Computational Statistics and Data Analysis 65 (2013) 80–97

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Robust tests in generalized linear models with missing responses

Ana M. Bianco [a,*], Graciela Boente [a,*], Isabel M. Rodrigues [b]

[a] *Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Argentina*

[b] *Departamento de Matemática and CEMAT, Instituto Superior Técnico, Technical University of Lisbon (TULisbon), Lisboa, Portugal*

ARTICLE INFO

Article history:
Received 14 August 2011
Received in revised form 6 May 2012
Accepted 10 May 2012
Available online 23 May 2012

Keywords:
Fisher-consistency
Generalized linear models
Influence function
Missing data
Outliers
Robust estimation
Robust testing

ABSTRACT

In many situations, data follow a generalized linear model in which the mean of the responses is modelled, through a link function, linearly on the covariates. Robust estimators for the regression parameter in order to build test statistics for this parameter, when missing data occur in the responses, are considered. The asymptotic behaviour of the robust estimators for the regression parameter is obtained, under the null hypothesis and under contiguous alternatives. This allows us to derive the asymptotic distribution of the robust Wald-type test statistics constructed from the proposed estimators. The influence function of the test statistics is also studied. A simulation study allows us to compare the behaviour of the classical and robust tests, under different contamination schemes. Applications to real data sets enable to investigate the sensitivity of the *p*-value to the missing scheme and to the presence of outliers.

1. Introduction

The generalized linear model (McCullagh and Nelder, 1989), GLM, is a popular technique for modelling a wide variety of data and assumes that the observations $(y_i, \mathbf{x}_i^{\mathrm{T}})$, $1 \leq i \leq n$, $\mathbf{x}_i \in \mathbb{R}^k$, are independent with the same distribution as $(y, \mathbf{x}^{\mathrm{T}}) \in \mathbb{R}^{k+1}$ such that the conditional distribution of $y|\mathbf{x}$ belongs to the canonical exponential family

$$\exp\left\{\left[y\theta(\mathbf{x}) - B\left(\theta(\mathbf{x})\right)\right]/A(\tau) + C(y,\tau)\right\}, \tag{1}$$

for known functions A, B and C. In this situation, if we denote by B' the derivative of B, the mean $\mu(\mathbf{x}) = \mathbb{E}(y|\mathbf{x}) = B'\left(\theta(\mathbf{x})\right)$ is modelled linearly through a known link function, g, i.e., $g(\mu\,(\mathbf{x})) = \theta(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. Robust procedures for generalized linear models have been considered, among others, by Stefanski et al. (1986), Künsch et al. (1989), Bianco and Yohai (1996), Cantoni and Ronchetti (2001), Croux and Haesbroeck (2003) and Bianco et al. (2005); see also, Maronna et al. (2006). Recently, robust tests for the regression parameter under a logistic model were considered by Bianco and Martínez (2009).

In practice, some response variables may be missing by design (as in two-stage studies) or by happenstance. As is well known, the methods proposed by the above mentioned authors are designed for complete data sets and problems arise when missing observations are present. Even if there are many situations in which both the response and the explanatory variables are missing, we will focus our attention on those cases in which missing data occur only in the responses. Actually, missingness of responses is very common in opinion polls, market research surveys, mail enquiries, social-economic investigations, medical studies and other scientific experiments, when the explanatory variables can be controlled. This pattern appears, for example, in the scheme of double sampling proposed by Neyman (1938), where first a complete sample

[*] Correspondence to: Instituto de Cálculo, FCEyN, UBA Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina. Tel.: +54 11 45763375; fax: +54 11 45763375.

E-mail addresses: abianco@dm.uba.ar (A.M. Bianco), gboente@dm.uba.ar, gboente@fibertel.com.ar (G. Boente), irodrig@math.ist.utl.pt (I.M. Rodrigues).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.05.008

is obtained and then, some additional covariate values are computed since, perhaps, this is less expensive than to obtain more response values. Hence, we will focus our attention on robust inference when the response variable may have missing observations, but the covariate $\mathbf{x}$ is totally observed.

In this paper, we consider the robust estimators for the regression parameter $\boldsymbol{\beta}$ introduced by Bianco et al. (2011a), under a GLM model. When there are no missing data, these estimators include the family of estimators previously studied by several authors such as Bianco and Yohai (1996), Cantoni and Ronchetti (2001), Croux and Haesbroeck (2003) and Bianco et al. (2005). It is shown that the robust estimators of $\boldsymbol{\beta}$ are asymptotically normally distributed which allows us to construct a robust procedure to test the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ versus $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$. The paper is organized as follows. The robust proposal is given in Section 2, the asymptotic distribution of the regression estimators and a robust Wald-type test for the regression parameter are provided in Section 3. The results of a Monte Carlo study are summarized in Section 4, while in Section 5 we investigate the empirical breakdown point of the different procedures. The proposed procedure is illustrated over two real data examples in Section 6 where we carried out a sensitivity study for the p-value. An expression for the influence function of the test is obtained in Section 7. Proofs are relegated to the Appendix.

2. Robust inference

2.1. Framework and the robust estimators

Suppose we obtain a random sample of incomplete data $\left(y_i, \mathbf{x}_i^{\mathrm{T}}, \delta_i\right)$, $1 \leq i \leq n$, of a generalized linear model where $\delta_i = 1$ if y_i is observed, $\delta_i = 0$ if y_i is missing and $(y_i, \mathbf{x}_i^{\mathrm{T}}) \in \mathbb{R}^{k+1}$ are such that the conditional distribution $F(\cdot, \mu_i, \tau)$ of $y_i|\mathbf{x}_i$ belongs to the canonical exponential family given in (1), with $\mu_i = H(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ and $\mathrm{VAR}(y_i|\mathbf{x}_i) = A^2(\tau)V(\mu_i) = A^2(\tau)B''\left(\theta(\mathbf{x}_i)\right)$ with B'' the second derivative of B. Let $(\boldsymbol{\beta}, \tau)$ denote the true parameter values and $\mathbb{E}_F$ the expectation under the true model; thus $\mathbb{E}_F(y|\mathbf{x}) = H(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. In a more general situation, we will think of τ as a nuisance parameter such as an additional scale or dispersion parameter or even, the tuning constant for the score function to be considered below. For instance, under a Gamma regression model τ is related to the shape parameter, while for Poisson and logistic regression, $\tau = 1$.

Let $(y, \mathbf{x}^{\mathrm{T}}, \delta)$ be a random vector with the same distribution as $\left(y_i, \mathbf{x}_i^{\mathrm{T}}, \delta_i\right)$. Bianco et al. (2011a) defined robust estimators of the regression parameter when missing responses occur under an ignorable missing mechanism. To be more precise, they assumed that y is missing at random (MAR), that is, δ and y are conditionally independent given $\mathbf{x}$, i.e.,

$$P\left(\delta = 1|(y, \mathbf{x})\right) = P\left(\delta = 1|\mathbf{x}\right) = p\left(\mathbf{x}\right). \tag{2}$$

A common assumption in the literature states that $\inf_{\mathbf{x}} p\left(\mathbf{x}\right) > 0$, meaning that at any value of the covariate response variables are observed.

For the sake of completeness, we remind the definition of the simplified estimators considered in Bianco et al. (2011a) where also a propensity score approach is considered. Through a heuristic argument based on the influence function, Bianco et al. (2011a) showed that in some situations, such as the Gamma model to be considered below, the asymptotic variance of the robust simplified estimators is smaller than that of the propensity score ones. For that reason, we will focus here on test statistics based on the robust simplified estimators.

Let $w_1 : \mathbb{R}^k \to \mathbb{R}$ be a weight function to control leverage points on the carriers $\mathbf{x}$ and $\rho : \mathbb{R}^3 \to \mathbb{R}$ a loss function. For any $\mathbf{b} \in \mathbb{R}^k$, $t \in \mathbb{R}$, let us define

$$S_n(\mathbf{b}, t) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x}_i), \tag{3}$$

$$S(\mathbf{b}, t) = \mathbb{E}_F\left[\delta\rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\right] = \mathbb{E}_F\left[p(\mathbf{x})\rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\right]. \tag{4}$$

In order to define Fisher-consistent estimators, Bianco et al. (2011a) assumed that $w_1(\cdot)$ and $\rho(\cdot)$ are such that, $S(\boldsymbol{\beta}, \tau) = \min_{\mathbf{b}} S(\mathbf{b}, \tau)$. As mentioned above, the parameter t in $S(\mathbf{b}, t)$ plays the role of a nuisance parameter.

Let $\widehat{\tau} = \widehat{\tau}_n$ be robust consistent estimators of τ, the *robust simplified estimator* $\widehat{\boldsymbol{\beta}}$ of the regression parameter is defined as

$$\widehat{\boldsymbol{\beta}} = \underset{\mathbf{b}}{\operatorname{argmin}}\, S_n(\mathbf{b}, \widehat{\tau}). \tag{5}$$

Under mild conditions the consistency of $\widehat{\boldsymbol{\beta}}$ is derived in Bianco et al. (2011a).

When ρ is continuously differentiable, if we denote by $\Psi\left(y, u, t\right) = \partial\rho(y, u, t)/\partial u$, then $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ satisfy the differentiated equations $S^{(1)}(\boldsymbol{\beta}, \tau) = \mathbf{0}_k$ and $S_n^{(1)}(\mathbf{b}, \widehat{\tau}) = \mathbf{0}_k$, respectively, where $S^{(1)}(\mathbf{b}, t) = \mathbb{E}_F\left(\Psi\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})p(\mathbf{x})\mathbf{x}\right)$ and $S_n^{(1)}(\mathbf{b}, t) = (1/n)\sum_{i=1}^{n} \delta_i\Psi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x}_i)\mathbf{x}_i$.

Remark 2.1.1. Two classes of loss functions ρ have been considered in the literature. One of them aims to bound the deviances, while the other one introduced by Cantoni and Ronchetti (2001) bounds the Pearson residuals. In both cases, the correction term needed to ensure Fisher-consistency is included in the function ρ. For a complete description,