Régression bayésienne sur profils d'exposition : application en épidémiologie des rayonnements ionisants

Marion Belloni [1] & Sophie Ancelet [2] & Chantal Guihenneuc [3]

[1] *Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PSE-SANTE/SESANE/LEPID, Fontenay-Aux-Roses, 92262, France, marion.belloni@irsn.fr*

[2] *Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PSE-SANTE/SESANE/LEPID, Fontenay-Aux-Roses, 92262, France, sophie.ancelet@irsn.fr*

[3] *EA 7537 , Faculté de Pharmacie de Paris, Université Paris Descartes, 4 avenue de l'Observatoire 75006 Paris, chantal.guihenneuc@parisdescartes.fr*

Résumé. En épidémiologie des rayonnements ionisants, les effets sanitaires des expositions professionnelles sont souvent étudiés séparemment pour chaque source de rayonnement. Or, les travailleurs sont exposés simultanément à plusieurs sources de rayonnements ionisants qui sont corrélées entre elles et également à certains agents chimiques et physiques. On s'intéresse dans ce travail au risque de décès par cancer du poumon dans la cohorte française des mineurs d'uranium. Ces mineurs sont exposés au radon, aux rayonnements gamma et aux poussières d'uranium, ainsi qu'à d'autres agents chimiques. Une méthode adaptée à la corrélation entre ces expositions est ici proposée. Si cette corrélation n'est pas prise en compte dans la régression multiple, on obtient des estimateurs instables, et donc non interprétables. On présente ici une approche hiérarchique bayésienne appelée "régression bayésienne sur profils d'exposition" qui permet de traiter ce problème avec des variables explicatives continues et catégorielles. Cela consiste à regrouper les individus ayant des profils similaires, c'est-à-dire des caractéristiques proches, et d'estimer le risque associé à chaque groupe ainsi constitué. La répartition en groupes et l'estimation de risque se font conjointement sous le paradigme bayésien. L'inférence bayésienne du modèle proposé a été implémentée sous Python via un algorithme de type MCMC. Après un post-traitement des chaînes de Markov obtenues, l'identification et la caractérisation des groupes de mineurs d'uranium ayant des profils à haut risque et à bas risque de décès par cancer du poumon sont faites.

Mots-clés. Régression sur profils, Multicolinéarité, Inférence bayésienne, Cancer du poumon, Rayonnements ionisants

Abstract. In radiation epidemiology, the health effects of occupational exposures are often studied separatly for each source of radiation. But, workers are exposed to several sources of correlated exposures such as ionizing radiation, chemical and physical agents. In this work we focused on the risk of death by lung cancer in the French cohort

of uranium miners. These miners are exposed to radon, gamma rays and uranium dust, as well as other chemical agents. We propose a suitable method taking into account of exposures correlation. Indeed, if multicollinearity is not taken into account in multiple regression, unstable and therefore uninterpretable estimators are obtained. We present here a Bayesian hierarchical approach called the "Bayesian Profile Regression" which makes possible to treat the multicollinearity issue with continuous and categorical explanatory variables. This involves clustering individuals with similar profiles, that is, with close exposure characteristics, and estimating the associated risk for each group. Clustering and risk estimation are done jointly under the Bayesian paradigm. The Bayesian inference of the proposed model was implemented using Python via an MCMC type algorithm. After post-processing of the obtained Markov chains, groups of uranium miners with high-risk and low-risk of death by lung cancer profiles can be identified and characterized.

Keywords. Profile regression, Multicolinearity, Bayesian inference, Lung cancer, Ionizing radiation

1 Introduction

Dans le cas d'une régression multiple, la multicolinéarité est un problème fréquent qui survient lorsque plusieurs variables explicatives du modèle sont corrélées et mesurent une grande partie du phénomène modélisé. Les conséquences d'une multicolinéarité trop prononcée sont importantes. Tout d'abord, les variances des estimateurs des associations sont plus élevées. Malgré l'existence d'une réelle relation entre un prédicteur et une variable réponse, on peut ne pas détecter ce lien et conclure à une association non significative. De plus, les estimateurs deviennent instables, ce qui signifie qu'ils peuvent varier en fonction du sous-échantillon sur lequel les coefficients sont estimés mais également en cas de retrait du modèle d'au moins une des variables corrélées. Cette instabilité peut amener à un estimateur négatif (respectivement positif) dans le cas d'une association positive (respectivement négative) entre le prédicteur et la variable réponse. Ainsi, les coefficients estimés ne peuvent pas être interprétés de façon fiable en situation de multicolinéarité.

On adapte l'approche dite de régression bayésienne sur profils d'exposition présentée par Molitor *et al.* (2010). C'est un modèle hiérarchique composé de deux sous-modèles qui sont estimés conjointement. Cette approche hiérarchique est à deux niveaux qui sont estimés conjointement. Le premier sous-modèle décrit la distribution de probabilité des variables explicatives au sein d'un groupe d'individus ayant un profil d'exposition et un risque sanitaire proches. Les variables de classe sont inconnues (classification non supervisée). Le second sous-modèle décrit l'association entre le risque sanitaire d'intérêt et les groupes d'individus ayant des profils d'exposition similaires. Les résultats d'une telle approche permettent l'estimation du risque sanitaire associé à chaque groupe, l'identification des groupes à haut risque et à bas risque et enfin la caractérisation des groupes identifiés. La détermination de profils dits à haut risque et à bas risque est alors obtenue.

2 Cas d'étude

En épidémiologie des rayonnements ionisants, les effets sanitaires associés à des expositions radiologiques multiples sont souvent étudiés séparément, en modélisant l'association entre le risque sanitaire et une exposition unique. Ainsi, l'effet conjoint associé à des expositions multiples est peu connu. Les normes de radioprotection sont ainsi principalement fondées sur des analyses de risque basées sur un cadre d'exposition mono-factorielle. Il est néanmoins légitime de se demander dans quelle mesure l'estimation d'un risque faible en situation d'exposition unique à faibles doses aux rayonnements ionisants reste faible en situation de co-expositions à de multiples sources de rayonnements ionisants (voire à de possibles expositions chimiques, etc). Ces expositions multiples sont souvent corrélées car associées à un même scénario d'exposition (lieu, période, type de travail, etc).

Ainsi, dans le cadre de leur activité professionnelle, les mineurs d'uranium sont, de manière chronique, soumis à des expositions multiples qui sont potentiellement cancérigènes. Ils sont notamment exposés à trois sources de rayonnements ionisants à faibles doses: 1) le radon et ses descendants à vie courte par inhalation (contamination interne), 2) les poussières d'uranium par inhalation (contamination interne) 3) et les rayonnements gamma (exposition externe). Dans ce travail, nous nous interessons à la cohorte post-55 des mineurs d'uranium français. Elle comprend les 3377 mineurs d'uranium embauchés après le 31 décembre 1955. Ils ont été suivi en moyenne pendant 33 ans et 94 cas de décès par cancer du poumon ont été observés. Radon, poussières d'uranium et rayonnements gamma ont déjà été associées, séparément, à un risque significativement plus élevé de décès par cancer du poumon chez les mineurs d'uranium français (Rage (2015)).

Comme souligné par Vacquier *et al.* (2011), les données d'exposition aux rayonnements ionisants sont corrélées dans la cohorte post-55 car associées à un même phénomène qu'est la désintégration de l'uranium 238 présent dans la croûte terrestre. De précédentes analyses ont montré qu'on est bien dans une situation de multicolinéarité.

3 Méthode

3.1 Modèle

Le modèle de régression bayésienne sur profils d'exposition est composé de deux sous-modèles : le sous-modèle de maladie (association entre la maladie et les groupes et distribution de probabilité des mesures d'exposition au sein de chaque groupe) et le sous-modèle de répartition (attribution d'un mineur à un groupe).

Le sous-modèle de maladie est un modèle en excès de risque instantané (EHR par la suite) classiquement utilisé en épidémiologie des rayonnements ionisants. Le risque instantané de décès par cancer du poumon du mineur i au temps t, noté $h_i(t)$ est défini par $h_i(t) = h_0(t) \cdot (1 + \beta_{Z_i})$ avec $h_0(t)$ le risque instantané de base au temps t c'est-à-dire le risque dans le groupe de référence des mineurs non exposés aux rayonnements ionisants.

$h_0(t)$ est supposé être une fonction constante par morceaux définie par quatre paliers risque $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ correspondant chacun à une classe d'âge dans laquelle le risque de base est supposé constant. Z_i le label de groupe auquel appartient le mineur i et β_c l'excès de risque instantané de décès par cancer du poumon du groupe c. Ainsi, deux mineurs appartenant au même groupe c ont le même risque de décès par cancer du poumon.

Le sous-modèle d'attribution associe un mineur d'uranium i à un groupe Z_i et définit la distribution de probabilité des variables d'exposition conditionnellement à un groupe. Les différentes variables d'exposition considérées dans cette étude pour la caractérisation des groupes et profils d'exposition proches sont les suivantes :

- les mesures d'expositions professionnelles au radon X_i^R, aux rayonnements gamma X_i^G et aux poussières d'uranium X_i^P cumulées au cours de la période de suivi du mineur i (variables continues);
- le "job type" J_i du mineur i. Cette variable est un proxy pour les conditions d'exposition ainsi que d'éventuelles autres expositions professionnelles. C'est une variable catégorielle à cinq modalités: 1) foreurs avant mécanisation 2) foreurs après mécanisation 3) autre travail souterrain avant mécanisation 4) autre travail souterrain après mécanisation 5) travail en surface. La distinction entre avant et après mécanisation est importante à faire car l'arrivée des machines a permis de passer d'un travail manuel à un travail mécanisé (donc diminution de la pénibilité) mais dans le même temps, d'introduire une exposition supplémentaire au diesel;
- l'âge à la première exposition A_i du mineur i (variable continue);
- la localisation de la mine M_i. On distingue ici les mineurs ayant travaillé principalement dans l'Hérault par rapport aux autres mines. Cette distinction permet de différencier les mines en fonction du type de sol: sédimentaire pour la mine de l'Hérault et granitique pour les autres;
- le temps d'exposition T_i du mineur i, variable catégorielle dont les catégories ont été choisies afin d'avoir un nombre similaire de mineurs dans chacune d'entre elles. Ainsi, on distingue les mineurs qui ont été exposés 5 ans et moins, de 6 à 12 ans, de 13 à 18 ans et enfin ceux qui ont été exposés pendant au moins 19 années.

Chaque covariable suit une loi de probabilité dont les paramètres dépendent du groupe c. On a supposé des distributions lognormales pour les variables continues positives et des distributions multinomiales pour les variables catégorielles.

$$\begin{cases} X_i^R | Z_i = c, \mu_c^R, \sigma_c^R & \sim & LogN(\mu_c^R, \sigma_c^R) \\ X_i^G | Z_i = c, \mu_c^G, \sigma_c^G & \sim & LogN(\mu_c^G, \sigma_c^G) \\ X_i^P | Z_i = c, \mu_c^P, \sigma_c^P & \sim & LogN(\mu_c^P, \sigma_c^P) \\ A_i | Z_i = c, \mu_c^A, \sigma_c^A & \sim & LogN(\mu_c^A, \sigma_c^A) \\ J_i | Z_i = c, \boldsymbol{p_c^J} & \sim & Multinomiale(\boldsymbol{p_c^J}) \\ M_i | Z_i = c, \boldsymbol{p_c^M} & \sim & Multinomiale(\boldsymbol{p_c^M}) \\ T_i | Z_i = c, \boldsymbol{p_c^T} & \sim & Multinomiale(\boldsymbol{p_c^T}) \end{cases}$$

Le modèle d'attribution d'un mineur à un groupe dépend de la probabilité ϕ_c d'appartenir au groupe c. Soit C le nombre maximal de groupes de mineurs, $\boldsymbol{\phi} = (\phi_1, \phi_2, ..., \phi_C)$

définit le vecteur des probabilités d'appartenance des mineurs à chacun des C groupes. Le vecteur de paramètres $\boldsymbol{\phi}$ suit un processus de Dirichlet. Grâce au processus de Dirichlet, le nombre de groupes n'est pas fixé arbitrairement, il est donc également estimé. Seul un nombre maximal C de groupes doit être fixé. La construction de ces poids de mélange $\boldsymbol{\phi} = (\phi_c, c = 1, ..., C)$, avec C le nombre maximal de clusters, est la suivante (appelée "stick-breaking"). On définit $C-1$ variables aléatoires $V_1, ..., V_{C-1}$ telles que $V_c \sim Beta(1, \alpha)$ et, par relation déterministe, on obtient les poids d'attribution à chaque groupe: $\phi_c = V_c \cdot (1 - \sum_{k=1}^{c-1} \phi_k)$ pour $c = 1, ..., C-1$ et $\phi_C = 1 - \sum_{k=1}^{C-1} \phi_k$. Le nombre réel de groupes (c'est-à-dire non vides) est guidé par le paramètre α. Une petite valeur de α correspond un petit nombre réel de groupes.

3.2 Choix des lois *a priori* et inférence bayésienne

Des lois *a priori* normales centrées avec une large variance (10^6) ont été considérées pour les paramètres de risque $\beta_c, c = 1, ..., C$ et pour les moyennes géométriques de l'âge à la première exposition μ_c^A pour chaque groupe c. Des données externes sur la mortalité par cancer du poumon chez les hommes en France entre 1968 et 2005 ont été utilisées pour spécifier les lois *a priori* gamma informatives sur les paramètres λ_1, λ_2, λ_3 et λ_4 définissant le risque instantané de base de mortalité par cancer du poumon chez les mineurs d'uranium français (supposé constant par classe d'âge). Des lois uniformes sur l'intervalle [0,100] ont été considérées pour les paramètres d'écart-type géométrique des distributions lognormale σ_c^R, σ_c^G, σ_c^P et σ_c^A. Les lois *a priori* considérées pour les paramètres de moyenne géométrique des expositions aux rayonnements gamma μ_c^G, au radon μ_c^R et aux poussières d'uranium μ_c^P sont des lois normales dont les paramètres sont basés sur les niveaux d'exposition de la cohorte allemande des mineurs d'uranium. On considère des lois *a priori* de Dirichlet de paramètres égaux à 1/2 pour les paramètres $\boldsymbol{p_c^J}$, $\boldsymbol{p_c^M}$ et $\boldsymbol{p_c^T}$. Finalement, comme recommandé par Molitor *et al.* (2010), on utilise une loi uniforme sur l'intervalle [0.3,10] pour le paramètre α qui influence le nombre réel de groupes *a posteriori*.

Un package R existe pour traiter la régression sur profils d'exposition (PReMiuM, Liverani (2015)), mais le modèle de survie en EHR n'est pas une option possible pour le modèle de maladie sous PReMiuM. Ainsi le modèle hiérarchique bayésien a été inféré à l'aide d'un algorithme Monte-Carlo par Chaînes de Markov (MCMC) de type Metropolis-Hastings within Gibbs codé avec le langage Python. Les chaînes de Markov obtenues nécessitent un traitement supplémentaire afin de trouver la meilleure partition d'individus.

3.3 Post-traitement

Le post-traitement est fait après la réalisation de l'algorithme MCMC du modèle précédemment décrit. La première étape consiste à construire une matrice carrée S_k de dimension $n \times n$ (n le nombre de mineurs) à chaque itération k de l'algorithme. L'élément en position (i, j)

de la matrice S_k vaut 1 si les individus i et j appartiennent au même groupe à l'itération k, et 0 sinon. La moyenne des matrices S_1,...,S_K permet d'obtenir la matrice S qui contient les probabilités que deux mineurs appartiennent au même cluster. La meilleure partition $\boldsymbol{z}^{best}$ est celle qui minimise la distance des moindres carrés fondée sur S.

Les lois *a posteriori* des paramètres sont alors déduites de cette meilleure partition $\boldsymbol{z}^{best}$. Ainsi, un échantillon de la distribution a posteriori du paramètre θ du groupe c est obtenu à chaque itération k par $\bar{\theta}_{c,k} = \frac{1}{n_c} \sum_{i:z_i^{best}=c} \theta_{z_i^k,k}$ avec n_c le nombre de mineurs du groupe c, z_i^k le groupe auquel appartient le mineur i à l'itération k

4 Résultats

Après avoir ajusté le modèle aux données de la cohorte post-55 des mineurs d'uranium français et effectué le post-traitement, il ressort principalement huit profils de mineurs que l'on peut caractériser et dont le risque a été estimé. Les groupes sont caractérisés par le niveau d'exposition au radon, aux rayonnements gamma, aux poussières d'uranium, le "job type", la localisation de la mine dans laquelle travaillent les mineurs, l'âge à la première exposition et la durée d'exposition. La caractérisation des profils permet de mettre en évidence que le risque ne dépend pas uniquement des niveaux d'exposition aux trois sources de rayonnements ionisants, mais aussi de l'âge à la première exposition ainsi que de la localisation de la mine.

Bibliographie

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M., & Richardson, S. (2015). PReMiuM: An R package for profile regression mixture models using Dirichlet processes. Journal of statistical software, 64(7), 1.

Molitor, J., Papathomas, M., Jerrett, M., & Richardson, S. (2010). Bayesian profile regression with an application to the National Survey of Children's Health. Biostatistics, 11(3), 484-498.

Rage, E., Caër-Lorho, S., Drubay, D., Ancelet, S., Laroche, P., & Laurier, D. (2015). Mortality analyses in the updated French cohort of uranium miners (1946–2007). International archives of occupational and environmental health, 88(6), 717-730.

Vacquier, B., Rage, E., Leuraud, K., Caër-Lorho, S., Houot, J., Acker, A., & Laurier, D. (2011). The influence of multiple types of occupational exposure to radon, gamma rays and long-lived radionuclides on mortality risk in the French "post-55" sub-cohort of uranium miners: 1956–1999. Radiation research, 176(6), 796-806