MODÉLISATION PAR LES MODÈLES MULTI-ÉTATS

Une étude canadienne menée auprès de 125 enfants scolarisés en 6ème a porté sur leur comportement tabagique : après leur rentrée scolaire ($t_0 = 0$), ils ont été suivis aux moments suivants (en années) : $t_1 = 0.15$, $t_2 = 0.75$, $t_3 = 1.10$ et $t_4 = 1.90$ (voir [2]) [1]. Trois "états" tabagiques ont été considérés :

— État 1 : l'enfant n'a jamais fumé
— État 2 : l'enfant fume actuellement
— État 3 : l'enfant a déjà fumé, mais ne fume plus actuellement

À chaque visite, on a dénombré le nombre d'enfants qui sont dans chaque état en fonction de leur état à la visite précédente et on comptabilise ainsi les transitions (les nombres entre parenthèses sont des nombres attendus utiles par la suite) :

État à t_0 / État à t_1	1	2	3
1	93 (96.0)	3 (1.7)	2 (0.3)
2	0 (0)	8 (12.9)	10 (5.1)
3	0 (0)	1 (0.5)	8 (8.5)

TABLE 1. Transitions des états à la date t_0 aux états à la date t_1

État à t_1 / État à t_2	1	2	3
1	89 (85.7)	2 (4.2)	2 (3.1)
2	0 (0)	7 (4.0)	5 (8.0)
3	0 (0)	5 (2.8)	15 (17.2)

TABLE 2. Transitions des états à la date t_1 aux états à la date t_2

État à t_2 / État à t_3	1	2	3
1	83 (84.9)	3 (2.9)	3 (1.2)
2	0 (0)	9 (6.8)	5 (7.2)
3	0 (0)	2 (2.3)	20(19.7)

TABLE 3. Transitions des états à la date t_2 aux états à la date t_3

État à t_3 / État à t_4	1	2	3
1	76 (74.5)	3 (4.3)	4 (4.3)
2	0 (0)	6 (3.7)	8 (10.3)
3	0 (0)	0 (4.3)	28(23.7)

TABLE 4. Transitions des états à la date t_3 aux états à la date t_4

On définit pour un individu donné l'état $X(t)$ dans lequel il se trouve à la date t et les probabilités de transition $p_{ij} = \mathbb{P}\big(X(t) = j | X(s) = i\big)$

1. L'analyse qui suit fait appel à une modélisation complexe comportant plusieurs paramètres d'intérêt. Le but est de pouvoir interpréter la démarche suivie par les auteurs. Il ne sera en particulier fait aucun calcul de variance asymptotique d'estimateurs qui aurait nécessité des calculs de covariances (au delà du cours). Néanmoins, ce type de modélisation communément appelée "modèle multi-états", se rencontre dans différentes pathologies : cancer [3], sida [4], complications ophtalmologique du diabète [5] et asthme [1].

pour $0 \leq s \leq t$ et $i = 1, 2, 3, j = 1, 2, 3$. On propose une modélisation markovienne des données, où la distribution de probabilité de l'état futur $X(t)$, connaissant l'état présent $X(s)$ ne dépend pas du passé. On note $\omega_\ell, \quad \ell = 1, 2, 3, 4$ les temps écoulés $\omega_\ell = t_\ell - t_{\ell-1}$.

A. On note $P(\ell)$ la matrice des probabilités de transition $p_{ij}(\omega_\ell) = p_{ij}\big(t_{\ell-1}, t_\ell\big)$ où $\ell = 1, 2, 3, 4$.

— Combien de paramètres inconnus contient en général et dans notre cas spécifique la matrice $P(\ell)$ et combien y a-t-il de matrices $P(\ell)$ [2] ? Cela fait combien de paramètres à estimer au total dans notre problème ? Dans la suite, on s'intéressera à la vraisemblance de ces paramètres.

— On considère par exemple un enfant qui passe de l'état 1 en t_0 à l'état 2 en t_1 pour y rester ensuite. En utilisant la propriété markovienne de notre contexte, écrire la contribution à la vraisemblance de cet enfant conditionnellement à son état en t_0.

— Écrire la vraisemblance et la log-vraisemblance en fonction de ces paramètres et des comptages de transitions $n_{ij\ell}$ observés (conditionnelles à la distribution des enfants dans les états en t_0).

— Écrire les dérivées partielles par rapport à chacun de ces paramètres (on trouvera une formule commune) : cela consiste à dériver par rapport à un paramètre en considérant les autres paramètres comme des constantes. On rappelle que la dérivée de la fonction $\ln(x)$ est $\frac{1}{x}$.

— On obtient le système des équations du maximum de vraisemblance en posant l'égalité à zéro de chacune de ces dérivées partielles. Résoudre ce système pour obtenir les estimateurs du maximum de vraisemblance des paramètres.

— Quels sont les nombres attendus dans chaque cellule avec ces estimations ?

B. On suppose de plus que notre modélisation est un processus de Markov homogène : les probabilités de transition ne dépendent que de l'intervalle de temps écoulé, c'est-à-dire que $p_{ij}\big(s, s+u\big) = p_{ij}(0, u) = p_{ij}(u)$ pour i, j, s et u. On définit les intensités (ou forces instantanées de passage) données par

$$\lambda_{ij} = \lim_{\Delta t \to 0} \frac{p_{ij}(\Delta t)}{\Delta t} \quad \text{pour} \quad i \neq j \quad \text{et} \quad \lambda_{ii} = -\sum_{i \neq j} \lambda_{ij}.$$

La matrice des intensités considérée ici est

$$Q = \begin{pmatrix} -\theta_1 & \theta_1 & 0 \\ 0 & -\theta_2 & \theta_2 \\ 0 & \theta_3 & -\theta_3 \end{pmatrix} \quad \text{où} \quad \theta_i > 0, i = 1, 2, 3.$$

— Représenter les possibilités de passage entre les différents états avec un graphique où la présence d'une flèche entre un état et un autre, surmontée de la valeur de l'intensité, indique la possibilité de passage instantané.

— On admettra que les probabilités de transition peuvent s'écrire à partir des intensités, par exemple $p_{11}(\omega) = e^{-\theta_1 \omega}$ (les autres sont un peu plus complexes). On peut alors écrire la vraisemblance en fonction de ces nouveaux paramètres et les estimer par maximisation

2. On oubliera pas qu'il s'agit de probabilités et on considèrera de plus, les transitions possibles et impossibles

de la vraisemblance. Les estimations obtenues sont (on admettra) :

$$\widehat{Q} = \begin{pmatrix} -0.136 & 0.136 & 0 \\ 0 & -2.28 & 2.28 \\ 0 & 0.47 & -0.47 \end{pmatrix}$$

ce qui conduit à des estimations des nombres attendus $e_{ij\ell}$ données entre parenthèses au début de l'énoncé. Exprimer les probabilités de transition estimées $p_{ij}(\omega_\ell)$ en fonction de $e_{ij\ell}$.

— Avec ces estimations et ce modèle, à combien peut-on prédire le nombre d'enfants qui n'ayant jamais fumé à l'entrée de l'étude ne fumeront pas trois ans plus tard ?

— On souhaite tester la validité de cette modélisation markovienne (adéquation du modèle) et pour cela comparer la vraisemblance optimisée sans modèle (partie **A.**) avec la vraisemblance optimisée en modélisant les probabilités de transition à partir des intensités. Quelle est la statistique du rapport de vraisemblance ? Donner sa valeur numérique (pour une cellule où $n_{ij\ell} = 0$, compter "zéro").

— Soient m_0 le nombre de paramètres à estimer dans le modèle sous l'hypothèse nulle (adéquation du modèle) et m_1 le nombre de paramètres estimés sous l'hypothèse alternative, les deux hypothèses étant "emboîtées", c'est à dire que H_0 peut être considérée comme un cas particulier de H_1. Donner m_0 et m_1. La statistique du rapport de vraisemblance est distribuée sous H_0 suivant la loi du χ^2 dont le nombre de degrés de liberté est la différence $m_1 - m_0$. Conclure sur la validité du modèle. Commenter.

Références

[1] T. Boudemaghe and J. Daures. Modeling asthma evolution by a multi-state model. *Revue d'épidémiologie et de santé publique*, 48(3) :249–255, 2000.

[2] J. D. Kalbfleisch and J. F. Lawless. The Analysis of Panel Data under a Markov Assumption. *Journal of the American Statistical Association*, 80(392) :863–871, 1985.

[3] R. Kay. A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies. *Biometrics*, 42(4) :855–865, 1986.

[4] I. M. Longini, W. S. Clark, R. H. Byers, J. W. Ward, W. W. Darrow, G. F. Lemp, and H. W. Hethcote. Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in medicine*, 8(7) :831–843, 1989.

[5] G. Marshall and R. H. Jones. Multi-state models and diabetic retinopathy. *Statistics in medicine*, 14(18) :1975–1983, 1995.