Riccardo De Bin

Department of Mathematics, Uni Oslo, Norway

Title: Detection of influential points as a byproduct of resampling-based variable selection procedures

Abstract: Influential points can cause severe problems when deriving a multivariable regression model. A novel approach to check for such points is proposed, based on the variable inclusion matrix, a simple way to summarize results from resampling-based variable selection procedures. These procedures rely on the variable inclusion matrix, which reports whether a variable (column) is included in a regression model fitted on a pseudo-sample (row) generated from the original data (e.g., bootstrap sample or subsample). The variable inclusion matrix is used to study the variable selection stability, to derive weights for model averaged predictors and in others investigations. Concentrating on variable selection, it also allows understanding whether the presence of a specific observation has an influence on the selection of a variable. From the variable inclusion matrix, indeed, the inclusion frequency (I-frequency) of each variable can be computed only in the pseudo-samples (i.e., rows) which contain the specific observation. When the procedure is repeated for each observation, it is possible to check for influential points through the distribution of the I-frequencies, visualized in a boxplot, or through a Grubbs' test. Outlying values in the former case and significant results in the latter point to observations having an influence on the selection of a specific variable and therefore on the finally selected model. This novel approach is illustrated in two real data examples.