Régression multibloc sur classes latentes. Application en épidémiologie vétérinaire.

Stéphanie Bougeard*[1], Claire Chavin[1], Gilbert Saporta[2], and Ndeye Niang[2]

[1]Anses, Epidemiology, Health and Welfare, Laboratory of Ploufragan-Plouzané-Niort – Epidemiology, Health and Welfare, Laboratory of Ploufragan-Plouzané-Niort – France
[2]CNAM Paris – Conservatoire National des Arts et Métiers (CNAM) – France

Résumé

L'analyse statistique des données d'épidémiologie vétérinaire vise à déterminer les facteurs de risque d'une maladie ou d'un problème de santé publique vétérinaire. Pour répondre à cet objectif, les modèles linéaires généralisés (e.g., régression logistique) sont généralement utilisés. Pour le cas où les observations proviennent de différentes sous-populations, sans que les affiliations à celles-ci soient connues, les modèles linéaires généralisés existent sous forme de modèles sur classes latentes, aussi connus sous le nom de modèles de mélange (Wedel et DeSarbo, 1995). Ces modèles linéaires généralisés locaux postulent l'existence de variables inobservables (i.e., les sous-populations inconnues d'observations) dont les effets sont mesurables (i.e., les liens entre la variable à expliquer et les variables explicatives diffèrent selon la sous-population). Cependant, en épidémiologie vétérinaire notamment, ces méthodes présentent deux principales limites : (i) le nombre d'observations dans une sous-population doit être plus grand que nombre de variables, (ii) et les variables doivent présenter une distribution multi-normale, ces deux hypothèses étant rarement vérifiées en pratique.
Nous proposons une extension des modèles de mélange pour le cas d'un grand nombre de variables ne vérifiant pas d'hypothèse distributionnelle. Ces variables peuvent présenter de plus la particularité d'être organisées en blocs thématiques. La méthode proposée est appelée régression multibloc sur classes latentes (Bougeard et al., 2017, 2018). Elle combine la recherche simultanée de sous-populations au sein des observations, ainsi que de modèles de régression (multibloc) locaux associés à chacune de ces sous-populations. Un test - basé sur la minimisation de l'erreur de prédiction - propose à l'utilisateur une partition optimale en sous-populations.

La régression multibloc sur classes latentes est appliquée - pour illustration - à la recherche de marqueurs de risque de l'utilisation des antibiotiques dans 113 élevages français de lapins. Les marqueurs de risques potentiels sont organisés en quatre blocs thématiques relatifs aux pratiques de gestion et d'hygiène (bloc 1, 8 variables), aux problèmes sanitaires (bloc 2, 7 variables), à la structure de l'exploitation (bloc 3, 5 variables) et aux pratiques thérapeutiques (bloc 4, 7 variables). La méthode propose une séparation optimale des observations en deux sous-populations (N1=52 élevages, N2=61 élevages). Le coefficient de détermination du modèle est nettement amélioré ; sa valeur est R2=0,25 pour le modèle comportant toutes les observations, et R12=0,56 (sous-population 1), R22=0,65 (sous-population 2) pour le modèle comportant deux sous-populations d'élevages. Quelle que soit la sous-population, les pratiques de gestion et d'hygiène (bloc 1) sont importantes. Cependant, la structure de

*Intervenant

l'exploitation (bloc 3) est importante pour les élevages de la sous-population 1, alors que ce sont les problèmes sanitaires (bloc 2) et les pratiques thérapeutiques (bloc 4) qui importent le plus pour les élevages de la sous-population 2.
Cette nouvelle méthode permet d'optimiser les modèles de régression (i.e., amélioration de l'explication et de la prédiction), leurs interprétations (i.e., les marqueurs de risque significatifs diffèrent selon les sous-populations) et ainsi de conduire à de nouvelles pistes de recherche et d'action.