Estimation de la fonction de risque en présence de données incomplètes

B. LIQUET[1] P. JOLY[2] D. COMMENGES[2]

[1]Laboratoire de Statistique et Analyse des Données (Grenoble)

[2]Equipe de Biostatistique EMI 03-38 (Bordeaux)

FIMA 20 octobre 2005

Plan de la présentation

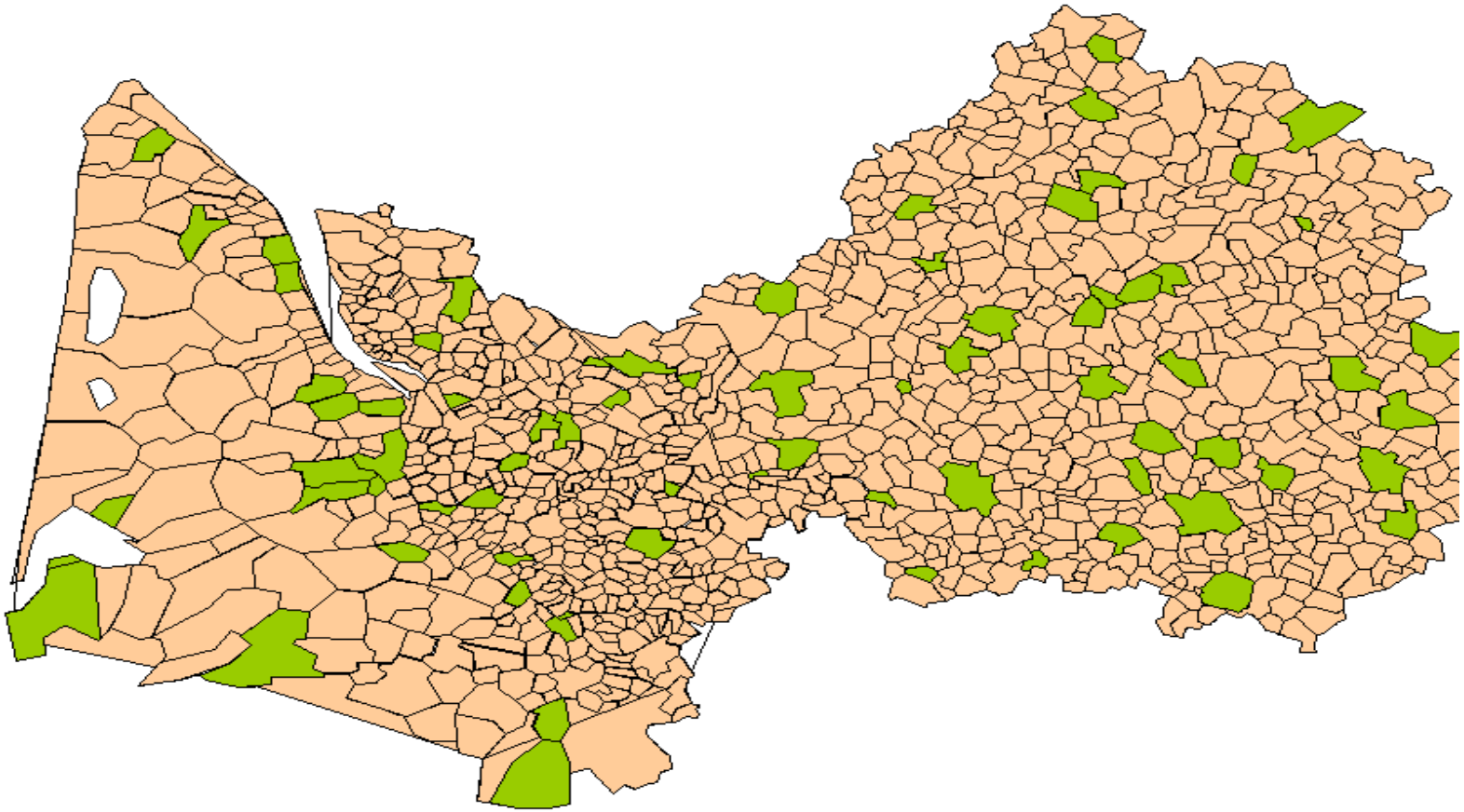
Plan

les données PAQUID : Personnes Agées Quid

Cohorte de personnes agées étudiant le vieillissement cérébral normal et pathologique

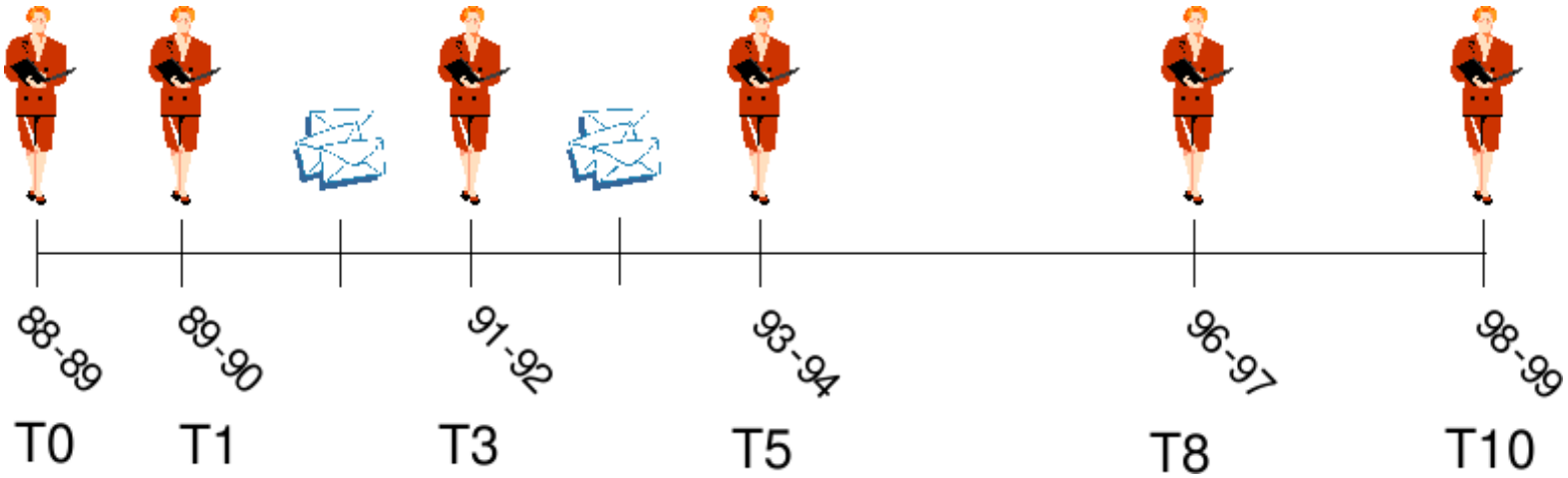
- 3675 sujets de 65 ans et plus vivant en Gironde et en Dordogne, recrutés en 1988
- 2133 femmes et 1542 hommes
- variables explicatives : sexe (S) et niveau d'étude (E) etc ...

Recrutement des sujets : 75 communes



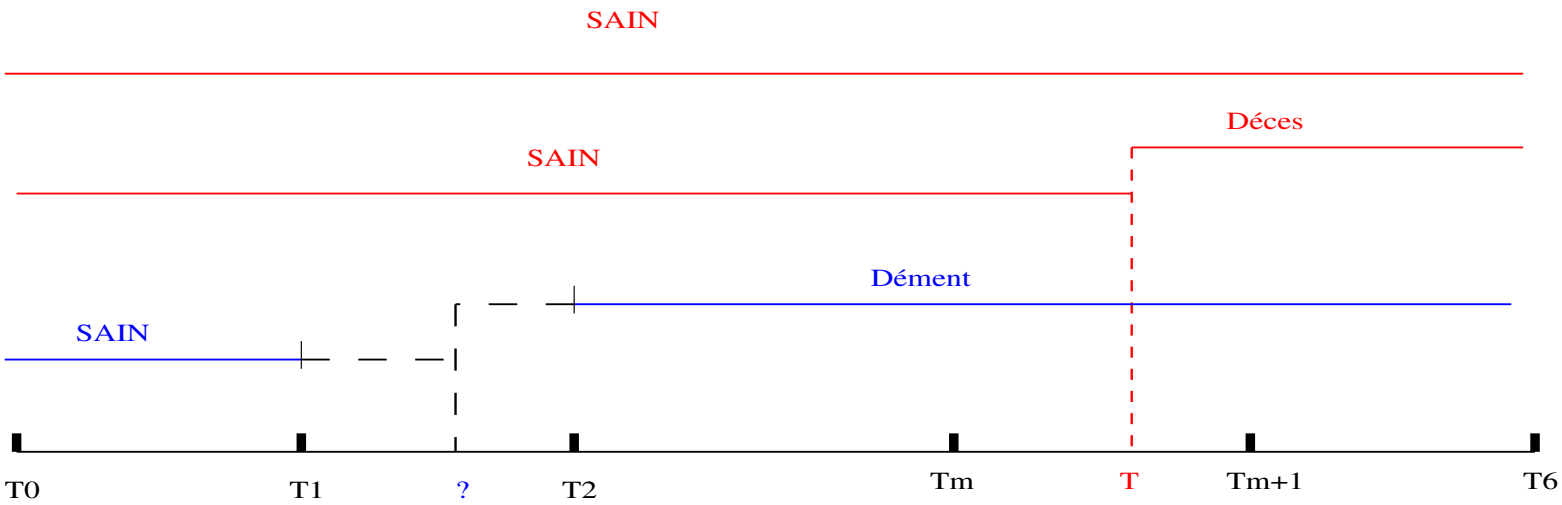
- tirés au sort sur les listes électorales
- vivant à Domicile en Gironde et en Dordogne (Sud Ouest de la France)

Suivi des sujets et Receuil des données



- Entretien au domicile par des psychologues
  - caractéristiques socio-démographiques
  - conditions de vie et habitudes
  - état de santé/autonomie
  - batterie de tests psychométriques
- Identification des sujets déments en 2 étapes
  - critères DSM III de démence complétés par la psychologue en fin d'entretien
  - visite à domicile par un neurologue

Données censurées



- censure à droite :
  - sujet non dément à sa dernière visite
  - sujet décédé à T et non dément à Tm
- censure par intervalle : dément entre T1 et T2
- troncature à gauche : sujets sont non-déments à l'entrée dans l'étude

objectifs

- Modéliser le risque de démence en fonction de l'âge
  - le temps de base sera l'âge des individus
- obtenir une estimation lisse
- prise en compte du phénomène de censure
- prise en compte des variables explicatives (par ex : difference H/F)

Plan

Modélisation du risque de démence dans la cohorte PAQUID

- fonction de risque (risque instantané de démence)

$$\lambda(t) = \frac{f(t)}{S(t)} = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T > t)}{dt}$$

  probabilité d'etre atteint de démence à l'âge t conditionnellement au fait que le sujet n'est pas dément avant t.

- estimation non-paramétrique :
  - Ramlau-Hansen : $\widehat{\lambda}_h(t) = \frac{1}{h} \int K\left(\frac{t-u}{h}\right) d\widehat{A}(u)$ où $\widehat{A}$ est l'estimateur de Nelson-Aalen
  - vraisemblance pénalisé : $p\mathcal{L}_h = \log \mathcal{L} - h \int \lambda''^2(u) du$ où $\mathcal{L}$ est la vraisemblance
- choix du paramètre de lissage h

Estmateur par la vraisemblance pénalisée

$$p\mathcal{L}_h(\lambda) = \log \mathcal{L}(\lambda) - h \int \lambda''^2(u)du \quad (1)$$

- la vraisemblance $\mathcal{L}(\lambda) = \prod L_i$
  - individu censuré à droite à T_m et troncature à gauche : $L_i = S(T_m)/S(A_i)$ où A_i est l'age du patient au début de l'étude
  - individu censuré par intervalle $[T_m; T_{m+1}]$ et troncature à gauche : $L_i = (S(T_m) - S(T_{m+1}))/S(A_i)$
  - $\log \mathcal{L}(\lambda) = \sum_I \log\left(\frac{S(T_m) - S(T_{m+1})}{S(A_i)}\right) + \sum_C \log\left(\frac{S(T_m)}{S(A_i)}\right)$
- Maximisation de (1) $\Longrightarrow \widehat{\lambda}_h(t)$
  - Approximation par une base de splines :
    - $\lambda(t) = \sum \theta_i M_i(t)$
    - M_i spline d'ordre 3(polynome a support compact degré 2)
  - Algorithme de type Newton-Raphson (Marquard)

Estmateur par la vraisemblance pénalisée

$$p\mathcal{L}_h(\lambda) = \log \mathcal{L}(\lambda) - h \int \lambda''^2(u)du \quad (1)$$

- la vraisemblance $\mathcal{L}(\lambda) = \prod L_i$
  - individu censuré à droite à T_m et troncature à gauche : $L_i = S(T_m)/S(A_i)$ où A_i est l'age du patient au début de l'étude
  - individu censuré par intervalle $[T_m; T_{m+1}]$ et troncature à gauche : $L_i = (S(T_m) - S(T_{m+1}))/S(A_i)$
  - $\log \mathcal{L}(\lambda) = \sum_I \log\left(\frac{S(T_m) - S(T_{m+1})}{S(A_i)}\right) + \sum_C \log\left(\frac{S(T_m)}{S(A_i)}\right)$
- Maximisation de (1) $\Longrightarrow \widehat{\lambda}_h(t)$
  - Approximation par une base de splines :
    - $\lambda(t) = \sum \theta_i M_i(t)$
    - M_i spline d'ordre 3(polynome a support compact degré 2)
  - Algorithme de type Newton-Raphson (Marquard)

Estmateur par la vraisemblance pénalisée

$$p\mathcal{L}_h(\lambda) = \log \mathcal{L}(\lambda) - h \int \lambda''^2(u)du \quad (1)$$

- la vraisemblance $\mathcal{L}(\lambda) = \prod L_i$
  - individu censuré à droite à T_m et troncature à gauche : $L_i = S(T_m)/S(A_i)$ où A_i est l'age du patient au début de l'étude
  - individu censuré par intervalle $[T_m; T_{m+1}]$ et troncature à gauche : $L_i = (S(T_m) - S(T_{m+1}))/S(A_i)$
  - $\log \mathcal{L}(\lambda) = \sum_I \log\left(\frac{S(T_m) - S(T_{m+1})}{S(A_i)}\right) + \sum_C \log\left(\frac{S(T_m)}{S(A_i)}\right)$
- Maximisation de (1) $\Longrightarrow \widehat{\lambda}_h(t)$
  - Approximation par une base de splines :
    - $\lambda(t) = \sum \theta_i M_i(t)$
    - M_i spline d'ordre 3(polynome a support compact degré 2)
  - Algorithme de type Newton-Raphson (Marquard)

Estmateur par la vraisemblance pénalisée

$$p\mathcal{L}_h(\lambda) = \log \mathcal{L}(\lambda) - h \int \lambda''^2(u)du \quad (1)$$

- la vraisemblance $\mathcal{L}(\lambda) = \prod L_i$
  - individu censuré à droite à T_m et troncature à gauche : $L_i = S(T_m)/S(A_i)$ où A_i est l'age du patient au début de l'étude
  - individu censuré par intervalle $[T_m; T_{m+1}]$ et troncature à gauche : $L_i = (S(T_m) - S(T_{m+1}))/S(A_i)$
  - $\log \mathcal{L}(\lambda) = \sum_I \log\left(\frac{S(T_m) - S(T_{m+1})}{S(A_i)}\right) + \sum_C \log\left(\frac{S(T_m)}{S(A_i)}\right)$
- Maximisation de (1) $\Longrightarrow \widehat{\lambda}_h(t)$
  - Approximation par une base de splines :
    - $\lambda(t) = \sum \theta_i M_i(t)$
    - M_i spline d'ordre 3(polynome a support compact degré 2)
  - Algorithme de type Newton-Raphson (Marquard)

Estmateur par la vraisemblance pénalisée

$$p\mathcal{L}_h(\lambda) = \log \mathcal{L}(\lambda) - h \int \lambda''^2(u)du \quad (1)$$

- la vraisemblance $\mathcal{L}(\lambda) = \prod L_i$
  - individu censuré à droite à T_m et troncature à gauche : $L_i = S(T_m)/S(A_i)$ où A_i est l'age du patient au début de l'étude
  - individu censuré par intervalle $[T_m; T_{m+1}]$ et troncature à gauche : $L_i = (S(T_m) - S(T_{m+1}))/S(A_i)$
  - $\log \mathcal{L}(\lambda) = \sum_I \log\left(\frac{S(T_m) - S(T_{m+1})}{S(A_i)}\right) + \sum_C \log\left(\frac{S(T_m)}{S(A_i)}\right)$
- Maximisation de (1) $\Longrightarrow \widehat{\lambda}_h(t)$
  - Approximation par une base de splines :
    - $\lambda(t) = \sum \theta_i M_i(t)$
    - M_i spline d'ordre 3(polynome a support compact degré 2)
  - Algorithme de type Newton-Raphson (Marquard)

Objectifs : choix d'un estimateur pour la fonction de risque

- Estimation non-paramétrique :
  - Ramlau-Hansen : $\widehat{\lambda}_h(t) = \frac{1}{h} \int K\left(\frac{t-u}{h}\right) d\widehat{A}(u)$
  - vraisemblance pénalisé :$p\mathcal{L}_h = \log \mathcal{L} - h \int \lambda''^2(u)du$
- Choix du paramètre de lissage h
- Critère d'information
- Notation : $\mathcal{W} = (W_1, \ldots, W_n)$
  - censure à droite : $W_i = (\tilde{T}_i, \delta_i)$; $\tilde{T}_i = \min(T_i, C_i)$, $\delta_i = \mathbb{I}_{[T_i \leq C_i]}$
    $T_i \sim F, f, S$ et $\lambda(t) = \frac{f(t)}{S(t)}$

Objectifs : choix d'un estimateur pour la fonction de risque

- Estimation non-paramétrique :
  - Ramlau-Hansen : $\widehat{\lambda}_h(t) = \frac{1}{h} \int K\left(\frac{t-u}{h}\right) d\widehat{A}(u)$
  - vraisemblance pénalisé :$p\mathcal{L}_h = \log \mathcal{L} - h \int \lambda''^2(u)du$
- Choix du paramètre de lissage h
- Critère d'information
- Notation : $\mathcal{W} = (W_1, \ldots, W_n)$
  - censure à droite : $W_i = (\tilde{T}_i, \delta_i)$; $\tilde{T}_i = \min(T_i, C_i)$, $\delta_i = \mathbf{I}_{[T_i \leq C_i]}$
    $T_i \sim F, f, S$ et $\lambda(t) = \frac{f(t)}{S(t)}$

Critère de sélection

Information de Kullback-Leibler :

$$\begin{aligned} I(f,\widehat{f}^{\mathcal{W}}) &= \int f(x)\log\frac{f(x)}{\widehat{f}^{\mathcal{W}}(x)}dx \\ &= \int f(x)\log f(x)dx - \int f(x)\log\widehat{f}^{\mathcal{W}}(x)dx \end{aligned}$$

f densité du vrai modèle (inconnu), $\widehat{f}^{\mathcal{W}}$ densité du modèle essayé.

- $I(f,\widehat{f}^{\mathcal{W}}) \geq 0$
- $I(f,\widehat{f}^{\mathcal{W}}) = 0 \Longleftrightarrow \widehat{f}^{\mathcal{W}} \equiv f$

Mesure la perte d'information lorsque $\widehat{f}^{\mathcal{W}}$ est utilisée pour approcher f.

$$\text{minimiser } I(f,\widehat{f}^{\mathcal{W}}) \Longrightarrow \text{maximiser } KL(\mathcal{W}) = \int f(x)\log\widehat{f}^{\mathcal{W}}(x)dx$$

Critères proposés

- Information de Kullback-Leibler :

$$\mathrm{KL}(\mathcal{W}) = \mathrm{E}\left\{\log \widehat{f}_h^{\mathcal{W}}(T')|\mathcal{W}\right\}$$

où $\mathcal{W} = (W_1, \ldots, W_n)$ et $T' \sim F$

- L'espérance de la log-vraisemblance (ELL) :

$$\mathrm{ELL} = \mathrm{E}\left\{\log \mathcal{L}^{\widehat{\lambda}_h(\mathcal{W})}(\mathcal{W}')\right\}$$

où $\mathcal{W}' \overset{d}{=} \mathcal{W}$ et $\mathcal{L}^{\widehat{\lambda}_h(\mathcal{W})}(\mathcal{W}')$ est la fonction de vraisemblance de l'estimateur $\widehat{\lambda}_h(\cdot; \mathcal{W})$ pour les observations $\mathcal{W}'$

Estimateurs du $\mathrm{ELL} = \mathrm{E}\left\{\log \mathcal{L}^{\widehat{\lambda}_h(\mathcal{W})}(\mathcal{W}')\right\}$

- Estimation par bootstrap
  - biais corrigé par bootstrap ELL_{bboot}

$$\mathrm{ELL}_{bboot} = \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}) - \widehat{b}(\mathcal{W})$$

  - bootstrap direct

$$\mathrm{ELL}_{boot} = \mathrm{E}_* \left\{\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(\mathcal{W}'^*)\right\} \quad \mathcal{W}^* \stackrel{d}{=} \mathcal{W}'^*,\ W_j^* \sim \widehat{F}_W$$

- Likelihood cross-validation
  - $\mathrm{LCV} = \sum_{i=1}^{n} \log \mathcal{L}^{\widehat{\lambda}_h(\mathcal{W}^{-i})}(W_i)$, $\mathrm{E}\left[\mathrm{LCV}\{\hat{\lambda}_h(\mathcal{W})\}\right] \simeq \mathrm{ELL}\{\widehat{\lambda}_h(\mathcal{W})\}$
  - $\mathrm{LCVa}(\mathcal{W}_n) = \sum_{i=1}^{n} \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(W_i) - mdf$,

    où $mdf = \mathrm{trace}([\widehat{H} - 2h\Omega]^{-1}\widehat{H})$ et $(\Omega_{ij}) = \int M_i''(u) M_j''(u) du$

Estimateurs du $\mathrm{ELL} = \mathrm{E}\left\{\log \mathcal{L}^{\widehat{\lambda}_h(\mathcal{W})}(\mathcal{W}')\right\}$

- Estimation par bootstrap
  - biais corrigé par bootstrap ELL_{bboot}

$$\mathrm{ELL}_{bboot} = \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}) - \widehat{b}(\mathcal{W})$$

  - bootstrap direct

$$\mathrm{ELL}_{boot} = \mathrm{E}_* \left\{\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(\mathcal{W}'^*)\right\} \quad \mathcal{W}^* \overset{d}{=} \mathcal{W}'^*,\ W_j^* \sim \widehat{F}_W$$

- Likelihood cross-validation
  - $\mathrm{LCV} = \sum_{i=1}^{n} \log \mathcal{L}^{\widehat{\lambda}_h(\mathcal{W}^{-i})}(W_i)$, $\mathrm{E}\left[\mathrm{LCV}\{\widehat{\lambda}_h(\mathcal{W})\}\right] \simeq \mathrm{ELL}\{\widehat{\lambda}_h(\mathcal{W})\}$
  - $\mathrm{LCVa}(\mathcal{W}_n) = \sum_{i=1}^{n} \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(W_i) - mdf$,

    où $mdf = \mathrm{trace}([\widehat{H} - 2h\Omega]^{-1}\widehat{H})$ et $(\Omega_{ij}) = \int M_i''(u) M_j''(u) du$

Estimateurs du $\mathrm{ELL} = \mathrm{E}\left\{\log \mathcal{L}^{\widehat{\lambda}_h(\mathcal{W})}(\mathcal{W}')\right\}$

- Estimation par bootstrap
  - biais corrigé par bootstrap ELL_{bboot}

$$\mathrm{ELL}_{bboot} = \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}) - \widehat{b}(\mathcal{W})$$

  - bootstrap direct

$$\mathrm{ELL}_{boot} = \mathrm{E}_* \left\{\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(\mathcal{W}'^*)\right\} \quad \mathcal{W}^* \stackrel{d}{=} \mathcal{W}'^*,\ W_j^* \sim \widehat{F}_W$$

- Likelihood cross-validation
  - $\mathrm{LCV} = \sum_{i=1}^{n} \log \mathcal{L}^{\widehat{\lambda}_h(\mathcal{W}^{-i})}(W_i)$, $\mathrm{E}\left[\mathrm{LCV}\{\widehat{\lambda}_h(\mathcal{W})\}\right] \simeq \mathrm{ELL}\{\widehat{\lambda}_h(\mathcal{W})\}$
  - $\mathrm{LCVa}(\mathcal{W}_n) = \sum_{i=1}^{n} \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(W_i) - mdf$,

    où $mdf = \mathrm{trace}([\widehat{H} - 2h\Omega]^{-1}\widehat{H})$ et $(\Omega_{ij}) = \int M_i''(u) M_j''(u) du$

Estimateurs du $\mathrm{ELL} = \mathrm{E}\left\{\log \mathcal{L}^{\widehat{\lambda}_h(\mathcal{W})}(\mathcal{W}')\right\}$

- Estimation par bootstrap
  - biais corrigé par bootstrap ELL_{bboot}

$$\mathrm{ELL}_{bboot} = \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}) - \widehat{b}(\mathcal{W})$$

  - bootstrap direct

$$\mathrm{ELL}_{boot} = \mathrm{E}_* \left\{\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(\mathcal{W}'^*)\right\} \quad \mathcal{W}^* \stackrel{d}{=} \mathcal{W}'^*,\ W_j^* \sim \widehat{F}_W$$

- Likelihood cross-validation
  - $\mathrm{LCV} = \sum_{i=1}^{n} \log \mathcal{L}^{\widehat{\lambda}_h(\mathcal{W}^{-i})}(W_i)$, $\mathrm{E}\left[\mathrm{LCV}\{\hat{\lambda}_h(\mathcal{W})\}\right] \simeq \mathrm{ELL}\{\widehat{\lambda}_h(\mathcal{W})\}$
  - $\mathrm{LCVa}(\mathcal{W}_n) = \sum_{i=1}^{n} \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(W_i) - mdf$,

    où $mdf = \mathrm{trace}([\widehat{H} - 2h\Omega]^{-1}\widehat{H})$ et $(\Omega_{ij}) = \int M_i''(u) M_j''(u) du$

Estimateurs du $\mathrm{ELL} = \mathrm{E}\left\{\log \mathcal{L}^{\widehat{\lambda}_h(\mathcal{W})}(\mathcal{W}')\right\}$

- Estimation par bootstrap
  - biais corrigé par bootstrap ELL_{bboot}

$$\mathrm{ELL}_{bboot} = \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}) - \widehat{b}(\mathcal{W})$$

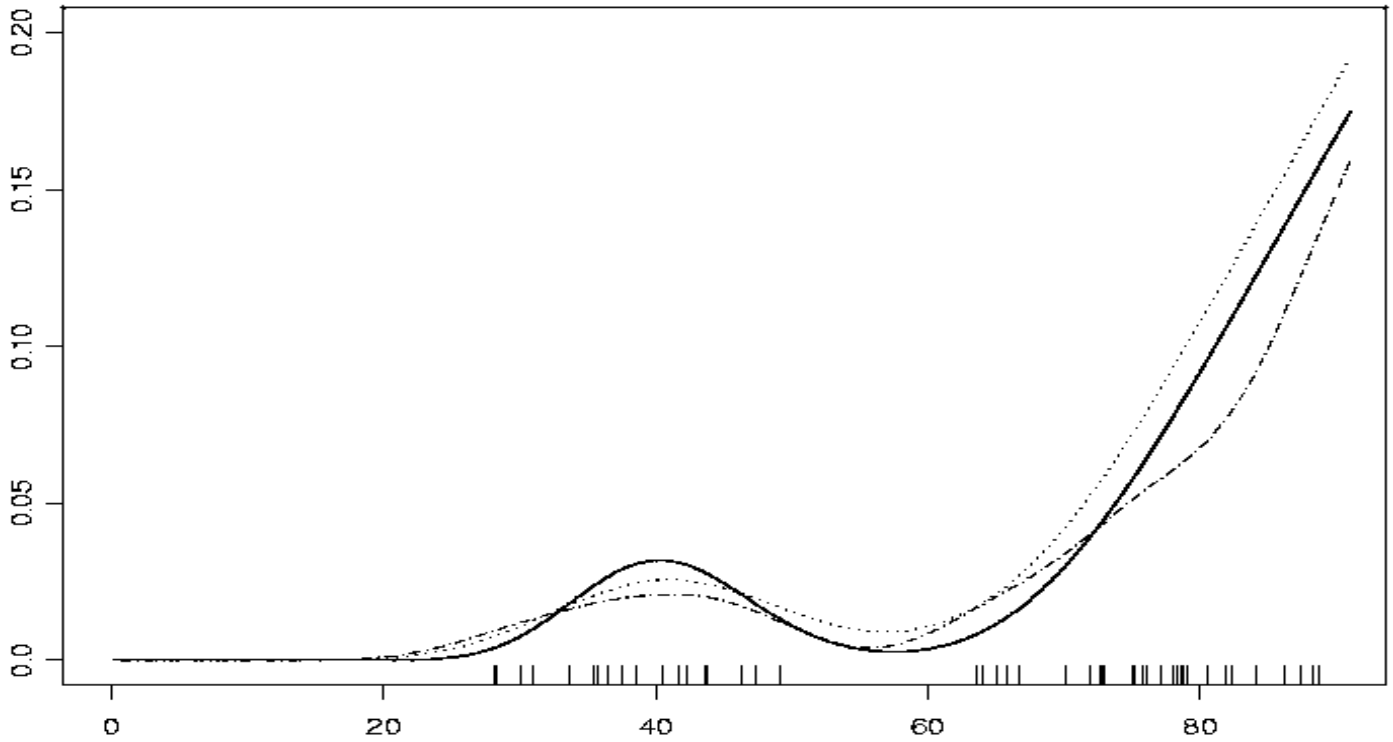
  - bootstrap direct

$$\mathrm{ELL}_{boot} = \mathrm{E}_* \left\{\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(\mathcal{W}'^*)\right\} \quad \mathcal{W}^* \stackrel{d}{=} \mathcal{W}'^*,\ W_j^* \sim \widehat{F}_W$$

- Likelihood cross-validation
  - $\mathrm{LCV} = \sum_{i=1}^{n} \log \mathcal{L}^{\widehat{\lambda}_h(\mathcal{W}^{-i})}(W_i)$, $\mathrm{E}\left[\mathrm{LCV}\{\hat{\lambda}_h(\mathcal{W})\}\right] \simeq \mathrm{ELL}\{\widehat{\lambda}_h(\mathcal{W})\}$
  - $\mathrm{LCVa}(\mathcal{W}_n) = \sum_{i=1}^{n} \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(W_i) - mdf,$

    où $mdf = \mathrm{trace}([\widehat{H} - 2h\Omega]^{-1}\widehat{H})$ et $(\Omega_{ij}) = \int M_i''(u)M_j''(u)du$

Simulation



Simulation : estimateur à noyau (1)

	$-\mathrm{KL}(\widehat{\lambda}_h^{\mathcal{W}})$ for kernel estimators	
n	KL	ELL
15% censoring		
30	3.96(0.005)	3.96(0.005)
50	3.98(0.003)	3.99(0.003)
100	4.01(0.002)	4.01(0.002)
25% censoring		
30	3.89(0.004)	3.91(0.005)
50	3.93(0.004)	3.93(0.004)
100	3.95(0.002)	3.95(0.002)
50% censoring		
30	3.81(0.02)	3.92(0.04)
50	3.80(0.009)	3.84(0.02)
100	3.80(0.005)	3.80(0.005)

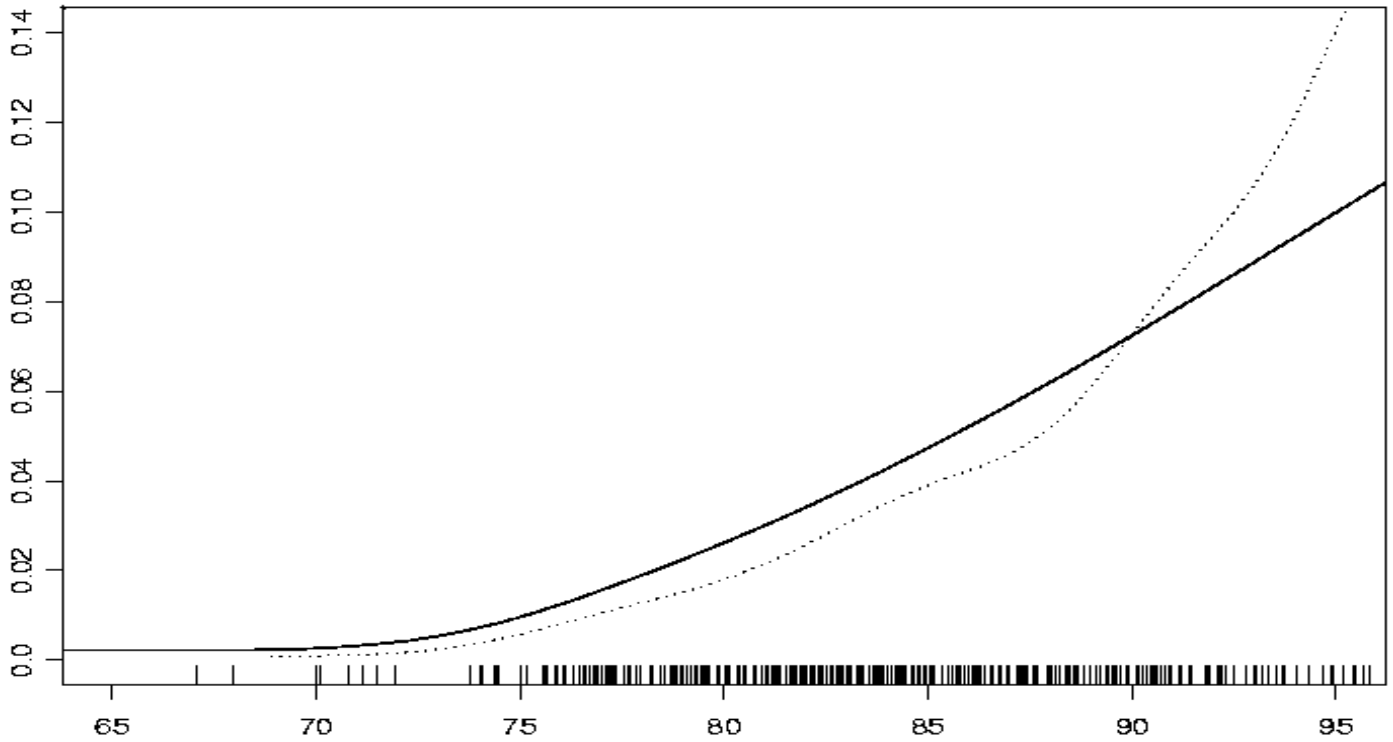
Simulation : estimateur à noyau (2)

	$-\mathrm{KL}(\widehat{\lambda}_h^{\mathcal{W}})$ for kernel estimators					
n	ELL	ELL_{bboot}	LCV	ELL_{iboot}	ELL_{boot}	RH
15% censoring						
30	3.96(0.005)	4.00(0.009)	4.01(0.02)	3.98(0.005)	4.04(0.01)	4.19(0.06)
50	3.99(0.003)	4.00(0.006)	4.00(0.008)	4.00(0.005)	4.04(0.01)	4.22(0.06)
100	4.01(0.002)	4.02(0.002)	4.02(0.002)	4.02(0.002)	4.05(0.005)	4.12(0.02)
25% censoring						
30	3.91(0.005)	3.94(0.009)	3.96(0.01)	3.92(0.006)	3.98(0.01)	4.26(0.08)
50	3.93(0.004)	3.95(0.007)	3.96(0.01)	3.94(0.006)	3.99(0.01)	4.2(0.06)
100	3.95(0.002)	3.96(0.002)	3.96(0.002)	3.96(0.002)	3.99(0.007)	4.10(0.03)
50% censoring						
30	3.92(0.04)	3.99(0.07)	4.04(0.07)	4.01(0.07)	4.02(0.08)	4.36(0.1)
50	3.84(0.02)	3.85(0.02)	3.91(0.03)	3.85(0.02)	3.88(0.03)	4.18(0.09)
100	3.80(0.005)	3.81(0.005)	3.83(0.02)	3.80(0.005)	3.84(0.008)	3.95(0.03)

Simulation : vraisemblance pénalisée

	$-\mathrm{KL}(\widehat{\lambda}_h^{\mathcal{W}})$
n=50 and 15% censoring	
ELL	3.99(0.003)
LCV	4.00(0.005)
LCV_a	4.00(0.005)
ELL_{bboot}	4.00(0.006)
ELL_{iboot}	4.06(0.01)
n=50 and 25% censoring	
ELL	3.93(0.006)
LCV	3.98(0.006)
LCV_a	3.99(0.009)
ELL_{bboot}	4.01(0.02)
ELL_{iboot}	4.08(0.02)
n=50 and 50% censoring	
ELL	3.96(0.008)
LCV	4.00(0.02)
LCV_a	4.07(0.03)
ELL_{bboot}	4.06(0.03)
ELL_{iboot}	4.32(0.06)

Application : risque de démence chez les femmes



Prise en compte des variables explicatives

- Données de survie (censure à droite)
  - $\mathcal{W} = (W_1, \ldots, W_n)$, $W_i = (\tilde{T}_i, \delta_i, X_i)$; $\tilde{T}_i = \min(T_i, C_i)$, $\delta_i = \mathbf{I}_{[T_i \leq C_i]}$
  - $T_i \sim F, f, S$ et $\lambda(t) = \frac{f(t)}{S(t)} = \lim_{dt \to 0} \frac{P(t \leq T < t+dt | T > t)}{dt}$
- modèle à risques proportionnels $\mathcal{M}_0$

$$\lambda(t|X_i) = \lambda^0(t) \exp \beta X_i \quad i = 1, \ldots, n$$

- modèle stratifié $\mathcal{M}_{\{1\}}$

$$\lambda(t|X_i) = \begin{cases} \lambda^0(t) & \text{if } X_i = 0 \\ \lambda^1(t) & \text{if } X_i = 1 \end{cases}$$

Objectif

- Choix d'un estimateur de la fonction de risque
  - modèle stratifié
  - modèle à risque proportionnel
- Construction d'un critère de Sélection :
  - choix du paramètre de lissage
  - choix entre modèle stratifié et risque proportionnel

Familles d'estimateurs

Procédure : vraisemblance pénalisée $\Longrightarrow$ famille d'estimateurs

- modèle à risques proportionnels : $\widehat{\lambda}_h^0(\cdot)$ et $\widehat{\beta}$

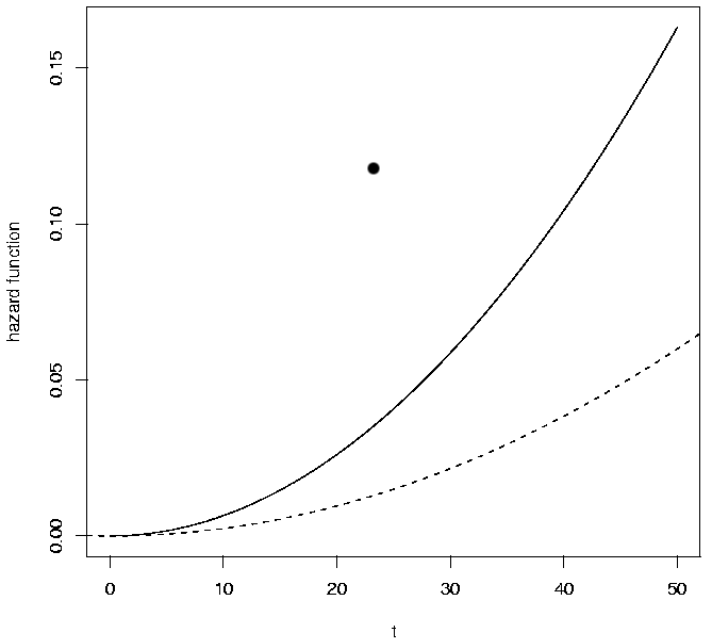
$$p\mathcal{L}_h(\mathcal{W}) = \log \mathcal{L}^{\lambda^0,\beta}(\mathcal{W}) - h \int {\lambda^0}''^2(u)du$$

$\longrightarrow$ famille d'estimateurs à 1 hyper-paramètre h

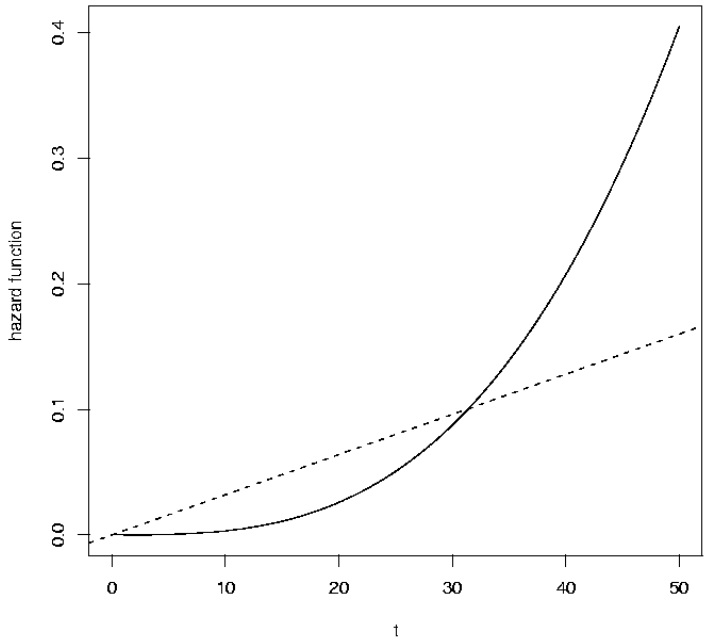
- modèle stratifié : $\widehat{\lambda}_h^0(\cdot)$ et $\widehat{\lambda}_h^1(\cdot)$

$$p\mathcal{L}_h(\mathcal{W}) = \log \mathcal{L}^{\lambda^0,\lambda^1}(\mathcal{W}) - h \int {\lambda^0}''^2(u)du - h \int {\lambda^1}''^2(u)du$$

$\longrightarrow$ famille d'estimateurs à 1 hyper-paramètre h



$\mathcal{M}_0$: Modèle à risques proportionnels



$\mathcal{M}_{\{1\}}$: Modèle stratifié

Modèle simulé $\mathcal{M}_0$

$\simeq$ 10%	n= 50		n= 100		n= 500	
censoring	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$
KL	837	163	897	103	962	38
LCVa	840	160	907	93	933	67
LCVa≡KL	701	24	809	5	899	4
$\simeq$ 25%	n= 50		n= 100		n= 500	
censoring	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$
KL	788	212	877	123	955	45
LCVa	828	172	892	108	932	68
LCVa≡KL	637	21	780	11	888	1
$\simeq$ 50%	n= 50		n= 100		n= 500	
censoring	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$
KL	811	189	862	138	949	51
LCVa	770	230	878	122	929	71
LCVa≡KL	616	35	748	8	878	0

Modèle simulé $\mathcal{M}_{\{1\}}$

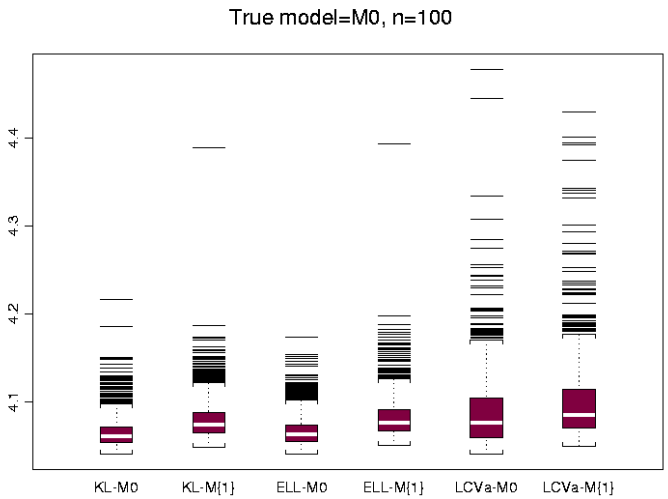
$\simeq$ 10%	n= 50		n= 100		n= 500	
censoring	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$
KL	4	996	0	1000	0	1000
LCVa	407	593	242	758	3	997
LCVa$\equiv$KL	0	589	0	758	0	997
$\simeq$ 25%	n= 50		n= 100		n= 500	
censoring	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$
KL	1	999	0	1000	0	1000
LCVa	442	558	293	707	6	994
LCVa$\equiv$KL	1	558	0	707	0	994
$\simeq$ 50%	n= 50		n= 100		n= 500	
censoring	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$
KL	44	956	10	990	0	1000
LCVa	441	559	319	681	8	992
LCVa$\equiv$KL	10	525	0	671	0	992

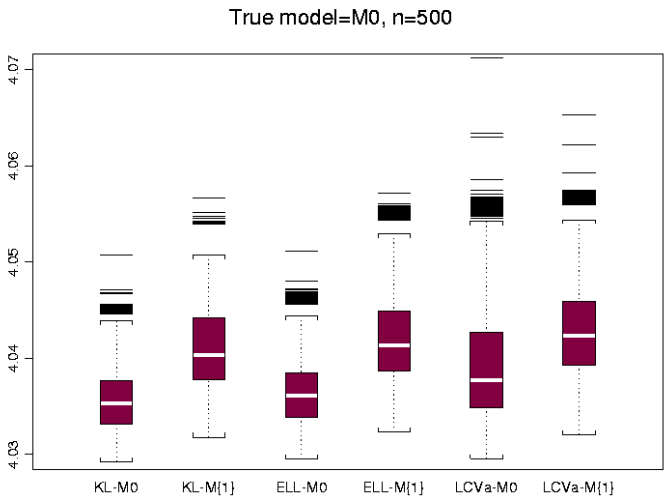
Modèle simulé $\mathcal{M}_0$

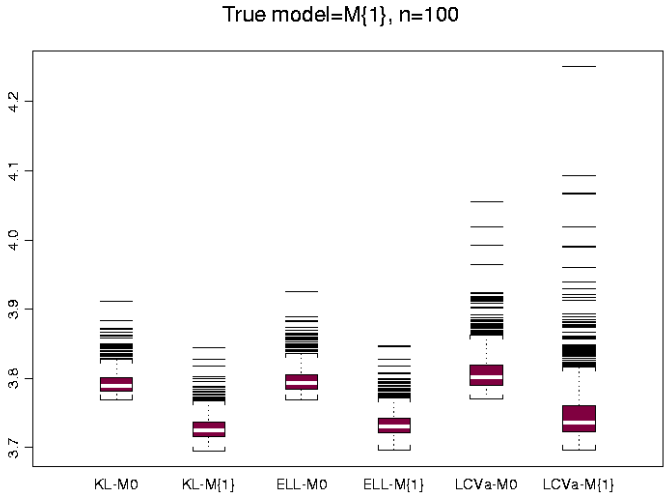
$\simeq$ 10% censoring	n= 50 $-\overline{KL}$	n= 100 $-\overline{KL}$	n= 500 $-\overline{KL}$
KL	4.061(0.0007)	4.053(0.0004)	4.040(0.0001)
ELL	4.070(0.0011)	4.056(0.0004)	4.040(0.0001)
LCVa	4.125(0.0034)	4.072(0.0011)	4.044(0.0002)
$\simeq$ 25% censoring	n= 50 $-\overline{KL}$	n= 100 $-\overline{KL}$	n= 500 $-\overline{KL}$
KL	4.074(0.0008)	4.064(0.0005)	4.036(0.0001)
ELL	4.086(0.0014)	4.067(0.0005)	4.036(0.0001)
LCVa	4.162(0.0050)	4.088(0.0013)	4.040(0.0002)
$\simeq$ 50% censoring	n= 50 $-\overline{KL}$	n= 100 $-\overline{KL}$	n= 500 $-\overline{KL}$
KL	4.114(0.0018)	4.079(0.0009)	4.053(0.0003)
ELL	4.137(0.0027)	4.098(0.0011)	4.056(0.0003)
LCVa	4.297(0.0119)	4.124(0.0027)	4.062(0.0006)

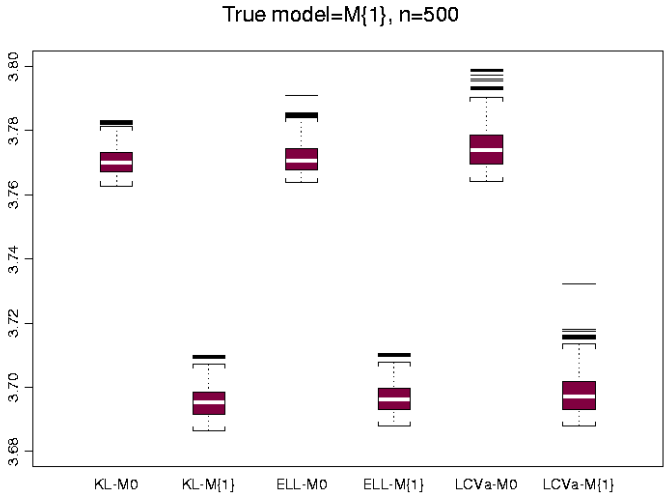
Modèle simulé $\mathcal{M}_{\{1\}}$

$\simeq$ 10% censoring	n= 50 $-\overline{KL}$	n= 100 $-\overline{KL}$	n= 500 $-\overline{KL}$
KL	3.734(0.0008)	3.715(0.0004)	3.695(0.0001)
ELL	3.740(0.0009)	3.719(0.0004)	3.696(0.0001)
LCVa	3.827(0.0039)	3.751(0.0016)	3.697(0.0002)
$\simeq$ 25% censoring	n= 50 $-\overline{KL}$	n= 100 $-\overline{KL}$	n= 500 $-\overline{KL}$
KL	3.740(0.0010)	3.729(0.0006)	3.695(0.0001)
ELL	3.746(0.0011)	3.734(0.0006)	3.697(0.0002)
LCVa	3.863(0.0103)	3.769(0.0017)	3.699(0.0003)
$\simeq$ 50% censoring	n= 50 $-\overline{KL}$	n= 100 $-\overline{KL}$	n= 500 $-\overline{KL}$
KL	3.781(0.0020)	3.750(0.0011)	3.713(0.0003)
ELL	3.795(0.0021)	3.772(0.0022)	3.717(0.0003)
LCVa	3.978(0.0092)	3.821(0.0043)	3.719(0.0006)

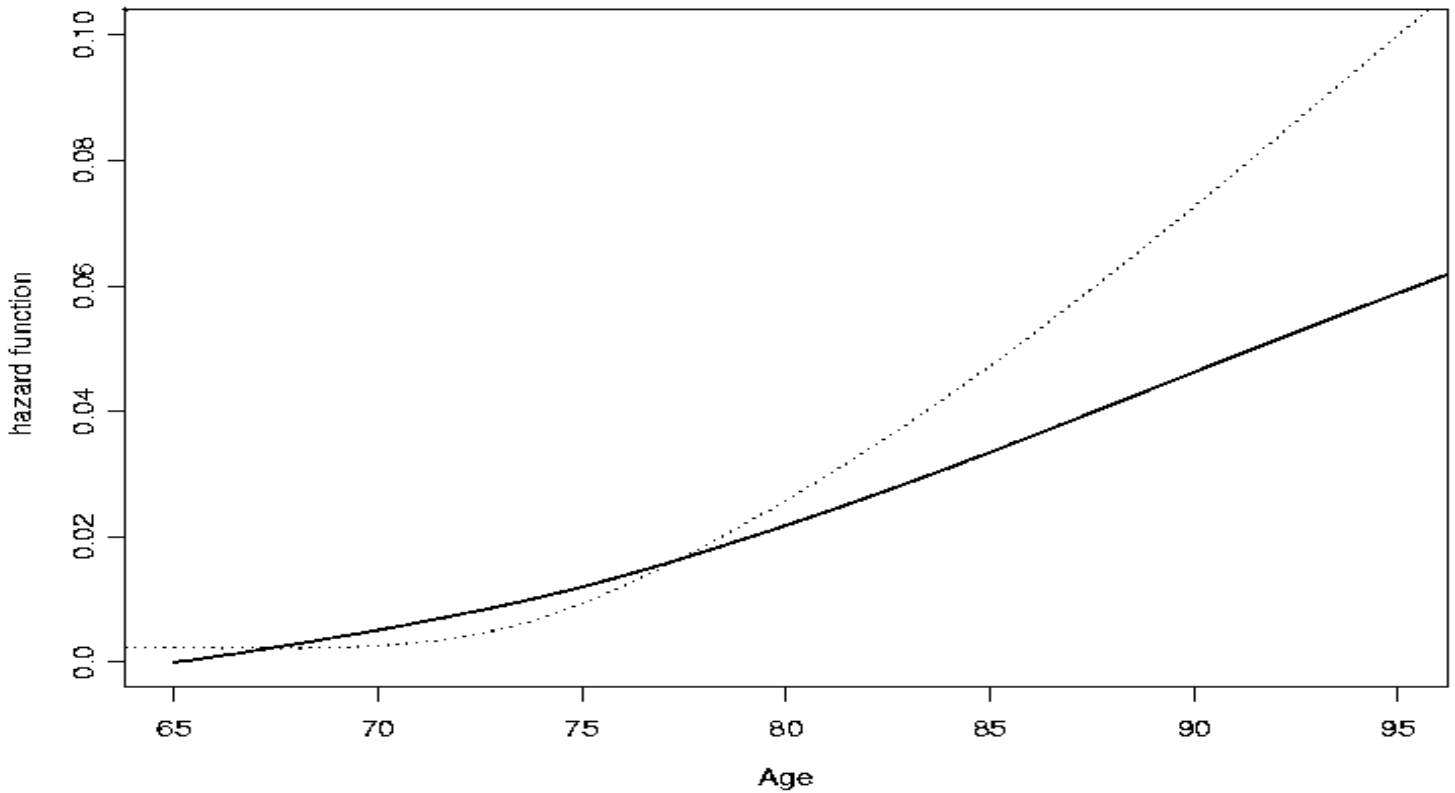

True model=M0, n=100
4.1
4.2
4.3
4.4
KL-M0
KL-M{1}
ELL-M0
ELL-M{1}
LCVa-M0
LCVa-M{1}


True model=M0, n=500
4.03
4.04
4.05
4.06
4.07
KL-M0
KL-M{1}
ELL-M0
ELL-M{1}
LCVa-M0
LCVa-M{1}


True model=M{1}, n=100
3.7
3.8
3.9
4.0
4.1
4.2
KL-M0
KL-M{1}
ELL-M0
ELL-M{1}
LCVa-M0
LCVa-M{1}


True model=M{1}, n=500
3.68
3.70
3.72
3.74
3.76
3.78
3.80
KL-M0
KL-M{1}
ELL-M0
ELL-M{1}
LCVa-M0
LCVa-M{1}

Application : Modélisation du risque de démence dans la cohorte PAQUID selon le sexe



modèle à risques proportionnels (Modèle B) ELL_{bboot} =-1517.04 et LCV_a=-1519.92

modèle stratifié (Modèle A) ELL_{bboot} =-1515.17 et LCV_a=-1517.45

Risque de démence selon le sexe et le niveau d'étude

- modèle à risques proportionnels stratifié sur le sexe (c) :

$$\lambda_i(t) = \begin{cases} \lambda_0^h(t) \exp \beta E_i & \text{si } S_i = 0 \text{ (femme)} \\ \lambda_1^h(t) \exp \beta E_i & \text{si } S_i = 1 \text{(homme)} \end{cases}$$

- modèle à risques proportionnels sur le niveau d'étude (D) :

$$\lambda(t|S_i, E_i) = \begin{cases} \lambda_h^0(t) \exp \beta_0 E_i & \text{if } S_i = 0 \text{ (women)}, \\ \lambda_h^1(t) \exp \beta_1 E_i & \text{if } S_i = 1 \text{ (men)}. \end{cases}$$

- modèle stratifié sur le sexe et le niveau d'étude (E) :

$$\lambda_i(t) = \begin{cases} \lambda_{0,0}^h(t) & \text{si } S_i = 0 \text{ et } E_i = 0 \\ \lambda_{1,0}^h(t) & \text{si } S_i = 1 \text{ et } E_i = 0 \\ \lambda_{0,1}^h(t) & \text{si } S_i = 0 \text{ et } E_i = 1 \\ \lambda_{1,1}^h(t) & \text{si } S_i = 1 \text{ et } E_i = 1 \end{cases}$$

Risque de démence selon le sexe et le niveau d'étude

- modèle à risques proportionnels stratifié sur le sexe (c) :

$$\lambda_i(t) = \begin{cases} \lambda_0^h(t) \exp \beta E_i & \text{si } S_i = 0 \text{ (femme)} \\ \lambda_1^h(t) \exp \beta E_i & \text{si } S_i = 1 \text{(homme)} \end{cases}$$

- modèle à risques proportionnels sur le niveau d'étude (D) :

$$\lambda(t|S_i, E_i) = \begin{cases} \lambda_h^0(t) \exp \beta_0 E_i & \text{if } S_i = 0 \text{ (women)}, \\ \lambda_h^1(t) \exp \beta_1 E_i & \text{if } S_i = 1 \text{ (men)}. \end{cases}$$

- modèle stratifié sur le sexe et le niveau d'étude (E) :

$$\lambda_i(t) = \begin{cases} \lambda_{0,0}^h(t) & \text{si } S_i = 0 \text{ et } E_i = 0 \\ \lambda_{1,0}^h(t) & \text{si } S_i = 1 \text{ et } E_i = 0 \\ \lambda_{0,1}^h(t) & \text{si } S_i = 0 \text{ et } E_i = 1 \\ \lambda_{1,1}^h(t) & \text{si } S_i = 1 \text{ et } E_i = 1 \end{cases}$$

Risque de démence selon le sexe et le niveau d'étude

- modèle à risques proportionnels stratifié sur le sexe (c) :

$$\lambda_i(t) = \begin{cases} \lambda_0^h(t) \exp \beta E_i & \text{si } S_i = 0 \text{ (femme)} \\ \lambda_1^h(t) \exp \beta E_i & \text{si } S_i = 1 \text{(homme)} \end{cases}$$

- modèle à risques proportionnels sur le niveau d'étude (D) :

$$\lambda(t|S_i, E_i) = \begin{cases} \lambda_h^0(t) \exp \beta_0 E_i & \text{if } S_i = 0 \text{ (women)}, \\ \lambda_h^1(t) \exp \beta_1 E_i & \text{if } S_i = 1 \text{ (men)}. \end{cases}$$

- modèle stratifié sur le sexe et le niveau d'étude (E) :

$$\lambda_i(t) = \begin{cases} \lambda_{0,0}^h(t) & \text{si } S_i = 0 \text{ et } E_i = 0 \\ \lambda_{1,0}^h(t) & \text{si } S_i = 1 \text{ et } E_i = 0 \\ \lambda_{0,1}^h(t) & \text{si } S_i = 0 \text{ et } E_i = 1 \\ \lambda_{1,1}^h(t) & \text{si } S_i = 1 \text{ et } E_i = 1 \end{cases}$$

Résultat de la sélection par le LCV

	model A	model B	model C	model D	model E
LCVa	-1517.45	-1519.92	-1496.28	-1497.18	-1498.42
ELL_{bboot}	-1515.17	-1517.04	-1492.79	-1494.03	-1495.35

- modèle à risques proportionnels stratifié sur le sexe :

$$\lambda_i(t) = \begin{cases} \lambda_0^h(t) \exp \beta E_i & \text{si } S_i = 0 \text{ (femme)} \\ \lambda_1^h(t) \exp \beta E_i & \text{si } S_i = 1 \text{(homme)} \end{cases}$$

- risque relatif pour le niveau d'éducation : 1.97 ($IC_{95\%} = [1.63; 2.37]$)

Résultat de la sélection par le LCV

	model A	model B	model C	model D	model E
LCVa	-1517.45	-1519.92	-1496.28	-1497.18	-1498.42
ELL_{bboot}	-1515.17	-1517.04	-1492.79	-1494.03	-1495.35

- modèle à risques proportionnels stratifié sur le sexe :

$$\lambda_i(t) = \begin{cases} \lambda_0^h(t) \exp \beta E_i & \text{si } S_i = 0 \text{ (femme)} \\ \lambda_1^h(t) \exp \beta E_i & \text{si } S_i = 1 \text{(homme)} \end{cases}$$

- risque relatif pour le niveau d'éducation : 1.97 ($IC_{95\%} = [1.63; 2.37]$)

Résultat de la sélection par le LCV

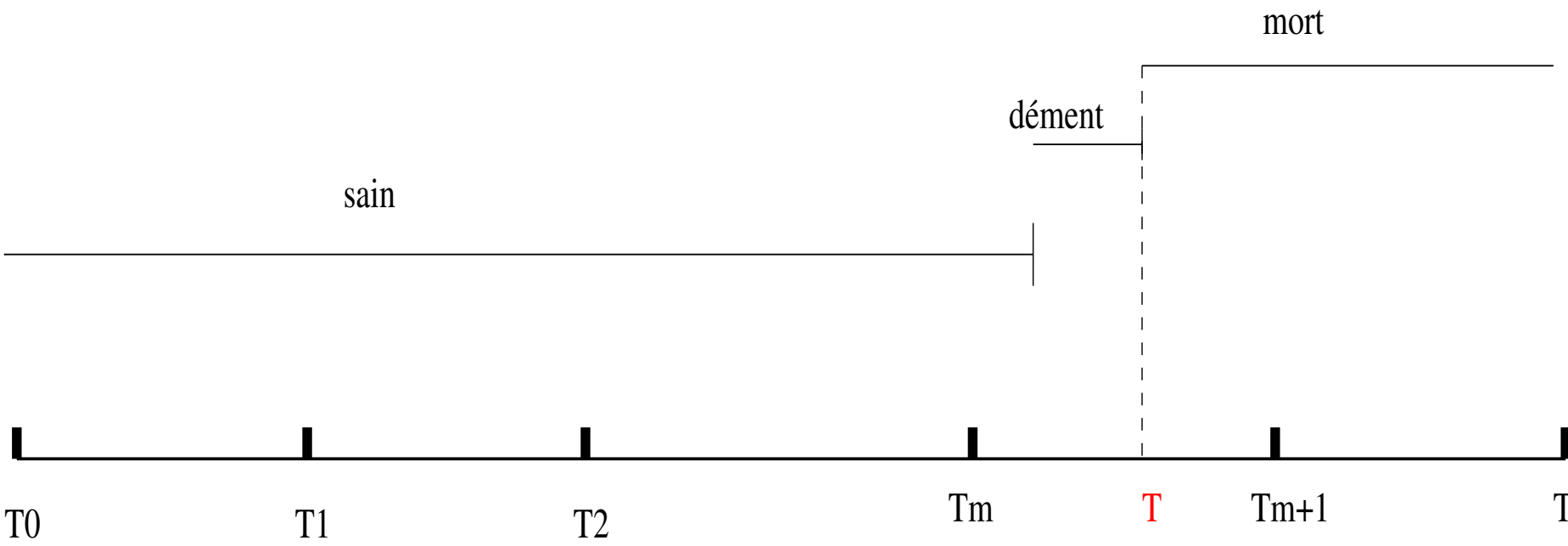
	model A	model B	model C	model D	model E
LCVa	-1517.45	-1519.92	-1496.28	-1497.18	-1498.42
ELL_{bboot}	-1515.17	-1517.04	-1492.79	-1494.03	-1495.35

- modèle à risques proportionnels stratifié sur le sexe :

$$\lambda_i(t) = \begin{cases} \lambda_0^h(t) \exp \beta E_i & \text{si } S_i = 0 \text{ (femme)} \\ \lambda_1^h(t) \exp \beta E_i & \text{si } S_i = 1 \text{(homme)} \end{cases}$$

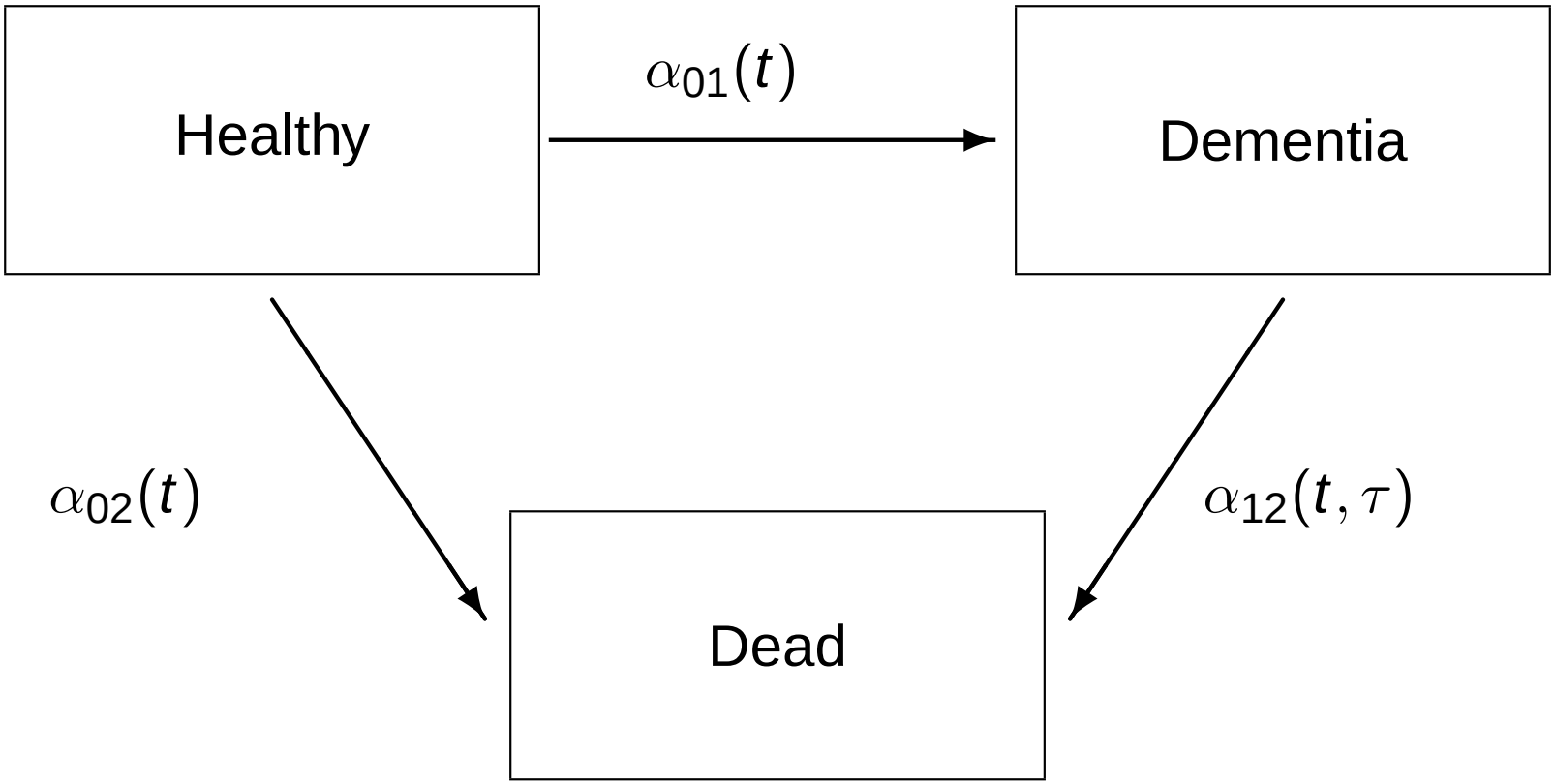
- risque relatif pour le niveau d'éducation : 1.97 ($IC_{95\%} = [1.63; 2.37]$)

Sous-estimation du risque de démence



- sujet considéré comme censuré à droite à T_m
- Démence et mort $\Longrightarrow$ risque compétitif
- censure par intervalle et forte mortalité pour les déments $\Longrightarrow$ sous-estimation

The “illness-death” model



- Les fonctions de transitions α_{01} et α_{02} dépendent de l’age
- α_{12} dépend de l’age où/et du temps passé dans l’état dément

Modèle et données

- Le temps passé dans l'état sain peut être censuré à droite, par intervalle où tronqué à gauche
- Le temps passé dans l'état dément est censuré à droite.
- Les dates de transition vers l'état mort sont connus exactement.
- $\alpha_{12}(t,\tau) = \alpha_{12}(t) \Longrightarrow$ modèle de Markov non-homogène
- $\alpha_{12}(t,\tau) = \alpha_{12}(\tau) \Longrightarrow$ est modèle semi-Markov .

Vraisemblance pénalisée

$$pl(\alpha_{01}, \alpha_{12}, \alpha_{02}) = l(\alpha_{01}, \alpha_{12}, \alpha_{02}) - \kappa_1 \int \alpha_{01}''^2(u)du$$

$$-\kappa_2 \int \alpha_{12}''^2(u)du - \kappa_3 \int \alpha_{02}''^2(u)du$$

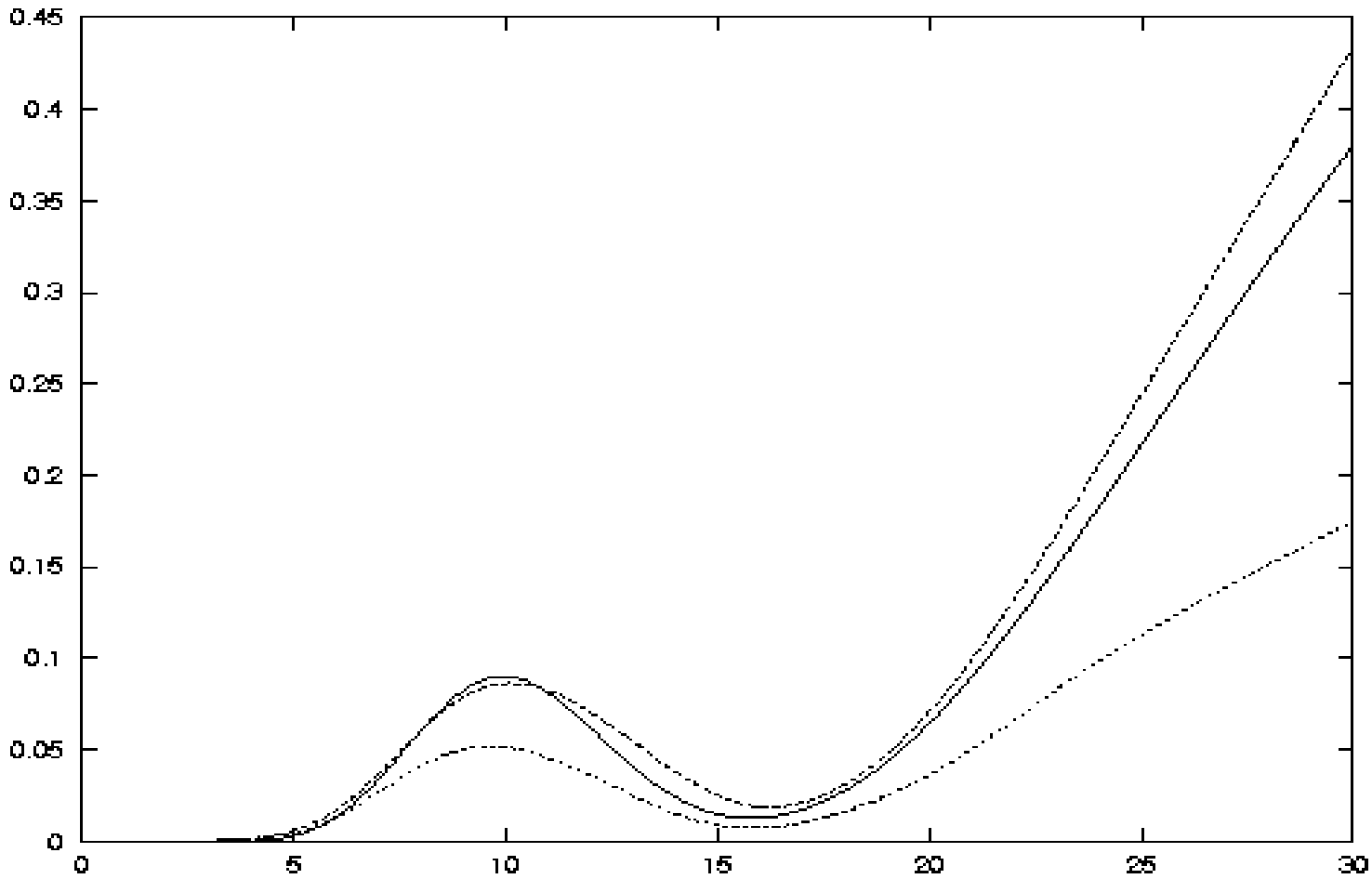
où l est la log-vraisemblance, κ_1, κ_2 et κ_3 paramètres de lissage

- estimateurs non-paramétriques $\Longrightarrow$ maximise $pl(\alpha_{01}, \alpha_{12}, \alpha_{02})$
- approximation par des splines
- une base de spline pour chaque transition $(\alpha_{01}, \alpha_{02}, \alpha_{12})$

Simulations

- $\alpha_{01}(t)$: mélange de Gamma
- $\alpha_{12}(t, \tau)$ et $\alpha_{02}(t)$ Weibull
- censure par intervalle et censure à droite
- 1000 sujets, 253 déments et 466 mort sans etre avoir été dément

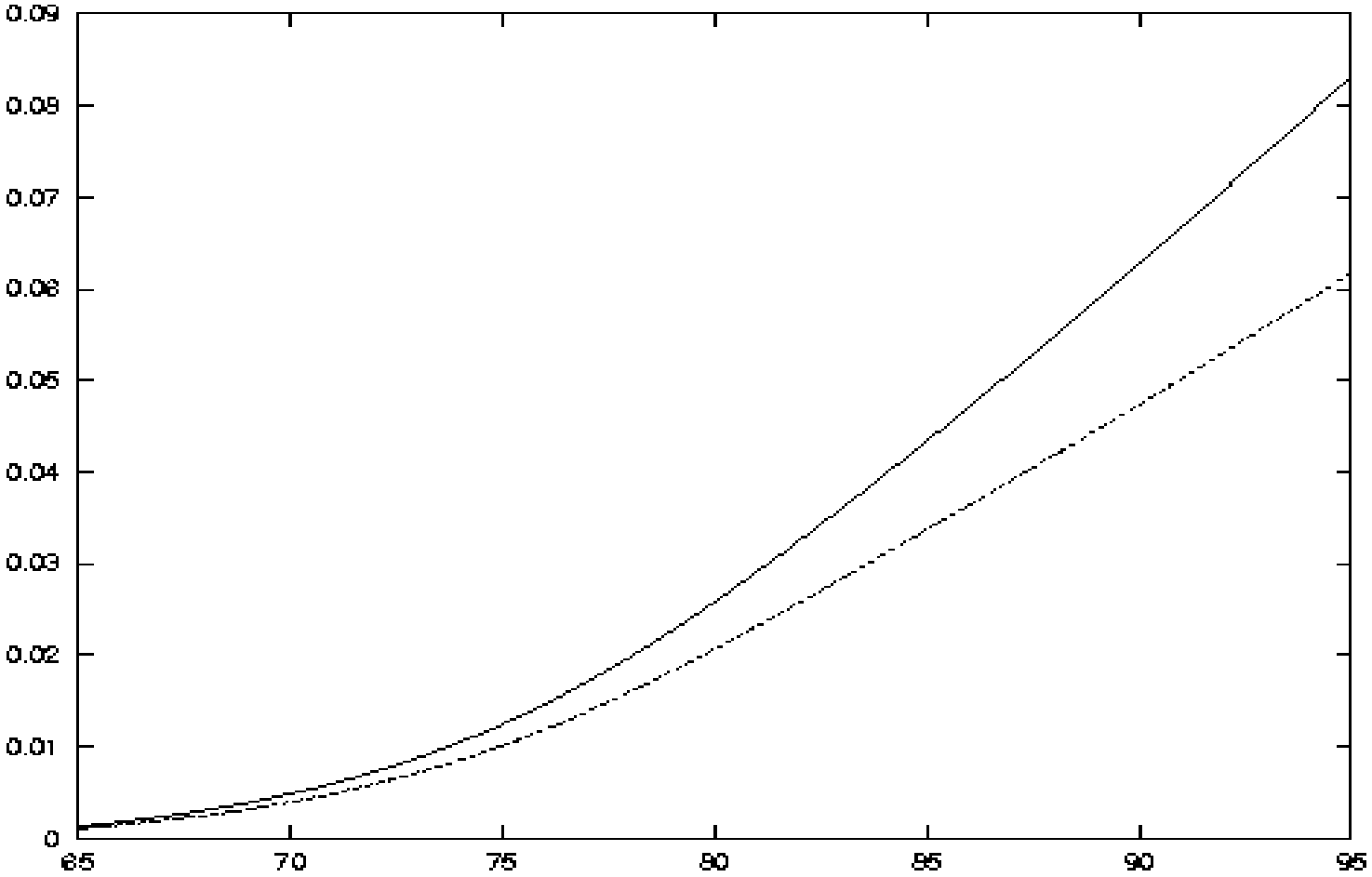
Simulations



— α_{01} simulé, - - - $\hat{\alpha}_{01}$ modèle 3 états ,
..... $\hat{\alpha}_{01}$ modèle de survie.

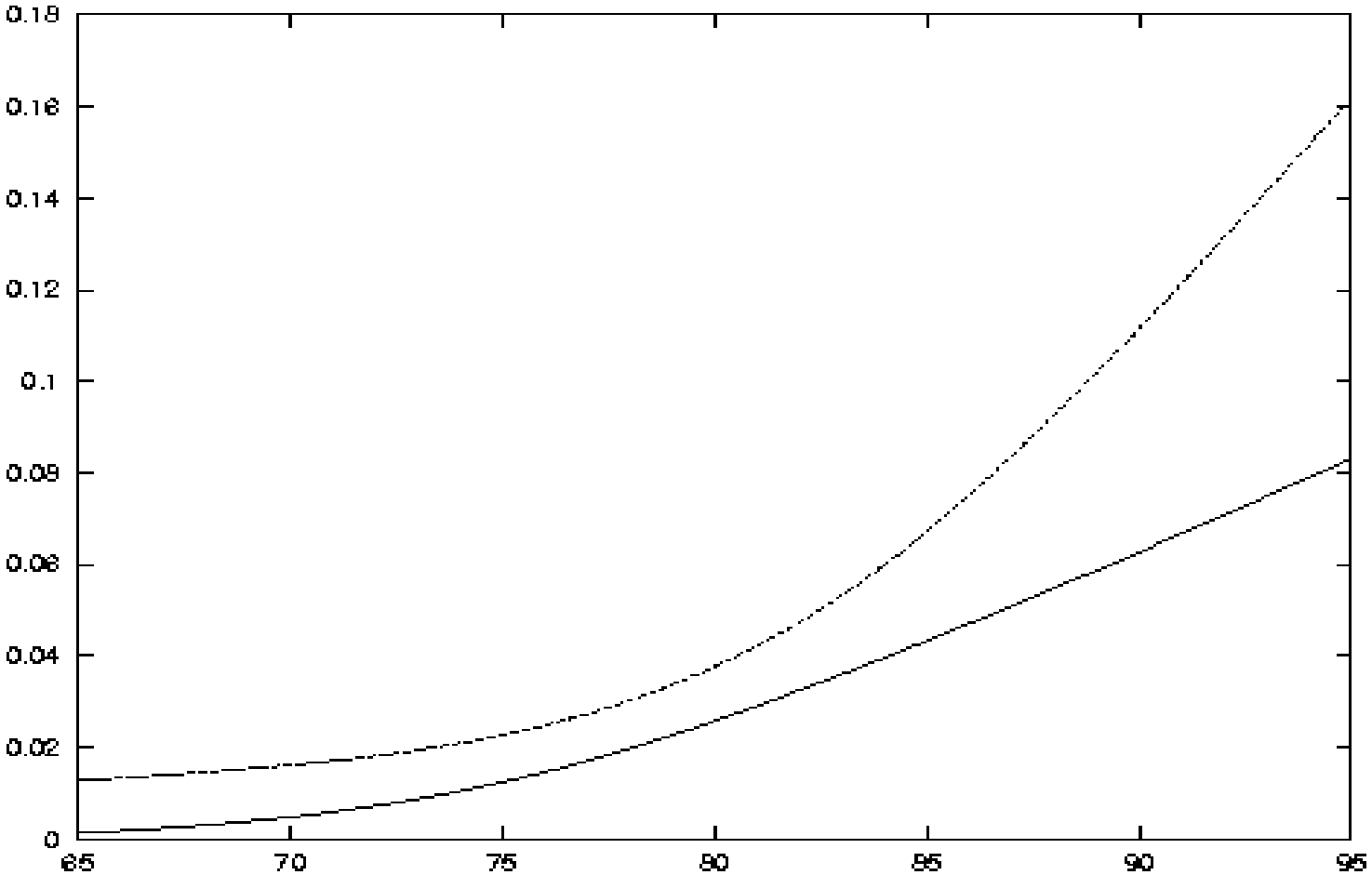
Données PAQUID

3675 sujets, 281 déments dont 11 déces. 1077 sujets mort sans être déments.

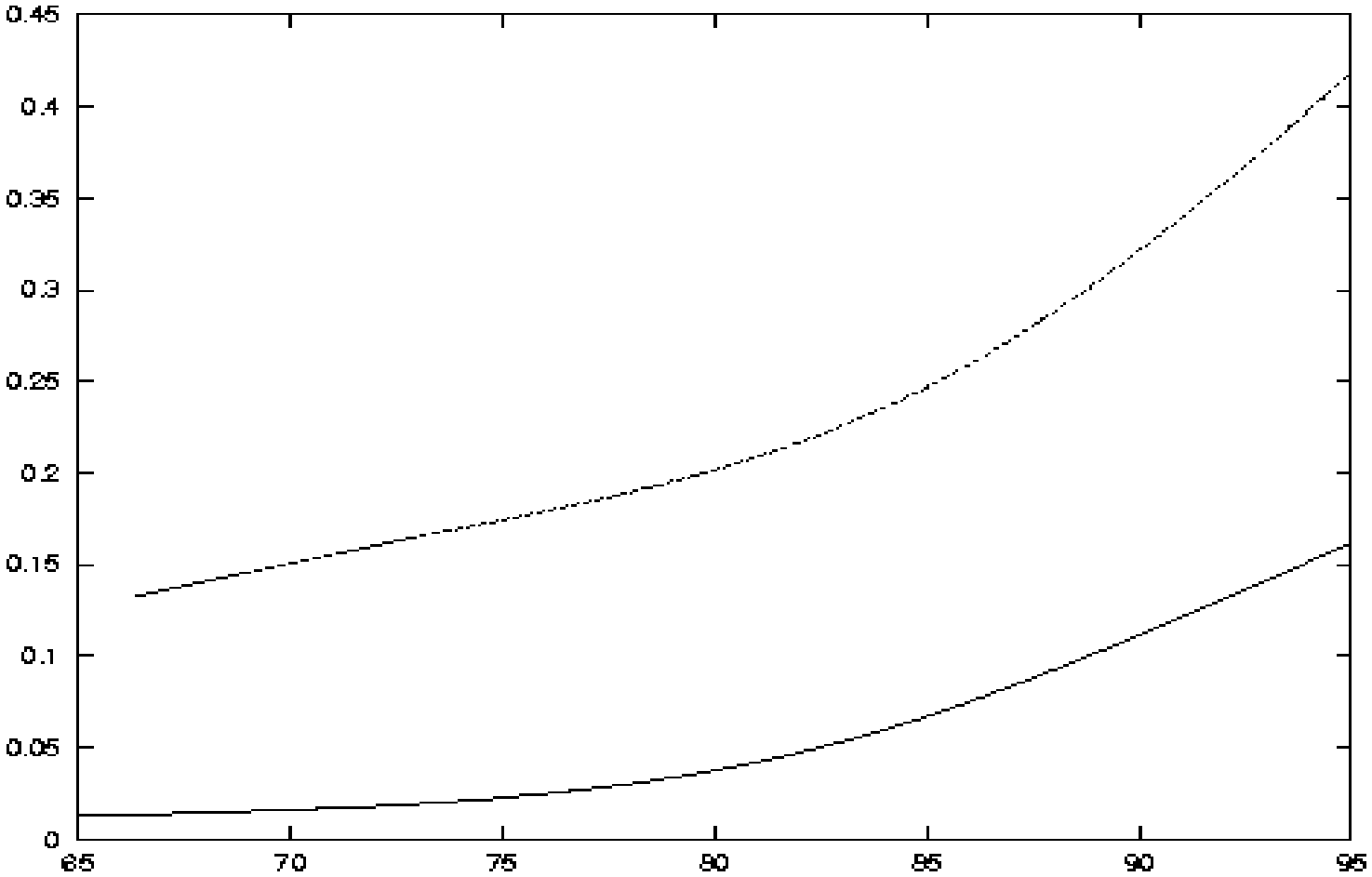


$\widehat{\alpha}_{01}$: —— modèle 3 états , - - - modèle de survie.

Données PAQUID



$\widehat{\alpha}_{01}$ (——)
et $\widehat{\alpha}_{02}$ (- - -)



$\widehat{\alpha}_{02}$ (——) et $\widehat{\alpha}_{12}$ (- - -).

Bibliographie I

P. Joly, D. Commenges, C. Helmer, D. Commenges,
A penalized likelihood approach for an illness-death model with interval-censored data : application to age-specific incidence of dementia
Biostatistics , 3 :433–443. 2002.

B. Liquet, C. Sakarovitch, D. Commenges,
Bootstrap choice of estimators in non-parametric families
Biometrics , 59 :172–178. 2003.

B. Liquet, D. Commenges,
Estimating the expectation of the log-likelihood with censored data for estimator selection.
Lifetime Data Analysis, 10 :351–367. 2004.

B. Liquet, S. Sarraco, D. Commenges,
Selection between proportional and stratified hazards models based on expected log-likelihood.
Computational Statistic, 2005