Modèles d’analyse des biographies en temps discret
Exemple d’utilisation

Jean-Marie Le Goff
Centre Lines
Pôle National de recherche Lives
Université de Lausanne

Plan

- **Deux types de données discrètes**
- **Modèles à temps discret**
  - **Modèle logistique**
  - **Modèle complémentaire log-log**
- **Préparation d'une base de données**
- **Exemple: Diffusion d'internet en Suisse (panel suisse des ménages)**
- **Conclusion**

Deux types de temps discret

- **Modèles d'analyse des biographies s'appuient sur l'idée d'un temps continu**
  - **modèle de Cox, modèles paramétriques**
  - **Peu de personnes connaissent l'événement d'intérêt à un même moment.**
- **Cette approche en temps continu n'est pas toujours adéquate**

Deux types de temps discret

- **1) Evénements ont lieu à certains moments réguliers dans le temps**
  - **Exemples**
    - **Passage pour des élèves ou des étudiants dans une année supérieure (septembre)**
    - **Début d'une activité professionnelle ou cessation (début ou fin de mois)**
  - **« Vrai » temps discret**

Deux types de temps discret

- **2) Mesure et unité de temps**
  - **Les événements interviennent bien de manière continue**
  - **Les intervalle de temps sont longs (une année)**
  - **Beaucoup de personnes connaissent l'événement durant un intervalle (*Tied data*)**
  - **Données de panel (changement d'une vague à une autre)**

Deux types de temps discret

- **Modèles de Cox sensibles à ces *tied data* (ainsi que modèles paramétriques)**

- **Méthodes d'approximation**
  - **Breslow**
  - **Efron**
  - **Peu efficaces lorsque le nombre d'événements ayant lieu à un même moment est élevé**

Estimer des modèles à temps discret (Allison, 1982)

- **Yamaguchi (1991):**
  - **Mise en œuvre de ces modèles dès lors que 5% des personnes connaissent l'événement d'intérêt dans un intervalle de temps ou à un moment donné**

- **Est modélisée la probabilité de connaitre l'événement à un temps (intervalle) donné sachant qu'on ne l'a pas connu auparavant (risque en temps discret):**

$$P(t_l) = P(T = t_l \mid T \geq t_l)$$

Premier modèle: modèle logit à temps discret

$$P(t, x_t) = \frac{1}{1 + \exp\left[-\left(\alpha_t + \beta x_t\right)\right]}$$

$$\log\left[\frac{P(t, x_t)}{1 - P(t, x_t)}\right] = \alpha_t + \beta x_t$$

- **α_t: fonction du temps**
- **« Vrai » temps discret**
- **Si $P(t,x_t)$ proche de 0, $1-P(t,x_t)$ est proche de 1: modélisation de la probabilité conditionnelle**
- **Alternative: modèle probit (Box-Stephensmeier & Jones, 1997)**

Deuxième modèle: modèle complémentaire loglog

$$P(t, x_t) = 1 - \exp\left[-\exp(\alpha_t + \beta x_t)\right]$$

$$\log\left[-\log\left(1 - P(t, x_t)\right)\right] = \alpha_t + \beta x_t$$

- **Coefficients *β* estimés correspondent à ceux qui seraient estimés en temps continu avec un modèle de Cox (Courgeau et Lelièvre, 1989, annexe 1)**
- **Modèle adéquat si le processus sous-jacent est un processus continu**

De fait, dans la littérature, c'est plutôt le modèle logistique à temps discret qui est utilisé quel que soit le type de données discrète

Préparation d'une base de données (Allison, 1982)

- **Equation du log de vraisemblance pour les modèles en temps discret se simplifie en une équation du log de vraisemblance d'une variable dichotomique (cf. annexe**

- **Fichier personne-période (personne-année)**
  - **Un individu est représenté par un nombre de lignes égal au nombre de périodes (d'années) de présence avant de connaitre l'événement ou de sortir d'observation**
    - **La variable dichotomique est alors 0 pour toutes les lignes sauf la dernière**
    - **1 (si l'individu a connu l'événement) ou 0 (s'il ne l'a pas connu) pour la dernière ligne**
  - **Grande facilité à gérer les variables dépendantes du temps**

Exemple fictif

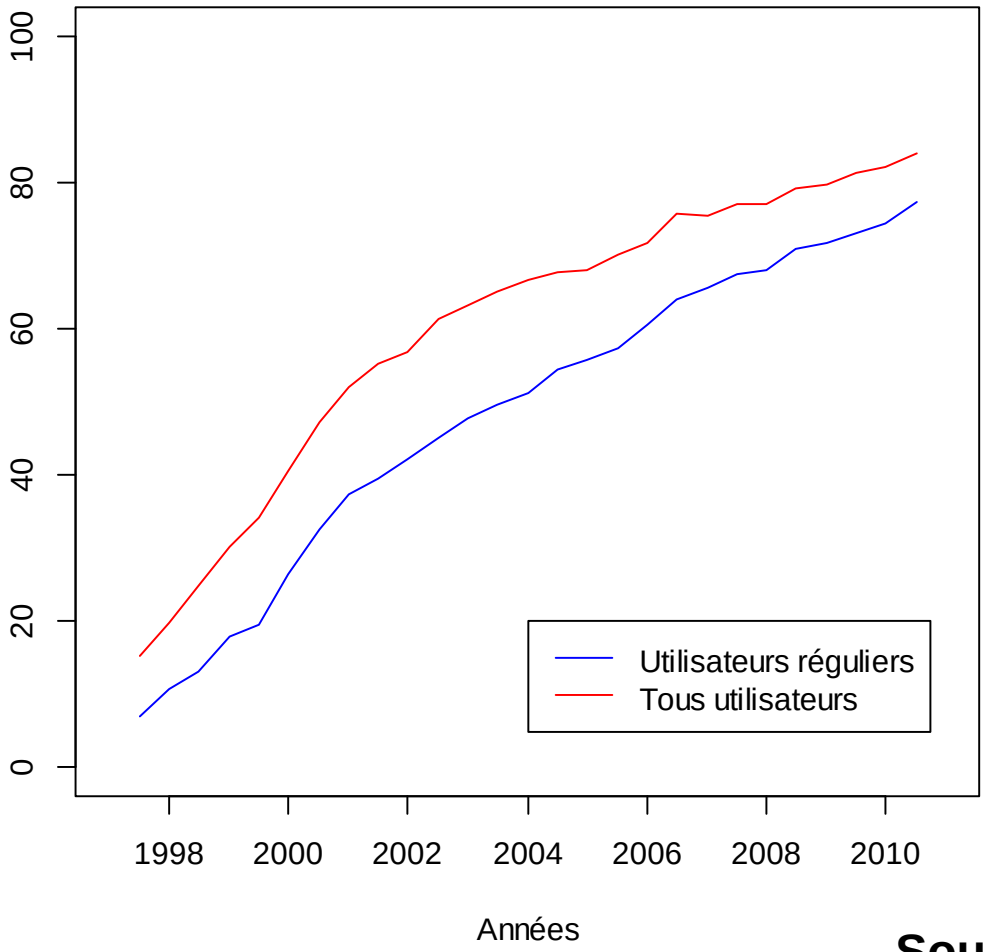
Individual	Time	Censor	x	y(t)
1	1	0	2	0
1	2	0	2	0
1	3	0	2	0
1	4	0	2	0
1	5	0	2	1
1	6	0	2	1
1	7	1	2	1
2	1	0	5	0
2	2	0	5	0
2	3	0	5	1

Un exemple: Diffusion de l'utilisation d'internet chez les séniors

- **Panel suisse des ménages (www.swisspanel.ch)**
  - **Echantillon 1999**
  - **Toutes les personnes âgées de plus de 14 ans dans un ménages sont interrogées**
  - **Question sur l'usage d'internet**
  - **Diffusion dans les ménages (entre partenaires)**

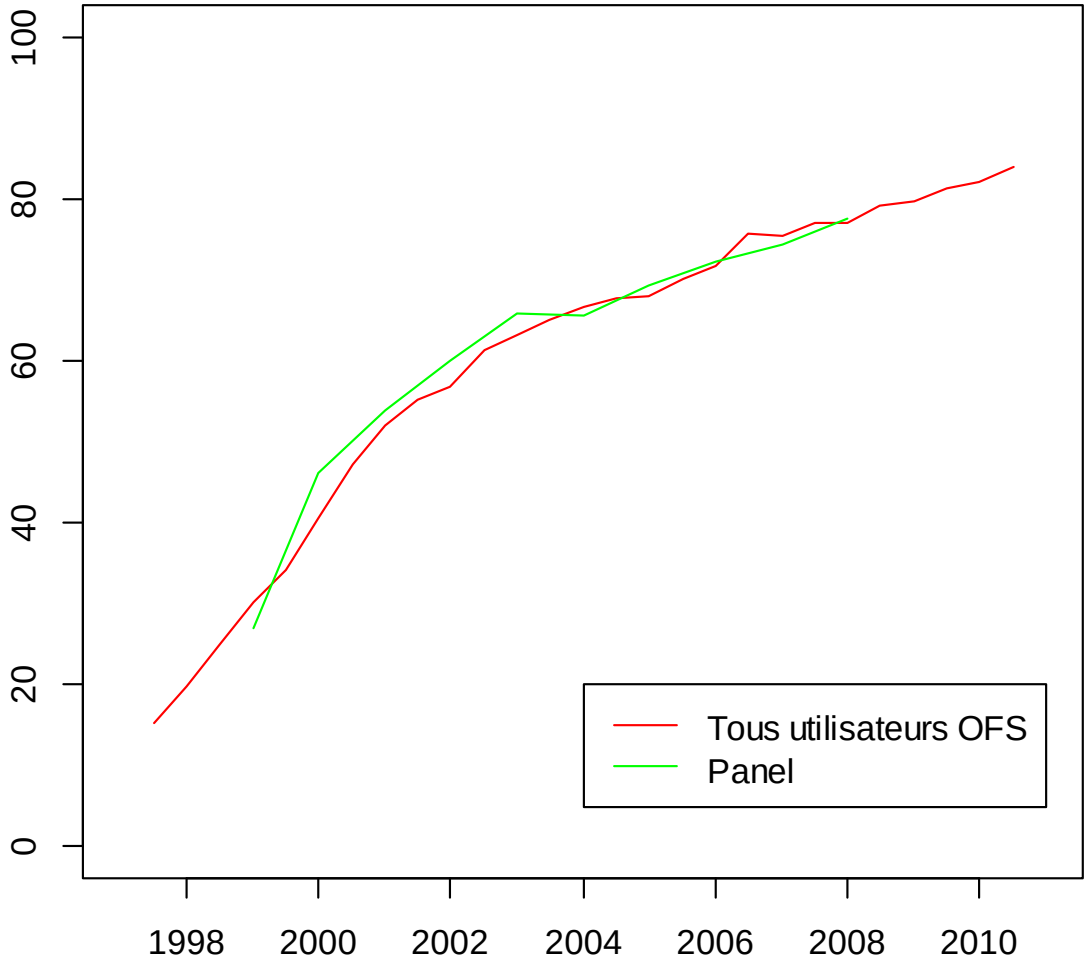
Diffusion d'Internet en Suisse

(en % de la population agée de plus de 14 ans)



Source: OFS/SUKO

Usage d'internet dans le SHP



Source: OFS/SUKO et SHP

Données (SHP)

- **Couples sans enfant vivant dans le ménage**
- **Partenaires âgés de 50 ans ou plus**
- **Interviewés depuis 1999**
- **Partenaire pouvant avoir déjà adopté pour certains d’entre eux**
- **284 hommes and 324 femmes**

Femmes

	Complémentaire log log				Logit			
	Modèle 1a		Modèle 2a		Modèle 1b		Modèle 2b	
Constante	-4.53 (-0.461)	***	-4.65 (0.469)	***	-4.53 (0.468)	***	-4.66 (0.478)	***
Années								
1999-2001	Ref		Ref		Ref		Ref	
2002-2004	0.49 (0.289)		0.38 (0.295)		0.51 (0.299)		0.39 (0.305)	
2005-2008	0.18 (0.348)		0.04 (0.354)		0.17 (0.359)		0.03 (0.366)	
Education								
Niveau bas	Ref		Ref		Ref		Ref	
Niveau moyen	1.24 (0.432)	**	1.22 (0.432)	**	1.26 (0.438)	**	1.25 (0.438)	**
Niveau élevé	1.48 (0.559)	**	1.53 (0.558)	**	1.53 (0.574)	**	1.57 (0.575)	**
Activité professionnelle								
Non	Ref		Ref		Ref		Ref	
Oui	0.7 (0.251)	**	0.73 (0.251)	**	0.72 (0.260)	**	0.76 (0.261)	**
Partenaire utilise Internet								
Non			Ref.				Ref	
Oui			0.62 (0.268)	*			0.64 (0.280)	*
-2LMV	482.22		477.21		482.17		477.19	
AIC	494.22		491.21		494.17		491.19	

*: Significatif au seuil de 5%; **: 1%; ***: 0,1%

Hommes

	Complémentaire log log				Logit			
	Modèle 1a		Modèle 2a		Modèle 1b		Modèle 2b	
Constante	-3.23 (0.511)	***	-3.21 (0.511)	***	-3.2 (0.521)	***	-3.19 (0.521)	***
Années								
1999-2001	Ref		Ref		Ref		Ref	
2002-2004	-0.55 (0.253)	*	-0.58 (0.256)	*	-0.59 (0.268)	*	-0.62 (0.271)	*
2005-2008	-1.37 (0.431)	**	-1.43 (0.441)	**	-1.43	**	-1.5 (0.454)	**
Education								
Niveau bas	Ref		Ref		Ref		Ref	
Niveau moyen	1.13 (0522)	*	1.11 (0.524)	*	1.18 (0.442)	*	1.15 (0.535)	*
Niveau élevé	1.2 (0.533)	*	1.19 (0.533)	*	1.24 (0.534)	*	1.22 (0.547)	*
Activité professionnelle								
Non	Ref		Ref		Ref		Ref	
Oui	0.77 (0.245)	**	0.74 (0.246)	**	0.84 (0.546)	**	0.82 (0.270)	**
Partenaire utilise Internet								
Non			Ref.				Ref	
Oui			0.32 (0.414)				0.36 (0.447)	
-2LMV	522.53		521.97		522.48		559.1	
AIC	534.53		535.97		534.48		521.87	

*: Significatif au seuil de 5%; **: 1%; ***: 0,1%

Conclusion

- **Modèles simples à mettre en œuvre**
- **Evénements concurrents: Régressions multinomiales**
- **Possibilité de prendre en compte des intervalles de temps variés (Allison, 2010)**
- **Multi-niveaux (Barber et al, 2000, Allison, 2010)**

Annexe (Cf Allison, 1982; Singer & Wilett, 1985; Le Goff & Forney, 2013)

- Dans le cadre d'un modèle de l'analyse des biographies, un individu *i* contribue à la vraisemblance par *f(ti)* s'il connaît l'événement et *S(ti)* 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* :

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

- Où δi 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) :

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

Annexe

Or comme:

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

Et:

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

En substituant et en passant par le logarithme de la vraisemblance:

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

Annexe

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,
le logarithme de la vraisemblance devient :

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

Cette dernière formulation indique que 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.

Références

- Allison P. (1982). Discrete time methods for the analysis of event histories. *Sociological Methodology*. 13: 61-98.
- Allison P. (2010). *Survival Analysis Using SAS. A practical Guide. Second Edition*. Cary: SAS Institute Inc.
- Barber J. S., Murphy S., Axinn W.G., & Maples J. (2000). DiscreteTime Multilevel Hazard Analysis. *Sociological Methodology.* 30:201-235.
- Box-Stephensmeier J. & Jones B.S. (2004). *Event history modelling. A Guide for Social Scientists*. Cambridge: Cambridge University Press.
- Courgeau & Lelièvre (1989). *Analyse démographique des biographies*. Paris: Ined.
- Le Goff J.-M. & Forney Y. (2013). Analyse des événements d'histoire de vie. Estimation de modèles logistiques à temps discret. Lausanne. Université de Lausanne. *Cahiers recherche et méthodes*. 3.
- Singer J. & Wilett J. B. (1985). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*. 18(2): 155-195.
- Yamaguchi K. (1991). *Event History Analysis*. Newbury-Park: Sage.