Stage de M2 Biostatistiques / Epidémiologie

Les méthodes de sélection en grande dimension pour les études épidémiologiques cas-témoins

Tandis qu'une littérature abondante a été consacrée aux développements de méthodes de sélection de biomarqueurs à partir de données de grande dimension depuis la méthode Lasso proposée par Tibshirani [1] en 1996 avec des applications aussi bien sur des données d'essais cliniques ou de cohortes, les développements méthodologiques et les applications dans le cadre d'études épidémiologiques de types cas-témoins semblent, après une brève revue de la littérature, beaucoup plus rares. Ces études cas-témoins sont mises en place pour étudier l'association entre un facteur d'exposition et la survenue d'un événement rare. La particularité de l'analyse des données d'une étude cas-témoins est l'appariement d'un cas (i.e. patient avec événement d'intérêt) avec 1 ou plusieurs témoins (i.e. patients sans évènement d'intérêt) et la méthode statistique classique pour analyser ce type de données est la régression logistique conditionnelle [2]. Bien souvent, les méthodes de sélection en grande dimension pour une étude cas-témoins se limitent à des analyses univariées ou aux méthodes de sélection automatiques (backward, forward ou encore stepwise). L'approche Lasso adaptée à la régression logistique conditionnelle a été développée par Reid *et al.* en 2014 [3]. Récemment, des approches non paramétriques basées sur l'utilisation d'arbres de décision ont également été proposées [4, 5, 6]. De par leur nature non-paramétrique, ces dernières approches ne posent aucune hypothèse sur la forme de la relation entre les facteurs d'exposition et l'événement d'intérêt. De plus, ces méthodes non paramétriques reposent sur leur capacité à modéliser des relations complexes comme des interactions. Ainsi, l'utilisation de telles méthodes dans le cadre de l'analyse de données cas-témoins en grande dimension semble pertinente. L'objectif de ce stage est d'approfondir cette revue de la littérature et de comparer les performances des différentes méthodes existantes à travers une large étude de simulation. Les méthodes seront ensuite appliquées à une étude cas-témoin nichée dans la cohorte French Childhood Survivor Study (FCCSS) dont l'objectif était d'identifier des biomarqueurs de l'expression génétique associés aux pathologies cardiaques après traitement d'un cancer dans l'enfance. L'identification de biomarqueurs prédictifs de l'effet des traitements (recherche d'interaction) sur le développement des pathologies cardiaques sera aussi évaluée.

Mots clefs : étude cas-témoin appariée, procédure de sélection, grande dimension, méthodes basées sur les arbres de décision, régression logistique conditionnelle

Profil recherché

Ce stage s'adresse à un(e) étudiant(e) de Master 2 dans l'un de ces domaines : science des données, statistique appliquée, biostatistique / épidémiologie. Une maîtrise de la programmation en R est requise. Une connaissance des données de santé et des méthodes avancées en apprentissage statistique sera un plus.

Environnement de travail

Le travail sera réalisé dans le service de Biostatistiques et d'Epidémiologie de Gustave Roussy (Villejuif, Paris) sous la supervision de Gwénaël Le Teuff (Gustave Roussy), Audrey Poterie (Laboratoire de Mathématiques Bretagne Atlantique, Université Bretagne Sud, Vannes) et Nadia Haddy (ANSM).

La durée du stage envisagée est de 6 mois, avec une date de début comprise entre février et avril 2025 suivant la disponibilité de l'étudiant(e).

Contact

Les candidat(e)s intéressé(e)s doivent postuler en envoyant un CV et une lettre de motivation à :

Gwenael.leteuff@gustaveroussy.fr et Audrey.Poterie@univ-ubs.fr

Références :

[1] R Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B, vol. 58, no1,1996, p.267-288.

[2] NE Breslow, NE Day, KT Halvorsen, RL Prentice, C Sabai. Estimation of multiple relative risk functions in matched case-control studies. Am J Epidemiol. 1978 Oct;108(4):299-307

[3] Reid S, Tibshirani R: Regularization Paths for Conditional Logistic Regression: The clogitL1 295 Package. J Stat Softw 58:12, 2014

[4] NS Zadeh, S Lin, GC Runger. Matched Forest: Supervised Learning for High-Dimensional Matched Case–Control Studies. 2020. Bioinformatics 36(5):1570-76

[5] Schauberger G, Tanaka LF, Berger M. A tree-based modeling approach for matched case-control studies. Stat in Med, 2023, 42 :676-692

[6] Shi, Haolun, et Guosheng Yin. 2018. « Boosting conditional logit model ». Journal of Choice Modelling 26:48-63.