Computational Statistics and Data Analysis 53 (2009) 4095–4105

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust testing in the logistic regression model

Ana M. Bianco [a,b,*], Elena Martínez [a]

[a] Instituto de Cálculo, F. C. E. y N., Universidad de Buenos Aires, Argentina

[b] CONICET, Argentina

ARTICLE INFO

Article history:
Received 19 August 2008
Received in revised form 23 April 2009
Accepted 24 April 2009
Available online 5 May 2009

ABSTRACT

We are interested in testing hypotheses that concern the parameter of a logistic regression model. A robust Wald-type test based on a weighted Bianco and Yohai [Bianco, A.M., Yohai, V.J., 1996. Robust estimation in the logistic regression model. In: H. Rieder (Ed) Robust Statistics, Data Analysis, and Computer Intensive Methods In: Lecture Notes in Statistics, vol. 109, Springer Verlag, New York, pp. 17–34] estimator, as implemented by Croux and Haesbroeck [Croux, C., Haesbroeck, G., 2003. Implementing the Bianco and Yohai estimator for logistic regression. Computational Statististics and Data Analysis 44, 273–295], is proposed. The asymptotic distribution of the test statistic is derived. We carry out an empirical study to get a further insight into the stability of the p-value. Finally, a Monte Carlo study is performed to investigate the stability of both the level and the power of the test, for different choices of the weight function.

1. Introduction

In the binomial regression model we assume that the response variable Y has a Bernoulli distribution such that

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta}), \tag{1}$$

where F is a strictly increasing cumulative distribution function, $\mathbf{X} \in \Re^p$ is the vector of explanatory variables and $\boldsymbol{\beta} \in \Re^p$ is the unknown regression parameter. When

$$F(t) = \frac{\exp(t)}{1 + \exp(t)} \tag{2}$$

we have the logistic regression model, which is the model we will consider from now on. However, our results can be extended to other link functions.

It is well known that the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ can be severely affected by outlying observations. Croux et al. (2002) discuss the breakdown behavior of the MLE in the logistic regression model and show that the MLE breaks down to zero when severe outliers are added to a data set. In the last few decades, a lot of work has been done in order to obtain robust estimates of the parameter in this model and also in the more general framework of generalized linear models. Among others, we can mention the proposals given by Pregibon (1982), Stefanski et al. (1986), Künsch et al. (1989), Morgenthaler (1992), Carroll and Pederson (1993), Christmann (1994) and Bianco and Yohai (1996) and more recently Croux and Haesbroeck (2003) and Bondell (2005, 2008).

* Corresponding address: Instituto de Cálculo, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina. Tel.: +54 11 45763375; fax: +54 11 45763375.

E-mail address: abianco@dm.uba.ar (A.M. Bianco).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2009.04.015

We are interested in testing parametric hypotheses about the regression parameter of the logistic regression model. Robust testing in this setting has received much less attention than robust estimation. Testing procedures based on classical estimates inherit the sensitivity of these estimators to atypical data, in the sense that a small amount of outlying observations can affect the level or the power of the tests. Testing procedures that, under contamination, retain a stable level and also a good power under specified alternatives, are desirable. The works of Heritier and Ronchetti (1994) and Cantoni and Ronchetti (2001) go in this direction. Heritier and Ronchetti (1994) introduce robust tests for a general parametric model, which includes logistic regression. Cantoni and Ronchetti (2001) define robust deviances based on generalizations of quasi–likelihood functions and propose a family of test statistics for model selection in generalized linear models. They also investigate the stability of the asymptotic level under contamination.

In this paper we propose a Wald–type statistic based on a weighted version of the Bianco and Yohai (1996) estimator introduced by Croux and Haesbroeck (2003). Our proposal is a natural robustification of the classical Wald–type test, in the sense that the statistic of the test is a quadratic form based on robust estimators of the regression parameter and its asymptotic covariance matrix. We show that the asymptotic behavior of the proposed test is the same as that of its classical counterpart, that is, central χ^2 under the null hypothesis and noncentral χ^2 under contiguous alternatives.

This paper is organized as follows. In Section 2 we briefly review some estimators related to the weighted estimator introduced by Croux and Haesbroeck (2003) and in Section 3 we state its asymptotic properties. In Section 4 we define the test statistic and we state its asymptotic distribution. In Section 5 we analyze the behavior of the p-values of the classical and proposed statistics when an outlying observation with increasing leverage is added to a data set. By means of a simulation study, we illustrate in Section 6 the performance of the proposed test in terms of both level and power. Finally, in Section 7 we provide some concluding remarks.

2. Preliminaries

Consider the sample $\mathcal{Z}_n = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$, with $\mathbf{Z}_i = (\mathbf{X}_i', Y_i)'$ and

$$P(Y_i = 1|\mathbf{X}_i = \mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta}), \tag{3}$$

where F is given by (2). The maximum likelihood estimator of $\boldsymbol{\beta}$ is defined as

$$\widehat{\boldsymbol{\beta}}_{ML} = \arg\max_{\mathbf{b}} \log L(\mathbf{b}, \mathcal{Z}_n) = \arg\min_{\mathbf{b}} \sum_{i=1}^{n} D(Y_i, \mathbf{X}_i'\mathbf{b}),$$

where $L(\mathbf{b}, \mathcal{Z}_n)$ is the likelihood function and $D(Y_i, \mathbf{X}_i'\mathbf{b}) = -Y_i \ln(F(\mathbf{X}_i'\mathbf{b})) - (1-Y_i)\ln(1-F(\mathbf{X}_i'\mathbf{b}))$ is the deviance component of the i-th observation.

Pregibon (1982) proposes to modify the goal function using a monotone loss function ρ in order to give less weight to those observations which are poorly predicted by the model. Since Y_i are indicator variables, the proposed estimators can be defined as

$$\arg\min_{\mathbf{b}} \sum_{i=1}^{n} Y_i\rho(-\ln(F(\mathbf{X}_i'\mathbf{b}))) + (1-Y_i)\rho(-\ln(1-F(\mathbf{X}_i'\mathbf{b}))).$$

He suggests to use a loss function $\rho(t)$ in the Huber type family, given by

$$\rho(t) = \begin{cases} t & \text{if } t \le c \\ 2(tc)^{\frac{1}{2}} - c & \text{if } t > c, \end{cases}$$

where c is a positive constant. However, the resulting estimators are not consistent. Later on, Bianco and Yohai (1996) introduce a class of M-estimators which are a more robust version of them, and consistent, by including a bias correction term G and using a bounded loss function. These estimators are defined by the minimization of

$$\sum_{i=1}^{n} \phi(Y_i, \mathbf{X}_i'\mathbf{b}), \tag{4}$$

where

$$\phi(y, t) = y\rho(-\ln(F(t))) + (1-y)\rho(-\ln(1-F(t))) + G(F(t)) + G(1-F(t)) - G(1) \tag{5}$$

with ρ a bounded, differentiable and nondecreasing function. The function G is defined as

$$G(t) = \int_0^t \psi(-\ln u)\mathrm{d}u,$$

with $\psi(t) = \rho'(t)$.