Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Statistical Planning and Inference 136 (2006) 982–1006

ELSEVIER

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Testing in logistic regression models based on ϕ-divergences measures☆

Julio Angel Pardo, Leandro Pardo*, María del Carmen Pardo

Department of Statistics and O.R., Complutense University of Madrid, Plaza de Ciencias, 3, Madrid 28040, Spain

Received 29 June 2003; accepted 17 August 2004
Available online 8 October 2005

Abstract

In this paper, we consider inference based on very general divergence measures under assumptions of a logistic regression model. We use the minimum ϕ-divergence estimator in a ϕ-divergence statistic, which is the basis of some new statistics, for solving the classical problems of testing in a logistic regression model. A diagnostic analysis is developed based on the new estimators and test statistics.

MSC: 62B10; 62H15; 62J12; 62J20

Keywords: Logistic regression model; ϕ-divergence measure; Goodness-of-fit tests; Model diagnostics

1. Introduction

A common name for a regression model for categorical response variables is logistic regression model since the regression part of the models, i.e., a linear combination of the values of the explanatory variables and the regression coefficients, is a logistic transformation of the probabilities of the response categories. The usefulness of this transformation lies in the fact, that it transforms the interval between 0 and 1 on to the real axis $(-\infty, \infty)$. The logistic regression model assumes that we have a vector of independent random variables, $Y_1, \ldots, Y_I$, such that Y_i is distributed as a Binomial with parameters n_i and π_i, $i = 1, \ldots, I$.

☆ This work was supported partially by Grant DGI (BMF2003-00892).

* Corresponding author. Tel.: +34 91394 4425; fax: +34 91394 4606.
E-mail address: leandro_pardo@mat.ucm.es (Leandro Pardo).

doi:10.1016/j.jspi.2004.08.008

We assume that π_i depends on $k+1$ explanatory variables and $k+1$ regression coefficients, i.e.,

$$\pi_i = \frac{\exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}, \quad i = 1, \ldots, I, \tag{1}$$

where π_i is the inverse to the common logit function

$$\operatorname{logit}(\pi_i) = \log(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

Let $\beta = (\beta_0, \ldots, \beta_k)^{\mathrm{T}}$ be a $(k+1) \times 1$ vector of unknown parameters with $\beta_i \in (-\infty, \infty)$. We will assume $x_{i0} = 1$, $i = 1, \ldots, I$ and we denote by $x_i = (x_{i0}, \ldots, x_{ik})$ and by X the $I \times (k+1)$ matrix with rows x_i, $i = 1, \ldots, I$. We also shall assume that $\operatorname{rank}(X) = k+1$.

If we denote by $n_{11}, \ldots, n_{I1}$ the observed values of the random variables $Y_1, \ldots, Y_I$ and $\pi(x_i^{\mathrm{T}}\beta) = \pi_i$ it is well-known that the likelihood function for the logistic regression model is given by

$$L\left(\beta_0, \ldots, \beta_k\right) = \prod_{i=1}^{I} \binom{n_i}{n_{i1}} \pi\left(x_i^{\mathrm{T}}\beta\right)^{n_{i1}} (1 - \pi\left(x_i^{\mathrm{T}}\beta\right))^{n_i - n_{i1}}$$

so the point maximum likelihood estimator (MLE), $\widehat{\beta}$, is obtained minimizing almost surely over

$$\Theta = \{\left(\beta_0, \ldots, \beta_k\right) : \beta_i \in (-\infty, \infty),\ i = 0, \ldots, k\}$$

the expression

$$\sum_{j=1}^{2} \sum_{i=1}^{I} \frac{n_{ij}}{N} \log \frac{\frac{n_{ij}}{N}}{\pi_{ij} \frac{n_i}{N}},$$

where $\pi_{i1} = \pi(x_i^{\mathrm{T}}\beta)$, $\pi_{i2} = 1 - \pi(x_i^{\mathrm{T}}\beta)$, $n_{i2} = n_i - n_{i1}$, $(i = 1, \ldots, I)$ and $N = \sum_{i=1}^{I} n_i$. Therefore, the MLE, see Pardo et al. (2004), can be defined by

$$\widehat{\beta} = \arg \min_{\beta_0, \beta_1, \ldots, \beta_k} D_{\text{Kullback}}\left(\widehat{p}, p(\beta)\right),$$

where

$$\widehat{p} = \left(\frac{n_{11}}{N}, \frac{n_{12}}{N}, \frac{n_{21}}{N}, \frac{n_{22}}{N}, \ldots, \frac{n_{I1}}{N}, \frac{n_{I2}}{N}\right)^{\mathrm{T}},$$

$$\begin{aligned} p(\beta) &\equiv (p_{11}(\beta), p_{12}(\beta), \ldots, p_{I1}(\beta), p_{I2}(\beta))^{\mathrm{T}} \\ &= \left(\pi(x_1^{\mathrm{T}}\beta)\frac{n_1}{N}, (1 - \pi(x_1^{\mathrm{T}}\beta))\frac{n_1}{N}, \ldots, \pi(x_I^{\mathrm{T}}\beta)\frac{n_I}{N}, (1 - \pi(x_I^{\mathrm{T}}\beta))\frac{n_I}{N}\right)^{\mathrm{T}} \end{aligned} \tag{2}$$

and

$$D_{\text{Kullback}}(p, q) = \sum_{i=1}^{k} p_i \log \frac{p_i}{q_i}$$