clogitLasso: an R package for L1 penalized estimation of conditional logistic regression models

Marta Avalos, Hélène Pouyes

clogitLasso: an R package for L^1 penalized estimation of conditional logistic regression models

M. Avalos[a,b] and **H. Pouyes**[a,c]

[a]INSERM, ISPED, Centre INSERM U897–Epidemiologie–Biostatistique,
F–33000 Bordeaux, France
[b]Univ. Bordeaux, ISPED, Centre INSERM U897–Epidemiologie–Biostatistique,
F–33000 Bordeaux, France
marta.avalos@isped.u-bordeaux2.fr

[a]INSERM, ISPED, Centre INSERM U897–Epidemiologie–Biostatistique,
F–33000 Bordeaux, France
[c]Univ. de Pau,
Pau, France
helene.pouyes@isped.u-bordeaux2.fr

Keywords: lasso, penalized conditional likelihood, matching, epidemiology.

The conditional logistic regression model is the standard tool for the analysis of epidemiological studies in which one or more cases (the event of interest), are individually matched with one or more controls (not showing the event). These situations arise, for example, in matched case–control studies and self–matched case–only studies (such as the case–crossover [1], the case–time–control [2] or the case–case–time–control [3] designs).

Usually, odds ratios are estimated by maximizing the conditional log–likelihood function and variable selection is performed by conventional manual or automatic selection procedures, such as stepwise. These techniques are, however, unsatisfactory in sparse, high-dimensional settings in which penalized methods, such as the lasso (*least absolute shrinkage and selection operator*) [4], have emerged as an alternative. In particular, the lasso and related methods have recently been adapted to conditional logistic regression [5].

The R package *clogitLasso* implements, for small to moderate sized samples (less than $3,000$ observations), the algorithms discussed in [5], based on the stratified discrete-time Cox proportional hazards model and depending on the *penalized* package [6]. For large datasets, *clogitLasso* computes the highly efficient procedures proposed in [7, 8], based on an IRLS (iteratively reweighted least squares) algorithm [9] and depending on the *lassoshooting* package [10]. The most common situations that involve 1:1, 1:M and N:M matching are available.

The talk outlines the statistical methodology behind *clogitLasso* as well as its practical application by means of three real data examples arising from Epidemiology.

References

[1] Maclure, M. (1991). The case–crossover design: a method for studying transient effects on the risk of acute event. *American journal of epidemiology* **133**, 144-153.
[2] Suissa, S. (1995).The case-time-control design. *Epidemiology* **6**, 248-53.
[3] Wang, S., Linkletter, C., Maclure, M., Dore, D., Mor, V., Buka, S., Wellenius, GA. (2011).

Future cases as present controls to adjust for exposure trend bias in case-only studies. *Epidemiology* **22**, 568-74.
[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.
[5] Avalos, M., Grandvalet, Y., Duran-Adroher, N., Orriols, L., Lagarde, E. (2012). Analysis of multiple exposures in the case-crossover design via sparse conditional likelihood. *Stat Med* **15**.
[6] Goeman, J. (2010). L^1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, **52**:70–84.
[7] Avalos, M., Pouyes, H., Grandvalet, Y., Wittkop, L., Orriols, L., Letenneur, L., Lagarde, E. (2012). High–dimensional variable selection in individually matched case-control studies. Technical Report. ISPED, Univ Bordeaux Segalen. Bordeaux, France.
[8] Avalos, M., Orriols, L., Pouyes, H., Grandvalet, Y., Lagarde, E. (2012). Variable selection in the case–crossover design via Lasso with application to a registry-based study of medicinal drugs and driving. Technical Report. ISPED, Univ Bordeaux Segalen. Bordeaux, France.
[9] Green, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, **46**:149–192.
[10] Jörnsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T., Nordlander, B., Sander, C., Gennemark, P., Funa, K., *et al.*. (2011). Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Molecular Systems Biology*, **7**:486.