Avoiding the pitfalls of S-estimators with categorical predictors

M. Koller[1]

[1] Seminar für Statistik, Department of Mathematics, ETH Zürich, Switzerland

Keywords: Robust, Regression, robustbase, lmrob, categorical.

Abstract

One major drawback of the initial estimator—an S-estimator—used to calculate the MM-estimate in lmrob (a function in the R-package *robustbase*) was its failure to cope with data with categorical predictors. The algorithm used for the S-estimate, the *fast S* algorithm (M. Salibian-Barrera and V. J. Yohai, 2006), relies on subsampling the data to generate candidates for initial estimates. When some levels of factors have low frequencies, most of the random subsamples will naturally be singular and have to be discarded. Thus an excessively large number of subsamples is required to calculate the S-estimate.

The fast S algorithm, in a nutshell, takes a random sample of the observations of size equal to the number of predictors, solves a least squares problem on this reduced data set and refines the resulting parameter estimate using a redescending M-estimate of regression with simultaneous scale on the whole dataset. This is repeated for a pre-specified number of times. The final S-estimate is then taken to be the one that resulted in the smallest scale estimate.

In case of a singular subsample, the algorithm fails at solving the least squares problem and then has to start again by taking a new subsample. We avoid this by checking the subsample for singularity while building it. We exploit the fact that a quadratic least squares problem is equivalent to solving the linear system of equations and calculate the LU-factorization of the transpose of the matrix. A suitable version of the factorization algorithm is used that processes column after column and detects singularities immediately. In case of singularity, we therefore can simply discard this observation / column and proceed to the next one without having to repeat any of the prior steps. Thus, for generating a random subsample, we first shuffle the observations and then calculate the factorization as just described above, until we reach the desired size of the subsample. We call this scheme *nonsingular subsampling*. This procedure will always yield a nonsingular, random subsample if the design matrix is of full rank. If the random subsample is nonsingular in the first place, then the algorithm is exactly as fast as the simple variant (including the least squares part) and much faster otherwise, because it does not need the restarts.

lmrob, as of version 0.9-0 of robustbase, implements this nonsingular subsampling scheme. This version also adds the possibility of using an M-S estimator (R. A. Maronna and V. J. Yohai, 2000) as initial estimator, which estimates the parameters for the categorical predictors separately by an L1-estimator.

References

M. Koller (to appear). Avoiding the pitfalls of S-estimators with categorical predictors.

M. Salibian-Barrera and V. J. Yohai (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, 15(2), 414–427.

R. A. Maronna and V. J. Yohai (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, 89, 197–214.