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

des processus d'estimation et de sélection de variables et, d'autre part, une réduction de la variance (Bickel et al. (2009), pour le modèle de régression linéaire, Bunea (2008), pour le modèle de régression logistique).

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 deux exemples. Le code R est téléchargeable à l'adresse http://biostat.isped.u-bordeaux2.fr.

2 Régression logistique conditionnelle

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. Les données peuvent alors être analysées à l'aide d'une régression logistique, ce qui permet d'ajuster sur les facteurs de confusion potentiels, et donc, si le modèle logistique est correcte, d'en éliminer leur effet. Néanmoins, dans certaines situations, le poids des facteurs de confusion est tellement important, qu'un simple ajustement ne suffit pas pour garantir une interprétation simple des résultats. Une stratégie consiste à imposer expérimentalement un certain degré de similitude entre les groupes à comparer, en appariant chaque cas avec un nombre préfixé de témoins comparables, en termes d'exposition à ces facteurs. La régression logistique conditionnelle est adaptée aux études cas-témoin sur séries stratifiées, dans lesquelles cas et témoins sont segmentés en petits groupes à l'intérieur desquels l'exposition à d'éventuels facteurs de confusion est constante (Falissard et Lellouch, 2005).

2.1 Modèle

Considérons le vecteur aléatoire $(X_1, \ldots, X_p, Y)$, où Y est une variable binaire, codée 0–1. Nous nous intéressons à la relation entre la variable réponse Y et plusieurs variables explicatives $X = (X_1, \ldots, X_p)$. Les observations sont des groupes d'individus (strates, notées $k = 1, \ldots, K$), constitués d'un cas ($Y_{1k} = 1$) et M témoins ($Y_{ik} = 0$, $i = 2, \ldots, M+1$), chacun d'entre eux ayant une valeur de X : pour l'individu i de la strate k, nous avons le vecteur d'observations $\mathbf{x}_{ik} = (x_{ik1}, \ldots, x_{ikp})$, $i = 1, \ldots, M+1$, $k = 1, \ldots, K$.

Soit P_{ik} la probabilité (non conditionnelle au mode d'échantillonnage) 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 :

$$\begin{aligned} P_{ik} = P(Y_{ik} = 1|\mathbf{x}_{ik}) &= \frac{\exp(\alpha_0 + \sum_{l=1}^{K} \alpha_l 1_l + \sum_{j=1}^{p} \beta_j x_{ikj})}{1 + \exp(\alpha_0 + \sum_{l=1}^{K} \alpha_l 1_l + \sum_{j=1}^{p} \beta_j x_{ikj})} \\ &= \frac{\exp(\alpha_0 + \alpha_k + \mathbf{x}_{ik}\boldsymbol{\beta})}{1 + \exp(\alpha_0 + \alpha_k + \mathbf{x}_{ik}\boldsymbol{\beta})}, \end{aligned} \tag{1}$$

où $\mathbf{1}_l$ est une fonction indicatrice qui vaut 1 si l'individu appartient à la strate l et 0 autrement ; α_0 est la proportion de cas chez les sujets non exposés, les α_k sont les coefficients représen-

tant l'effet des variables d'appariement sur la réponse, qui traduisent les différences entre les strates ; et les coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ représentent les effets des variables explicatives ou, de façon équivalente, le log–rapport de cotes. Cette relation entre la probabilité de survenue de l'événement (ou le risque de survenue de l'événement puisque $Y_{ik}|\mathbf{x}_{ik}$ sont distribuées suivant une loi de Bernoulli et donc, $P_{ik} = \mathbb{E}[Y_{ik}|\mathbf{x}_{ik}]$) et les valeurs des variables explicatives peut également s'exprimer à l'aide de la transformation logit :

$$\text{logit}[P_{ik}] = \log \frac{P(Y_{ik} = 1|\mathbf{x}_{ik})}{1 - P(Y_{ik} = 1|\mathbf{x}_{ik})} = \alpha_0 + \alpha_k + \mathbf{x}_{ik}\boldsymbol{\beta}. \tag{2}$$

2.2 Estimation

Supposons que les différences entre les strates ne sont pas d'intérêt (au sens où les variables de stratification sont des variables potentiellement confondantes, et non pas les facteurs de risque potentiels auxquels on s'intéresse principalement). 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 :

$$(1 - P_{1k}) \ldots (1 - P_{i-1k}) P_{ik} (1 - P_{i+1k}) \ldots (1 - P_{M+1k}) = \frac{P_{ik}}{1 - P_{ik}} \prod_{j=1}^{M+1} (1 - P_{jk}). \tag{3}$$

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

$$\frac{\frac{P_{1k}}{1-P_{1k}} \prod_{j=1}^{M+1}(1 - P_{jk})}{\sum_{j=1}^{M+1} \frac{P_{jk}}{1-P_{jk}} \prod_{j=1}^{M+1}(1 - P_{jk})} = \frac{\frac{P_{1k}}{1-P_{1k}}}{\sum_{j=1}^{M+1} \frac{P_{jk}}{1-P_{jk}}} = \frac{\exp(\mathbf{x}_{1k}\boldsymbol{\beta})}{\sum_{j=1}^{M+1} \exp(\mathbf{x}_{ik}\boldsymbol{\beta})}, \tag{4}$$

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

$$l(\boldsymbol{\beta}, D) = \sum_{k=1,\ldots,K} \Big[\mathbf{x}_{1k}\boldsymbol{\beta} - \ln \Big(\sum_{l=1}^{M+1} \exp(\mathbf{x}_{lk}\boldsymbol{\beta}) \Big) \Big]. \tag{5}$$

Pour l'estimation du vecteur de paramètres $\boldsymbol{\beta}$, la méthode généralement utilisée est la méthode du maximum de vraisemblance conditionnelle :

$$\widehat{\boldsymbol{\beta}}_D^{\text{MV}} = \underset{\boldsymbol{\beta}}{\text{argmax}} \big(l(\boldsymbol{\beta}, D) \big). \tag{6}$$

Notons que si tous les sujets d'une strate k ont le même vecteur d'observations $\mathbf{x}_{1k} = \ldots = \mathbf{x}_{M+1k}$, alors la strate k n'intervient pas dans l'estimation de $\boldsymbol{\beta}$ car la contribution de cette strate à la vraisemblance est indépendante de $\boldsymbol{\beta}$.

3 Régression logistique conditionnelle pénalisée

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 net-

tement plus grand que le nombre de variables explicatives ou en cas de colinéarité (Greenland, 2000; Corcoran et al., 2001; Bull et al., 2007; Hansson et Khamis, 2008).

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. Le point de vue du praticien est qu'aucune stratégie de sélection s'est montrée meilleure que la stratégie consistant à inclure toutes les variables explicatives dans le modèle (après un filtrage préliminaire), car ces méthodes élimineraient ou ignoreraient facilement des facteurs importants (Greenland, 2008). Une approche alternative est donnée par les méthodes de pénalisation.

3.1 La méthode lasso

Le lasso est une méthode de pénalisation et de sélection de variables initialement proposée pour la régression linéaire (Tibshirani, 1996). Dans ce cadre, elle consiste à estimer le vecteur de paramètres par minimisation du critère quadratique des moindres carrés sous une contrainte sur la somme des valeurs absolues des coefficients ou, de façon équivalente, par minimisation du critère quadratique pénalisé par la norme L_1 des coefficients. 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.

Le lasso a ensuite été généralisé à d'autres types de critères, tels que ceux basés sur la vraisemblance (par exemple, Tibshirani (1997); Klinger (2001)), et d'autres contextes, tel que le non–paramétrique (par exemple, Lin et Zhang (2006); Avalos et al. (2007); Ravikumar et al. (2007)). Des versions adaptées à certains problèmes ont également été proposées (par exemple, Zou et Hastie (2005); Zou (2006); Yuan et Lin (2006); Meinshausen (2007)), et des problèmes propres aux données issues d'une étude cas–témoin (non appariés) analysés (Bunea et Barbu, 2009). Les avancées théoriques (par exemple, Efron et al. (2004); Zhao et Yu (2006); Meinshausen et Yu (2009)), les développements d'algorithmes efficaces (par exemple, Efron et al. (2004); Park et Hastie (2007); Goeman (2008)) ainsi que la disponibilité de ces algorithmes sur des logiciels accessibles ont collaboré à la récente popularité de la méthode.

3.2 La méthode lasso appliquée à la régression logistique conditionnelle

Le lasso appliqué à la régression logistique conditionnelle consiste à maximiser la fonction (5) pénalisée par la norme L_1 du vecteur de coefficients inconnus :

$$\widehat{\boldsymbol{\beta}}_D^{\mathrm{L}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left(l(\boldsymbol{\beta}, D) - \lambda \|\boldsymbol{\beta}\|_1 \right) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, l_{p1}\big(\boldsymbol{\beta}, D\big), \tag{7}$$

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.

3.3 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 (pénalisé) stratifié discret (cf. annexe pour une introduction du modèle de Cox). Il suffit donc de structurer le fichier des données afin d'utiliser les méthodes qui ajustent le modèle de Cox (pénalisé) 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) (Chardon et al., 2008).

Plusieurs algorithmes ont été proposés pour la résolution du lasso pour le modèle de Cox. Tibshirani (1997) propose un programme quadratique ; Gui et Li (2005) et Park et Hastie (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 les deux derniers 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 $\widehat{\boldsymbol{\beta}}_D^{\mathrm{L}}(\lambda)$.

3.4 D'autres types de pénalisation

Dans ce travail, nous nous focalisons sur le problème de la sélection de variables via le lasso. Néanmoins, l'application d'autres types de pénalisation à la régression logistique conditionnelle sont envisageables. En particulier, des manipulations similaires à celles effectuées pour obtenir l'estimateur lasso, à l'aide de la librairie *penalized*, nous permettent d'obtenir l'estimateur *ridge* :

$$\widehat{\boldsymbol{\beta}}_D^{\mathrm{R}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left(l(\boldsymbol{\beta}, D) - \lambda \|\boldsymbol{\beta}\|_2 \right) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, l_{p2}\big(\boldsymbol{\beta}, D\big), \tag{8}$$

où λ est un paramètre de régularisation, $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} |\beta_j^2|$ est la norme L_2 des coefficients au carré et $l_{p2}(\boldsymbol{\beta}, D)$ indique la log–vraisemblance conditionnelle pénalisée par la norme L_2 des coefficients. Plus généralement, nous pouvons obtenir l'estimateur *elastic net* :

$$\widehat{\boldsymbol{\beta}}_D^{\mathrm{EN}}(\lambda_1, \lambda_2) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left(l(\boldsymbol{\beta}, D) - \lambda_1 \|\boldsymbol{\beta}\|_1 - \lambda_2 \|\boldsymbol{\beta}\|_2 \right) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, l_{p12}\big(\boldsymbol{\beta}, D\big), \tag{9}$$

où $l_{p12}(\boldsymbol{\beta}, D)$ indique la log–vraisemblance conditionnelle pénalisée par les normes L_1 et L_2 des coefficients, et λ_1 et λ_2 sont des paramètres de régularisation. L'obtention de ce dernier estimateur demande néanmoins un nombre de calculs plus important. En effet, l'optimisation du paramètre λ_1 par validation croisée, oblige à fixer le paramètre λ_2 sur une grille de valeurs et vice-versa.

4 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. Le paramètre de régularisation λ peut être estimé par une méthode de rééchantillonnage telle que la validation croisée. Des intervalles de confiance de λ (et des coefficients de régression) peuvent être estimés par bootstrap.

4.1 Choix du paramètre de régularisation

Nous estimons le paramètre λ par la valeur qui maximise le critère de validation croisée à L ensembles appliqué à la vraisemblance conditionnelle, respectant l'appariement des données (Verweij et Houwelingen, 1993; Van der Laan et al., 2004; Goeman, 2008). 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 (l'ensemble de test) 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 est :

$$\begin{aligned} CVl(\lambda) &= \frac{1}{L}\sum_{l=1}^{L} l\big(\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{D\setminus D_l}(\lambda), D_l\big) \\ &= \frac{1}{L}\sum_{l=1}^{L} l\big(\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{D\setminus D_l}(\lambda), D\big) - l\big(\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{D\setminus D_l}(\lambda), D\setminus D_l\big). \end{aligned} \tag{10}$$

L'objectif principal est ici la sélection de variables, avant la prédiction ou l'estimation. Ce critère fait intervenir le terme de pénalité, ainsi les grandes dimensions sont pénalisées, favorisant l'élimination des variables considérées moins pertinentes.

4.2 Mesure de la stabilité

Des résultats théoriques ont montré que le lasso retient, en générale, les variables pertinentes, mais aussi quelques variables non pertinentes (Meinshausen et Yu, 2009). Nous souhaitons disposer d'un critère supplémentaire nous permettant de mesurer la stabilité des résultats et de repérer les variables pour lesquelles l'élimination ou sélection dans le modèle n'est pas nette.

Des intervalles de confiance basés sur des approximations de la matrice de variances–covariances des coefficients estimés ont été déduits pour le lasso et méthodes dérivées (Tibshirani, 1996, 1997; Zou, 2006). Néanmoins, ces intervalles de confiance sont peu utilisés dans les applications de ces méthodes. Une possible explication est que leur interprétation diffère de celle de méthodes basées sur une vraisemblance classique. Par exemple, une borne inférieur de l'intervalle égale à 0 peut traduire l'élimination de la variable dans $\alpha\%$ des échantillons bootstrap. Une autre possible explication est la largeur importante, par rapport aux intervalles de confiance des méthodes basées sur une vraisemblance classique (Pötscher, 2007; Pötscher et Schneider, 2008).

Ici, nous construisons des intervalles de confiance des coefficients estimés par un bootstrap non paramétrique (méthode percentile de Efron et Tibshirani (1993)), adapté à l'appariement

des données. Nous générons B échantillons bootstrap indépendants, où les observations sont les strates, constituées d'un cas et M témoins. Pour chaque échantillon bootstrap, le paramètre de régularisation est estimé par validation croisée, et les coefficients associés sont calculés : $\widehat{\boldsymbol{\beta}}_*^{\mathrm{L}}(\lambda)_1, \ldots, \widehat{\boldsymbol{\beta}}_*^{\mathrm{L}}(\lambda)_B$. Soit $\widehat{\boldsymbol{\beta}}_*^{\mathrm{L}}(\lambda)^{(\alpha)}$ le $100\times\alpha$-ième percentile empirique des valeurs $\widehat{\boldsymbol{\beta}}_*^{\mathrm{L}}(\lambda)_b$, c'est-à-dire, la $B \times \alpha$-ième valeur de la liste ordonnée des B répétions $\widehat{\boldsymbol{\beta}}_*^{\mathrm{L}}(\lambda)$. De façon similaire, soit $\widehat{\boldsymbol{\beta}}_*^{\mathrm{L}}(\lambda)^{(1-\alpha)}$ le $100 \times (1-\alpha)$-ième percentile empirique. L'intervalle percentile à $1-2\alpha$ est approximativement : $[\widehat{\boldsymbol{\beta}}_*^{\mathrm{L}}(\lambda)^{(\alpha)}, \widehat{\boldsymbol{\beta}}_*^{\mathrm{L}}(\lambda)^{(1-\alpha)}]$. Si $B \times \alpha$ n'est pas un entier, soit $k = [(B+1) \times \alpha]$, le plus grand entier $\leq (B+1) \times \alpha$ (supposant $\alpha \leq 0.5$). Alors on définit les α et $1-\alpha$ fractiles par la k-ième et $(B+1-k)$-ième plus grandes valeurs.

5 Exemple

Nous utilisons deux jeux de données réels que nous avons complétés avec des données simulées, afin d'illustrer l'utilisation des méthodes.

5.1 Exemple 1

Le premier jeu, issu d'une enquête cas-témoin sur l'infertilité (provenant de la librairie *survival* de R), nous permet d'illustrer l'utilisation de la méthode lasso et des intervalles de confiance bootstrap pour des données appariées en présence de variables explicatives non pertinentes. Chaque cas est apparié à 2 témoins pour l'âge, le niveau socio-économique et la parité, donnant lieu à 83 strates. 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 10 variables générées aléatoirement (loi Bernoulli de probabilité 0,8), indépendantes entre elles, des autres variables explicatives, et indépendantes de la variable réponse.

La figure 1 montre en haut 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 qui maximise le critère de validation croisée à 10 ensembles (la moyenne de la fonction de vraisemblance dont les coefficients sont estimés sur les ensembles d'apprentissage et évaluée sur les ensembles de test). Ce critère est représenté dans le graphique inférieur. Le lasso sélectionne les deux facteurs de risque. Parmi les variables qui sont indépendantes de la réponse, une est sélectionnée et neuf sont éliminées du modèle choisi par validation croisée.

Le tableau 1 montre les valeurs des coefficients estimés par la méthode lasso, pour la valeur du paramètre de régularisation estimé par validation croisée à 10 ensembles, ainsi que des intervalles de confiance obtenus à partir de 1000 échantillons bootstrap, pour de différents niveaux de confiance. 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 la variable non pertinente sélectionnée. Cela ne semble pas dépendre du niveau de confiance choisi.

5.2 Exemple 2

Le deuxième jeu de données, issu d'une étude rétrospective sur les risques de maladie coronaire dans une région de l'Afrique du Sud (provenant de la librairie *glmpath* de R), nous permet

Variable	$\widehat{\beta}_j$	$\text{IC}_{99\%}$	$\text{IC}_{95\%}$	$\text{IC}_{80\%}$
induced	0.680	$[0.008, 1.896]$	$[0.296, 1.568]$	$[0.478, 1.297]$
spontaneous	1.089	$[0.693, 2.423]$	$[0.765, 1.999]$	$[0.920, 1.682]$
1	0.073	$[-0.059, 0.694]$	$[0.000, 0.530]$	$[0.000, 0.392]$
2	0.000	$[-0.461, 0.218]$	$[-0.325, 0.071]$	$[-0.225, 0.000]$
3	0.000	$[-0.573, 0.162]$	$[-0.421, 0.000]$	$[-0.272, 0.000]$
4	0.000	$[-0.404, 0.333]$	$[-0.281, 0.190]$	$[-0.101, 0.068]$
5	0.000	$[-0.505, 0.514]$	$[-0.344, 0.312]$	$[-0.156, 0.130]$
6	0.000	$[-0.310, 0.651]$	$[-0.149, 0.400]$	$[-0.022, 0.200]$
7	0.000	$[-0.312, 0.437]$	$[-0.175, 0.270]$	$[-0.018, 0.137]$
8	0.000	$[-0.223, 0.560]$	$[-0.003, 0.444]$	$[0.000, 0.296]$
9	0.000	$[-0.292, 0.363]$	$[-0.209, 0.249]$	$[-0.052, 0.132]$
10	0.000	$[-0.473, 0.427]$	$[-0.292, 0.282]$	$[-0.121, 0.137]$

TAB. 1 – *Coefficients et intervalles de confiance bootstrap estimés par la méthode lasso.*

d'étudier le comportement des méthodes lasso et ridge pour des données appariées, ainsi que du critère de validation croisée, en présence de colinéarité. La base comporte 9 variables explicatives : la pression artérielle systolique en mmHg (sbp), la consommation cumulative en tabac en kg (tobacco), les lipoprotéines de base densité ou "mauvais" cholestérol en mmol/l (ldl), le niveau des tissus gras (adiposity), la présence ou absence d'antécédents familiaux de maladie coronaire (famhist2), le score de comportement de stress de type A (typea), l'indice de masse corporelle en kg/m^2 (obesity), la consommation cumulative d'alcool en l (alcohol), l'âge en début d'étude en années (age). La variable réponse est la présence ou absence de maladie coronaire (chd). Nous rajoutons à cette base des variables générées en additionnant du bruit normal à des variables déjà existantes : tobacco2, ldl2, adiposity2, typea2, obesity2, alcohol2, de telle sorte que la corrélation entre la variable initiale et la variable générée est approximativement 0.75. Du fait de la grande différence d'âge entre, par exemple, le groupe de sujets avec des antécédents familiaux de maladie coronaire et le groupe de sujets sans antécédents, un modèle logistique classique pourrait être très fragile : le moindre écart aux hypothèses du modèle (par exemple, la linéarité avec le logit) pourrait expliquer à lui seul un effet apparent des antécédents familiaux. Chaque cas est apparié à un témoin de façon aléatoire comparable en termes d'âge, (appartenance à la même classe d'âge, considérant 4 classes d'âge basées sur les quartiles). Pour 8 des hommes plus âgés il n'a pas été possible de leur attribuer un pair, car il ne restait plus de sujets n'ayant pas connu la maladie coronaire dans la même tranche d'âge. Le jeu de données utilisé comporte 152 pairs cas-témoin.

La figure 2 montre, en haut, les estimations lasso des coefficients et, en bas, les estimations ridge des coefficients en fonction du paramètre de régularisation. La valeur du paramètre de régularisation est estimée par validation croisée à 5, 10, 50 et 152 ensembles, ce dernier correspondant à une validation croisée *leave-one-out*. Pour la validation croisée à 5, 10 et 50 ensembles, les valeurs indiquées sur les graphiques correspondent aux moyennes sur 100 validations croisées. Les différents critères donnent lieu à un choix de modèle différent, ainsi, en ce qui concerne le lasso, le modèle choisi par validation croisée à 5 ensembles contient seulement six coefficients non nuls, étant les variables typea et tobacco2 les plus influentes. Quand le cri-

tère de sélection est la validation croisée *leave-one-out*, le modèle contient 12 coefficients non nuls et les facteurs de risque les plus influents sont famhist2, ldl et tobacco2. Nous avons également remarqué une grande variabilité du critère validation croisée, un problème soulevé dans la littérature (Breiman, 1996). La méthode la ridge ne sélectionne pas de variables, néanmoins les courbes des coefficients étant plus lisses, les différences entre les modèles sélectionnés par les différents critères sont moins importantes.

6 Conclusion

L'analyse des facteurs associés au risque de survenue de l'événement d'intérêt par régression logistique conditionnelle permet de contrôler les facteurs de confusion potentiels. Néanmoins, cette méthode peut fournir des estimateurs imprécis quand le modèle contient un nombre élevé de paramètres. Nous avons proposé une adaptation de la méthode lasso à la régression logistique conditionnelle, consistant à pénaliser la vraisemblance conditionnelle par la norme L_1 des coefficients. Cette méthode permet l'analyse de données appariées quand le nombre de variables est important. Dans notre premier exemple, nous remarquons que le lasso pour la régression logistique conditionnelle conserve bien les variables pertinentes. En revanche, il n'élimine pas toutes les variables non pertinentes. Cette observation (relevée dans toutes les répétitions que nous avons mené à terme de cet exemple) concorde avec des résultats théoriques obtenus sur le lasso (Meinshausen et Yu, 2009). Le calcul d'intervalles de confiance permet d'affiner l'élimination de ces variables non pertinentes. Dans cette étude, nous nous sommes focalisés sur le problème de sélection de variables. Cependant, la pénalisation L_2, ou plus généralement la pénalisation $L_1 + L_2$, peuvent également être appliquées à des jeux de données appariées. L'ensemble de méthodes de pénalisation constituent une alternative aux méthodes classiques de sélection automatique de sous-ensembles, fort discréditées dans des domaines d'application tels que l'épidémiologie (Greenland, 2008).

Annexe

Le modèle de Cox exprime une relation entre la fonction de risque associée à la survenue d'un événement et le vecteur des p variables explicatives $\mathbf{x} = (x_1, \ldots, x_p)$:

$$\lambda(t, \mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}\boldsymbol{\beta}), \tag{11}$$

où $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ est le vecteur de coefficients de régression et $\lambda_0(t)$ est la fonction de risque de base. Soient $t_1 < t_2 < \ldots < t_n$ les différents temps d'événements observés et $(1), \ldots, (n)$ les indices des sujets ayant subi l'événement respectivement en $t_1, \ldots, t_n$. En scindant la vraisemblance en deux parties, afin de ne conserver que la partie concernant les coefficients de régression, et sous l'hypothèse de risques proportionnels, la probabilité conditionnelle que le sujet i subisse l'événement en t_i sachant qu'il est à risque au temps t_i et qu'il n'y a qu'un seul événement en t_i parmi les sujets à risque au temps t_i est égale à :

$$P_i = \frac{\lambda_0(t_i) \exp(\mathbf{x}_{(i)}\boldsymbol{\beta})}{\sum_{l:t_l \geq t_i} \lambda_0(t_l) \exp(\mathbf{x}_{(l)}\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}_{(i)}\boldsymbol{\beta})}{\sum_{l:t_l \geq t_i} \exp(\mathbf{x}_{(l)}\boldsymbol{\beta})}. \tag{12}$$

La vraisemblance partielle est le produit des probabilités conditionnelles calculées à chaque temps d'événements. La fonction de log-vraisemblance partielle, évaluée en $\boldsymbol{\beta}$ s'écrit :

$$l(\boldsymbol{\beta}) = \sum_{i=1,\ldots,n} \Big[\mathbf{x}_{(i)}\boldsymbol{\beta} - \ln \Big(\sum_{l:t_l \geq t_i} \exp(\mathbf{x}_{(l)}\boldsymbol{\beta}) \Big) \Big]. \tag{13}$$

La vraisemblance partielle du modèle de Cox n'est pas une vraisemblance dans le sens statistique du terme, mais il a été établi qu'elle peut être utilisée comme telle pour estimer les coefficients de régression $\boldsymbol{\beta}$ et tester l'influence de variables explicatives sur la fonction de risque. Cette vraisemblance partielle nécessite l'hypothèse de données continues, c'est à dire qu'il n'y a pas plusieurs évènements (d'*ex aequo*) à la même date. Pour des données réelles cette hypothèse n'est pas toujours vérifiée, on approche alors la vraisemblance partielle. L'approximation la plus utilisée est celle de Breslow :

$$\tilde{l}(\boldsymbol{\beta}) = \sum_{i=1,\ldots,n} \Big[\mathbf{s}_i\boldsymbol{\beta} - \ln \Big(\sum_{l:t_l \geq t_i} \exp(\mathbf{x}_{(l)}\boldsymbol{\beta}) \Big)^{m_i} \Big], \tag{14}$$

où $\mathbf{s}_i$ est la somme des vecteurs des variables explicatives des m_i sujets ayant subi l'événement au temps t_i (Joly et al., 2009).

Quand les données sont stratifiées, le risque de base peut être différent dans chaque strate :

$$\lambda(t, \mathbf{x}) = \lambda_{0k}(t) \exp(\mathbf{x}\boldsymbol{\beta}), \tag{15}$$

$k = 1, \ldots, K$. La log-vraisemblance est alors calculée comme la somme des fonctions log-vraisemblance dans chacune des strates.

Références

Avalos, M., Y. Grandvalet, et C. Ambroise (2007). Parsimonious additive models. *Computational Statistics & Data Analysis 51*, 2851–2870.

Bickel, P., Y. Ritov, et A. Tsybakov (2009). Simultaneous analysis of lasso and dantzing selector. *Annals of Statistics 37*, 1705–1732.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist. 24*, 2350–2383.

Bull, S., J. Lewinger, et S. Lee (2007). Confidence intervals for multinomial logistic regression in sparse data. *Stat. Med. 26*, 903–18.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via l_1 and $l_1 + l_2$ penalization. *Electron. J. Statist. 2*, 1153–1194.

Bunea, F. et A. Barbu (2009). Dimension reduction and variable selection in case control studies via regularized likelihood optimization.

Chardon, B., S. Host, G. Pedrono, et I. Gremy (2008). Contribution of case-crossover design to the analysis of short-term health effects of air pollution : reanalysis of air pollution and health data. *Rev Epidemiol Sante Publique. 56*, 31–40.

Corcoran, C., C. Mehta, N. Patel, et P. Senchaudhuri (2001). Computational tools for exact conditional logistic regression. *Stat. Med. 20*, 2723–39.

Efron, B., T. Hastie, I. Johnstone, et R. Tibshirani (2004). Least angle regression. *Ann. Statist. 32*, 407–499.

Efron, B. et R. J. Tibshirani (1993). *An introduction to the Bootstrap*. New York: Chapman and Hall.

Falissard, B. et J. Lellouch (2005). *Comprendre et utiliser les statistiques dans les sciences de la vie*. Issy-les-Moulineaux: Masson.

Goeman, J. (2008). An efficient algorithm for l_1 penalized estimation. Technical report, Department of Medical Statistics and BioInformatics, Leiden University Medical Center.

Greenland, S. (2000). Small-sample bias and corrections for conditional maximum-likelihood odds-ratio estimators. *Biostatistics 1*, 113–22.

Greenland, S. (2008). Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol. 167*, 523–9.

Gui, J. et H. Li (2005). Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics 21*, 3001–8.

Hansson, L. et H. Khamis (2008). Matched samples logistic regression in case-control studies with missing values: when to break the matches. *Stat Methods Med Res. 17*, 595–607.

Joly, P., A. Alioum, D. Commenges, M. L. Goff, et B. Liquet (2009). Master sciences, technologies, santé, mention santé publique, 2009-2010, sta201 : Méthodes d'analyses de données de survie. Polycopié de cours, Université Victor Segalen Bordeaux 2, ISPED.

Klinger, A. (2001). Inference in high dimensional generalized linear models based on soft thresholding. *J. Royal. Statist. Soc B. 63*, 377–392.

Lin, Y. et H. Zhang (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics 34*, 2272–2297.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis 52*, 374–393.

Meinshausen, N. et B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics 37*, 246–270.

Park, M. et T. Hastie (2007). l_1-regularization path algorithm for generalized linear models. *J. Royal. Statist. Soc B. 69*, 659–677.

Pötscher, B. et U. Schneider (2008). Confidence sets based on penalized maximum likelihood estimators. Technical report, Department of Statistics and Decision Support Systems, University of Vienna.

Pötscher, J. (2007). Confidence sets based on sparse estimators are necessarily large. Technical report, Department of Statistics and Decision Support Systems, University of Vienna.

Ravikumar, P., H. Liu, J. Lafferty, et L. Wasserman (2007). Spam: Sparse additive models. In *Advances in Neural Information Processing Systems*, Volume 20.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B. 58*, 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Stat Med. 16*, 385–95.

Van der Laan, M., S. Dudoit, et S. Keles (2004). Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology 3*, Art. 4.

Verweij, P. et H. V. Houwelingen (1993). Cross-validation in survival analysis. *Stat Med. 12*, 2305–14.

Yuan, M. et Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B 68*, 49–67.

Zhao, P. et B. Yu (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research 7*, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Zou, H. et T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B 67*, 301–320.

Summary

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. 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 two examples.

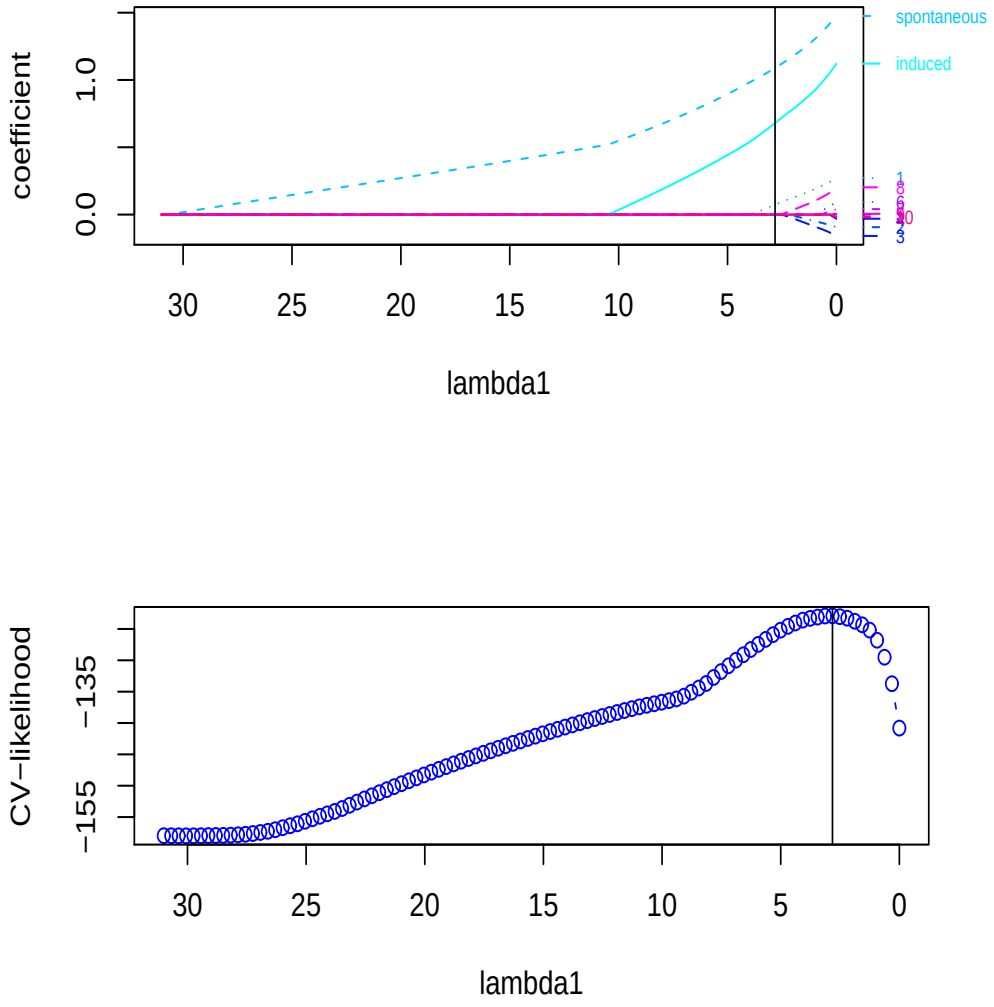

FIG. 1 – *Valeurs du critère validation croisée en fonction de λ (graphique inférieur). La ligne verticale indique la valeur de λ qui maximise le critère. Estimations des coefficients en fonction de λ, par la méthode lasso (graphique supérieur). La ligne verticale indique la valeur des coefficients pour la valeur de λ sélectionnée par validation croisée. Les données correspondent à une enquête cas-témoin sur l'infertilité.*

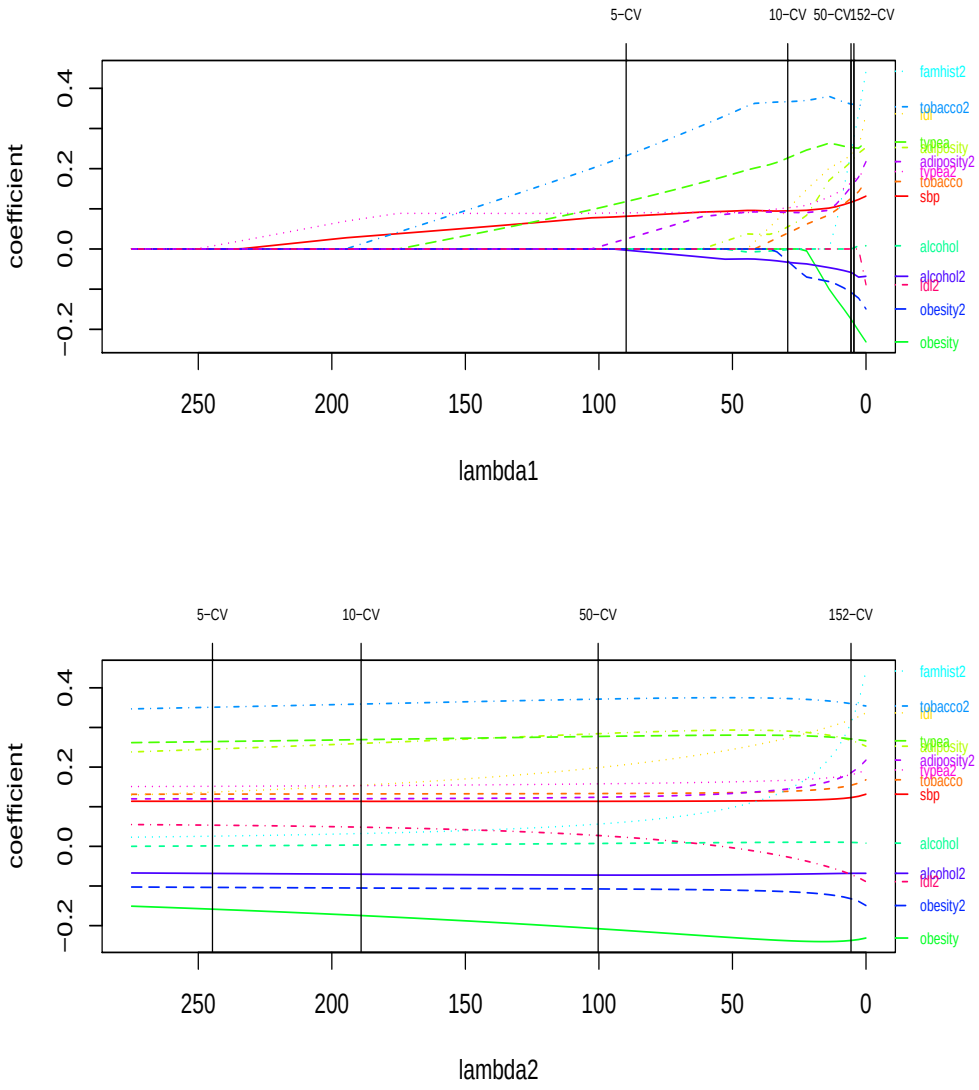

FIG. 2 – *Estimations des coefficients en fonction de λ par la méthode lasso, en haut, et par la méthode ridge, en bas. Les lignes verticales indiquent les valeurs des coefficients pour la valeur moyenne de λ sélectionnée par validation croisée à 5, 10, 50 et 152 ensembles, respectivement. Les données sont extraites d'une étude sur les risques de maladie coronaire.*