HAL
archives-ouvertes.fr

# Penalized additive logistic regression for cardiovascular risk prediction

Marta Avalos, Yves Grandvalet, Christophe Ambroise

▶ **To cite this version:**

Marta Avalos, Yves Grandvalet, Christophe Ambroise. Penalized additive logistic regression for cardiovascular risk prediction. 2004, pp.301. inserm-00149854

**HAL Id: inserm-00149854**
**https://www.hal.inserm.fr/inserm-00149854**

Submitted on 11 May 2010

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL author manuscript
Statistics in Health Sciences (2004) 301



# PENALIZED ADDITIVE LOGISTIC REGRESSION FOR CARDIOVASCULAR RISK PREDICTION

**Author's names:** **Avalos, M.* & Grandvalet, Y. & Ambroise, C.**

**Affiliation(s):** HeuDiaSyC Laboratory UMR CNRS 6599, Compiègne University of Technology, France

**Email:** {avalos, grandvalet, ambroise}@hds.utc.fr

**Phone:** +33 (0)3 44 23 44 23; **Fax:** +33 (0)3 44 23 44 77

**Keywords:** Model selection, function estimation, hypertension.

**Topic area of the submission:** Model selection and non parametric hypothesis testing.

Predicting individual risk is needed to target preventive interventions toward people with the highest probability of benefit over a given time period [1]. Any estimate of cardiovascular risk is currently based on the use of statistical models inferred from cohort data with methods such as logistic regression. Logistic regression is a linear classification method that has become a standard analyzing tool for binary responses in medical statistics. Although attractively simple, the logistic model fails in some situations.

1) If the number of prognostic factors is large (with respect to the number of observations) or if they are highly correlated, then the variance of coefficient estimates may be high, leading to prediction inaccuracy.
Subset selection is extensively used to address this difficulty. However, to avoid bias many researchers may be more inclined to use full models which includes well–known predictors. Another way to overcome these obstacles consists in imposing a penalty on large fluctuations of the estimated parameters. The $l_1$–based penalizer is known as the *lasso* (least absolute shrinkage and selection operator) [2]. The lasso estimates a vector of linear regression coefficients by minimizing the residual sum of squares subject to a constraint on the $l_1$–norm of coefficient vector. An interesting feature of the $l_1$–norm constraint is that it shrinks some coefficients and sets others to exactly zero. On the other hand, the smooth form of the constraint leads to a less variable model than that provided by subset selection [3].

2) In real life, effects are generally not linear. When the study exposure is continuous, linear models may not accurately characterize the exposure–response curve.
A generalization of the standard logistic model is the additive logistic model [4]. Suppose we have data $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ are the predictor variables and $y_i$ are the responses. Additive logistic regression replaces $\sum_j x_j\beta_j$ with $\sum_j f_j(x_j)$, where $\beta_j$ is the ordinary least squares estimate and $f_j$ is a non parametric function. In these models, the exposure of interest is modeled flexibly. Furthermore, additivity is preserved and hence, we can examine the predictor effects separately so that, the model is simple to interpret. Nevertheless, since non parametric methods are used to fit $f_j$, variable selection

HAL author manuscript inserm-00149854, version 1

HAL author manuscript inserm-00149854, version 1

becomes more complex than in the linear case: We not only need to select which terms to include in the model, but also how smooth they should be.

The aim of this study is to model parsimoniously the relationship between a binary response and several continuous covariates in the case of possible nonlinearities in the effect of the covariates. We present a new method for variable selection and function estimation in non parametric additive logistic models fitted by cubic smoothing splines: penalized additive logistic regression. The method is based on a generalization of the lasso [5]. Our proposal maximizes the log–likelihood subject to two constraints: one on the $l_1$–norm of linear components of cubic splines coefficients and the other one on the (generalized) $l_1$–norm of nonlinear components of cubic splines coefficients. Discrete covariates appear in the form of linear parametric functions and are penalized by one constraint. Because of their nature, these constraints shrink linear and nonlinear coefficients, some of them going exactly to zero. Hence, they give parsimonious models, select significant variables, and reveal nonlinearities in the effects of predictors.

Our strategy for computing the estimates is to express the usual Newton–Raphson update as an iterative reweighted least squares (IRLS) step, combined with a fixed point algorithm, to resolve the constrained optimization problem, and the backfitting algorithm to fit additive models.

Penalized additive logistic regression is applied to predict the risk of cardiovascular disease in a real database from the INDANA project (Individual Data Analysis of Antihypertensive Intervention Trials) [1]. The data includes hypertensive participants from the control group of 1 controlled trial (SHEP). The extracted dataset consists in 2230 subjects, described by several clinical characteristics and prospectively followed during at least 6 years for incidence of cardiovascular outcomes. The outcome considered here is the 6–year incidence of the endpoint defined by occurrence of cardiovascular death. It is represented in the dataset by a binary variable: occurrence or no occurrence of the outcome event. The performance of penalized additive logistic regression is evaluated using cross–validation and compared to standard logistic regression.

**References:**

[1] Gueyffier, F.; Boutitie, F.; Boissel, J.P; Coope, J.; Cutler, J.; Ekbom, T.; Fagard, R.; Friedman, L.; Perry, H.M., Pocock, S. INDANA: a meta–analysis on individual patient data in hypertension. Protocol and preliminary results. Therapie, 1995, 50(4), 353–362.

[2] Tibshirani, R.J. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, B, 1995, 58(1), 267-288.

[3] Steyerberg, E.W.; Eijkemans, M.J.C.; Harrell, F.E. Jr.; Habbema, J.D.F. Pronostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Statistics in Medicine, 2000, 19, 1059–1079.

[4] Hastie, T.; Tibshirani, R.J. Generalized additive models for medical research. Statistical Methods in Medical Research, 1995, 4, 187–196.

[5] Avalos, M.; Grandvalet, Y.; Ambroise, C. Regularization methods for additive models. In: M.R. Berthold, H.J. Lenz, E. Bradley, R. Kruse and C. Borgelt, Eds.; LNCS Springer; 5th International Symposium on Intelligent Data Analysis, 2003; 509-520.

HAL author manuscript inserm-00149854, version 1