Modèles multi-états Markoviens et semi-Markoviens

Philippe Saint Pierre

IMT, Equipe de Statistique et Probabilités

Université Paul Sabatier - Toulouse III

Ecole CIMPA, Lomé, Septembre 2018

ANALYSE DE SURVIE MULTIVARIÉE

PLUSIEURS DURÉES DÉPENDANTES

- Plusieurs durées pour un individu
  - Durée sans trouble oculaire avec un traitement différent dans chaque oeil
  - Fiabilité de deux composants d'une machine
  - Événements récurrents : crise d'asthme
  - Événements concurrents : sortie d'une période sans emploi
- Une durée pour plusieurs individus d'un même groupe
  - Durées de vie jumeaux ou d'un couple

CONTEXTE

- Données censurées à droite
- Pas d'indépendance : $\mathbb{P}(T > t, U > u) \neq \mathbb{P}(T > t) \times \mathbb{P}(U > u)$
- On ne peut pas analyser ces données avec les modèles de survie classiques

ANALYSE DE SURVIE MULTIVARIÉE

- Hougaard (2000).

ESTIMATION NON-PARAMÉTRIQUE

- Estimation de la fonction de survie multivariée

$$S(t,u) = \mathbb{P}(T > t, U > u)$$

- Aucune hypothèse sur la forme de la fonction de survie
- Dabrowska (1988), Prentice, Moodie et Wu (2004), Lopez et Saint-Pierre (2012), ...
- Méthodologie, implémentation et interprétation compliquées

COPULE DE SURVIE

- Estimation de la fonction de survie multivariée

$$S(t,u) = \mathbb{P}(T > t, U > u)$$

- On montre que

$$\begin{aligned} S(t,u) &= S_1(t) + S_2(u) - 1 + F(t,u) \\ &= S_1(t) + S_2(u) - 1 + C(1 - S_1(t), 1 - S_2(u)) \\ &= \mathscr{C}(S_1(t), S_2(u)) \end{aligned}$$

où C est une copule et $\mathscr{C}$ est une copule de survie

- Estimation survie marginales avec l'estimateur de Kaplan-Meier
- Estimation du paramètres de copule par maximum de vraisemblance ou méthodes bayésiennes

MODÈLES DE FRAGILITÉ

- Introduction d'un terme de fragilité dans le modèle de Cox

$$\alpha(t|Z_{ij}, \omega_i) = \alpha_0(t)\omega_i exp(\beta Z_{ij})$$

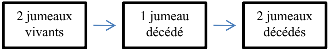
Z_{ij} covariable de l'individu j du groupe i
ω_i variable aléatoire *i.i.d.*

- L'effet aléatoire est commun aux individus d'un même groupe et différent d'un groupe à l'autre
- Permet de prendre en compte la dépendance entre les temps d'événement des individus d'un même groupe
- Ajustement sur les variables qui ne sont pas incluses dans le modèle (non mesurées ou non observées)
- Cas bivariée correspond à des groupes de deux individus

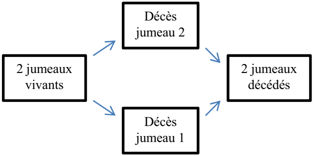
MODÈLES MULTI-ÉTATS

Survie multivariée : durées de vie de deux jumeaux

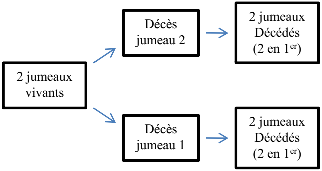
Modèle 1



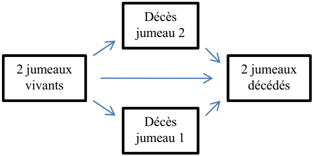
Modèle 2



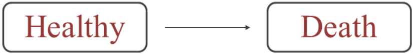
Modèle 3



Modèle 4

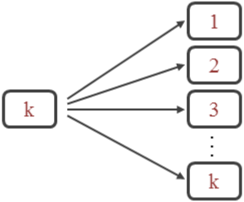


Analyse de survie

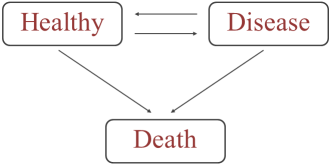


Événements répétés

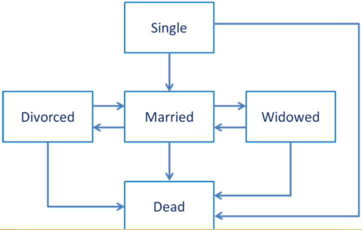
Événements concurrents



Épidémiologie : représenter l'évolution d'un patient à travers les différents stades d'une maladie

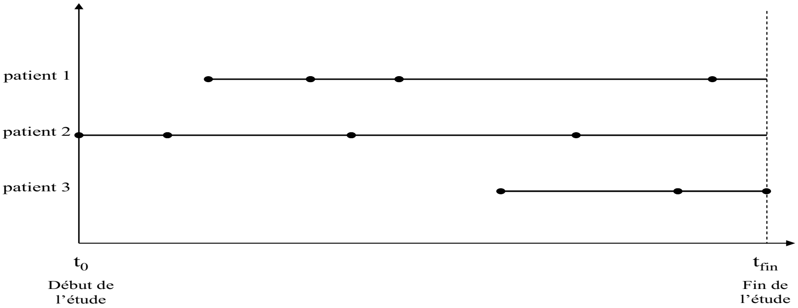


Science sociale : représenter l'évolution de la situation professionnelle ou familiale



Type de données

- Données de cohorte, données longitudinales
- Mesures répétées dans le temps : la même variable est mesurée plusieurs fois pour un même individu
- Mesures effectuées à des moments quelconques (spécifiques à chaque individu)



MODÈLES MULTI-ÉTATS

- Les modèles multi-états permettent de modéliser ce type de données
- Notions d'état et de processus pour représenter l'évolution d'un phénomène
- L'analyse consiste à estimer les intensités de transition entre les états

- Propriété de Markov : l'information sur les états précédents est résumée par l'état présent
- Les intensités de transition α peuvent dépendre de deux échelles de temps :
  - la durée du suivi t (ou le temps calendaire, l'âge),
  - le temps de séjour d

$\alpha(t,d) = \alpha$	$\rightarrow$	Modèle de Markov homogène
$\alpha(t,d) = \alpha(t)$	$\rightarrow$	Modèle de Markov non-homogène
$\alpha(t,d) = \alpha(d)$	$\rightarrow$	Modèle semi-Markovien homogène
$\alpha(t,d) = \alpha(t,d)$	$\rightarrow$	Modèle semi-Markovien non-homogène

Modèle de Markov

Processus de Markov

- $\{X(t), t \in [0, +\infty[\}$ un processus de Markov à temps continu et à espace d'états fini $S = \{1, \ldots, s\}$
- Probabilités de transition $\mathbf{P}(s,t) = \{p_{hj}(s,t)\}$

$$p_{hj}(s,t) = \mathbb{P}(X(t) = j \mid X(s) = h)$$

- Intensités de transition entre les états, $\mathbf{Q}(t) = \{\alpha_{hj}(t)\}$

$$\begin{aligned}
\alpha_{hj}(t) &= \lim_{\Delta t \to 0} \frac{p_{hj}(t, t+\Delta t)}{\Delta t}, \qquad h \neq j \\
\alpha_{hh}(t) &= -\sum_{j \neq h} \alpha_{hj}(t)
\end{aligned}$$

- Considérons $X_1(.), \ldots, X_n(.)$, n copies indépendantes de $X(.)$

MODÈLE DE MARKOV HOMOGÈNE

- La matrice des probabilités de transition

$$\mathbf{P}(s, s+\Delta t) = \mathbf{P}(0, \Delta t) = \exp(\mathbf{Q} \times \Delta t)$$

- Intensités de transition : $\alpha_{hj}(t) = \alpha_{hj}$
- Modèle à risques proportionnels

$$\alpha_{hj} = \alpha_{hj0} \exp(\beta_{hj}^T \mathbf{Z}), \qquad h \neq j$$

α_{hj0} intensité de transition de base

$\mathbf{Z} = (Z_1, ..., Z_k)^T$ vecteur de k covariables associé à l'individu

$\beta_{hj} = (\beta_{hj,1}, ..., \beta_{hj,k})^T$ vecteur de coefficients de régression associé à la transition de l'état h vers l'état j

Modèle de Markov homogène par période

- Intensités de transition définies sur $r+1$ intervalles $[t_{k-1}, t_k[$

$$\begin{array}{ll} \alpha_{hj}(t) = \alpha_{hj0} & \text{si } t_0 \leq t < t_1 \\ \alpha_{hj}(t) = \alpha_{hjk} = \alpha_{hj0} \exp(\sum_{v=1}^{k} \gamma_{hj,v}) & \text{si } t_k \leq t < t_{k+1} \end{array}$$

$\forall k = 1, \ldots, r$ et $t_{r+1} = +\infty$

α_{hjk} intensités de base dans l'intervalle $[t_k, t_{k+1}[$

$\gamma_{hj} = (\gamma_{hj,1}, \ldots, \gamma_{hj,k})$ vecteur de coefficients de régression

- Les intensités sont constantes sur chaque intervalle
- Estimation des paramètres par maximum de vraisemblance

Modèle de Markov non-homogène

- Intensités de transition $\alpha_{hj}(t)$ dépendent de la durée du suivi

- $\mathbf{N} = (N_{hj}; h \neq j, h,j = 1,\ldots,s)$ un processus de comptage multivarié

  $N_{hj}(t)$ compte le nombre de transitions (dans toute la population) observées de h vers j dans $[0,t]$

- $\lambda = (\lambda_{hj}; h \neq j)$, intensité du processus $\mathbf{N}$

$$\lambda_{hj}(t) = \alpha_{hj}(t) Y_h(t), \ \forall h \neq j$$

  $Y_h(t) = \sum_{i=1}^{n} \mathbb{1}_{\{X_i(t^-)=h\}}$ est le nombre total d'individus observés dans h juste avant t (individus à risque)

Modèle de Markov non-homogène

- $\mathbf{A(t)} = \{A_{hj}(t)\}$, la matrice des intensités cumulées

$$A_{hj}(t) = \int_0^t \alpha_{hj}(s)ds$$

- La matrice des probabilités de transition $\mathbf{P}(s,t) = \{p_{hj}(s,t)\}$ est définie par le produit intégral

$$\mathbf{P}(s,t) = \prod_{u \in]s,t]} (\mathbf{Id} + d\mathbf{A}(u))$$

$\mathbf{Id}$ matrice identité

Estimation non-paramétrique

- Estimateur de Nelson-Aalen des intensités cumulées

$$\widehat{A}_{hj}(t) = \int_0^t \frac{J_h(u)}{Y_h(u)} dN_{hj}(u), \qquad \forall h \neq j$$

$J_h(t) = \mathbb{1}_{\{Y_h(t)>0\}}$

- Estimateur de Aalen-Johansen de la matrice des probabilités de transition

$$\widehat{\mathbf{P}}(s,t) = \prod_{u \in]s,t]} \left(\mathbf{Id} + d\widehat{\mathbf{A}}(u)\right), \qquad 0 < s \leq t$$

$\widehat{\mathbf{A}}$ estimateur de Nelson-Aalen

- Comparaison des intensités obtenues dans différents groupes (test du Log-rank)

Estimation semi-paramétrique

- Les intensités de transition suivent un modèle à risques proportionnels

$$\alpha_{hji}(t \mid \mathbf{Z}_i) = \alpha_{hj0}(t)\exp(\beta_{hj}^T \mathbf{Z}_i) \qquad ,h \neq j$$

- Sachant β, l'estimateur de Breslow

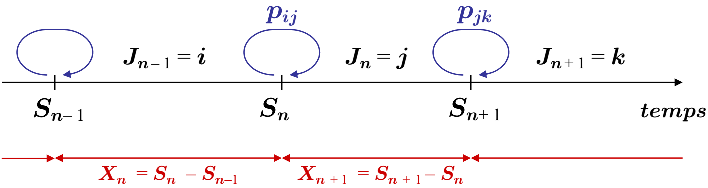
$$\hat{A}_{hj0}(t) = \int_0^t \frac{J_h(u)}{\sum_{i=1}^n \exp(\beta_{hj}^T \mathbf{Z}_i) Y_{hi}(t)} dN_{hj}(u),$$

- β estimé par maximisation de la vraisemblance partielle de Cox

$$\mathscr{L}_{Cox}(\beta) = \prod_t \prod_{i=1} \prod_{h \neq j} \left[\frac{Y_{hi}(t)\exp(\beta_{hj}^T \mathbf{Z}_i)}{\sum_{i=1}^n \exp(\beta_{hj}^T \mathbf{Z}_i) Y_{hi}(t)} \right]^{\Delta N_{hji}(t)}$$

MODÈLE SEMI-MARKOVIEN HOMOGÈNE

- Modèle Markovien inadapté $\Longrightarrow$ Approche semi-Markovienne
- Soit $(J_n, S_n)_{n\geq 0}$ un processus semi-Markovien



  - $(J_n)_{n\geq 0}$ est une chaîne de Markov homogène
  - $X_n = S_n - S_{n-1}$, le temps de séjour dans l'état J_{n-1}
  - $J_n \neq J_{n+1}$

- Probabilités de transition de la chaîne de Markov $(J_n)_{n\geq 0}$

$$p_{hj} = \mathbb{P}(J_{n+1} = j \mid J_n = h)$$

- Distribution du temps de séjour (dans l'état h avant d'aller dans l'état j)

$$F_{hj}(d) = \mathbb{P}(X_{n+1} \leq d \mid J_n = h, J_{n+1} = j)$$

$\Longrightarrow S_{hj}(x), f_{hj}(x)$, $\alpha_{hj}(x)$ (fonctions de survie, de densité et d'intensité correspondantes)

- Intensités de transition du processus semi-Markovien

$$\lambda_{hj}(d) = \lim_{\Delta d \to 0} \frac{1}{\Delta d} \mathbb{P}(J_{n+1} = j, d < X_{n+1} \leq d + \Delta d \mid J_n = h, X_{n+1} > d)$$

$$= \begin{cases} \frac{p_{hj} f_{hj}(d)}{S_h(d)} & \text{si } S_h(d) = \sum_{j=1}^{s} p_{hj} S_{hj}(d) > 0 \\ 0 & \text{sinon} \end{cases}$$

VRAISEMBLANCE

- Contributions à la vraisemblance
  - Transition de $h \to j$ observée : $S_h(d)\lambda_{hj}(d) = p_{hj}f_{hj}(d)$
  - Censure à droite dans l'état h : $S_h(d)$
- Vraisemblance

$$L = \prod_{i=1}^{n} \left\{ \prod_{k=1}^{n_i} p_{J_{i,k-1}J_{i,k}} f_{J_{i,k-1}J_{i,k}}(X_{i,k}) \right\} \times \left[S_{J_{i,n_i}}(U_i) \right]^{\delta_i}$$

$\delta_i = 1$ si l'individu i est censuré et 0 sinon

U_i le temps de séjour censuré pour l'individu i si $\delta_i = 1$

n_i le nombre de transitions pour l'individu i

ESTIMATION PARAMÉTRIQUE

- Modélisation des distributions des temps de séjour par des lois paramétriques
- Modèle à risques proportionnels pour les intensités des temps de séjour

$$\alpha_{hj}(d \mid \mathbf{Z}) = \alpha_{hj0}(d) \exp(\beta_{hj}^T \mathbf{Z})$$

α_{hj0} risque d'une loi de Weibull, Weibull généralisée, ...

$\mathbf{Z}$ vecteur des covariables

β_{hj} vecteur des coefficients de régression

- Estimation des paramètres par maximisation de la vraisemblance

APPLICATION À L'ASTHME

PACKAGES DU LOGICIEL R

- CRAN Task View : *Survival Analysis*
- Modèle Markovien homogène : *msm*
- Modèle Markovien non homogène : *etm, mstate, msSurv*
- Modèle semi-Markovien homogène : *SemiMarkov*

BASE DE DONNÉES

- 406 patients asthmatiques

  4.15 consultations par patient en moyenne

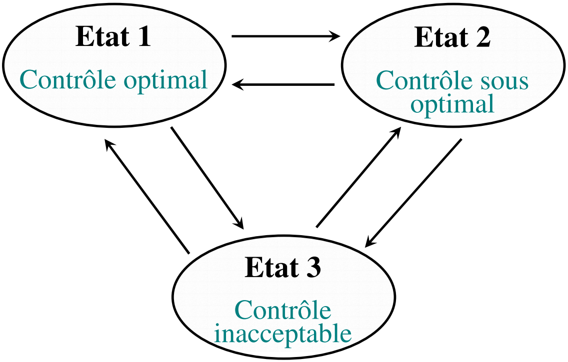
  Reculs de 3 mois à quatre ans

- Notion de contrôle : jugement global du médecin sur l'activité de la maladie

- Etude de l'Indice de Masse Corporelle (IMC)
  - $IMC = poids/(taille)^2$
  - 2 modalités : $IMC < 25$ et $IMC \geq 25$ (surpoids)

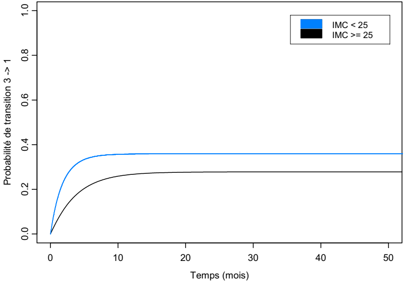
Modèle

- Modèle à 3 états de contrôle
- Toutes les transitions entre les états sont possibles



Modèle homogène

- Probabilités de transition $3 \rightarrow 1$ pour la covariable IMC



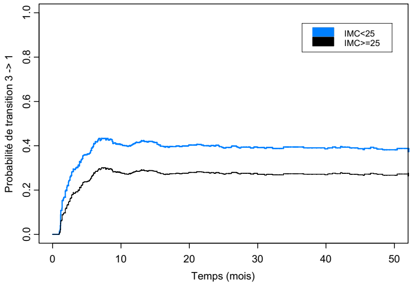
- Modèle homogène par période (2 périodes)

$\Rightarrow$ Meilleur ajustement aux données

$\Rightarrow$ L'hypothèse d'homogénéité semble trop restrictive

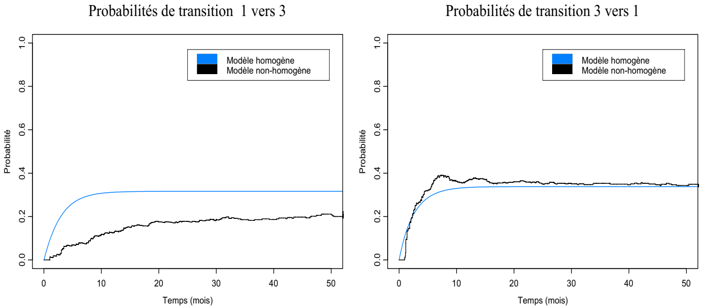
Modèle non-homogène

- Probabilités de transition $3 \rightarrow 1$ pour la covariable IMC



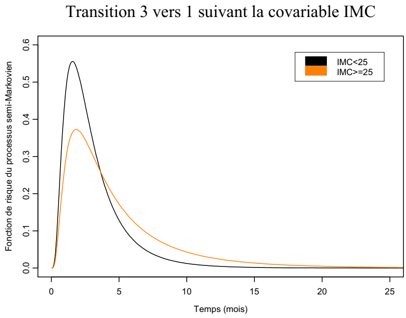
Modèle non-homogène

- Estimations dans un modèle homogène et non-homogène



Modèle semi-Markovien

- Estimations paramétriques des intensités $3 \rightarrow 1$ du processus semi-Markovien pour la covariable IMC



Coefficients de régression de l'IMC

- Coefficients de régression pour la covariable IMC

Transition	Modèle MH			Modèle MNH		
	$\hat{\beta}$	$E.T.$	p-value	$\hat{\beta}$	$E.T.$	p-value
$1 \rightarrow 2$	-0.409	0.383	0.02	-0.248	0.236	0.28
$1 \rightarrow 3$	-0.122	0.269	0.71	0.655	0.303	0.03
$2 \rightarrow 1$	0.542	0.364	$<$0.01	-0.030	0.203	0.88
$2 \rightarrow 3$	0.041	0.378	0.87	0.204	0.249	0.42
$3 \rightarrow 1$	**-1.170**	**0.370**	$<$**0.01**	**-0.595**	**0.220**	$<$**0.01**
$3 \rightarrow 2$	-0.561	0.293	$<$0.01	-0.234	0.191	0.21

- Effet de l'IMC également significatif pour la transition $3 \rightarrow 1$
  - par stratification
  - avec l'IMC dépendant du temps
  - avec le modèle semi-Markovien

Coefficients de régression de l'IMC

- Modèle Markov homogène avec 2 états de contrôle

Transition	Covariables	Modèle univarié			Modèle multivarié		
		$\hat{\beta}$	$E.T.$	p-value	$\hat{\beta}$	$E.T.$	p-value
	IMC	-0.129	0.247	0.60	-0.174	0.278	0.53
$1 \rightarrow 2$	Sévérité	0.665	0.272	0.04	0.820	0.305	$<$0.01
	Corticoïdes Oraux	0.110	0.269	0.68	-0.422	0.305	0.17
	Antécédents Corticoïdes	0.651	0.262	0.01	0.498	0.299	0.10
	IMC	**-0.801**	**0.184**	$<$**0.01**	**-0.637**	**0.219**	$<$**0.01**
$2 \rightarrow 1$	Sévérité	-0.726	0.203	$<$0.01	-0.062	0.255	0.81
	Corticoïdes Oraux	-1.002	0.209	$<$0.01	-0.693	0.248	$<$0.01
	Antécédents Corticoïdes	-0.852	0.212	$<$0.01	-0.312	0.266	0.24

$\Rightarrow$ Résultats ajustés

DISCUSSION

- Méthodes adaptées pour modéliser des données de survie multivariée
- Résultats facilement interprétables en pratique
- Packages R
- Hypothèses sous-jacentes (Markov, homogénéité, risques proportionnels, observation continue)

- Méthodes prenant en compte une censure par intervalles
- Méthodes prenant en compte des effets aléatoires
- Développer des méthodes adaptées à une censure informative