Projet 1

Hearts : A mixture model for count data

Thomas DUPONT, Enzo MORI, Gregory PERRET, Simeo **POTI**RON

03/2022

Méthodes bayésiennes et Modèles hiérarchiques

1 Description du jeu de données

Nous disposons d'un tableau de données concernant les effets d'un médicament sur des patients atteints de contractions ventriculaires prématurées (CVP) au niveau du coeur.

PVC's per minute			
PatientNumber(i)	Pre-drug (x_i)	Post-drug (y_i)	Decrease
1	6	5	1
...	...	...	...
12	51	0	51

Dans le tableau, un 0 correspond à un patient soigné, ou à une erreur sur un patient ayant un nombre anormalement élevé de CVP. Le modèle initial est le suivant :

1. $x_i \sim$ Poisson(λ_i), pour tous les patients *(avant la prise du médicament)*
2. $y_i \sim$ Poisson($\beta.\lambda_i$), pour les patients non soignés *(après la prise du médicament)*
3. P(cure)=θ

Pour éliminer la nuisance λ_i, on introduit la distrbution y_i conditionnellement à $t_i = x_i + y_i$, qui est équivalente à une distribution binomiale pour y_i de paramètre t_i et de probabilité p = $\frac{\beta}{1+\beta}$. Le modèle de mélange final peut ainsi être exprimé de la façon suivante :

$P(y_i = 0|t_i) = \theta + (1-\theta)(1-p)^{t_i}$

$P(y_i|t_i) = (1-\theta)\frac{t_i!}{y_i!(ti-yi)!}p_i^{y_i}(1-p_i)^{t_i-yi} \qquad y_i = 1, 2, ..., t_i$

2 Justification du modèle mathématique

Nous avons fait le choix de traiter le problème avec un algorithme de Metropolis-Hastings associé à un échantilloneur de Gibbs.

D'après notre énoncé, on a : $p = \frac{\beta}{1+\beta}$. Dans notre modèle, on cherche à déterminer β et θ. On introduit alors α et δ tels que :

$$\left.\begin{aligned}\beta = exp(\alpha) \Rightarrow \alpha = sigmoid(p)\\ \theta = sigmoid(\delta)\end{aligned}\right\}$$ On introduit α et δ car on connaît leur loi a priori.

On garde tout de même θ et p dans la formule pour que le calcul reste lisible.

On commence par calculer la loi a posteriori :

$\pi(\alpha, \delta|(y_i, t_i)) \propto \prod_{i,y_i=0}(\theta + (1-\theta)(1-p)^{t_i}) \prod_{i,y_i\neq 0}(\binom{t_i}{y_i}(1-\theta)(1-p)^{t_i-y_i}p^{y_i})\pi(\alpha)\pi(\delta)$

On a de plus : $\alpha \sim \mathcal{N}(0, 10^{-4})$ et $\delta \sim \mathcal{N}(0, 10^{-4})$, d'où : $\pi(\alpha) = \exp\left(-\frac{\alpha^2}{2*10^{-4}}\right)$ et $\pi(\delta) = \exp\left(-\frac{\delta^2}{2*10^{-4}}\right)$

Lorsque la loi à posteriori n'est pas classique il est plus judicieux d'effectuer un passage au logarithme à l'étape suivante (cela permet également de réduire les erreurs de calcul) :

$\log(\pi(\alpha, \delta|(y_i, t_i))) \propto cste + \sum_{i,y_i=0} log(\theta + (1-\theta)(1-p)^{t_i}) + \sum_{i,y_i\neq 0}(log(1-\theta) + y_i log(p)$

$+(t_i - y_i)log(1-p)) - \frac{\alpha^2}{2.10^{-4}} - \frac{\delta^2}{2.10^{-4}}$

On isole les termes en fonction des variables :

- Si on garde θ/δ :

  $cste + \sum_{i,y_i=0} log(\theta + (1-\theta)(1-p)^{t_i}) + \sum_{i,y_i\neq 0} log(1-\theta) - \frac{\delta^2}{2.10^{-4}}$

- Si on garde $\beta/\alpha/p$:

  $cste + \sum_{i,y_i=0} log(\theta + (1-\theta)(1-p)^{t_i}) + \sum_{i,y_i\neq 0} y_i log(p) + (t_i - y_i)log(1-p) - \frac{\alpha^2}{2.10^{-4}}$

3 Résultats et analyse

Nous vérifions que nos chaînes de Markov (α_t) et (δ_t) sont cohérentes graphiquement :

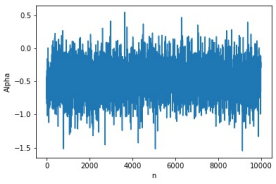

Figure 1: *Chaîne de Markov α_t pour t=0 à 10000*

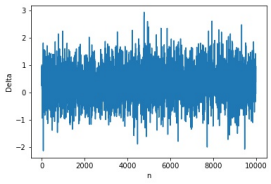

Figure 2: *Chaîne de Markov δ_t pour t=0 à 10000*

Ainsi par la suite, pour une chaine de longueur 10000 et un temps de chauffe de 1000 on obtient les résultats suivants :

Résultats					
	α	β	δ	θ	p
Moyenne	-0.47	0.65	0.32	0.57	0.39
Ecart-Type	0.27	0.18	0.61	0.14	0.06
Quantile 2.5%	-1.00	0.37	-0.86	0.30	0.27
Médiane	-0.47	0.62	0.32	0.58	0.38
Quantile 97.5%	0.05	1.05	1.56	0.83	0.51

Nous avons des résultats sensiblement similaires à ceux fournis par l'énoncé :

Résultats				
	α	β	δ	θ
Moyenne	-0.48	0.64	0.31	0.57
Ecart-Type	0.28	0.18	0.62	0.14
Quantile 2.5%	-1.04	0.35	-0.89	0.29
Médiane	-0.48	0.31	0.32	0.58
Quantile 97.5%	0.07	1.07	1.55	0.83

En conclusion, notre travail nous a permis d'estimer β et θ. Avec les données initiales sur x_i et y_i, on obtient que : $\frac{E(y_i)}{E(x_i)} = \beta$

On obtient de plus θ=0.57, ce qui signifie que les patients disposent d'une probabilité de 57% de guérir de la maladie.

Considérant ceux ne guérissant pas du traitement, on reprend le rapport vu ci-dessus : $\frac{E(y_i)}{E(x_i)} = 0.65 < 1$, donc il apparaît que ce traitement fait en sorte de diminuer le nombre de CVP par minute, et ce même si le patient n'est pas totalement guéri.