Projet Bayes : Oxford - Smooth Fit to Log-Odds Ratios

Oussama Hadder - Marwane Fahim - Salimou Moussa - Achraf Drissi

Option mathématiques et applications
Parcours sciences des données

Tuteur:
Mention : Mathieu Ribatet
Mathieu.Ribatet@ec-nantes.fr

Présentation des données

Dans les années 1993, Breslow et Clayton ont ré-analysé des tableaux de contingence de 2 par 2, qui présentaient les cas de décès dus au cancer infantile ainsi que les témoins, en fonction de l'exposition maternelle aux rayons X. Au total, 120 combinaisons d'âge (de 0 à 9 ans) et d'année de naissance (de 1944 à 1964) ont été étudiées, et chaque combinaison a donné lieu à un tableau. Les données peuvent être organisées sous la forme suivante.

Strata	Exposure : X-ray /Total Cases	Controls	Age	year -1954
1	3/28	0/28	9	-10
...				
120	7/32	1/32	1	10

Modèle mathématique

Considérons les variables suivantes : r_{0i} représente le nombre de personnes exposées aux rayons X parmi les n_{0i} individus non atteints de cancer (témoins) dans une catégorie i donnée, et r_{1i} représente le nombre de personnes exposées aux rayons X parmi les n_{1i} individus atteints de cancer (cas) dans cette même catégorie i. Si nous définissons p_{0i} comme la probabilité qu'un individu non atteint de cancer ait été exposé aux rayons X et p_{1i} comme la probabilité qu'un individu atteint de cancer ait été exposé, il est alors raisonnable de choisir des lois binomiales pour r_{0i} et r_{1i} de la manière suivante :

$$r_{0i} \sim \text{Binomial}(p_{0i}, n_{0i}) \quad r_{1i} \sim \text{Binomial}(p_{1i}, n_{1i})$$

Plutôt que d'étudier directement les probabilités p_{0i} et p_{1i}, nous nous intéressons à leurs cotes (Odds), calculées comme le rapport entre le nombre d'événements produisant un résultat et le nombre d'événements ne le produisant pas. Ainsi, nous avons respectivement $\frac{p_{0i}}{1-p_{0i}}$ et $\frac{p_{1i}}{1-p_{1i}}$, et nous appliquons la fonction logarithme pour obtenir une symétrie dans les résultats.

Ensuite, nous nous intéressons au rapport des cotes logarithmiques (log-odds ratio), c'est-à-dire $\log(\Psi) = \log\left(\frac{\frac{p_{1i}}{1-p_{1i}}}{\frac{p_{0i}}{1-p_{0i}}}\right) = (p_{1i}) - (p_{0i})$ (intuitivement, $p_{1i} \geq p_{0i}$). Cette quantité mesure l'association entre les événements A_i et B_i, tels que $P(A_i) = p_{0i}$ et $P(B_i) = p_{1i}$.

Nous nous intéressons alors au log-cote ratio, c'est-à-dire $\log(\Psi) = \log \frac{\frac{p_{1i}}{1-p_{1i}}}{\frac{p_{0i}}{1-p_{0i}}} = \log(\frac{p_{1i}(1-p_{0i})}{p_{0i}(1-p_{1i})}) = (p_{1i}) - (p_{0i})$ (intuitivement, $p_{1i} \geq p_{0i}$). Cette quantité mesure l'association entre les événements A_i et B_i, tels que $P(A_i) = p_{0i}$ et $P(B_i) = p_{1i}$.

Nous posons donc :

$$(p_{0i}) = \log\left(\frac{p_{0i}}{1-p_{0i}}\right) = \mu_i \quad (p_{1i}) = \log\left(\frac{p_{1i}}{1-p_{1i}}\right) = \mu_i + \log(\Psi_i)$$

Le modèle utilisé est le suivant :

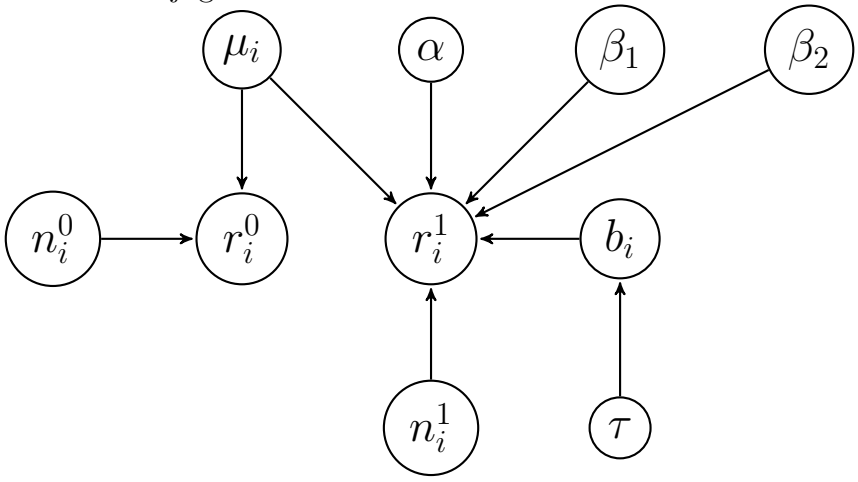
$$\log(\Psi_i) = \alpha + \beta_1 \text{year}_i + \beta_2(\text{year}_i^2 - 22) + b_i$$

La modélisation de Ψ_i est basée sur une régression logistique avec coefficients aléatoires pour étudier l'effet des variables explicatives (*year* et $year_2$) car on souhaite connaître l'influence de l'année de naissance des sujets sur leur mortalité. α est l'intercept et b représente les résidus du modèle. Les coefficients de ce dernier sont β_1 et β_2 qui représentent l'impact respectif de *year* et $year_2$.

Les lois a priori du modèle sont :

Paramètres	Lois a priori
α	$N(0, 10^6)$
β_1	$N(0, 10^6)$
β_2	$N(0, 10^6)$
τ	Gamma$(10^{-3}, 10^{-3})$
b_i	$N\left(0, \frac{1}{\tau}\right)$
μ_i	$N(0, 10^6)$

La loi a priori de τ est une Gamma car τ est strictement positif et en plus, la loi Gamma est conjuguée.



Les lois conditionnelles

— Loi conditionnelle de α :

$$\pi(\alpha \mid r_1, \mu, \beta_1\beta_2, b, \sigma^2) \propto \pi(\alpha) \cdot \prod_{i=1}^{N} \pi(r_i^1 \mid \alpha, \mu_i, \beta_1, \beta_2, b)$$

$$\pi(\alpha \mid r_1, \mu, \beta_1, \beta_2, b, \sigma^2) \propto \exp\left(-\frac{\alpha^2}{2 \cdot 10^6} \prod_{i=1}^{N} p_i^1 r_i^1 (1 - p_i^1) n_i^1 - r_i^1\right)$$

— De la même façon, on trouve les lois conditionnelles de β_1 et β_2 :

$$\pi(\beta_i \mid r_1, \mu, \alpha, \beta_j, b, \sigma^2) \propto \exp\left(-\frac{\beta_i^2}{2 \cdot 10^6} \prod_{i=1}^{N} p_i^1 r_i^1 (1 - p_i^1) n_i^1 - r_i^1\right)$$

— La loi conditionnelle de τ est une loi conjuguée, on trouve directement :

$$\tau \mid r_1, \mu, \alpha, \beta_1, \beta_2, b \sim \text{Gamma}\left(10^{-3} + \frac{N}{2}, 10^{-3} + \frac{1}{2}\sum_{i=1}^{N} b_i^2\right)$$

— Loi conditionnelle de μ_i, $\forall i \in \{1 : n\}$:

$$\pi(\mu_i \mid \cdots) \propto \pi(\mu_i)\pi(r_i^0 \mid \cdots)\pi(r_i^1 \mid \cdots)$$

$$\pi(\mu_i \mid \cdots) \propto \exp\left(-\frac{\mu_i^2}{2 \cdot 10^6} p_i^0 r_i^0 (1 - p_i^0) p_i^0 - r_i^0 p_i^1 r_i^1 (1 - p_i^1) n_i^1 - r_i^1\right)$$

— Loi conditionnelle de b_i, $\forall i \in \{1 : n\}$:

$$\pi(b_i \mid \cdots) \propto \pi(b_i \mid \sigma^2)\pi(r_i^1 \mid \cdots)$$

$$\pi(b_i \mid \cdots) \propto \exp\left(-\frac{b_i^2}{2 \cdot \sigma^2} p_i^0 r_i^0 (1 - p_i^0) p_i^0 - r_i^0 p_i^1 r_i^1 (1 - p_i^1) n_i^1 - r_i^1\right)$$

Analyse des résultats

Méthode utilisée :

Pour estimer les variables aléatoires listées précedemment, nous avons utilisé pour celà l'algorithme de Gibbs. Cet algorithme est une méthode d'échantillonnage de Monte Carlo par chaînes de Markov (MCMC) qui permet de générer des échantillons à partir de la distribution a posteriori des paramètres étant donné les données.Afin d'assurer la convergence de l'algorithme, nous avons utilisé une phase de "burn-in" de 1000 échantillons. Durant cette phase, les échantillons générés par l'algorithme sont écartés et non utilisés pour estimer les paramètres. Cette étape permet de réduire l'influence des conditions initiales sur les résultats finaux.

Résultats et discussion :

Nous obtenons , aprés échantillonnage, les chaînes de Markov et les densités des différents paramètres que l'on souhaite estimer (α, β_1, β_2 et σ^2) :

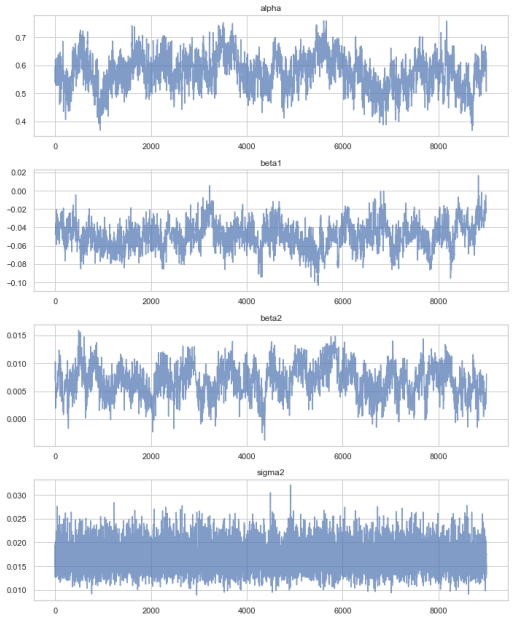


Figure 1 – Chaînes de Markov des paramètres estimés.

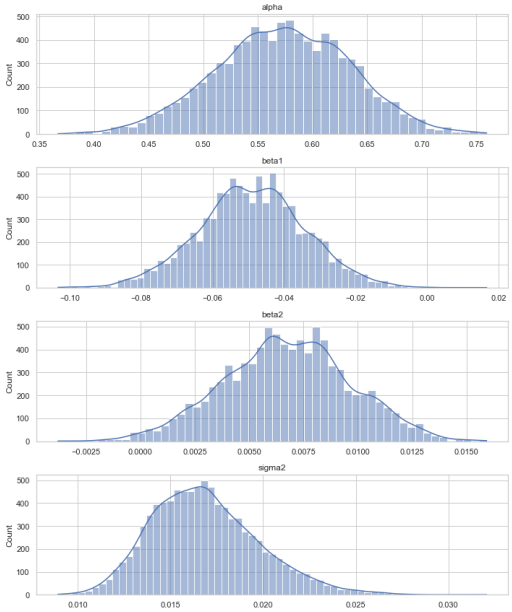


Figure 2 – Densité des paramètres estimés.

Nous pouvons donc comparer les moyennes et les ecarts-types des paramètres estimés avec ceux de l'enoncé :

	Moyenne		Écart-type	
	Estimation	Réference	Estimation	Réference
α	0.577929	0.579	0.060648	0.062
β_1	-0.046333	-0.045	0.014021	0.01553
β_2	0.006984	0.0070	0.003163	0.003084
σ	0,130088	0.09697	0.0538	0.06011

Table 1 – Résultats et énoncés des paramètres

En mettant en parallèle nos résultats avec les valeurs de référence, on peut voir que nos estimations se rapprochent de ce qu'on attendait. Ça nous montre que notre mise en place de l'algorithme de Gibbs fonctionne bien et que notre modèle est adapté pour analyser les données de cette recherche.

Il faut quand même préciser que, même si nos estimations se rapprochent des valeurs de référence, il y a toujours une incertitude qui vient avec l'utilisation de

l'algorithme de Gibbs pour estimer les paramètres. Cette incertitude se voit dans les intervalles de crédibilité, qui nous donnent une idée de l'éventail où les vrais paramètres pourraient se trouver avec une certaine probabilité.

Interpretation :

Nous avons ensuite tracé la courbe des log-Odds en fonction des années pour pouvoir interpreter les résultats de nos estimations , voici la courbe :

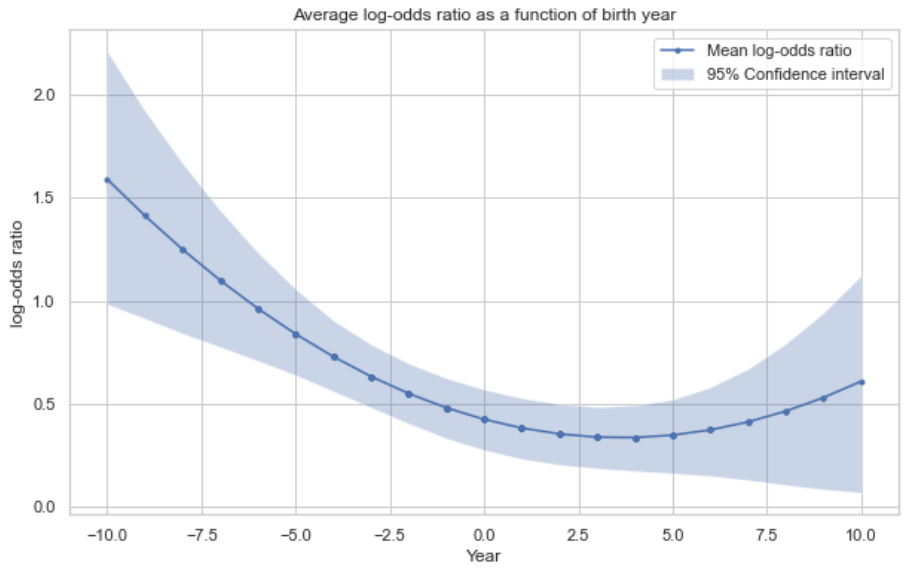


Figure 3 – Courbe des log-Odds ($log(\Psi)$) en fonction des années, avec une représentation de l'intervalle de crédibilité à 95%.

La courbe obtenue montre l'évolution moyenne du log-odds ratio en fonction de l'année de naissance. La courbe montre une décroissance du $\log(\Psi)$ avec l'année de naissance, ce qui indique que le rapport des cotes d'exposition aux rayons X chez les enfants atteints de cancer par rapport aux témoins diminue au fil du temps. Cette tendance pourrait suggérer une réduction du risque associé à l'exposition aux rayons X pour les enfants nés plus récemment, possiblement en raison de changements dans les pratiques médicales ou de sensibilisation aux risques.

Cette découverte est importante car elle suggère que les efforts visant à réduire l'exposition des femmes enceintes et des enfants aux rayons X ont peut-être été couronnés de succès. Cela pourrait indiquer une réduction du risque associé à l'exposition aux rayons X pour les enfants nés plus récemment, possiblement en raison de changements dans les pratiques médicales ou de sensibilisation aux risques. Cette tendance positive est encourageante et souligne l'importance de continuer à travailler à la réduction de l'exposition inutile aux rayons X dans les établissements médicaux. Cependant , nous constatons qu'il y'a un commencement de tendance croissante à

partir de l'année 1960 , mais malheureusement nous ne pouvons pas en conclure grand chose par manque de données courante , il faut donc récolter plus de données récentes pour pouvoir valider les résultats.