

Splines multidimensionnelles pénalisées pour modéliser le taux de survenue d'un événement. Application au taux de mortalité en excès et à la survie nette en épidémiologie des maladies chroniques

Mathieu Fauvernier¹⁻², Laurent Roche¹⁻², Zoé Uhry^{3,1-2}, Nadine Bossard¹⁻², Laurent Remontet¹⁻²

¹ Hospices Civils de Lyon, Pôle Santé Publique, Service de Biostatistique - Bioinformatique, Lyon, France.

² Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Équipe Biostatistique-Santé, Villeurbanne, France.

³ Département des Maladies Non-Transmissibles et des Traumatismes, Santé Publique France, Saint-Maurice, France.

Résumé

L'étude du temps de survenue d'un événement représente un champ très important des statistiques. Lorsque l'événement étudié est le décès, on cherche à décrire la survie des individus ainsi que leur taux de mortalité, c'est-à-dire la « force de mortalité » qui s'applique à un instant donné.

Les patients atteints d'une maladie chronique présentent en général un excès de mortalité par rapport à une population ne présentant pas la maladie en question. En épidémiologie, l'étude du taux de mortalité en excès des patients, et notamment de l'impact des facteurs pronostiques sur celui-ci, représente donc un enjeu majeur de santé publique.

D'un point de vue statistique, la modélisation du taux de mortalité (en excès) implique de prendre en compte les effets potentiellement non-linéaires et dépendants du temps des facteurs pronostiques ainsi que les interactions.

Les splines de régression, polynômes par morceaux paramétriques et flexibles, sont des outils particulièrement bien adaptés pour modéliser des effets d'une telle complexité.

Toutefois, la flexibilité des splines de régression comporte un risque de sur-ajustement. Pour éviter ce risque, les splines de régression pénalisées ont été proposées dans le cadre des modèles additifs généralisés (GAM). Leur principe est le suivant : à chaque spline peuvent être associés un ou plusieurs termes de pénalité contrôlés par des paramètres de lissage. Les paramètres de lissage représentent les degrés de pénalisation souhaités. En pratique, ils sont inconnus et doivent être estimés tout comme les paramètres de régression.

Dans le cadre de ce travail, nous avons développé une méthode permettant de modéliser le taux de mortalité (en excès) à l'aide de splines de régression multidimensionnelles pénalisées. Des splines cubiques restreintes ont été utilisées comme splines unidimensionnelles ou bien comme bases marginales afin de former des splines multidimensionnelles par produits tensoriels. Le processus d'optimisation s'appuie sur deux algorithmes de Newton-Raphson emboîtés. L'estimation des paramètres de lissage est effectuée en optimisant un critère de validation croisée (LCV) ou bien la vraisemblance marginale des paramètres de lissage (LAML) par un algorithme de Newton-Raphson dit externe. A paramètres de lissage fixés, les paramètres de régression sont estimés par maximisation de la vraisemblance pénalisée par un algorithme de Newton-Raphson dit interne.

Les bonnes propriétés de cette approche en termes de performances statistiques et de stabilité numérique ont ensuite été démontrées par simulation.

La méthode a été implémentée au sein d'un package R *survPen*.

Enfin, la méthode a été appliquée sur des données réelles pour répondre à deux questions épidémiologiques. L'une concernait l'impact de la défavorisation sociale sur la mortalité en excès des patients atteints d'un cancer du col de l'utérus. L'autre proposait d'étudier l'impact de l'âge courant sur la mortalité en excès des patients atteints de sclérose en plaques.

Mots clés : Splines pénalisées, survie, taux de mortalité, taux en excès, survie nette, épidémiologie, maladies chroniques