{1-P(t)}\right) = b'_0\,(duree\,0\,à\,12) + b'_1\,(duree\,12\,à\,23) + b'_2\,(duree\,24\,à\,35) + b'_3\,(duree\,36\,à\,59)$$
$$+ b'_4\,(duree\,60\,à\,119) + b'_5\,(duree\,120\,et\,plus) \qquad (4.2)$$

L'ensemble des coefficients b'_1 à b'_5 n'exprimeront plus alors l'augmentation ou la diminution du logit par rapport à celui correspondant au premier intervalle de temps, mais indiqueront directement la valeur des logit des intervalles de temps auxquels ils se réfèrent. En code-syntaxe SPSS, le modèle est spécifié ainsi :

```
LOGISTIC REGRESSION VAR=censurdp
  /METHOD=ENTER dur012 dur1224 dur2436 dur3660 dur60120 dur120p
  /ORIGIN.
```

L'ajout de l'option origine revient à demander à ce que ne soit pas estimée de constante. On retrouve les résultats du modèle précédent (même valeur pour –2LMV). La différence porte sur la valeur des coefficients en ce sens que l'on a directement les coefficients du logit des probabilités conditionnelles pour chaque intervalle de temps (tableau 6).

Tableau 6 : valeur des coefficients du modèle sans constante (extrait de listing SPSS)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	DUR012	-3,590	,055	4276,037	1	,000	,028
	DUR1224	-3,505	,067	2742,546	1	,000	,030
	DUR2436	-3,681	,087	1810,694	1	,000	,025
	DUR3660	-3,892	,084	2167,774	1	,000	,020
	DUR60120	-4,736	,103	2111,965	1	,000	,009
	DUR120P	-5,815	,167	1214,665	1	,000	,003

a. Variable(s) entered on step 1: DUR012, DUR1224, DUR2436, DUR3660, DUR60120, DUR120P.

4.1.3 Comparaison de différentes spécifications de la dépendance au temps

Le modèle de Gompertz se spécifie ainsi dans SPSS (cf. équation 2.12) :

```
LOGISTIC REGRESSION VAR=censurdp
  /METHOD=ENTER dureep1
```

Où *dureep1* est égale à la variable *dureep* plus 1 : on considère ici que les individus sont sortis durant le *n-ième* mois de leur activité professionnelle. De même, le modèle de Weibull se spécifie de la manière suivante (cf. équation 2.13) :

```
LOGISTIC REGRESSION VAR=censurdp
  /METHOD=ENTER lnduree
```

Où « *lnduree* » correspond au logarithme de *dureep1*.

Les résultats des différents modèles sont synthétisés dans le tableau 7. La valeur du paramètre *a* est négative, aussi bien dans le cas du modèle de Gompertz que dans le cas du modèle de Weibull (les coefficients étant tous les deux significatifs). En d'autres termes, les paramètres obtenus indiquent que la probabilité conditionnelle de départ d'emploi diminue au cours du temps. Dans les deux cas, le test du maximum de vraisemblance indique aussi que le modèle est amélioré par rapport à un modèle dans lequel n'interviendrait qu'une constante (modèle 0). Toutefois, la valeur du maximum de vraisemblance obtenue (ici, l'opposé de son double) est plus grande (ici, plus petite) dans le cas du modèle de Gompertz que dans le cas du modèle de Weibull. Qui plus est, elle se rapproche fortement de la valeur du maximum de vraisemblance obtenue dans le modèle de type Piecewise. Ceci tend à montrer qu'une relation linéaire entre le logit de la probabilité conditionnelle et le temps est plus adéquate qu'une relation linéaire entre ce logit et le logarithme du temps. Cette meilleure adéquation du modèle de Gompertz à nos données par rapport au modèle de Weibull peut être observée de manière plus concrète par l'intermédiaire d'un graphique qui retrace l'évolution de la probabilité conditionnelle au cours du temps. Pour estimer la distribution de ces probabilités au cours du temps, il est nécessaire de partir des valeurs du logit des probabilités conditionnelles. Ainsi, si au temps *t* :

$$\frac{P(t)}{1-P(t)} = d \qquad (4.3)$$

alors :

$$P(t) = \frac{d}{1+d} \qquad (4.4)$$

Les distributions de probabilité de départ d'emploi pour chacun des quatre modèles sont représentées sur la figure 2. Les modèles de Gompertz, Weibull et Piecewise montrent des probabilités qui diminuent au cours du temps, à l'exception du modèle Piecewise en tout début de période d'observation. Il est vrai néanmoins que chacun des modèles a été estimé en considérant que les sorties d'emploi pouvaient avoir lieu dès le premier mois alors que, rappelons-le, nos données concernent des emplois qui ont duré plus de trois mois. En toute

rigueur, il aurait sans doute fallu prendre en compte cet aspect. Néanmoins, on peut considérer que le modèle Piecewise reflète assez bien la distribution de la probabilité conditionnelle au cours du temps. Comme on peut le voir, la distribution obtenue à partir du modèle de Gompertz est proche de celle obtenue avec le modèle Piecewise. En revanche, la distribution obtenue à partir du modèle de Weibull est plus éloignée en ce sens que cette modélisation accentue la diminution de la probabilité conditionnelle durant les premiers mois d'observation, mais surévalue les probabilités conditionnelles en fin de période.

Tableau 7 : valeurs des différents paramètres des modèles géométriques, Piecewise, Gompertz et Weibull

	Géométrique		Piecewise		Gompertz		Weibull	
	coef.	Sig.	Coef.	Sig.	Coef.	Sig.	Coef.	Sig.
DUREE 12 à 24			0.085					
DUREE 24 à 36			-0.091					
DUREE 36 à 60			-0.303	***				
DUREE 60 à 120			-1.146	***				
DUREE 120 et plus			-2.225	***				
a (Gompertz)					-0.130	***		
a (Weibull)							-0.285	***
Constante	-4.032	***	-3.590	***	-3.395	***	-0.376	***
-2LMV	9929.46		9517.47		9534.90		9771,49	

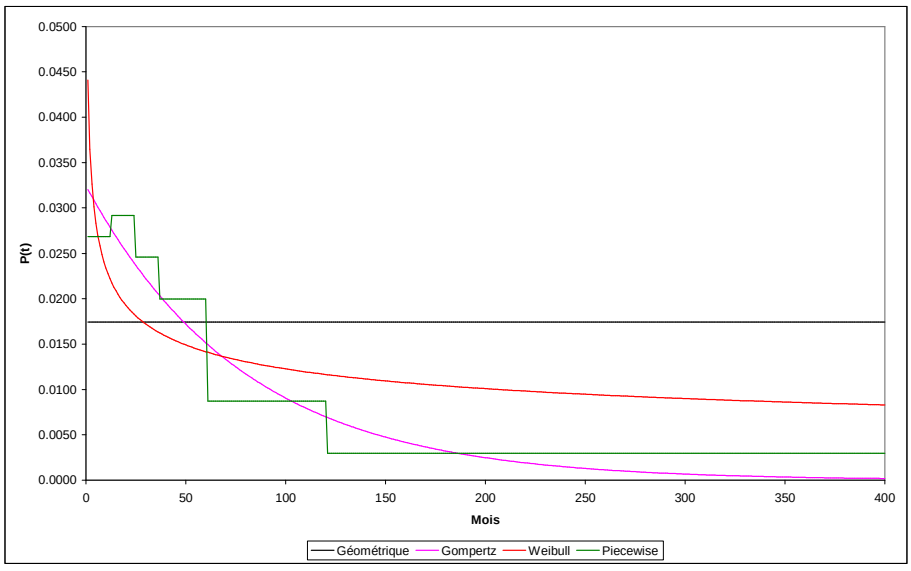
* significatif à un seuil de 10%

** significatif à un seuil de 5%

*** significatif à un seuil de 1%

Pour revenir à notre hypothèse de départ, il semble qu'au processus d'accumulation d'expérience professionnelle au fur et à mesure que la durée de l'emploi s'allonge correspond une relation linéaire entre le logit de la probabilité de sortie d'emploi et le temps. Toutefois, il est important de noter que ce résultat doit être pris avec prudence, notamment, parce que ces modèles ont été estimés sans prendre en compte les différentes caractéristiques individuelles ou de l'emploi.

Figure 2 : distribution de probabilités estimée à partir des différentes spécifications de la dépendance au temps



4.2 Introduction d'autres horloges temporelles

Dans le point précédent, l'horloge prise en compte était la durée de l'emploi. On peut, toutefois se demander si la probabilité conditionnelle ne dépend pas d'autres horloges. Dans le cas présent, nous allons porter notre intérêt sur les dépendances entre la probabilité conditionnelle et le temps calendaire (ou temps historique) ainsi que l'âge des individus.

4.2.1 Effet de période

La variable « *tmonth* », que nous avons construite lors de la phase d'élaboration du fichier personne-mois, indique le nombre de mois écoulé depuis janvier 1945. Il s'agit donc d'une variable temporelle dont l'échelle se situe sur le calendrier historique. Dans un grand nombre d'analyse démographique, sont prises en compte des variables de mois ou d'année en tant que caractéristiques fixes, le plus souvent pour mesurer *un effet de cohorte*. On pourrait ainsi prendre dans notre exemple la date de début d'emploi afin de mesurer l'impact de ce moment sur la probabilité de départ d'emploi, quelle que soit la durée d'emploi à laquelle on se situe. Dans le cas présent, on considèrera cette variable en tant que variable dépendante du temps, c'est-à-dire, que si l'embauche dans un emploi a eu lieu en janvier 1960, cette variable aura pour valeur 121[20] lors de l'embauche, 122 le mois suivant, (etc.). En d'autres termes, plutôt qu'à un effet de cohorte, nous nous intéressons plutôt à un *effet de période*. Dans nos données, le premier emploi a été occupé en 1960, si bien que le calendrier de notre analyse va porter sur une période comprise entre cette date et le milieu des années quatre-vingt-dix, date de l'enquête. Une partie de cette période concerne les moments qui suivent les chocs pétroliers qui, s'ils ne sont pas traduits par un chômage important en Suisse, ont correspondu à un fort ralentissement économique.

On peut ainsi faire l'hypothèse, qu'en raison de ce ralentissement économique, les individus ont été moins mobiles, notamment parce qu'il y a moins eu création ou libération de postes auxquels ces personnes auraient pu prétendre après leur premier emploi. Cette moindre mobilité peut s'être accentuée dans les années quatre-vingt-dix lors de l'apparition en Suisse d'un chômage important. En conséquence, notre hypothèse est que la probabilité conditionnelle de départ d'emploi a diminué au fil des ans. L'hypothèse alternative d'une augmentation de cette probabilité ne doit toutefois pas être négligée. A ces crises économiques correspond l'émergence de nouvelles techniques de management des personnels qui favorisent la mobilité professionnelle, voire la précarité de l'emploi, notamment pour les jeunes débutants. En conséquence, la probabilité de départ d'emploi peut avoir augmenté.

Nous allons ici estimer des effets de période en construisant des variables indiquant à quelle période du calendrier on se situe. L'ensemble de la période est découpé en tranche de cinq années, à l'exception des années soixante. Le modèle logit est spécifié de manière analogue au modèle de type Piecewise que nous avons développé précédemment. Ainsi, nous allons estimer des constantes pour chacun des intervalles de temps calendaire que nous venons de définir en considérant la première période des années soixante en tant que période de référence. Bien entendu, le modèle prend toujours en compte les effets de la durée d'emploi, si bien que ce dernier sera mesuré pour la période de référence, c'est-à-dire, les années soixante. Le modèle se présente alors ainsi :

[20] C'est-à-dire, le 121^{ième} mois depuis janvier 1945.

$$Log\left(\frac{P(t,x_t)}{1-P(t,x_t)}\right)=constante+....+\beta_6(cal\,70\,à\,74)+\beta_7(cal\,75\,à\,79)+\beta_8(cal\,80\,à\,84)$$
$$+\beta_9(cal\,85\,à\,89)+\beta_{10}(cal\,90\,à\,94) \qquad (4.5)$$

En syntaxe SPSS, le modèle se spécifie ainsi :

```
LOGISTIC REGRESSION VAR=censurdp
  /METHOD=ENTER dur1224 dur2436 dur3660 dur60120 dur120p
  /METHOD=ENTER cal7074 cal7579 cal8084 cal8589 cal9094.
```

L'ensemble des variables *cal.* (moment du calendrier) aurait pu être spécifié directement après les variables *dur* (durée de l'emploi). Toutefois, en dédoublant l'instruction *method=enter*, nous aurons dans le fichier de résultats ceux qui concernent successivement un modèles 0 (tous les coefficients sont égaux à 0) avec l'estimation d'une seule constante, un modèle 1 (tous les coefficients associés aux variables *cal* sont égaux à 0), ces deux modèles correspondant donc aux modèles 0 et 1 que nous avions estimés précédemment, et un modèle 2 (tous les coefficients sont différents de 0). La progression peut être examinée à partir du tableau « *omnibus* » (tableau 8). Nous avons vu que la valeur du -2LMV pour le modèle 1 était égale à 9517,465. Lorsqu'il y a estimation du modèle prenant en compte des effets de calendrier, cette valeur devient 9507,196. La différence entre les deux modèles correspond à 9517,465-9507,196=10,269. Cette différence est indiquée dans la première ligne du tableau 8, cette ligne étant consacrée à la progression du modèle entre deux étapes. Le χ^2 calculé est donc égal à 10,269, et le test apparaît seulement significatif au seuil de 10%. En d'autres termes, les effets de période sont très faibles, voire non significatifs. Signalons que la dernière ligne du tableau 8 indique la progression du maximum de vraisemblance entre le modèle 0 et le modèle 2. Le test est ici, bien évidemment significatif.

Tableau 8 : amélioration du modèle lorsque l'on introduit les effets de moment (extrait de listing SPSS)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	10,268	5	,068
	Block	10,268	5	,068
	Model	422,266	10	,000

Les résultats des coefficients montrent une diminution de la probabilité de sortie d'emploi qui est significative seulement durant la période qui va de 1970 à 1974[21]. En revanche, les rythmes de sortie redeviennent semblables à ce qu'ils étaient dans les années soixante dans les années quatre-vingt et quatre-vingt-dix. Ni l'hypothèse d'une diminution de la probabilité de départ d'emploi, ni l'hypothèse alternative ne semblent vérifiées. Il se peut, néanmoins, que les deux effets de la moindre mobilité et de l'augmentation de la précarité des jeunes aient joué simultanément, l'un ayant annulé l'autre. Dans les modèles suivants, nous garderons les variables se rapportant à l'effet de moment. Il est, en effet, possible que des caractéristiques deviennent significatives au fur et à mesure que l'on introduit différentes variables.

[21] Résultat non montré ici.

4.2.2 Effet d'âge

Dans ce point, notre intérêt porte sur l'effet de l'âge « *toutes choses égales par ailleurs* », cette dernière expression voulant dire dans le cas présent, *compte tenu de l'effet de la durée d'emploi et des effets de momen*t. Comme précédemment, nous nous intéressons à l'âge en tant que variable dépendante du temps et non en tant que variable fixe, par exemple, au moment de l'embauche dans l'emploi. Notre hypothèse est que la probabilité conditionnelle de départ d'emploi diminue avec l'âge. En effet, on peut supposer que les individus aspirent de moins en moins à la mobilité au fur et à mesure qu'ils vieillissent, notamment parce que ces individus connaissent le passage des différentes étapes de l'entrée dans la vie adulte et l'acquisition de l'autonomie économique durant cette phase de la vie.

Pour mesurer les effets de l'âge sur la probabilité conditionnelle de départ d'emploi, nous avons créé quatre groupes d'âge quinquennaux entre 15 et 34 ans, un cinquième groupe d'âge correspondant à 34 ans et plus. Le groupe d'âge 20-24 ans constitue la catégorie de référence dans notre modèle. Le modèle logit s'écrit alors (modèle 3) :

$$\log\left(\frac{P(t,x_t)}{1-P(t,x_t)}\right)=constante....+\beta_{11}(age\ 15\ à\ 19)+\beta_{12}(age\ 25\ à\ 29)+ \qquad (4.6)$$
$$\beta_{13}(age\ 30\ à\ 34)+\beta_{14}(age\ 34 et\ plus)$$

Dans la syntaxe SPSS, il suffit d'ajouter les quatre variables introduites. La mention *method=enter* nous permet d'estimer un modèle par étape de la même manière que précédemment :

```
LOGISTIC REGRESSION VAR=censurdp
  /METHOD=ENTER dur1224 dur2436 dur3660 dur60120 dur120p
  /METHOD=ENTER cal7074 cal7579 cal8084 cal8589 cal9094
  /METHOD=ENTER ag1519 ag2529 ag3034 ag34p.
```

La statistique de comparaison des maximum de vraisemblance entre les modèles 2 et 3 est égale à 18,276, et est significative au seuil de 1% (le nombre de degrés de liberté est cette fois-ci de 4). Les coefficients estimés montrent une tendance à une diminution de la probabilité de départ d'emploi en fonction de l'âge, ces coefficients devenant significatifs à partir de 25 ans (tableau 9). Ce résultat confirme cette fois-ci notre hypothèse. Il est à noter, pour terminer, que les coefficients associés aux variables relatives au temps calendaire ne deviennent pas davantage significatives.

4.3 Effet des autres caractéristiques

4.3.1 Introduction des caractéristiques de l'emploi

Ainsi que nous l'avons mentionné, les emplois sont distingués en fonction de trois caractéristiques. La première concerne la fonction exercée et a été regroupée en activité professionnelle de type primaire (agriculture, sylviculture), industrielle ou de service[22]. On

[22] Attention, il ne s'agit pas ici de la branche d'activité de l'employeur, mais de la fonction exercée. Ainsi, une activité de type employé de bureau, mais exercé dans une industrie sera classée en tant qu'activité de type service.

peut faire l'hypothèse que les activités de type primaire sont plus stables que les autres types d'activités. La seconde caractéristique concerne le statut occupé dans l'emploi, en distinguant les indépendants, les cadres, les collaborateurs (qui n'exercent aucune activité de direction ou de cadre dans leur entreprise), cette catégorie constituant alors le groupe de référence, et les autres statuts ou statuts inconnus. On peut supposer que les indépendants sont les plus stables dans leur activité professionnelle. La troisième variable décrit le temps de travail. Les individus de référence sont ceux qui travaillent à temps plein, et les deux autres modalités sont « travail à temps partiel » et « temps de travail indéterminé ». La deuxième modalité correspond, notamment, à des individus dont le temps de travail est irrégulier. Comme il s'agit de premiers emplois masculins, on peut supposer que le départ de l'emploi est à chaque moment plus fréquent lorsque l'emploi est occupé à temps partiel ou avec des horaires indéterminés. Les trois caractéristiques relatives à l'emploi ont été ajoutées une à une, selon le même principe que nous avons adopté précédemment (tableau 10).

Tableau 9 : Odds ratio et significativité (Effet de différentes horloges)

	Modèle 0		Modèle 1		Modèle 2		Modèle 3	
	Exp(B)	Sig.	Exp(B)	Sig.	Exp(B)	Sig.	Exp(B)	Sig.
DUREE 0 à 11			1		1		1	
DUREE 12 à 23			1.089		1.091		1.062	
DUREE 24 à 35			0.913		0.914		0.887	
DUREE 36 à 59			0.739	***	0.738	***	0.729	***
DUREE 60 à 119			0.318	***	0.317	***	0.343	***
DUREE 120 et plus			0.108	***	0.108	***	0.205	***
CAL 60 à 69					1		1	
CAL 70 à 74					0.738	***	0.726	***
CAL 75 à 79					0.822	*	0.815	*
CAL 80 à 84					0.898		0.888	
CAL 85 à 89					1.000		1.012	
CAL 90 à 94					0.810		0.975	
AGE 15 à 19 ans							0.871	
AGE 20 à 24 ans							1	
AGE 25 à 29 ans							0.864	**
AGE 30 à 34 ans							0.620	***
AGE 35 et plus							0.249	***
Constante	0.018	***	0.028	***	0.032	***	0.034	***
-2LMV	9929.463		9517.465		9507.196		9488,420	

* significatif à un seuil de 10%

** significatif à un seuil de 5%

*** significatif à un seuil de 1%

Les résultats montrent qu'il n'y a pas de différences selon le type d'activité. Les emplois de type primaire pourraient sembler moins fréquemment quittés, mais le coefficient est significatif seulement au seuil de 10% (modèle 4). Cet effet disparaît lorsque l'on introduit le statut d'emploi (modèle 5). Ainsi, les indépendants quittent moins souvent leur activité professionnelle, ce qui va dans le sens de notre hypothèse. On peut supposer que la disparition de l'effet associé à un travail de type agriculture est dû au fait que beaucoup d'emplois dans ce secteur d'activité sont des emplois indépendants. Notons, enfin, que les cadres ont eux aussi une moins grande probabilité conditionnelle de départ d'emploi. En ce qui concerne la troisième variable, indiquons seulement qu'aucune des caractéristiques concernant le temps de travail n'a d'effet sur la probabilité de départ d'emploi, infirmant ainsi l'hypothèse que nous avions formulée (modèle 6).

Tableau 10 : Odds ratio et significativité (avec les caractéristiques de l'emploi)

	Modèle 3		Modèle 4		Modèle 5		Modèle 6		Modèle 7	
	Exp(B)	Sig.	Exp (B)	Sig.	Exp(B)	Sig.	Exp(B)	Sig.	Exp(B)	Sig.
DUREE 0 à 11	1		1		1		1		1	
DUREE 12 à 23	1.062		1.074		1.075		1.075		1.080	
DUREE 24 à 35	0.887		0.904		0.909		0.909		0.923	
DUREE 36 à 59	0.729	***	0.748	***	0.751	***	0.752	***	0.772	**
DUREE 60 à 119	0.343	***	0.356	***	0.345	***	0.346	***	0.361	***
DUREE 120 et plus	0.205	***	0.214	***	0.198	***	0.199	***	0.216	***
CAL 60 à 69	1		1		1		1		1	
CAL 70 à 74	0.726	***	0.704	***	0.695	***	0.694	***	0.702	***
CAL 75 à 79	0.815	*	0.792	**	0.773	**	0.773	**	0.778	**
CAL 80 à 84	0.888		0.863		0.842		0.842		0.839	
CAL 85 à 89	1.012		0.979		0.967		0.966		0.954	
CAL 90 à 94	0.975		0.942		0.918		0.918		0.892	
AGE 15 à 19 ans	0.871		0.905		0.886		0.886		0.881	
AGE 20 à 24 ans	1		1		1		1		1	
AGE 25 à 29 ans	0.864	**	0.872		0.933		0.931		0.982	
AGE 30 à 34 ans	0.620	***	0.622	**	0.706		0.703		0.760	
AGE 35 et plus	0.249	***	0.251	***	0.291	***	0.289	***	0.322	***
AGRICULTURE			0.731	*	0.958		0.959		0.957	
INDUSTRIE			1.088		1.078		1.080		1.082	
SECONDAIRE					1		1		1	
INDEPENDANT					0.200	***	0.200	***	0.201	***
CADRE SUP.					0.701	**	0.701	**	0.705	**
EMPLOYE					1		1		1	
AUTRE					0.750		0.750		0.756	
TEMPS PLEIN							1		1	
TEMPS PARTIEL							1.017		1.021	
INDETERMINE							1.001		0.995	
MAR01									0.957	
MAR12									0.669	*
MAR25									0.790	
MAR5P									0.823	
Constante	0.034	***	0.033	***	0.035	***	0.035	***	0.035	***
-2LMV	9488,420		9481.545		9451.573		9451.555		9446.741	

* significatif à un seuil de 10%

** significatif à un seuil de 5%

*** significatif à un seuil de 1%

4.3.2 Le rôle joué par le mariage sur le départ d'emploi

La dernière caractéristique individuelle introduite dans notre modèle est le mariage. Non seulement, il s'agit d'une caractéristique dépendante du temps, mais nous allons essayer de voir si le rôle du mariage évolue au fur et à mesure que le temps s'écoule après celui-ci, par exemple, en s'effaçant. On va considérer le temps écoulé depuis le mariage sous la forme d'une nouvelle horloge. Cette horloge est particulière en ce sens qu'elle ne concerne que les seules personnes mariées. On parle alors *d'horloge conditionnelle*. On peut supposer qu'un mariage a pour effet immédiat de diminuer la probabilité conditionnelle de départ d'emploi, les hommes préférant peut-être resté stables dans leur emploi lorsqu'ils commencent à former une famille. On peut, en outre, supposer que le mariage va jouer un rôle à long terme, en ce sens que l'effet négatif du mariage sur la probabilité de départ d'emploi va se prolonger, même longtemps après le mariage. Nous allons ainsi subdiviser la période qui suit le mariage

en différents intervalles de temps (0-11 mois, 12-23 mois, 24-59 mois et 60 mois et plus). Dans le cas présent, toutes ces variables sont introduites dans le modèle car les individus de référence sont ici les personnes qui ne se sont pas mariées. Les résultats du modèle figurent dans le tableau 10 (modèle 7). Ils montrent que le mariage n'a pas d'effet sur les probabilités de départ d'emploi, ni à court, ni à long terme.

5. Conclusion

Les modèles à temps discret doivent être privilégiés par rapport à des modèles en temps continu dès lors que les données de durée dont on dispose s'appuient sur une unité de mesure du temps qui est large (l'année). Ils sont, en outre, fortement conseillés à utiliser lorsque l'on dispose de données dans lesquelles un grand nombre d'individus connaissent l'événement étudié à un même moment. Parmi les modèles à temps discret, les modèles logit à temps discret sont faciles à utiliser puisqu'ils nécessitent seulement la compréhension des procédures permettant d'estimer un modèle logit classique. Bien qu'il puisse paraître arbitraire de choisir ce type de modèle, il présente l'avantage d'estimer des coefficients auxquels correspondent des probabilités qui sont comprises entre 0 et 1 (Vermunt, 1997).Il est estimé sur des fichiers personnes-périodes. Si ces fichiers personnes-périodes correspondent à rendre indépendantes les unes des autres chaque unité de temps d'un individu, cela ne joue en rien un rôle sur l'estimation des coefficients associés à chacune des différentes caractéristiques introduites dans le modèle.

Dans cet article, nous nous sommes concentrés sur des aspects de modélisation se rapportant à l'analyse d'un événement simple. Des modèles similaires peuvent toutefois être estimés, en considérant plusieurs événements, tels que deux événements en concurrence, des échéances multiples, un événement selon sa cause. De tels modèles peuvent bien évidemment être estimés avec la procédure « LOGISTIC REGRESSION » de SPSS, en estimant un modèle pour chaque type d'échéance : lorsqu'il y a estimation d'un modèle pour une échéance donnée, les individus qui connaissent un autre type d'échéance doivent alors être censurés au moment de l'occurrence de cet autre évènement. Toutefois, on peut aussi envisager d'utiliser la procédure « NORMREG » de SPSS qui permet d'estimer des modèles de régression logistique multinomiale (Kleinbaum & Klein, 2002). Dans ce cas, on estime un ensemble de coefficients pour chacune des échéances possibles.

Références bibliographiques

Allison Paul D. (1982). "Discrete-Time Methods for the Analysis of Event Histories". In *Sociological Methodology*, (13): 61-98.

Allison Paul D. (1984). *Event History Analysis. Regression for Longitudinal Data. Series: Quantitative Applications in the Social Sciences*. Newbury Park: Sage Publications, Vol 46.

Allison Paul D. (1995). *Survival Analysis Using the SAS® System*. Cary: SAS Campus Drive.

Blossfeld Hans-Peter & Rohwer Goetz (2002). *Techniques of Event History Modeling. New Approaches to Causal Analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Courgeau Daniel & Lelièvre Eva (1989). *Analyse démographique des biographies*. Paris : INED.

Cox David R. (1972). "Regression models and life-tables". *Journal of the Royal Statistical Society*, Series B, 34:187-202.

Dieckmann Andreas (1990). "Diffusion and Survival Models for the Process of entry into Marriage" in Tuma Nancy and Mayer Karl-Ulrich (eds.), *Event History Analysis in Life course Research*. Madisson: University of Wisconsin Press: 170-183.

Gabadinho Alexis (1998), *L'enquête suisse sur la famille*, Berne, OFS.

Gabadinho Alexis & Wanner Philippe (1999), *Fertility and Family Surveys in Countries of the ECE Region. Standard Country Report. Switzerland*, Geneva, United Nations Economic Commission for Europe, United Nations Population Fund, Economics Studies n°10m.

Hosmer David W. & Lemeshow Stanley (2000), *Applied Logistic Regression*. New-York: John Wiley and Sons (2nd ed.).

Kleinbaum David G. (1996). *Survival Analysis. A Self-Learning Text*. Berlin-New-York: Springer-Verlag.

Kleinbaum David G. & Klein Mitchel (2002). *Logistic Regression. A Self-Learning Text*. Berlin-New-York: Springer Verlag (2nd ed.).

Le Goff Jean-Marie & Forney Yannic (2012). Méthodes non-paramétriques de l'analyse des événements du parcours de vie (Event history Analysis). Estimations avec SPSS. Méthode de Kaplan-Meier et méthode actuarielle. *Cahiers recherche et méthodes*, 2.

Menard Scott (2001). *Applied Logistic Regression Analysis*. New-Bury Park: Sage (2nd ed.).

Mills Melinda (1999). *Construction of Input Data for Log-Linear Models of Event Histories.* Groningen : Population Research Centre-University of Groningen. Working Paper n0 3.

Pressat Roland (1983), *L'analyse démographique*. Paris : PUF (3e ed.).

Preston Samuel, Heuveline Patrick & Guillot Michel (2001). *Demography. Measuring and Modeling Population Processes*. Oxford: Blackwell Publishers.

SPSS (1994). *SPSS 6.1 Syntax Reference Guide*. USA: SPSS Inc.

Therneau Terry M. & Grambsch Patricia (2000). *Modeling Survival Data. Extending the Cox Model*. New-York, Berlin: Springer Verlag. Statistics for Biology and Health

Vermunt Joeren K. (1997). *Log-Linear Models for Event Histories*. Newbury-Park: Sage publications.

Yamaguchi Kazuo (1991). *Event History Analysis*. Applied Social Research Methods Series. Newbury Park: Sage Publications, Vol. 28.