DIVERGENCE MEASURES AND LOGISTIC REGRESSION MODELS

Leandro Pardo

Abstract. In this paper we present a review of some results about inference based on ϕ-divergence measures, under assumptions of logistic regression model . The minimum ϕ-divergence estimator, which is seen to be a generalization of the maximum likelihood estimator is considered. This estimator is used in a ϕ-divergence measure which is the basis of new statistics for solving some important problems regarding logistic regression models: fitting the logistic regression model, residuals and dimensional reduction. Finally, an extension is presented when we consider a multinomial response instead a binary response.

Keywords: Chi-square distribution, Logistic regression models, Minimum ϕ-divergence estimator, Goodness-of fit tests

AMS classification: 62B10, 62J15

§1. Introduction

Let Y_i, $i = 1, .., I$, be independent binomial random variables with parameters n_i and $\boldsymbol{p}_i$, $i = 1, ..., I$, we denote by n_{i1} the number of "succes". We assume that $\boldsymbol{p}_i \equiv \pi\left(\boldsymbol{x}_i^T\beta\right)$, where $\boldsymbol{x}_i = (1, x_{i1,}..., x_{ik})^T$ is a vector of $k+1$ explanatory variables and the dependence of them with $\boldsymbol{p}_i$ is given by,

$$\pi\left(\boldsymbol{x}_i^T\beta\right) = \exp\left(\boldsymbol{x}_i\beta\right) / \left(1 + \exp\left(\boldsymbol{x}_i\beta\right)\right),$$

where $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$ is a $k+1$ vector of unknown regression coefficients. The parameter space is given by

$$\Theta = \{(\beta_0, \beta_1, ..., \beta_k) : \beta_i \in (-\infty, \infty),\ i = 0, ..., k\} \tag{1}$$

and the loglikelihood function for β by

$$\begin{aligned}\log L(\beta) = &\ \log \prod_{i=1}^{I} \tbinom{n_i}{n_{i1}} \pi\left(\boldsymbol{x}_i^T\beta\right)^{n_{i1}} \left(1 - \pi\left(\boldsymbol{x}_i^T\beta\right)\right)^{n_i - n_{i1}} \\ = &\ c - \left\{\sum_{i=1}^{I}\left(\frac{n_{i1}}{N}\log\frac{n_{i1}}{\pi\left(\boldsymbol{x}_i^T\beta\right)n_i} + \frac{n_i - n_{i1}}{N}\log\frac{n_i - n_{i1}}{\left(1 - \pi\left(\boldsymbol{x}_i^T\beta\right)\right)n_i}\right)\right\},\end{aligned}$$

where $N = n_1 + ... + n_I$.

We consider the two following probability vectors

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

and

$$\boldsymbol{p}(\beta) \equiv \left(\pi\left(\boldsymbol{x}_1^T\beta\right)\frac{n_1}{N}, \left(1-\pi\left(\boldsymbol{x}_1^T\beta\right)\right)\frac{n_1}{N}, ..., \pi\left(\boldsymbol{x}_I^T\beta\right)\frac{n_I}{N}, \left(1-\pi\left(\boldsymbol{x}_I^T\beta\right)\right)\frac{n_I}{N}\right)^T.$$

It is immediate to get

$$\log L(\beta) = c - D_{Kull}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\beta)\right),$$

where by $D_{Kull}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\beta)\right)$ we are denoting the Kullback-Leibler divergence between the probability vectors $\widehat{\boldsymbol{p}}$ and $\boldsymbol{p}(\beta)$. Then the maximum likelihood estimator of β can be defined as

$$\widehat{\beta} = \arg\min_{\beta_0,\beta_1,...,\beta_k} D_{Kullback}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\beta)\right). \tag{2}$$

§2. Minimum ϕ-divergence estimator

The Kullback-Leibler divergence measure is a particular case of the $\phi-$ divergence measures defined by Csiszár (1963) and Ali and Silvey (1966), as

$$\begin{aligned} D_\phi\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\beta)\right) = & \left\{\sum_{i=1}^{I}\left(\frac{\pi\left(\boldsymbol{x}_i^T\beta\right)n_i}{N}\phi\left(\frac{n_{i1}}{\pi\left(\boldsymbol{x}_i^T\beta\right)n_i}\right)\right.\right. \\ + & \left.\left.\frac{\left(1-\pi\left(\boldsymbol{x}_i^T\beta\right)\right)n_i}{N}\phi\left(\frac{n_i-n_{i1}}{\left(1-\pi\left(\boldsymbol{x}_i^T\beta\right)\right)n_i}\right)\right)\right\}, \end{aligned} \tag{3}$$

where $\phi \in \Phi^*$, being Φ^* the class of all convex functions $\phi(x)$, $x > 0$, such that at $x = 1, \phi(1) = 0, \phi''(1) > 0$, and at $x = 0$, $0\phi(0/0) = 0$ and $0\phi(p/0) = \lim_{u\to\infty}\phi(u)/u$. For every $\phi \in \Phi^*$ that is differentiable at $x = 1$, the function $\psi(x) \equiv \phi(x) - \phi'(1)(x-1)$ also belongs to Φ^*. Then we have $D_\psi\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\beta)\right) = D_\phi\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\beta)\right)$, and ψ has the additional property that $\psi'(1) = 0$. Because the two divergence measures are equivalent, we can consider the set Φ^* to be equivalent to the set $\Phi \equiv \Phi^* \cap \{\phi : \phi'(1) = 0\}$. In what follows, we give our theoretical results for $\phi \in \Phi$, but we often apply them to choices of functions in Φ^*.

Based on (2) and on the definition of $D_\phi\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\beta)\right)$, in Pardo et al (2003a) **,** it was defined and studied the minimum ϕ-divergence estimator of β. This estimator is defined by

$$\widehat{\beta}^\phi = \arg\min_{\alpha,\beta_1,...,\beta_k} D_\phi\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\beta)\right).$$

For $\phi(x) = x\log x - x + 1$ we obtain the maximum likelihood estimator and for $\phi(x) = \frac{1}{2}(x-1)^2$ the minimum chi-squared estimator.

If we denote by $\boldsymbol{X}$ the $I \times (k+1)$ matrix with rows $\boldsymbol{x}_i$, $i = 1, ..., I$ and we assume that $rank(\boldsymbol{X}) = k+1$, in the cited paper of Pardo et al (2003a) it was established that under the assumption that ϕ is twice continuously differentiable in a neighborhood of 1 that

$$\begin{aligned} \widehat{\beta}^\phi = & \beta^0 + \left(\boldsymbol{X}^T Diag\left(\left(\tfrac{n_i}{N}\right)\pi\left(\boldsymbol{x}_i^T\beta^0\right)\left(1-\pi\left(\boldsymbol{x}_i^T\beta^0\right)\right)_{i=1,...,I}\right)\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T \\ \times & Diag\left(\left(\boldsymbol{C}_i^T\right)_{i=1,...,I}\right) Diag\left(\boldsymbol{p}\left(\beta^0\right)^{-1/2}\right)\left(\widehat{\boldsymbol{p}} - \boldsymbol{p}\left(\beta^0\right)\right) + o\left(\left\|\widehat{\boldsymbol{p}} - \boldsymbol{p}\left(\beta^0\right)\right\|\right), \end{aligned}$$

where

$$\boldsymbol{C}_i = \left(\left(\frac{n_i}{N} \right) \pi \left(\boldsymbol{x}_i^T \beta^0 \right) \left(1 - \pi \left(\boldsymbol{x}_i^T \beta^0 \right) \right) \right)^{1/2} \begin{pmatrix} \left(1 - \pi \left(\boldsymbol{x}_i^T \beta^0 \right) \right)^{1/2} \\ -\pi \left(\boldsymbol{x}_i^T \beta^0 \right) \end{pmatrix}, i = 1, ..., I$$

and also, under the assumption that $n_i/n \to \lambda_i$ when $n_i \to \infty$, we have

$$\sqrt{N} \left(\widehat{\beta}^{\phi} - \beta^0 \right) \underset{N \to \infty}{\overset{L}{\to}} N \left(\mathbf{0}, \left(\boldsymbol{X}^T Diag \left(\lambda_i \pi \left(\boldsymbol{x}_i^T \beta^0 \right) \left(1 - \pi \left(\boldsymbol{x}_i^T \beta^0 \right) \right) \right)_{i=1,..,I} \boldsymbol{X} \right)^{-1} \right) \quad (4)$$

being $\lambda_i = \lim_{N \to \infty} n_i/N$.

From (4) it is clear that

$$\widehat{Cov \left(\widehat{\beta}^{\phi} \right)} \approx \frac{1}{N} \left(\boldsymbol{X}^T Diag \left(\lambda_i \pi \left(\boldsymbol{x}_i^T \widehat{\beta}^{\phi} \right) \left(1 - \pi \left(\boldsymbol{x}_i^T \widehat{\beta}^{\phi} \right) \right) \right)_{i=1,..,I} \boldsymbol{X} \right)^{-1}.$$

§3. Fitting the Logistic Regression Model: Residuals

We denote by $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$ the maximum Likelihood estimators of $\beta_0, \beta_1, ..., \beta_k$. If we estimate $\pi \left(\boldsymbol{x}_i^T \beta \right)$ by the maximum likelihood estimator $\pi \left(\boldsymbol{x}_i^T \widehat{\beta} \right)$, the classical Pearson chi-squared statistic is given by

$$X^2 = \sum_{i=1}^{I} \frac{\left(n_{i1} - n_i \pi \left(\boldsymbol{x}_i^T \widehat{\beta} \right) \right)^2}{n_i \pi \left(\boldsymbol{x}_i^T \widehat{\beta} \right) \left(1 - \pi \left(\boldsymbol{x}_i^T \widehat{\beta} \right) \right)}$$

and the likelihood ratio test statistic by

$$D^2 = \sum_{i=1}^{I} 2 \left\{ n_{i1} \log \frac{n_{i1}}{n_i \pi \left(\boldsymbol{x}_i^T \widehat{\beta} \right)} + (n_i - n_{i1}) \log \frac{n_i - n_{i1}}{n_i \left(1 - \pi \left(\boldsymbol{x}_i^T \widehat{\beta} \right) \right)} \right\}.$$

If now we consider the minimum ϕ_2-divergence estimators $\widehat{\beta}_0^{\phi_2}, \widehat{\beta}_1^{\phi_2}, ..., \widehat{\beta}_k^{\phi_2}$ of parameters $\beta_0, \beta_1, ..., \beta_k$, instead of the maximum likelihood estimators, we can estimate $\pi \left(\boldsymbol{x}_i^T \beta \right)$ by $\pi \left(\boldsymbol{x}_i^T \widehat{\beta}^{\phi_2} \right)$. Pardo et al (2003b) have considered and studied the ϕ_1-divergence test statistic based on the minimum ϕ_2-divergence estimator $\widehat{\beta}^{\phi_2}$, given by

$$\begin{aligned} T_{\phi_1,\phi_2} = & \frac{2N}{\phi_1''(1)} D_{\phi_1} \left(\widehat{\boldsymbol{p}}, \boldsymbol{p} \left(\widehat{\beta}^{\phi_2} \right) \right) \\ = & \frac{2}{\phi_1''(1)} \sum_{i=1}^{I} n_i \left\{ \pi \left(\boldsymbol{x}_i^T \widehat{\beta}^{\phi_2} \right) \phi \left(\frac{n_{i1}}{\pi \left(\boldsymbol{x}_i^T \widehat{\beta}^{\phi_2} \right) n_i} \right) \right. \\ + & \left. \left(1 - \pi \left(\pi \left(\boldsymbol{x}_i^T \widehat{\beta}^{\phi_2} \right) \right) \right) \phi \left(\frac{n_{i2}}{\left(1 - \pi \left(\boldsymbol{x}_i^T \widehat{\beta}^{\phi_2} \right) \right) n_i} \right) \right\}. \end{aligned} \quad (5)$$

It is interesting to observe that for

$$\phi_2(x) = x \log x - x + 1 \quad \text{and} \quad \phi_1(x) = \frac{1}{2}(x-1)^2,$$

we obtain that $T_{\phi_1,\phi_2} \equiv X^2$ and for

$$\phi_2(x) = x\log x - x + 1 \quad \text{and} \quad \phi_1(x) = x\log x - x + 1,$$

we get $T_{\phi_1,\phi_2} \equiv D^2$. The following theorem presents the asymptotic distribution of the family of test statistics T_{ϕ_1,ϕ_2}.

Theorem 1. *Suppose that the data Y_i, $i = 1, ..., I$ are binomially distributed with parameters n_i and $\pi\left(\boldsymbol{x}_i^T\beta\right)$. Choose functions ϕ_1 and $\phi_2 \in \Phi$ and twice continuously differentiable in a neighborhood of 1. Under the hypothesis $\boldsymbol{p} = \boldsymbol{p}(\beta)$ and assuming that $n_i/n \to \lambda_i > 0$ when $n_i \to \infty$, the test statistic T_{ϕ_1,ϕ_2} has a chi-squared distribution with $I - (k+1)$ degrees of freedom.*

Based on this theorem, if the sample sizes are large enough, one can use the asymptotic quantile $\chi^2_{I-(k+1),1-\alpha}$, defined by the equation $P\left(\chi^2_{M-1} \leq \chi^2_{M-1,1-\alpha}\right) = 1-\alpha$, to propose the decision rule:

"Reject $H_{Null} : \boldsymbol{p} = \boldsymbol{p}(\beta)$ if $T_{\phi_1,\phi_2} > \chi^2_{I-(k+1),1-\alpha}$".

The statistics X^2, D^2 and T_{ϕ_1,ϕ_2} provide a single number which summarizes the agreement of observed and fitted values. The advantage (as well as the disadvantage) of these statistics is that a single number is used to summarize considerable information. Additional diagnostic analyses are necessary to describe the nature of one lack of fit. Residual comparing observed and fitted counts are useful for this purpose. From a classical point of view we have the residual based on the Pearson's test statistic X^2, given by

$$e_i = \frac{n_{i1} - n_i\pi\left(\boldsymbol{x}_i^T\widehat{\beta}\right)}{\sqrt{n_i\pi\left(\boldsymbol{x}_i^T\widehat{\beta}\right)\left(1-\pi\left(\boldsymbol{x}_i^T\widehat{\beta}\right)\right)}},$$

and residual based in the likelihood ratio test given by

$$d_i = sig\left(n_{i1} - n_i\pi\left(\boldsymbol{x}_i^T\widehat{\beta}\right)\right)\sqrt{2\left\{n_{i1}\log\frac{n_{i1}}{n_i\pi\left(\boldsymbol{x}_i^T\widehat{\beta}\right)} + (n_i - n_{i1})\log\frac{n_i - n_{i1}}{n_i\left(1-\pi\left(\boldsymbol{x}_i^T\widehat{\beta}\right)\right)}\right\}}.$$

In the same way we can define ϕ_1-residuals based on minimum ϕ_2-divergence estimator by

$$\begin{aligned} c_i^{\phi_1,\phi_2} = \; & sig\left(n_{i1} - n_i\pi\left(\boldsymbol{x}_i^T\widehat{\beta}^{\phi_2}\right)\right)\sqrt{\frac{2n_i}{\phi_1''(1)}}\left\{\pi\left(\boldsymbol{x}_i^T\widehat{\beta}^{\phi_2}\right)\phi_1\left(\frac{n_{i1}}{\pi\left(\boldsymbol{x}_i^T\widehat{\beta}^{\phi_2}\right)n_i}\right)\right. \\ + \; & \left.\left(1-\pi\left(\boldsymbol{x}_i^T\widehat{\beta}^{\phi_2}\right)\right)\phi_1\left(\frac{n_{i2}}{\left(1-\pi\left(\boldsymbol{x}_i^T\widehat{\beta}^{\phi_2}\right)\right)n_i}\right)\right\}^{1/2}. \end{aligned}$$

In the following theorem we present its asymptotic distribution.

Theorem 2. *Suppose that the data Y_i, $i = 1, ..., I$ are binomially distributed with parameters n_i and $\pi\left(\boldsymbol{x}_i^T\beta\right)$. Choose functions ϕ_1 and $\phi_2 \in \Phi$ twice continuously differentiable in a neighborhood of 1. Assuming that $n_i/n \to \lambda_i > 0$ when $n_i \to \infty$, we have*

$$c_j^{\phi_1,\phi_2} \underset{N\to\infty}{\overset{L}{\to}} \mathcal{N}\left(0, \tau_j^2\right)$$

where

$$\tau_j^2 = 1 - \lambda_j \pi\left(\boldsymbol{x}_j^T\beta^0\right)\left(1 - \pi\left(\boldsymbol{x}_j^T\beta^0\right)\right)\boldsymbol{x}_j \boldsymbol{I}_F\left(\beta^0\right)^{-1}\boldsymbol{x}_j^T$$

being $\boldsymbol{I}_F\left(\beta^0\right) = \boldsymbol{X}^T Diag\left(\lambda_i \pi\left(\boldsymbol{x}_j^T\beta^0\right)\left(1 - \pi\left(\boldsymbol{x}_j^T\beta^0\right)\right)_{j=1,...,I}\right)\boldsymbol{X}$ the Fisher Information matrix associated with the Logistic Regression Model.

This result is important because we can define the ϕ_1-standarized residuals based on minimum ϕ_2-divergence estimator by

$$\left(c_j^{\phi_1,\phi_2}\right)^* = \frac{c_j^{\phi_1,\phi_2}}{\widehat{\tau}_j^{\phi_2}}$$

being

$$\left(\widehat{\tau}_j^{\phi_2}\right)^2 = 1 - \frac{n_j}{N}\pi\left(\boldsymbol{x}_j^T\widehat{\beta}^{\phi_2}\right)\left(1 - \pi\left(\boldsymbol{x}_j^T\widehat{\beta}^{\phi_2}\right)\right)\boldsymbol{x}_j \boldsymbol{I}_F\left(\widehat{\beta}^{\phi_2}\right)^{-1}\boldsymbol{x}_j^T.$$

§4. Dimensionality reduction

By dimensionality reduction we understand the procedure to determine if the independent variables in the model are "significantly" related to the outcome variable. To choose this logistic regression model we use a backward deletion procedure. Any stepwise procedure for deletion of variables from a model is based on a statistical algorithm which checks for the "importance" of a variable is defined in terms of a measure of the statistical significance of the coefficient for the variable, i.e., we must carry out the hypothesis test

$$H_{Null} : \beta_j = 0 \text{ against } H_{Alt} : \beta_j \neq 0 \ \ (j = 1, ..., k) \tag{6}$$

In the first stage, we consider $j = 1, ..., k$, i.e., the logistic regression model with all available explanatory variables. Then, we delete the explanatory variable associated with the regression parameter, β_{k_1}, if the associated $p-$value for testing, (6) with $j = k_1$, p_{k_1}, is the highest. In the second stage, we rename the parameters $\beta_{k_1}, ..., \beta_k$ as $\beta_{k_1+1}, ..., \beta_{k-1}$, respectively. Therefore, the logistic regression model with all the explanatory variables except the corresponding to the parameter β_{k_1} is considered. So, after testing (6) for $j = 1, ..., k-1$, we delete the explanatory variable with the highest associated $p-$value and so on. Finally, we stop the procedure when the maximum $p-$value associated with the logistic regression model with the remaining expanatory variables is sufficiently small.

In Pardo et al. (2003b) in order to solve the problem considered in (6) the following families of tests statistics were considered

$$S_{\phi_1,\phi_2}^{\beta_j,t} = \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\boldsymbol{p}\left(\widehat{\beta}^{\phi_2,t}\right), \boldsymbol{p}\left({}^j\widehat{\beta}^{\phi_2,t}\right)\right),$$

where $\widehat{\beta}^{\phi_2,t}$ is the minimum ϕ_2- divergence estimator of $\beta_1, ..., \beta_{k+1-t}$ and ${}^j\widehat{\beta}^{\phi_2}$ is the minimum ϕ_2-divergence estimator of $(\beta_0, ..., \beta_{j-1}, 0, \beta_{j+1}, ..., \beta_{k+1-t})$ in the stage t, and

$$T^{\beta_j,t}_{\phi_1,\phi_2} = \frac{2N}{\phi_1''(1)}\left(D_{\phi_1}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}\left(\widehat{\beta}^{\phi_2,t}\right)\right) - D_{\phi_1}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}\left({}^j\widehat{\beta}^{\phi_2,t}\right)\right)\right).$$

In the following theorem we present the asymptotic distribution of the tests statistic $S^{\beta_j,t}_{\phi_1,\phi_2}$ and $T^{\beta_j,t}_{\phi_1,\phi_2}$.

Theorem 3. *Suppose that the data Y_i, $i = 1, ..., I$ are binomially distributed with parameters n_i and $\pi\left(\boldsymbol{x}_i^T\beta\right)$. Choose functions ϕ_1 and $\phi_2 \in \Phi$ twice continuously differentiable in a neighborhood of 1. Then for testing*

$$H_{Null} : \beta_j = 0 \text{ versus } H_{Alt} : \beta_j \neq 0,$$

the test statistics $S^{\beta_j,t}_{\phi_1,\phi_2}$ and $T^{\beta_j,t}_{\phi_1,\phi_2}$ are asymptotically distributed as a a chi-squared distribution with 1 degree of freedom, under the assumption that $n_i/n \to \lambda_i > 0$ when $n_i \to \infty$.

Based on this theorem, if the sample sizes are large enough, one can use the asymptotic quantile $\chi^2_{1,1-\alpha}$, defined by the equation $P\left(\chi^2_{M-1} \leq \chi^2_{M-1,1-\alpha}\right) = 1-\alpha$, to propose the decision rule:

"Reject H_{Null} if $S^{\beta_j,t}_{\phi_1,\phi_2} > \chi^2_{M-1,1-\alpha}$ (or $T^{\beta_j,t}_{\phi_1,\phi_2} > \chi^2_{1,1-\alpha}$)".

§5. Polytomous Logistic Regression Model

We consider a response random variable Y belonging to one of the J distinct categories $C_1, ..., C_J$, which is observed together with $p+1$ explanatory variables $\boldsymbol{x}^T = (1, x_1, ..., x_p) \in \mathbb{R}^{p+1}$. For convenience $x_0 = 1$. Let $\pi_j(\boldsymbol{x}) = P(Y \in C_j/\boldsymbol{x})$, $j = 1, ..., J$, denote the probability that the observation of the random variable Y belongs to the category C_j, $j = 1, ..., J$, when the explanatory variable is $\boldsymbol{x}^T$. More specifically suppose here, that the dependence between Y and $\boldsymbol{x}^T$ can be modeled by using the logistic assumption

$$\pi_j(\boldsymbol{x}) \equiv \pi\left(\boldsymbol{x}^T\beta_j\right) = \exp\left(\boldsymbol{x}^T\beta_j\right) / \sum_{l=1}^{J}\exp\left(\boldsymbol{x}^T\beta_j\right), \; j = 1, ..., J, \tag{7}$$

where $\beta_j^T = (\beta_{0j}, ..., \beta_{pj})$, $j = 1, ..., J-1$, is a vector of unknown parameters and $\beta_J^T = \left(0, .^{(p+1}.., 0\right)$, for convenience. The vector $\beta^T = \left(\beta_1^T, ..., \beta_{J-1}^T\right)$ is ν-dimensional with $\nu = (J-1)(p+1)$. The model described in (7) is the classical *Polytomous Logistic Regression Model* (PLRM) or *Multinomial Logistic Regression Model*. For more details about this model see Amemiya (1981), Anderson (1972, 1982, 1984), Lesaffre (1986), Lesaffre and Albert (1986, 1989), Mantel (1966), Theil (1969), McCullag (1980) Agresti (2002), Engel (1988) and references there in. In the following we shall denote by

$$\Theta = \left\{\beta_j^T = (\beta_{0j}, ..., \beta_{pj}), j = 1, ..., J-1 : \beta_{st} \in \mathbb{R}, s = 0, .., p; \; t = 1, ..., J-1\right\}$$

We assume that N different values of the vector of explanatory variables,

$$\boldsymbol{x}_i^T = (x_{i0}, x_{i1}, ..., x_{ip}), \quad i = 1, ..., N,$$

are available. Let $n(\boldsymbol{x}_i)$ be the number of observations considered when the explanatory variable $\boldsymbol{x}^T$ has the value $\boldsymbol{x}_i^T$, so that if $\boldsymbol{x}^T$ is fixed at $\boldsymbol{x}_i^T$ we have a multinomial distribution with parameters

$$\left(n(\boldsymbol{x}_i); \pi\left(\boldsymbol{x}_i^T\beta_1\right), ..., \pi\left(\boldsymbol{x}_i^T\beta_J\right)\right).$$

Vectors of probabilities are denoted by $\pi(\boldsymbol{x}_i) = \left(\pi\left(\boldsymbol{x}_i^T\beta_1\right), ..., \pi\left(\boldsymbol{x}_i^T\beta_J\right)\right)^T$ and total sample size by $n = n(\boldsymbol{x}_1) + ... + n(\boldsymbol{x}_N)$. Given the explanatory variable $\boldsymbol{x}_i^T$, we denote the number of observations in the class C_s by y_{si}. It is clear that $n(\boldsymbol{x}_i) = \sum_{s=1}^{J} y_{si}$. To estimate β_{js} $(j = 0, ..., p; s = 1, ..., J-1)$ we maximize the loglikelihood function

$$\log \prod_{i=1}^{N}\prod_{l=1}^{J} \frac{n(\boldsymbol{x}_i)!}{y_{1i}!\times ...\times y_{Ji}!}\pi\left(\boldsymbol{x}_i^T\beta_l\right)^{y_{li}} \approx \log \prod_{i=1}^{N}\prod_{l=1}^{J} \pi\left(\boldsymbol{x}_i^T\beta_l\right)^{y_{li}} \equiv L(\beta)$$

with $y_{Ji} = n(\boldsymbol{x}_i) - \sum_{s=1}^{J-1} y_{si}$.

It is not difficult to establish that

$$\left(\frac{\partial^2 L(\beta)}{\partial\beta^2}\right)_{(p+1)(J-1)\times(p+1)(J-1)} = -n\boldsymbol{X}^T\boldsymbol{V}_n(\beta)\boldsymbol{X},$$

where

$$\boldsymbol{X}^T = \left(\boldsymbol{X}_1^T, ..., \boldsymbol{X}_N^T\right)_{(p+1)(J-1)\times(J-1)N},$$

being

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_i^T & \mathbf{0}^T & ... & \mathbf{0}^T \\ \mathbf{0}^T & \boldsymbol{x}_i^T & ... & \mathbf{0}^T \\ . & . & . & . \\ \mathbf{0}^T & \mathbf{0}^T & ... & \boldsymbol{x}_i^T \end{pmatrix}_{(J-1)\times(J-1)(p+1)} \tag{8}$$

and the matrix $\boldsymbol{V}_n(\beta)$ is defined by

$$\boldsymbol{V}_n(\beta) = diag\left(\frac{n(\boldsymbol{x}_1)}{n}\boldsymbol{V}_1(\beta), ..., \frac{n(\boldsymbol{x}_N)}{n}\boldsymbol{V}_N(\beta)\right)_{N(J-1)\times N(J-1)}$$

with

$$\boldsymbol{V}_i(\beta)_{(J-1)\times(J-1)} = \left(\pi\left(\boldsymbol{x}_i^T\beta_s\right)\left(\delta_{st} - \pi\left(\boldsymbol{x}_i^T\beta_t\right)\right)\right)_{s,t=1,...,J-1}, \; i = 1, ..., N,$$

and δ_{st} is the Kronecker delta.

The Fisher information matrix is given by

$$\boldsymbol{I}_{F,n}(\beta) = \boldsymbol{X}^T\boldsymbol{V}_n(\beta)\boldsymbol{X} = \sum_{j=1}^{N}\frac{n(\boldsymbol{x}_j)}{n}\boldsymbol{X}_j^T\boldsymbol{V}_j(\beta)\boldsymbol{X}_j$$

and

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \underset{n\to\infty}{\overset{L}{\to}} N\left(\mathbf{0}, \boldsymbol{I}_{F,\lambda}(\beta_0)\right),$$

where $\widehat{\beta}$ is the maximum likelihood estimator of β and β_0 is the true value of the parameter β, $\boldsymbol{I}_{F,\lambda}(\beta_0) = \sum_{j=1}^{N}\lambda_j\boldsymbol{X}_j^T\boldsymbol{V}_j(\beta_0)\boldsymbol{X}_j$ and $\lambda_j = \lim\frac{n(\boldsymbol{x}_j)}{n}$, $j = 1, ..., N$.

Let us introduce the two following probability vectors,

$$\widehat{\boldsymbol{p}} = \left(\frac{y_{11}}{n}, ..., \frac{y_{J1}}{n}, \frac{y_{12}}{n}, ..., \frac{y_{J2}}{n}, ..., \frac{y_{1N}}{n}, ..., \frac{y_{JN}}{n} \right)^T,$$

and

$$\begin{array}{rcl} \boldsymbol{p}(\beta) = & \left(\frac{n(\boldsymbol{x}_1)}{n} \left(\pi \left(\boldsymbol{x}_1^T \beta_1 \right), ..., \pi \left(\boldsymbol{x}_1^T \beta_J \right) \right), ..., \frac{n(\boldsymbol{x}_N)}{n} \left(\pi \left(\boldsymbol{x}_N^T \beta_1 \right), ..., \pi \left(\boldsymbol{x}_N^T \beta_J \right) \right) \right)^T \\ = & \left(\frac{n(\boldsymbol{x}_1)}{n} \pi (\boldsymbol{x}_1)^T, ..., \frac{n(\boldsymbol{x}_N)}{n} \pi (\boldsymbol{x}_N)^T \right)^T. \end{array} \tag{9}$$

The Kullback-Leibler divergence measure between the probability vectors $\widehat{\boldsymbol{p}}$ and $\boldsymbol{p}(\beta)$ is

$$\begin{array}{rcl} D_{Kullback}(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\beta)) = & \sum_{l=1}^{J} \sum_{i=1}^{N} \frac{y_{ji}}{n} \log \frac{\frac{y_{li}}{n}}{\pi (\boldsymbol{x}_i^T \beta_l) \frac{n(\boldsymbol{x}_i)}{n}} \\ = & kte - \frac{1}{n} \log \prod_{i=1}^{N} \prod_{l=1}^{J} \pi \left(\boldsymbol{x}_i^T \beta_l \right)^{y_{li}} \\ \approx & -L(\beta). \end{array}$$

Then the MLE of parameter β can be equivalently defined by the condition

$$\widehat{\beta} = \arg\min D_{Kullback}(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\beta)),$$

and its extension to the minimum ϕ-divergence estimator, is given by

$$\widehat{\beta}^{\phi} \equiv \arg \min_{\beta_{01}, .., \beta_{pJ-1}} D_{\phi}(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\beta)). \tag{10}$$

In relation to the asymptotic properties of the minimum ϕ-divergence estimator we have that the most important result is

$$\sqrt{n} \left(\widehat{\beta}^{\phi} - \beta_0 \right) \underset{n \to \infty}{\overset{L}{\to}} N \left(\boldsymbol{0}, \boldsymbol{I}_{F,\lambda}(\beta_0) \right),$$

and then

$$\widehat{Cov \left(\widehat{\beta}^{\phi} \right)} \approx \frac{1}{N} \left(\boldsymbol{X}^T \boldsymbol{V}_n \left(\widehat{\beta}^{\phi} \right) \boldsymbol{X} \right)^{-1}.$$

After estimating the unknown parameters, we would like to know how effective the model we have is in describing the outcome variable. This is referred to as goodness-of-fit. We will conclude that the model fits if (a) summary measures of the distance between the observed sample values and the values predicted by the model are small and (b) the contribution of each pair (observed, predicted) to these summary measures is unsystematic and is small relative to the error structure of the model. Thus, a complete assessment of the fitted model will involve both (a) computation and evaluation of overall measures of fit, (b) examination of the individual components of these measures. It is possibe to present families of statistics based on ϕ-divergence measures to solve (a) in a similar way to the families of test statistics given in (5). Additional diagnosis analyses are necessary to describe the nature of one lack of fit. A family of residuals based on ϕ-divergences that is a generalization of the classical residuals it is possible to define on the basis of ϕ-divergence measures.

References

[1] ALI, S.M. AND SILVEY, S.D. A general class of coefficient of divergence of one distribution from another. *Journal of Royal Statistical Society, Series B,* 286 (1966), 131–142.

[2] AGRESTI, A. *Categorical Data Analysis* (Second Edition), John Wiley & Sons, 2002.

[3] AMEMIYA, T. Qualitative response models: a survey. *Journal of Economic Literature,* 19 (1981), 1483-1536.

[4] ANDERSON, J. A. Separate sample logistic discrimination. *Biometrika,* 59 (1972), 19-35.

[5] ANDERSON, J. A. Logistic discrimination. In *Handbook of Statistics,* P. R. Krishnaiah and L. N. Kanal, Eds., North-Holland Publ. Comp. , 1982, 169-191.

[6] ANDERSON, J. A. Regression and ordered categorical variables. *Journal of Royal Statistical Society, Series B,* 46 (1984), 1-30.

[7] CSISZAR, I.. Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of Hungarian Academy of Sciences, Series A,* 8 (1963), 85–108.

[8] ENGEL, J. Polytomous logistic regression. *Statistica Neerlandica,* 42 (1988), 233-252.

[9] KULLBACK, S. Kullback information. In *Encyclopedia of Statistical Sciences, Volume 4,* editors S. Kotz and N. L. Johnson, John Wiley & Sons, New York, 1985, pp. 421-425.

[10] LESAFFRE, E. Logistic discrimination analysis with aplication in electrocardiography, *P.H. Dissertation.* University of Leuven, 1986.

[11] LESAFFRE, E. AND ALBERT, A. Multiple-group Logistic Regression Diagnostic. *Applied Statistics,* 38 (1989), 425-440.

[12] MCCULLAG, P. Regression Models for Ordinary Data. *Journal of the Royal Statistical Society,* B 42 (1980), 109-142.

[13] MANTE, N. Models for complex contingency tables and polychotomous dosage response curves. *Biometrics*, 22 (1966), 83-95.

[14] PARDO, J. A., PARDO, L. AND PARDO, M. C. Testing in logistic regression models based on ϕ-divergences measures. Submitted, (2003)

[15] PARDO, J. A., PARDO, L. AND PARDO, M. C. Minimum ϕ-divergence estimator in logistic regression models. Submitted, (2003)

Leandro Pardo
Department of Statistics and O.R.,
Complutense University of Madrid.
telephone: 34-913944425 fax: 34-913944607
Leandro_Pardo@Mat.ucm.es