Journal of Statistical Planning and Inference 139 (2009) 3038 – 3049

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Selecting and checking a binomial response generalized linear model using phi-divergences☆

J.A. Pardo*, M.C. Pardo

Department of Statistics and O.R. (I), Complutense University of Madrid, Spain

ARTICLE INFO

Article history:
Received 18 January 2008
Received in revised form
15 February 2009
Accepted 17 February 2009
Available online 5 March 2009

Keywords:
Generalized linear model
Binary data
ϕ-Divergence
Backward procedure
Diagnostics

ABSTRACT

This work introduces specific tools based on phi-divergences to select and check generalized linear models with binary data. A backward selection criterion that helps to reduce the number of explanatory variables is considered. Diagnostic methods based on divergence measures such as a new measure to detect leverage points and two indicators to detect influential points are introduced. As an illustration, the diagnostics are applied to human psychology data.

1. Introduction

The test procedures in the linear regression model are based on the normal distribution of the error variable and thus on a normal distribution of the endogenous variable Y. However, in many fields of application this assumption may not be true. Nelder and Wedderburn (1972) proposed a general approach to fit linear models when the random error (and hence the response Y) belongs to a general very flexible class of distributions—the exponential family. These generalized models consist of three components:

1. The random component, which specifies the probability distribution of the response variable, Y. Let $i = 1, \ldots, I$ be an index running over all distinct combinations of the explanatory variables, such that $\boldsymbol{x}_i = (x_{i0}, \ldots, x_{ik})$ is a typical combination of observed explanatory variables. In our case, we shall assume that for each value of the explanatory variable $\boldsymbol{x}_i = (x_{i0}, \ldots, x_{ik})$ we have a binomial random variable $Y_i \equiv \sum_{j=1}^{n_i} Z_j$ (the random variables $Z_1, \ldots, Z_{n_i}$ are a random sample from a binary random variable Z that takes either the value 1 or the value 0, generally referred to as "success" or "failure", respectively) with parameters n_i and $\pi_i = P(Z = 1/\boldsymbol{x}_i)$. The value y_{i1} will represent the number of "successes".
2. The systematic component, which specifies a linear function of the explanatory variables. This relates a vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_I)^T$ such that

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}.$$

$\boldsymbol{\eta}$ is called the linear predictor, $\boldsymbol{X}$ is the $I \times (k+1)$ matrix with rows $\boldsymbol{x}_i$, $i = 1, \ldots, I$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^T$ is a $(k+1) \times 1$ vector of unknown parameters. We shall also assume that $rank(\boldsymbol{X}) = k + 1$.

☆ This work was supported in parts by Grants MTM2006-06872 and UCM2007-910707.

* Corresponding author.
E-mail addresses: japardo@Mat.ucm.es (J.A. Pardo), mcapardo@Mat.ucm.es (M.C. Pardo).

0378-3758/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.02.014

3. The link function g, which describes a functional relationship between π_i and the $k+1$ explanatory variables $\boldsymbol{x}_i = (x_{i0}, \ldots, x_{ik})$ through the linear predictor

$$\eta_i \equiv g(\pi_i) = \sum_{j=0}^{k} x_{ij}\beta_j, \quad i = 1, \ldots, I,$$

where g is a monotonic and differentiable function. We shall assume $x_{i0} = 1$, $i = 1, \ldots, I$.

In this paper, we turn our attention to select and check a binomial response generalized linear model (GLM). There are two competing goals: the model should be complex enough to fit the data well, and, on the other hand, it should be simple to interpret, smoothing rather than overfitting the data. In Section 2, we discuss a backward strategy for model selection based on ϕ-divergence test statistics. These test statistics use the minimum ϕ-divergence estimator, a natural generalization of the maximum likelihood estimator (MLE) for the GLM, which is given by

$$\widehat{\boldsymbol{\beta}}_\phi \equiv \arg \min_{\beta_0,\beta_1,\ldots,\beta_k} D_\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})),$$

where

$$\widehat{\boldsymbol{p}} \equiv (\hat{p}_{11}, \hat{p}_{12}, \ldots, \hat{p}_{I1}, \hat{p}_{I2})^T = \left(\frac{y_{11}}{N}, \frac{y_{12}}{N}, \frac{y_{21}}{N}, \frac{y_{22}}{N}, \ldots, \frac{y_{I1}}{N}, \frac{y_{I2}}{N}\right)^T$$

and

$$\begin{aligned}\boldsymbol{p}(\boldsymbol{\beta}) &\equiv (p_{11}(\boldsymbol{\beta}), p_{12}(\boldsymbol{\beta}), \ldots, p_{I1}(\boldsymbol{\beta}), p_{I2}(\boldsymbol{\beta}))^T \\ &= \left(\pi(\boldsymbol{x}_1^T\boldsymbol{\beta})\frac{n_1}{N}, (1-\pi(\boldsymbol{x}_1^T\boldsymbol{\beta}))\frac{n_1}{N}, \ldots, \pi(\boldsymbol{x}_I^T\boldsymbol{\beta})\frac{n_I}{N}, (1-\pi(\boldsymbol{x}_I^T\boldsymbol{\beta}))\frac{n_I}{N}\right)^T,\end{aligned}$$

with $y_{11}, \ldots, y_{I1}$ the observed values of the binomial random variables $Y_1, \ldots, Y_I$, $y_{i2} = n_i - y_{i1}$, $N = \sum_{j=1}^{2}\sum_{i=1}^{I} y_{ij}$ and $\pi(\boldsymbol{x}_i^T\boldsymbol{\beta}) = \pi_i$. The ϕ-divergence is defined by

$$D_\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})) \equiv \sum_{j=1}^{2}\sum_{i=1}^{I} p_{ij}(\boldsymbol{\beta})\phi\left(\frac{\widehat{p}_{ij}}{p_{ij}(\boldsymbol{\beta})}\right), \quad \phi \in \Phi, \tag{1}$$

where Φ is the class of all convex functions such that $\phi(1) = \phi'(1) = 0$, $\phi''(1) > 0$, $0\phi(\frac{0}{0}) = 0$ and $0\phi(p/0) = p\lim_{u\to\infty}\phi(u)/u$. For more details about ϕ-divergence see Vajda (1989) and Pardo (2006).

Pardo and Pardo (2008a) established that, under the assumption that $\phi(t)$ is twice differentiable at $t > 0$, the asymptotic distribution of the minimum ϕ-divergence estimator for GLM, $\widehat{\boldsymbol{\beta}}_\phi$, is normal with mean zero and covariance matrix given by

$$(\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^0)\boldsymbol{X})^{-1}, \tag{2}$$

where

$$\boldsymbol{W}(\boldsymbol{\beta}) = Diag\left(\left(\frac{n_i}{\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})(1-\pi(\boldsymbol{x}_i^T\boldsymbol{\beta}))}\left(\frac{\partial\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\partial\eta_i}\right)^2\right)_{i=1,\ldots,I}\right)$$

and $\boldsymbol{\beta}^0$ is the true value of the parameter $\boldsymbol{\beta}$. By $Diag((a_i)_{i=1,\ldots,I})$ we are denoting the diagonal matrix with elements $(a_i)_{i=1,\ldots,I}$ in the diagonal.

Other well-known model selection criterion based on a divergence measure is the Akaike Information Criterion (AIC) introduced by Akaike (1973). Following the early work of Akaike, Karagrigoriou and Mattheou (2009) propose the Divergence Information Criterion (DIC) based on Basu et al. (1998) power divergence (Basu et al., 1998).

After choosing a preliminary model, model checking addresses whether systematic lack of fit exists. The regression diagnostics introduced by Pregibon (1981) for the dichotomous logistic model are extended, in Section 3, to GLM and generalized using the ϕ-divergence defined in (1) for model checking. We also develop some new diagnostics. In Section 4 we illustrate the diagnostic methods introduced with a numerical example.

2. Variable selection

Variable selection methods aim at determining submodels with a moderate number of parameters that still fit the data adequately. For large models and large data sets, stepwise procedures are an useful additional tool. We describe backward stepwise selection based on a new statistic. Starting from a maximal model M, i.e., the GLM with all available explanatory