Modélisation de l'effet de facteurs de risque sur la probabilité de devenir dément : approche par pseudo-valeurs

Camille Sabathé [1,2,*] & Pierre Joly [1,2]

[1] *Université de Bordeaux, ISPED, Centre Inserm U1219, 33000 Bordeaux*
[2] *Inserm, ISPED, Centre Inserm U1219, 33000 Bordeaux*
[*] *camille.sabathe@u-bordeaux.fr*

Résumé. L'objectif de ce travail est d'étudier l'effet de facteurs de risque sur la probabilité de devenir dément. Une approche par pseudo-valeurs issues d'estimateurs non paramétriques permet de modéliser directement ces effets sur cet indicateur de santé. Une extension de la méthode par pseudo-valeurs dans le cadre de données censurées par intervalles est proposée ici. Pour cela, la probabilité de tomber malade est estimée à partir des estimateurs du maximum de vraisemblance pénalisée d'après un modèle *illness-death*. L'illustration de la méthode a été faite sur les données de la cohorte Paquid, qui a inclus plus de 3000 individus non déments âgés de 65 ans et plus, avec un suivi des sujets pendant plus de 25 ans.

Mots-clés. pseudo-valeurs ; données censurées par intervalles ; modèle illness-death ; probabilité de devenir dément

Abstract. The objective of this work is to study the effect of risk factors on the probability of developing dementia. Non-parametric pseudo-values approach allows direct modeling of these effects on this health indicator. We propose an extension of the pseudo-values method for interval-censored data. Briefly, the probability of dementia is estimated through penalized likelihood estimators of an illness-death model. The method is applied to the French cohort Paquid which included more than 3,000 non-demented subjects, aged 65 years or older and followed for dementia over more than 25 years.

Keywords. pseudo-values ; interval-censored data ; illness-death model ; probability of become demented

1 Contexte et objectif

La démence est une neuropathologie qui affecte plus d'un million de personnes en France en 2010 (World Health Organization et Alzheimer's Disease International, 2012). Les prédictions estiment à environ 1 750 000 individus atteints de démence en 2030 (Jacqmin-Gadda *et al.*, 2013). Du point de vue de la Santé Publique, il est important de mieux

comprendre cette affection chronique et ainsi cibler les stratégies de prévention ou de surveillance des personnes les plus à risque.

La démence affecte les sujets âgés et le risque de décès chez les personnes âgées est élevé. Les méthodes statistiques employées pour étudier la démence se doivent de prendre en compte ce risque compétitif. L'observation de la démence se fait en temps discret car le diagnostic de la maladie se fait lors des visites de suivis. Ceci entraine donc une censure par intervalle. Le modèle illness-death permet de tenir compte de ce type de données. En particulier, il prend en considération la possibilité qu'un individu décédé vu non dément à sa dernière visite ait pu développer une démence avant son décès.

L'objectif de ce travail est de quantifier l'effet de facteurs de risque sur la probabilité de devenir dément. D'un point de vue statistique, les méthodes classiques ne sont pas adaptées pour répondre à cette attente. D'un coté, la modélisation des intensités de transition ne permet pas un lien direct entre variables explicatives et probabilité de tomber malade. De l'autre, les méthodes qui modélisent directement la probabilité de devenir malade par rapport aux variables explicatives ne sont pas adaptées à la censure par intervalle (aussi bien pour le modèle de Fine et Gray (1999) que pour les pseudo-valeurs issues d'un estimateur non paramétrique (Andersen *et al.*, 2003). Ce travail propose d'étendre l'approche par pseudo-valeurs à des données censurées par intervalle.

2 Méthode

2.1 Modèle illness-death

Un des modèles multi-états couramment utilisé pour modéliser ce type de données de santé est le modèle *illness-death*. Ce modèle, représenté par la figure 1, peut être défini grâce à un processus stochastique, noté $\{X(t), t \geq 0\}$, avec $X(t)$ qui désigne l'état occupé au temps t. Pour le modèle *illness-death* on a donc $X(t) \in \mathcal{S} = \{0, 1, 2\}$. Ce modèle définit aussi deux variables aléatoires T_0 et T qui correspondent respectivement au temps de sortie de l'état 0 et au temps d'entrée dans l'état 2 (Andersen et Keiding, 2012):

$$T_0 = inf\{t; X(t) \neq 0\}$$
$$T = inf\{t; X(t) = 2\}$$

De plus, ce modèle est défini par ses intensités de transition. L'intensité de transition entre l'état k et l'état l est :

$$\alpha_{kl}(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(X(t + \Delta t) = l \mid X(t) = k)}{\Delta t}$$

avec $kl \in \{01, 02, 12\}$. De plus, on note $A_{kl}(t)$ les intensités de transition cumulées, avec $A_{kl}(t) = \int_0^t \alpha_{kl}(u)\mathrm{d}u$.

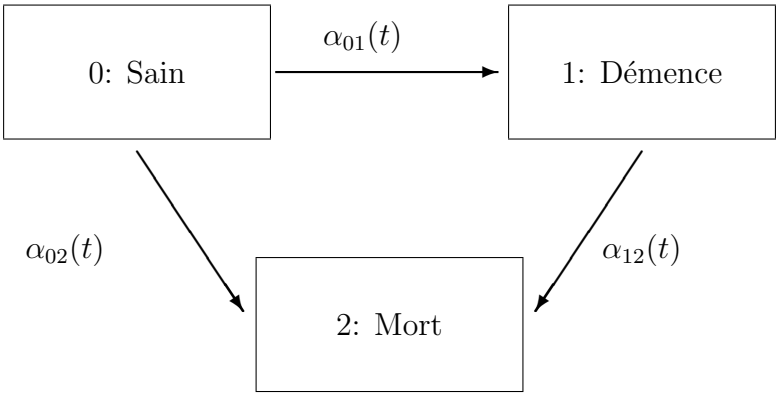

Figure 1: Modèle sain-malade-mort (*illness-death*)

Un indicateur de santé utilisé est la probabilité de devenir dément. Cette probabilité de devenir dément, aussi appelée fonction d'incidence cumulée, est :

$$\begin{aligned} F_{01}(t) &= \mathbb{P}(X(v) = 1 \mid X(0) = 0, 0 \leq v \leq t) \\ &= \int_0^t \exp\left[-A_{01}(u) - A_{02}(u)\right] \alpha_{01}(u)\mathrm{d}u \end{aligned}$$

2.2 Estimation de la probabilité de tomber malade

Une première étape consiste à estimer les paramètres des intensités de transition du modèle *illness-death*.

Pour tenir compte de la censure par intervalle, l'estimation de ces paramètres s'est fait par maximum de vraisemblance pénalisée en définissant la log-vraisemblance pénalisée du modèle, $pl(\alpha_{01}, \alpha_{02}, \alpha_{12})$, par :

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

avec $l(\alpha_{01}, \alpha_{02}, \alpha_{12})$ la log-vraisemblance, κ_{01}, κ_{02} et κ_{12} les paramètres de lissage pour chaque pénalisation. Le choix des paramètres de lissage est fait par validation croisée approchée (Joly *et al.* (2002)).

Après avoir estimé les 3 intensités de transition du modèle, il est possible d'estimer pour un temps, t, donné, la probabilité de tomber malade, c'est à dire $\hat{F}_{01}(t)$:

$$\hat{F}_{01}(t) = \int_0^t \exp\left[-\int_0^u \hat{\alpha}_{01}(v)\mathrm{d}v - \int_0^u \hat{\alpha}_{02}(v)\mathrm{d}v\right] \hat{\alpha}_{01}(u)\mathrm{d}u \quad (1)$$

$$= \int_0^t \exp\left[-\hat{A}_{01}(u) - \hat{A}_{02}(u)\right] \hat{\alpha}_{01}(u)\mathrm{d}u \quad (2)$$

2.3 Pseudo-valeurs

L'idée générale proposée par Andersen *et al.* (2003) est de modéliser l'effet de facteurs de risque sur une quantité d'intérêt grâce à des pseudo-valeurs. D'une manière globale, on peut dire qu'une pseudo-valeur est une statistique résumée de l'information apportée par les variables explicatives sur un estimateur. On définit la pseudo-valeur Y_i du sujet i par :

$$Y_i = n \times \hat{\theta} - (n-1) \times \hat{\theta}^{-i} \tag{3}$$

avec $\hat{\theta}$ l'estimation d'une quantité à partir d'un échantillon de n sujets , $\hat{\theta}^{-i}$ la même estimation à partir de l'échantillon sans le sujet i et n le nombre de sujets de l'échantillon. Pour rappel, les variables explicatives n'entrent jamais en compte pour le calcul d'une pseudo-valeur. Si l'on considère la probabilité de devenir dément entre 0 et t alors le calcul des pseudo-valeurs est :

$$Y_i(t) = n \times \hat{F}_{01}(t) - (n-1) \times \hat{F}_{01}^{-i}(t) \tag{4}$$

avec i=1,...,n.

2.4 Estimation des facteurs de risque influençant la probabilité de devenir dément

L'intérêt de cette approche est d'utiliser les pseudo-valeurs de chaque individu (calculées d'après la formule 4) comme variable réponse d'un modèle linéaire généralisé. Pour simplifier les notations, on pose $Y_{ij} = Y_i(t_j)$ avec $j = 1, ..., J$. Des études ont montré que 5 à 10 temps choisis de manière équidistante sur les différents temps d'évènements (c'est à dire que $5 \leq J \leq 10$) étaient suffisant pour capter assez d'information (Andersen *et al.*, 2003; Andersen et Klein, 2007; Andersen et Pohar Perme, 2010).

Les modèles utilisés dans la littérature sont de la forme suivante :

$$h\left(\mathrm{E}(Y_{ij})\right) = {Z_{ij}^{*}}^{\mathrm{T}} \beta^{*} \quad \text{avec } i = 1, ..., n \text{ et } j = 1, ..., J \tag{5}$$

avec h une fonction de lien, $Z_{ij}^{*} = (I(t_l = t_j), l = 1, .., J, Z_{ij})$ qui est le vecteur des p variables explicatives Z_{ij} auquel on a ajouté une indicatrice des temps, et β^{*} le vecteur de taille (p+J-1) des effets des variables du modèle. L'estimation des paramètres de régression se fait par équations d'estimation généralisées (GEE) et la matrice de travail avec une structure d'indépendance est la plus souvent utilisée dans la littérature.

Si on ne souhaite pas faire l'hypothèse que l'effet des variables explicatives sur notre quantité d'intérêt est le même au cours du temps, alors on peut inclure des interactions entre le temps et ces variables. On considère alors le modèle suivant :

$$h\left(\mathrm{E}(Y_{ij})\right) = {Z_{ij}^{+}}^{\mathrm{T}} \beta^{+} \quad \text{avec } i = 1, ..., n \text{ et } j = 1, ..., J \tag{6}$$

avec $Z_{ij}^{+} = (Z_{ij}, I(t_l = t_j), \tilde{Z}_{ij} \times I(t_l = t_j), l = 1, ..., J)$, Z_{ij} le p-vecteur de variable explicative et $\tilde{Z}_{ij}$ un vecteur de taille $q \leq p$ contenant les variables explicatives pour lesquelles on ne suppose pas un effet constant. On remarque que ce type de modélisation nécessite l'estimation de beaucoup de paramètres de régression.

Dans le cas particulier où les pseudo-valeurs sont calculées en un unique temps (c'est à dire $J = 1$), alors le modèle se restreint à :

$$h\left(\mathrm{E}(Y_{ij})\right) = Z_i^{\mathrm{T}}\beta \quad \text{avec } i = 1, ..., n \tag{7}$$

Lorsque le modèle 6 ne suppose pas un effet constant au cours du temps pour toutes les variables explicatives (c'est à dire $q = p$) alors cela revient à modéliser J modèles (7) où l'on ne se sert que des pseudo-valeurs du temps t_j considéré (et un utilisant une matrice de travail avec une structure d'indépendance).

Pour résumer, les différents modèles proposés ci-dessus donnent une utilisation simple et souple des pseudo-valeurs sur une quantité d'intérêt dans un modèle linéaire généralisé. Si l'on souhaite modéliser un effet plus global, on peut utiliser le modèle (5). Si on suppose que l'effet d'un facteur de risque change au cours du temps alors il vaut mieux utiliser le modèle décrit par l'équation 6.

3 Applications

Des simulations ont été menées pour regarder les performances de la méthode par pseudo-valeurs adaptées à des données censurées par intervalles (en terme de biais relatif et de taux de couverture). Une application aux données de la cohorte Paquid a été faite pour illustrer cette méthode sur des données réelles.

L'objectif principal de cette enquête était d'étudier le vieillissement cérébral, qu'il soit normal ou pathologique au travers par exemple des facteurs de risque (Dartigues *et al.*, 1992). L'étude Paquid a débuté en 1988 et 1989 en Gironde et Dordogne et a inclus 3 777 individus âgés de 65 ans et plus et vivant à domicile. Le suivi des individus s'est fait tous les 2 à 3 ans pendant plus de 25 ans avec une recherche active des cas de démence.

References

P. K. Andersen et N. Keiding : Interpretability and importance of functionals in competing risks and multistate models: Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, 31(11-12):1074–1088, 2012.

P. K. Andersen et J. P. Klein : Regression Analysis for Multistate Models Based on a Pseudo-value Approach, with Applications to Bone Marrow Transplantation Studies. *Scandinavian Journal of Statistics*, 2007.

P. K. Andersen, J. P. Klein et S. Rosthoj : Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003.

P. K. Andersen et M. Pohar Perme : pseudo observations in survival analysis. *Statistical Methods in Medical Research*, 2010.

J. F. Dartigues, M. Gagnon, P. Barberger-Gateau, L. Letenneur, D. Commenges, C. Sauvel, P. Michel et R. Salamon : The Paquid epidemiological program on brain ageing. *Neuroepidemiology*, 11 Suppl 1:14–18, 1992.

J. P. Fine et R. J. Gray : A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 1999.

H. Jacqmin-Gadda, A. Alperovitch, C. Montlahuc, D. Commenges, K. Leffondre, C. Dufouil, A. Elbaz, C. Tzourio, J. Ménard, J.-F. Dartigues et P. Joly : 20-Year prevalence projections for dementia and impact of preventive policy about risk factors. *European Journal of Epidemiology*, 2013.

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

World Health Organization et Alzheimer's Disease International, éditeurs. *Dementia : a public health priority.* Geneva, 2012.