Sélection de variables avec lasso dans la régression logistique conditionnelle

Marta Avalos

Equipe de Biostatistique de l'unité INSERM U897
Université de Bordeaux 2
146, rue Léo Saignat 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 entre celles-ci. 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 un exemple avec des données simulées.
Mots clés : choix de modèle, apprentissage statistique, régularisation L_1, vraisemblance conditionnelle pénalisée, études cas-témoin appariées.

Abstract

We propose a model selection procedure in the context of matched case-control studies and, more specifically, for the conditional logistic regression. The method is based on penalized conditional likelihood with an L_1-type penalty of the regression coefficients. This penalty, that automatically eliminates irrelevant covariates, is particularly adapted when the number of covariates is large (with respect to the number of events) or in case of collinearity between them. Resampling methods are applied for choosing the regularization term and for evaluating the stability of the selected model. The implementation of the method is illustrated by an example using simulated data.
Key words : model selection, statistical learning, L_1-regularization, penalized conditional likelihood, matched case-control studies.

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 (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). D'une part, les coefficients sont rétrécis vers 0 : l'introduction d'un biais entraîne une réduction de la variance. D'autre part, certains coefficients sont annulés exactement, par conséquent, l'estimation et la sélection de variables sont effectuées simultanément.

Nous étudions la pénalisation L_1 dans le cadre de la régression logistique conditionnelle. Les coefficients sont calculés par l'adaptation de l'algorithme proposé par Goeman (2008) pour le modèle de Cox pénalisé. Le terme de régularisation est sélectionné par une validation croisée qui tient compte de la nature dépendante des données. La stabilité des résultats est mesurée par des intervalles bootstrap, afin de prévenir des possibles conclusions erronées d'une modélisation automatique. La mise en œuvre de la méthode est illustrée par un exemple avec des données simulées.

Modélisation et Estimation

Nous nous intéressons à la relation entre une variable réponse Y, binaire (codée 0–1), et plusieurs variables explicatives $X = (X_1, ..., X_p)$. Les observations sont des groupes d'individus (strates), constitués d'un cas ($Y = 1$) et M témoins ($Y = 0$), chacun d'entre eux ayant une valeur de X : pour l'individu i de la strate k, nous avons le vecteur d'observations $\mathbf{x}_{ik} = (x_{ik1}, \ldots, x_{ikp})$, $i = 1, ..., M + 1$, $k = 1, ..., K$. Soit P_{ik} la probabilité (non conditionnelle) de survenue de l'événement pour le sujet i de la strate k. Considérons le modèle logistique, supposant que le risque varie d'un groupe à un autre :

$$\text{logit}[P(Y_{ik} = 1|\mathbf{x}_{ik})] = \log \frac{P(Y_{ik} = 1|\mathbf{x}_{ik})}{1 - P(Y_{ik} = 1|\mathbf{x}_{ik})} = \alpha_0 + \sum_{l=1}^{K} \alpha_l 1_l + \beta_1 x_{ik1} + \ldots + \beta_p x_{ikp} \tag{1}$$

où $\mathbf{1}_l$ est une fonction indicatrice qui vaut 1 si l'individu appartient à la strate l et 0 autrement ; les α_k sont les coefficients représentant l'effet des variables d'appariement sur la réponse, qui traduisent les différences entre les strates ; et les coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$

représentent les effets des variables explicatives ou, de façon équivalente, le log–rapport de cotes. Supposons que les différences entre les strates ne sont pas d'intérêt. On va donc se limiter à l'estimation de $\boldsymbol{\beta}$.

Considérons la strate k, la probabilité non conditionnelle d'observer la survenue de l'événement seulement chez l'individu i est :

$$(1-P_{1k})\dots(1-P_{i-1k})P_{ik}(1-P_{i+1k})\dots(1-P_{M+1k}) = \frac{P_{ik}}{1-P_{ik}}\prod_{j=1}^{M+1}(1-P_{jk}). \qquad (2)$$

La probabilité conditionnelle au mode d'échantillonnage (chaque strate a été constituée par 1 cas et M témoins), sous le modèle logistique, est donnée par :

$$\frac{\frac{P_{ik}}{1-P_{ik}}\prod_{j=1}^{M+1}(1-P_{jk})}{\sum_{j=1}^{M+1}\frac{P_{jk}}{1-P_{jk}}\prod_{j=1}^{M+1}(1-P_{jk})} = \frac{\frac{P_{ik}}{1-P_{ik}}}{\sum_{j=1}^{M+1}\frac{P_{jk}}{1-P_{jk}}} = \frac{\exp(\beta_1 x_{ik1}+\ldots+\beta_p x_{ikp})}{\sum_{j=1}^{M+1}\exp(\beta_1 x_{jk1}+\ldots+\beta_p x_{jkp})}, \qquad (3)$$

et la fonction de log-vraisemblance conditionnelle, évaluée en $\boldsymbol{\beta}$ (la vraie valeur des coefficients) et $D = \{(\mathbf{x}_{ik}, y_{ik})\}_{i=1,\dots,M+1;k=1,\dots,K}$ s'écrit :

$$l(\boldsymbol{\beta}, D) = \sum_{\substack{i=1,\dots,M+1\\k=1,\dots,K}}\Big[\mathbf{x}_{ik}\boldsymbol{\beta} - \ln\Big(\sum_{l=1}^{M+1}\exp(\mathbf{x}_{lk}\boldsymbol{\beta})\Big)\Big]. \qquad (4)$$

Les estimateurs basés sur la vraisemblance conditionnelle, utilisés pour l'estimation des risques, sont instables et ont une grande variance quand le nombre d'événements n'est pas nettement plus grand que le nombre de variables explicatives ou en cas de colinéarité entre celles-ci (Greenland, 2000 ; Bull et al., 2007). Les techniques classiques de sélection de sous-ensembles telles que la sélection progressive (*stepwise*) ou la sélection pas à pas descendante/ascendante sont également insatisfaisantes (Greenland, 2008). Une approche alternative est donnée par la méthode lasso, qui consiste à estimer le vecteur de paramètres $\boldsymbol{\beta}$ par le critère :

$$\widehat{\boldsymbol{\beta}}_D = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\big(l(\boldsymbol{\beta}, D) - \lambda\|\boldsymbol{\beta}\|_1\big) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, l_{p1}\big(\boldsymbol{\beta}, D\big), \qquad (5)$$

où λ est un paramètre de régularisation, $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\beta_j|$ est la norme L_1 des coefficients et $l_{p1}(\boldsymbol{\beta}, D)$ indique la log–vraisemblance conditionnelle pénalisée par la norme L_1 des coefficients et évaluée en $\boldsymbol{\beta}$ et D. L'introduction d'une pénalisation réduit la variabilité de l'estimation, améliorant ainsi la précision de prédiction. En outre, la pénalisation de type L_1 rétrécit certains coefficients, alors que les autres sont annulés exactement, aboutissant ainsi à des modèles parcimonieux.

Algorithmes

Notons que la vraisemblance conditionnelle (pénalisée) d'un modèle de régression logistique conditionnelle (pénalisée) peut s'écrire comme la vraisemblance partielle (pénalisée) d'un modèle à risques proportionnels de Cox (pénalisée) stratifié discret. Il suffit donc de structurer le fichier des données afin d'utiliser les méthodes qui ajustent le modèle de Cox pour effectuer la régression logistique conditionnelle (pénalisée). Premièrement, chaque strate est caractérisée par plusieurs lignes : une ligne correspondant au cas et une ou plusieurs lignes correspondant aux témoins. Deuxièmement, en plus des variables explicatives déjà existant, plusieurs variables doivent être définies : une variable identifiant les cas et les témoins, correspondant à la variable de censure dans le modèle de Cox (les cas ont tous connu l'événement, les témoins sont tous censurés) ; une variable identifiant la strate ; et une variable introduisant de manière fictive la notion de temps discret, nécessaire à la mise en œuvre du modèle de Cox discret (les cas subissent l'événement au temps 1, et les témoins sont censurés au temps 2)

Plusieurs algorithmes ont été proposés pour la résolution du lasso pour le modèle de Cox. Park et al., (2007) présentent une généralisation de l'algorithme lars-lasso (Efron et al., 2004) au modèle de Cox et Goeman, (2008) propose un algorithme basé sur une méthode d'ascension du gradient combinée avec un algorithme de Newton-Raphson. Pour ces deux algorithmes, des librairies R (*glmpath* et *penalized*, respectivement) ont été développées, mais elles n'admettent pas de stratification. Néanmoins, la librairie *penalized* permet de rentrer la variable indiquant le temps en tant que processus de comptage, ce qui permet l'introduction "manuelle" de la strate par une différentiation de la variable fictive temps dans chaque strate. Nous utilisons cette librairie pour obtenir les estimateurs.

Paramètre de régularisation

Le paramètre $\lambda \geq 0$ contrôle la complexité du modèle, de sorte que si $\lambda \to \infty$ aucune variable n'est retenue dans le modèle, alors que si $\lambda = 0$, la solution est celle obtenue par vraisemblance conditionnelle classique. La valeur optimale de λ est celle qui minimise l'erreur de prédiction. Cette erreur peut être estimée par une méthode de rééchantillonnage telle que la validation croisée.

Nous estimons le paramètre λ par la valeur qui maximise le critère de validation croisée à L ensembles appliqué à la vraisemblance conditionnelle pénalisée, respectant l'appariement des données (Van der Laan et al., 2006). Les données sont partitionnées en L blocs disjoints de la même taille K/L, où K est le nombre de strates (supposons, pour simplifier que K/L est un entier). Soit D_l le l–ième bloc et $D \setminus D_l$ l'ensemble d'apprentissage obtenu en ôtant les éléments du l–ième bloc. L'estimateur validation croisée sur la vraisemblance conditionnelle pénalisée est :

$$CVl_{p1}(\lambda) = \frac{1}{L}\sum_{l=1}^{L} l_{p1}\big(\widehat{\boldsymbol{\beta}}_{D\setminus D_l}, D_l\big) = \frac{1}{L}\sum_{l=1}^{L} l_{p1}\big(\widehat{\boldsymbol{\beta}}_{D\setminus D_l}, D\big) - l_{p1}\big(\widehat{\boldsymbol{\beta}}_{D\setminus D_l}, D\setminus D_l\big). \qquad (6)$$

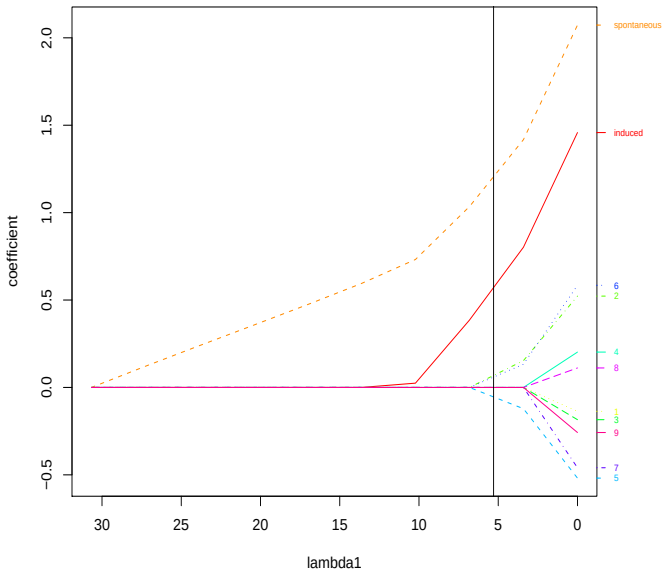

FIG. 1 – Estimations des coefficients en fonction de λ. La ligne verticale indique la valeur des coefficients pour la valeur de λ sélectionnée par validation croisée.

Afin de mesurer la stabilité des résultats, nous calculons des intervalles de confiance des coefficients estimés par un bootstrap non paramétrique, adapté à l'appariement des données.

Exemple

Nous utilisons un jeux de données réel issu d'une enquête cas-témoin sur l'infertilité (provenant de la librairie survival de R), dans laquelle 248 patientes ont été interrogées à propos de divers facteurs de risque. Chaque cas a été apparié à 2 témoins pour l'âge, le niveau socio-économique et la parité. Les variables explicatives disponibles sont les antécédents de fausses couches spontanées (0, 1, 2 ou plus) et provoquées (0, 1, 2 ou plus). Nous rajoutons à cette base, des 9 variables générées aléatoirement (loi Bernouilli de probabilité 0,5) et de façon indépendante à la variable réponse.

La figure 1 montre les estimations des coefficients associés à chaque variable en fonction de λ. La ligne verticale indique la valeur des coefficients pour la valeur de λ optimale, estimée par une validation croisée à 10 ensembles. Les deux facteurs de risque sont sélectionnés. Parmi les variables qui sont indépendantes de la réponse, trois sont sélectionnées et six sont éliminées du modèle choisi par validation croisée. Le tableau 1 montre les valeurs des coefficients et des rapports de cotes estimés (pour la valeur de λ

TAB. 1 – Coefficients et rapports de cotes estimés avec les intervalles de confiance bootstrap.

Variable	β_j	$IC_{95\%}$	RC	$IC_{95\%}$
induced	0,802	$[0,020; 2,645]$	2,230	[1,020 ; 14,080]
spontaneous	1,418	$[0,159; 3,531]$	4,129	$[1,173; 34,160]$
1	0,000	$[-0,833; 0,342]$	1,000	$[0,435; 1,410]$
2	0,154	$[0,000; 1,278]$	1,166	$[1,000; 3,590]$
3	0,000	$[-0,835; 0,326]$	1,000	$[0,434; 1,390]$
4	0,000	$[-0,424; 0,820]$	1,000	$[0,654; 2,270]$
5	-0,122	$[-1,242; 0,000]$	0,885	$[0,289; 1,000]$
6	0,132	$[0,000; 1,401]$	1,141	$[1,000; 4,060]$
7	0,000	$[-1,252; 0,195]$	1,000	$[0,286; 1,210]$
8	0,000	$[-0,528; 0,685]$	1,000	$[0,590; 1,980]$
9	0,000	$[-0,802; 0,175]$	1,000	$[0,448; 1,190]$

optimale) avec les intervalles de confiance obtenus à partir de 1000 bootstrap. Seulement les intervalles de confiance des coefficients des deux facteurs de risque ne contiennent pas 0, questionnant ainsi la pertinence de retenir dans le modèle les 3 variables sélectionnées parmi les variables indépendantes de l'infertilité.

Dans notre exemple, nous observons que le lasso pour la régression logistique conditionnelle conserve bien les variables pertinentes. En revanche, il élimine peu de variables non pertinentes. Le calcul d'intervalles de confiance permet d'affiner l'élimination de ces variables non pertinentes. Une étude plus importante mériterait d'être réalisée pour évaluer les propriétés de cette méthode dans le cadre de la régression logistique conditionnelle.

Bibliographie

[1] Bull S.B., Lewinger J.P. et Lee S.S. (2007). Confidence intervals for multinomial logistic regression in sparse data. *Stat Med.* 26(4) :903-18.

[2] Efron B., Hastie T., Johnstone I. et Tibshirani R. (2004). Least angle regression. *Ann. Statist.* 32(2) :407-499.

[3] Goeman J. (2008). An efficient algorithm for L_1 penalized estimation. *Technical report*, Department of Medical Statistics and BioInformatics, Leiden University Medical Center.

[4] Greenland S., Schwartzbaum J.A. et Finkle W.D. (2000) Problems due to small samples and sparse data in conditional logistic regression analysis. *Am J Epidemiol.*, 151(5) :531-9.

[5] Greenland S. (2008). Invited commentary : variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol.* 167(5) :523-9 ; discussion 530-1.

[6] Park, M.Y. et Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. Royal. Statist. Soc B.*, 69(4) :659-677

[7] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1) :267-288.
[8] Van der Laan M.J., Dudoit S. et Keles S. (2006). Asymptotic Optimality of Likelihood-Based Cross-Validation *Statistical Applications in Genetics and Molecular Biology* 3.1.