Sélection de variables avec lasso dans la régression logistique conditionnelle

Marta Avalos[*,**]

[*]Equipe de Biostatistique, INSERM U897, 33076 Bordeaux Cedex
marta.avalos@isped.u-bordeaux2.fr,
http://biostat.isped.u-bordeaux2.fr
[**]Université de Bordeaux 2, 33076 Bordeaux Cedex

Résumé. Nous proposons une procédure de sélection de modèle dans le cadre des études cas-témoin appariées et, plus précisément, pour la régression logistique conditionnelle. La méthode se base sur une pénalisation de type L_1 des coefficients de régression dans la vraisemblance conditionnelle. Cette pénalisation, permettant d'éliminer de façon automatique les variables considérées non pertinentes, est particulièrement adaptée aux problèmes où le nombre de variables explicatives est élevé (par rapport au nombre d'événements) ou en cas de colinéarité. Des méthodes de rééchantillonnage sont appliquées pour le choix du terme de régularisation ainsi que pour l'évaluation de la stabilité du modèle sélectionné. La mise en œuvre de la méthode est illustrée par deux exemples.

1 Introduction

Les enquêtes cas-témoin cherchent à mettre en évidence le lien entre une pathologie et des facteurs de risque, en tenant compte des possibles facteurs de confusion. Une stratégie, permettant de contrôler certains facteurs de confusion potentiels, consiste à apparier chaque cas avec un nombre préfixé de témoins comparables, en termes d'exposition à ces facteurs. En cas d'appariement, les observations ne sont plus indépendantes, une méthode adaptée à l'analyse est alors la régression logistique conditionnelle. Le modèle considéré est le même que dans la régression logistique classique, en revanche, la vraisemblance à maximiser doit être conditionnelle au mode d'échantillonnage.

Comme pour toute technique de modélisation, la sélection des variables qui doivent figurer dans le modèle est une étape clé en régression logistique conditionnelle. Si le nombre d'événements n'est pas nettement supérieur au nombre de variables explicatives ou si celles-ci sont corrélées, alors la variance des estimations sera importante, aboutissant à des prédictions imprécises (Greenland, 2000; Bull et al., 2007). Ce problème est habituellement abordé en utilisant des méthodes de sélection de sous-ensembles, qui sont néanmoins, connues pour son instabilité (Breiman, 1996; Greenland, 2008). Une approche différente, connue sous le nom de lasso (*least absolute shrinkage and selection operator*), consiste à maximiser la vraisemblance pénalisée par la norme L_1 des coefficients de régression (Tibshirani, 1996). Les coefficients sont alors rétrécis vers 0, certains d'entre eux étant annulés exactement. Le fait que certains coefficients de régression soient réduits à zéro a par conséquence, d'une part, la simultanéité