Computational strategies for deriving the best-subset regression models
Erricos J. Kontogiorghes\textsuperscript{1}, Cristian Gatu\textsuperscript{2}, Marc Hofmann\textsuperscript{2}, Petko Yanev\textsuperscript{3}
\textsuperscript{1} University of Cyprus and Birkbeck College University of London, UK
e-mail: erricos@dcs.bbk.ac.uk
\textsuperscript{2} University of Neuchâtel, Switzerland
\textsuperscript{3} CNR, Pisa, Italy

Riassunto: An efficient branch-and-bound algorithm for computing the best-subset regression models is proposed. The algorithm avoids the computation of the whole regression tree that generates all possible subset models. It is formally shown that if the branch-and-bound test holds, then the current subtree together with its right-hand side subtrees are cut. This reduces significantly the computational burden of the proposed algorithm when compared to an existing leaps-and-bounds method which generates two trees. Specifically, the proposed algorithm, which is based on orthogonal transformations, outperforms by \textit{O}(n^3) the leaps-and-bounds strategy. The criteria used in identifying the best subsets are based on monotone functions of the residual sum of squares (RSS) such as $R^2$, adjusted $R^2$, mean square error of prediction and $C_p$. Strategies and heuristics which improve the computational performance of the proposed algorithm are investigated. A computationally efficient heuristic version of the branch-and-bound strategy which decides to cut subtrees using a tolerance parameter is proposed. The heuristic algorithm derives models close to the best ones. However, it is shown analytically that the relative error of the RSS, and consequently the corresponding statistic, of the computed subsets is smaller than the value of the tolerance parameter which lies between zero and one. Extensions of the heuristic strategies are investigated. The tolerance is either a function of the submodel size or of the depth of the node in the regression tree. Several tolerance configurations are inspected and shown to perform very well. A weighted directed graph approach to compute the best subset model is proposed. Computational results and experiments on random and real data are presented and analyzed.

Keywords: Subset regression; Least squares; Branch-and-bound algorithm.

References


P. Yanev, C. Gatu, and E.J. Kontoghiorghes. Regression graphs and trees for regression subset selection (under review).