Recensement des sujets de stage à Santé Publique France pour des étudiants de Master – année universitaire 2017-2018

- **Stage proposé par**

Direction/ Cire : DATA

Maîtres de stage / personne contact :

Nom : Pelat Prénom : Camille

Téléphone : +33 (0) 1 41 79 68 12 Adresse email : camille.pelat@santepubliquefrance.fr

Co-encadrant : Edouard Chatignoux, (edouard.chatignoux@santepubliquefrance.fr)

- **Type de stage proposé**

☐ Master 1 ☐ Master 1 + sujet tutoré* ☐ Master 1 « observation »**

☒ Master 2 Professionnel ☒ Master 2 Recherche

Extension possible au delà de la période obligatoire ☐ Oui ☒ Non

* stage M1 + sujet tutoré (proposé par l'ISPED) : le stage complète un travail préalable de sujet tutoré de 2 mois et demi (revue de la littérature, analyse d'une base de données, etc.)
** stage M1 « observation » (proposé par l'ISPED) : le stagiaire contribue aux activités courantes de l'équipe sans être en charge d'un dossier particulier

Commentaires : Le stage s'adresse aux élèves ingénieurs ou de Master2, spécialisé en mathématiques et statistiques

- **Date proposée pour le stage et durée**

☒ pas de contrainte de date

☐ A partir de _____

Durée minimum en mois ___6 mois_____

- **Sujet proposé pour le stage**

Titre : Interprétation des modèles de régression généralisés multiniveaux.

Lieu : Santé publique France est un établissement national de santé publique chargé, entre autres, de surveiller l'état de santé de la population française et son évolution. L'étudiant-e sera intégré-e à la Direction Appui, Traitements et Analyses des données (DATA), qui regroupe les activités de support méthodologique, notamment en statistiques, pour les directions thématiques.

Contexte : La recherche de facteurs associés à des pathologies ou des comportements de santé (adhésion au dépistage ou à la vaccination) fait souvent appel à des modèles de régression généralisés multi-niveaux (GLMM - (Bolker et al., 2009)).
Dans ces modèles, se côtoient des effets fixes liés aux variables explicatives individuelles (premier niveau), ou aux variables explicatives de niveaux plus élevés (c'est-à-dire associées à la zone de résidence ou à tout autre « groupe » auquel appartient la personne), et des effets aléatoires.
Exemples :
Variables individuelles : âge de la personne, fumeur/non-fumeur…
Variables de niveau 2 : indice de défavorisation de la commune, revenu médian de la commune…

Variables de niveau 3 : existence d'un programme de dépistage départemental…

Les effets aléatoires, mesurent les liens, inexpliqués, entre la réponse individuelle (ex. vacciné/non vacciné) et l'appartenance à une modalité d'un niveau (par exemple, la couverture vaccinale moyenne peut être différente par département, même après ajustement sur les variables individuelles et de groupe, ou bien la réponse à un même traitement peut être meilleure chez tous les patients d'un même hôpital, sans que cela puisse être associé à l'une des variables explicatives introduites dans le modèle).

En termes de santé publique, il est important de pouvoir évaluer la contribution ou « taille d'effet » de chacune de ces trois composantes du modèle, de façon à aider le décideur à prendre les mesures appropriées.
Par exemple, le fait de participer au dépistage du cancer du sein est-il plus associé aux caractéristiques individuelles ou à celles du département ? Est-ce que l'effet aléatoire « département » explique plus la variabilité de l'adhésion au dépistage que les variables explicatives ?
Dans le premier cas, le programme de prévention pourra cibler les personnes aux caractéristiques individuelles défavorablement associées au dépistage. Dans le deuxième cas, le programme de prévention pourrait cibler les départements présentant les caractéristiques les plus défavorables. Dans le dernier cas, une inégalité territoriale de santé non expliquée par les variables fournies au modèle ressort comme le facteur le plus associé au dépistage. On pourra alors déployer plus d'efforts de prévention dans les départements où l'adhésion au dépistage (ajustée sur les variables explicatives) est la moins bonne et, parallèlement, essayer de comprendre l'origine des inégalités territoriales mises en lumière.

De la même façon, il est important de pouvoir dire si le modèle, dans son ensemble, explique bien la variable réponse, et quelle est la part de variance qui reste inexpliquée.

Toutes ces questions trouvent une réponse dans le cadre du modèle de régression linéaire multiple : le R2 quantifie la part de variance expliquée par le modèle, et la taille d'effet de chaque variable peut être mesurée par la part de variance expliquée (ANOVA). Cependant, dans le cadre des modèles linéaires généralisés multiniveaux, il n'y a pas de réponse toute faite.

Sujet : L'objectif du stage sera de proposer une méthode pour quantifier la part de variance expliquée par le modèle, et la contribution de chaque variable explicative, celle des différents niveaux et celle associée aux effets aléatoires. Il faudra pouvoir quantifier ces contributions d'une manière robuste et compréhensible, qui permette la prise de décision en santé publique.
La méthode choisie pourra être implémentée sous forme de fonction ou package R.

Un travail de synthèse de la littérature sera nécessaire, ainsi que des comparaisons de méthodes sur données réelles et/ou simulées. L'étudiant-e pourra notamment s'appuyer sur les travaux de Nakagawa (Nakagawa & Schielzeth, 2013), Merlo (Merlo et al., 2006; Merlo, Chaix, Yang, Lynch, & Råstam, 2005) et Browne (Browne, Subramanian, Jones, & Goldstein, 2005).

Une application de la méthode choisie pourra être faite sur une parmi trois études en cours à Santé publique France : dépistage de différents cancers, adhésion à la vaccination contre le papillomavirus ou déterminants météorologiques associés à la survenue des cas de légionellose.

Ce travail sera réalisé sous R.

Références

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–35. doi:10.1016/j.tree.2008.10.008

Browne, W. J., Subramanian, S. V., Jones, K., & Goldstein, H. (2005). Variance partitioning in multilevel logistic models that exhibit overdispersion. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(3), 599–613. doi:10.1111/j.1467-985X.2004.00365.x

Merlo, J., Chaix, B., Ohlsson, H., Beckman, A., Johnell, K., Hjerpe, P., … Larsen, K. (2006). A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology and Community Health*, *60*(4), 290–7. doi:10.1136/jech.2004.029454

Merlo, J., Chaix, B., Yang, M., Lynch, J., & Råstam, L. (2005). A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon. *Journal of Epidemiology and Community Health*, *59*(6), 443–9. doi:10.1136/jech.2004.023473

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R 2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. doi:10.1111/j.2041-210x.2012.00261.x

- **Prérequis**

☐ Aucun

☒ Compétences spécifiques (préciser) : mathématiques et statistiques

☒ Maîtrise d'un logiciel spécifique (préciser) : logiciel R

☐ Autre (préciser) :

- **Commentaires**