PRISE EN COMPTE DE LA CENSURE À GAUCHE DANS LA MODÉLISATION DE DONNÉES DE GRANDE DIMENSION

Nastasia Fouret [1,2] & Marta Avalos [1,3,4] & Linda Wittkop [1,3,4,5] & Rodolphe Thiébaut [1,3,4,5] & Daniel Commenges [1,3,4]

[1] *INSERM U897–Epidémiologie–Biostatistique, Univ. Bordeaux, ISPED*
[2] *Univ. Clermont–Ferrand*
[3] *Univ. Bordeaux, ISPED, INSERM U897–Epidémiologie–Biostatistique*
[4] *INRIA-SISTM, Bordeaux*
[5] *Pôle de Santé Publique CHU de Bordeaux*

Résumé. Dans ce travail, nous décrivons deux algorithmes d'estimation des modèles de régression pénalisée pour des données censurées à gauche. Le premier est basé sur une version paramétrique de l'estimateur de Buckley–James et une résolution itérative des moindres carrés pénalisés. Le deuxième est basé sur une fonction de vraisemblance pénalisée, qui différentie la contribution d'un individu selon que ses données soient censurées ou non. La méthode de Levenberg–Marquardt est appliquée afin de permettre l'obtention d'une solution numérique au problème de maximisation par rapport à plusieurs paramètres d'une fonction de vraisemblance non linéaire.

Afin d'évaluer l'apport de la prise en compte de la censure à gauche dans la prédiction de la réponse à partir des prédicteurs, ainsi que de comparer les différents algorithmes, une étude par simulation est développée.

Mots-clés. Pénalisation L^1, Buckley–James, Fonction de vraisemblance, Levenberg–Marquardt.

Abstract. In the present work, we describe two algorithms for estimating penalised regression models for left censoring data. The first one is based on a parametric version of the Buckley–James estimator and an iterative resolution of penalised least squares. The second one is based on a penalised likelihood function in which the contribution of an individual depends on whether his response was observed or censored. The Levenberg–Marquardt's algorithm is applied to resolve the maximisation problem with respect to several parameters of the nonlinear likelihood function.

A simulation study is developped to evaluate the interest of left censoring modelling in predicting the response from the predictors, as well as to compare the performance of different algorithms.

Keywords. L^1 penalisation, Buckley–James, Likelihood function, Levenberg–Marquardt.

1 Méthodes

Le modèle considéré suppose une réponse Y censurée à gauche et linéairement dépendante des variables explicatives $X = (X_1, \ldots, X_p)$. Les résidus, $\varepsilon = (\varepsilon_1, \ldots \varepsilon_n)$, sont supposés indépendant et identiquement distribués selon la loi $\mathcal{N}(0, \sigma^2)$ et indépendants de Y et X. On note c le seuil de censure (valeur fixée), $\mathcal{Y}$ la réponse non censurée, que l'on cherche à estimer.

1.1 Algorithmes

Nous présentons deux méthodes d'estimation des paramètres d'un modèle de régression linéaire, possédant un grand nombre de variables explicatives et une variable réponse censurée à gauche.

Vraisemblance La vraisemblance est décomposée en vraisemblance observée et vraisemblance censurée.

$$L(\beta, \sigma^2) = \prod_{\text{i observed}} \mathbb{P}(Y_i | X_i, Y_i > c) \times \prod_{\text{i censored}} \mathbb{P}(Y_i | X_i, Y_i \leq c) \tag{1}$$

Vraisemblance pénalisée L1 Nous considérons le problème de maximisation de la fonction de log–vraisemblance avec une contrainte de type L_1 sur le vecteur de coefficients linéaires :
$\mathcal{L}(\beta, \sigma^2) - \lambda ||\beta||_1$.

Buckley-James Buckley et James ont proposé dans les années soixante–dix un estimateur compensant la perte d'information due à la censure à droite. On note $\mathcal{Y}^*$ l'estimation de la réponse non censurée. L'estimateur de la réponse, $\mathcal{Y}^*$, correspond l'imputation de la réponse censurée par son espérance conditionnelle.

$$\mathcal{Y}_i^* = \delta_i Y_i + \beta X_{i.} + \int_{Y_i - \beta X_{i.}}^{\infty} \frac{y dF(y)}{1 - F(Y_i - \beta X_{i.})} \tag{2}$$

où $F(.)$ est la fonction de répartition des résidus, ε_i

La distribution des résidus F, généralement inconnue est substituée par un par un estimateur. Buckley et James proposent, par exemple d'utiliser l'estimateur de Kaplan-Meier, ou Cai et al., 2009 proposent la fonction de poids de Gehan. Lorsque la réponse est la charge virale, les résidus sont classiquement supposés gaussiens. Nous recherchons à estimer simultanément la réponse non censurée par l'estimateur Buckley–James, et les coefficients de régression.

Buckley-James pénalisé Pour estimer la réponse non censurée, nous utilisons les coefficients de régression β obtenus à partir d'une estimation LASSO. Les deux estimations sont effectuées successivement puis itérativement dans une boucle. Les conditions d'arrêts doivent être choisies de manière à garder la variance faible et ainsi limiter les variations d'estimation. L'algorithme est le suivant :

1. Estimation des coefficients de régressions $\hat{\beta}_{(0)}$ utilisant l'opérateur LASSO sans prendre en compte la censure.
2. Estimation de la réponse non censurée :
   $\mathcal{Y}_i^* = \delta_i Y_i + \beta X_{i.} + \left(\int_{-\infty}^{c} f_{\mathcal{N}}(u)du\right)^{-1} \int_{-\infty}^{c} u f_{\mathcal{N}}(u)du$
   où $f_{\mathcal{N}}$ est la densité de la loi gaussienne centrée réduite
3. Estimation des coefficients $\hat{\beta}_{(k)} = \min_{\beta\in\mathbb{R}^p} -\mathcal{L}(\beta, \mathcal{Y}^*, X, \sigma^2) + \lambda||\beta||_1$
4. Répéter les étapes 2 et 3 jusqu'à obtenir $||\hat{\beta}_{(k)} - \hat{\beta}_{(k-1)}|| <$ tolérance.

1.2 Simulations

Protocole 1 Dans cette première simulation quatre scénarios différents sont mis en place, faisant varier la taille de la matrice de données X (100×50 et 100×200) et la sporadicité des coefficients linéaire (un coefficient non nul et la moitié des coefficients non nuls), les valeurs des coefficients non nuls sont choisis aléatoirement. La matrice X est construite en concaténant $p \in \{50, 200\}$ vecteurs simulés indépendamment suivant une loi Bernoulli de paramètre $0,5$. Les résidus sont simulés indépendamment suivant une loi normale centrée réduite.

Protocole 2 Cette simulation joue sur 3 paramètres, la taille de la matrice de données X (100×50 et 100×200), le nombre de coefficients non nuls (10 % et 20 %) et le pourcentage d'observations censurées (20% et 30%). La matrice X est simulée de façon à ce que la probabilité d'avoir une mutation soit de $0,25$ et que la matrice de covariance de X se présente comme suit :

$$\begin{pmatrix} C_1 & 0 & \dots & \dots & 0 \\ 0 & C_2 & 0 & \dots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & C_j & 0 \\ 0 & \dots & \dots & 0 & I_{p_{ind}} \end{pmatrix} \qquad C_i = \begin{pmatrix} 1 & 0.4 & \dots & 0.4 \\ 0.4 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.4 \\ 0.4 & \dots & 0.4 & 1 \end{pmatrix}$$

Avec C_i le i–ème cluster, p_{ind} le nombre de prédicteurs indépendants, et $I_{p_{ind}}$ la matrice identité $p_{ind} \times p_{ind}$.

Critères de sélection et comparaison Pour les méthodes utilisant l'estimateur de Buckley et James combiné à LASSO ou Ridge la pénalité est choisie avec l'AIC. Le critère de comparaison des méthodes est l'erreur de prédiction : $||\mathcal{Y} - X\hat{\beta}||_2$, où $\mathcal{Y}$ est le vecteur réponse non censuré.

2 Résultats

1 Le tableau 1 résume les résultat obtenus. L'estimateur de Buclkey et James combiné à LASSO donne la plus grande erreur de prédiction. Alors que l'estimateur de Buckley et James combiné à Ridge a la plus petite erreur pour les deux scénarios à 50 variable et pour le scénario avec 200 variables et 50 prédicteurs pertinents. Pour le scénario avec 200 variables et 1 prédicteur pertinent LASSO et Ridge obtiennent de meilleurs performances.

TABLE 1 – Résultats simulation 1

	Méthodes	$\|\|\mathcal{Y} - X\hat{\beta}\|\|_2^2$		$\|\|\mathcal{Y} - X\hat{\beta}\|\|_2^2$
p=50	Buckley et James avec Ridge	0.577	**p=50**	0.636
1	Ridge	0.619	**25**	0.659
prédicteurs	Buckley et James avec LASSO	0.679	**prédicteurs**	0.731
	LASSO	0.619		0.659
p=200	Buckley et James avec Ridge	0.274	**p=200**	0.360
1	Ridge	0.093	**50**	0.372
prédicteur	Buckley et James avec LASSO	0.341	**prédicteurs**	0.986
	LASSO	0.093		0.372

2 Les erreurs de prédictions sont pour les 8 scénarios plus faibles pour l'éstimateur de Buckley et James combiné à RIDGE. L'éstimateur de Buckley et James combiné à LASSO obtient une erreur plus faible que LASSO, mais pour les deux derniers scénarios l'erreur obtenu avec Ridge seule est meilleur.

Nombre de variables	% sparsité	% censure	Buckley et James Ridge	Ridge	Buckley et James LASSO	LASSO
50	10%	20%	1.18	1.32	1.42	1.30
50	10%	30%	1.20	1.57	1.43	1.43
50	20%	20%	5.36	5.67	6.82	7.18
50	20%	30%	5.22	5.96	7.26	7.81
200	10%	20%	21.43	25.70	29.00	34.77
200	1 10%	30%	21.52	34.61	30.53	143.27
200	20%	20%	137.22	138.00	209.45	220.97
200	20%	30%	139.47	140.60	232.44	49.06

3 Conclusions

Compte tenu de la complexité de la vraisemblance modélisée, effectuer une régression LASSO à partir de son expression entraine tout de suite des problèmes de minimisations. L'éstimateur de Buckley–James semble une alternative moins couteuse en temps de calculs. Cet estimateur combiné à l'opérateur LASSO donne les résultats les moins intéressant, même dans les simulations où le nombre de coefficients non nuls est faible. On peut expliquer cela par l'algorithme qui permet à l'opérateur après chaque nouvelle estimation de la réponse non censurée de sélectionner de nouveaux coefficients linéaires sans prendre en compte ceux trouvés lors de l'itération précédente. En revanche, l'éstimateur de Buckley–James combiné à Ridge obtient les meilleurs résultats dans les scénarios de la seconde simulation présentée et trois des scénarios de la première. Pour conclure, ces résultats suggèrent que l'estimateur de Buckley–James combiné à une régression linéaire reconstruit la perte d'information due à la censure.

4 Remerciements

Ce travail a été effectué dans le cadre du stage de Master 2 de Nastasia Fouret au sein de l'équipe Biostatistique de l'INSERM U897.

Bibliographie

[1] Aziz N,Wang DQ (2009), A renovated Cook's distance based on the Buckley-James estimate in censored regression.*World Academy of Science, Eigineering and Technology* , 29.
[2] Buckley J, James I (1979), Linear regression with censored data. *Biometrika*, 66(3) :429–436.
[3] Cai T, Huang J, Tian L (2009), Regularized Estimation for the Accelerated Failure Time Model. *Statistical Applications in Genetics and Molecular Biology*,65 :394–404.
[4] Datta S, Le-Rademacher J, Datta S (2007), Predicting Patient Survival from Microarray Data by accelerated Failure Time Modeling Using Partial Least Squares and LASSO. *The International Biometric Society* 63 :259–271.
[5] Huang J, Ma S (2010), Variable selection in the accelerated failure time model via the bridge method. Lifetime Data Anal. *Lifetime Data Anal* 16 :176–195.
[6] Jacqmin-Gadda H, Thiébaut R, Chêne G, Commenges D. (2000), Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics*, 1 :355-368.
[7] Jin Z (2006), On least-squares regression with censored data. *Biometrika*.
[8] Johnson BA (2008), Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society, Series B*.
[9] Johnson BA (2009), On Lasso for censored data. *Electronic Journal of Statistics*.
[10] Johnson BA (2009), Supplementary material to rank based estimation in the l1-regularized partly linear model for censored outcomes with application to integrated analyses of clinical

predictors and gene expression data. *Biostatistics*.
[11] Li G (1995) On nonparametric likeliood ratio estimation of survival probabilities for censored data. *Statistics probability letters*.
[12] Marquardt DW (1963), An algorithm for Least-Squares Estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*.
[13] Pan XR, Zhon M (1999) Using one-parameter sub-family of distributions in empirical likelihood ratio with censored data. *Journal of Statistical Planning and Inference*.
[14] Wang S, Nan B, Zhu J, Beer DG (2008), Doubly Penalized Buckley-James Method for Survival Data with High-Dimensional Covariates. *Biometrics*, 64(1) : 132–140.
[15] Wang X, Song L (2011), Adaptive Lasso Variable Selection for the Accelerated Failure Models. *Communication in Statistics–Theory and Methods*, 40 : 4372–4386.
[16] Zhou M, Li G. (2007) Empirical likelihood analysis of Buckley-James estimator. *Journal of multivariate Analysis*.