On the Bianco-Yohai estimator for high breakdown logistic regression

C. Croux[1], and G. Haesbroeck[2]

[1] K.U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium.
[2] University of Liège (B37), Grande Traverse 12, B-4000 Liège, Belgium

Keywords: Breakdown point, Logistic regression, Robustness.

1 Summary

Bianco and Yohai (1996) proposed a highly robust procedure for estimation of the logistic regression model. The results they obtain were very promising. We complement there study by providing a fast and stable algorithm to compute this estimator. Moreover, we discuss the problem of the existence of the estimator. We make a comparison with other robust estimators by means of a simulation study and examples. A discussion of the breakdown point of robust estimators for the logistic regression model will also be given.

2 The Bianco-Yohai estimator

Most robust estimators for logistic regression are defined as solutions of “M-equations”, robustified versions of the first-order condition of the maximum likelihood estimator (e.g. Kordzakhia, Mishra, and Reiersolmoen 2001). Another approach was taken by Pregibon (1982) who defines a robust estimator as the minimizer of a certain loss function of the sum of deviances of the observations. Taking the identity function as loss function, yields the classical Maximum Likelihood (ML) estimator again. Bianco and Yohai (1996) proposed a corrected version of Pregibons estimator, which they showed to be consistent and asymptotically normal. We believe that this approach is very appealing, and analogous to high breakdown procedures in linear regression.

We propose a stable and quite fast algorithm to compute this estimator. It is based on local improvement steps on the unit sphere. A program for computing the Bianco-Yohai estimator and the associated standard-errors is available. For details we refer to the technical report on www.econ.kuleuven.ac.be/christophe.croux. A feature of this algorithm is that it will also warn us if the estimator is not existing. Non-existence of the estimator depends on the amount of overlap (as defined in Christmann and Rousseeuw 2001) in the data, and can already occur for the ML-estimator, see Albert and Anderson (1984). We derive criteria for the existence of the Bianco-Yohai estimator. We also discuss the problem of the selection of the loss-function.

The Bianco-Yohai estimator has an unbounded influence function in the design variable (which is also true for popular high breakdown estimators, like the LTS estimator, in linear regression). To obtain a bounded influence estimator, weights to downweight leverage points can be introduced, as suggested by Caroll and Pederson (1993). We will work with weights obtained from the Minimum Covariance determinant estimator of Rousseeuw and Van Driessen (1999).

3 Breakdown point for logistic regression

One aim in robust statistics is to build high breakdown point estimators. In linear regression models, the breakdown points of many robust estimators have been calculated. This is not the case for the logistic regression model, where breakdown values are not well established.

Christmann (1994) showed that any sensible estimator in the logistic model, robust or not, will tend to infinity if one *replaces* a certain number of observations to well chosen positions. The *replacement* breakdown point seems therefore not to be appropriate for measuring robustness of estimators in logistic regression. Only in the logistic regression model with large strata it can make sense to compute replacement breakdown points, see e.g. Müller and Neykov (2001). Therefore Künsch, Stefanski and Carroll (1989) proposed to look at what is happening when outliers are *added* to a sample. For the classical ML estimator we could show that it stays uniformly bounded if one adds outliers to the original sample. On the other hand, the norm of the ML-estimator can tend to zero, when adding only a few badly placed outlying observations.

This motivates another definition of the finite sample breakdown point for an estimator in the logistic regression model: one could speak about breakdown not only when the estimator tends to infinity, but also when it tends to zero. Then the ML-estimator has a zero implosion breakdown point. It might be a bit strange to speak of breakdown when the estimator tends to a central point in the parameter space. But a similar phenomenom is seen in the autoregressive model of order one, where the Least Squares estimator is driven to zero in presence of badly placed outliers. This example motivated Genton and Lucas (2000) to introduce a very general notion of breakdown point, which depends on the type of outlier constellation one considers and on a certain badness measure. Applying their definition to the logistic regression model, yields an expression equivalent to the implosion breakdown point.

Exact computation of the breakdown point of the Bianco-Yohai estimator is difficult, but upper and lower bounds can be obtained.

References

A. Albert, and J.A. Anderson (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika,* 71, 1–10.

A. M. Bianco , and V.J. Yohai (1996). Robust Estimation in the Logistic Regression Model. In H. Rieder, editor, *Robust Statistics, Data Analysis, and Computer Intensive Methods*, pp. 17–34; *Lecture Notes in Statistics* **109**, Springer Verlag: New York.

R.J. Carroll, and S. Pederson (1993). On Robust Estimation in the Logistic Regression Model. *Journal of the Royal Statistical Society B,* 55, 693–706.

A. Christmann (1994). Least Median of Weighted Squares in Logistic Regression with Large Strata. *Biometrika,* 81, 413–417.

A. Christmann, and P.J. Rousseeuw (2001). Measuring overlap in binary regression. *Computational Statistics & Data Analysis,* 37, 65–75.

M.G. Genton, and A. Lucas (2000). Comprehensive Definitions of Breakdown Points for Independent and Dependent Observations. Tinbergen Institute, Discussion paper TI 2000-40/2.

N. Kordzakhia, G.D. Mishra, and L. Reiersolmoen, L (2001). Robust estimation in the logistic regression model. *Journal of Statistical Planning and Inference,* 98, 211–223.

H.R. Künsch, L.A. Stefanski, and R.J. Carroll (1989). Conditionally Unbiased Bounded Influence Estimation in General Regression Models, with Applications to Generalized Linear Models. *Journal of the American Statistical Association*, 84, 460–466.

C.H. Müller, and N. Neykov (2001). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference*, to appear.

D. Pregibon (1982). Resistant Fits for some commonly used Logistic Models with Medical Applications. *Biometrics*, 38, 485–498.

P.J. Rousseeuw, and K. Van Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223.