Journal of Applied Probability & Statistics Vol. **3**, No. 2, 195–209

A Comparative Study of Variable Selection Procedures Applied in High Dimensional Medical Problems

C. Koukouvinos

Department of Mathematics, National Technical University of Athens, Zografou, Athens, 15773, Greece. Email: ckoukouv@math.ntua.gr

K. Mylona

Department of Mathematics, National Technical University of Athens, Zografou, Athens, 15773, Greece. Email: kmylona@central.ntua.gr

F. Vonta

Department of Mathematics, National Technical University of Athens, Zografou, Athens, 15773, Greece and Department of Mathematics and Statistics, University of Cyprus, CY-1678 Nicosia, Cyprus. Email: vonta@ucy.ac.cy

Abstract

In health studies, many potential factors are usually introduced to determine an outcome variable. In our study, different statistical methods are applied to analyze trauma annual data, collected by 30 General Hospitals in Greece. The first dataset consists of 1681 observations and 76 factors and the second of 6334 observations and 131 factors, that include demographic, transport and intrahospital data. The statistical methods employed in this work were the nonconcave penalized likelihood methods, SCAD, LASSO, and Hard, the generalized linear logistic regression, and the best subset variable selection, used to detect possible risk factors of death. A variety of different statistical models are considered, with respect to the combinations of factors and the number of observations. A comparative survey reveals differences between results and execution times of each method, and the analysis produces models that identify the significant prognostic factors affecting death from trauma.

Keywords: Variable selection, generalized linear model, nonconcave penalized likelihood, high-dimensional dataset, trauma.

2000 Mathematics Subject Classification: 62P10, 62-07.