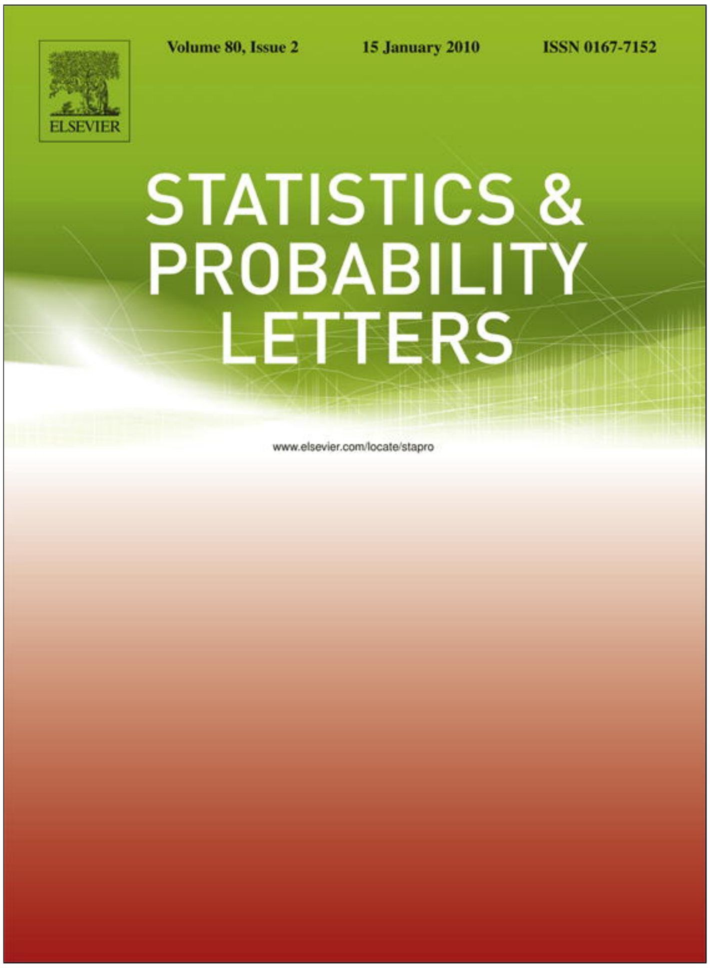
Volume 80, Issue 2
15 January 2010
ISSN 0167-7152
ELSEVIER
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

Statistics and Probability Letters 80 (2010) 89–95

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A generalized regression model for a binary response

Maria Kateri[a,*], Alan Agresti[b]

[a] Department of Statistics and Insurance Science, University of Piraeus, 185 34, Piraeus, Greece

[b] Department of Statistics, University of Florida, USA

ARTICLE INFO

Article history:
Received 4 May 2009
Received in revised form 12 August 2009
Accepted 19 September 2009
Available online 6 October 2009

ABSTRACT

Logistic regression is the closest model, given its sufficient statistics, to the model of constant success probability in terms of Kullback–Leibler information. A generalized binary model has this property for the more general ϕ-divergence. These results generalize to multinomial and other discrete data.

1. Introduction

Consider regression modeling of the effects of k explanatory variables, $x_1, \ldots, x_k$, on a binary response Y (success–failure). For observation i in a sample of size n, let Y_i be the response, let $\boldsymbol{x}_i = (x_{i1}, \ldots x_{ik})$ be the values of the explanatory variables, and let $p_i = \Pr(Y_i = 1)$. We assume that $Y_1, Y_2, \ldots, Y_n$ are independent.

In this framework, one's interest normally is in modeling p_i in terms of the explanatory variables. The best known binary regression model is the logistic regression model. In this article, we show a property that this model satisfies: It is the closest model, given its sufficient statistics, to the model of constant success probability. The distance employed in this result is Kullback–Leibler information (entropy).

We show alternative results for other models with other distance measures. These results can be put in the context of the many attempts that have been made to generalize binary regression models to provide families that include standard models such as logistic regression and probit regression models. Since the introduction of the generalized linear model (GLM), most such attempts have been motivated by the idea of the link function and the replacement of the logit or probit link by more general families of link functions, such as in Aranda-Ordaz (1981) and Stukel (1988). In this article, we approach the development of a class of models based on a generalized family of link functions from a different point of view, clarifying issues regarding the role of the link function. The familiar formula

$$\Pr(Y_i = 1|\mathbf{x}) = F^{-1}(\beta_0 + \beta_1 x_{i1} + \cdots \beta_k x_{ik}),$$

for some link function F, is not our starting point (as is usually the case) but rather our ending point.

As mentioned above, we view binary regression models in terms of their departure from the simple model of constant success probability, $\Pr(Y_i = 1|\mathbf{x}) = F^{-1}(\beta_0)$. Section 2 states the result about ordinary logistic regression model being the closest model to the model of constant success probability in terms of the Kullback–Leibler information, under conditions that correspond to the likelihood equations for the logistic model. Section 3 introduces a generalized class of binary regression models based on measuring the distance between two models using the ϕ-divergence (Csiszàr, 1963), which includes the Kullback–Leibler (KL) information as a special case. Section 4 discusses interpretational aspects of these models and introduces some characteristic special cases. Section 5 presents an example fitted by various members of this family of models. Section 6 presents generalizations for multinomial models, nested models for categorical data, and other generalized linear models.

[*] Corresponding author.
E-mail address: mkateri@unipi.gr (M. Kateri).

0167-7152/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2009.09.016

2. A property of the logistic regression model

Since our approach is based on distances between probability distributions, it is more convenient to express the logistic regression model in terms of p_i,

$$p_i = \Pr(Y_i = 1) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}, \quad i = 1, \ldots, n. \tag{2.1}$$

The corresponding model of constant success probability, by which none of the explanatory variables has an effect, is

$$p_i^{(0)} = \Pr(Y_i = 1) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}, \quad i = 1, \ldots, n, \tag{2.2}$$

that is, model (2.1) under H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$.

Denote the common value of $p_i^{(0)}$ under the common success probability model by q. For observation i, in terms of the KL information, the distance of the probability distribution $\boldsymbol{p}_i = (p_i, 1 - p_i)$ from the probability distribution $\boldsymbol{p}_0 = (q, 1 - q)$ is

$$KL(\boldsymbol{p}_i : \boldsymbol{p}_0) = p_i \log\left(\frac{p_i}{q}\right) + (1 - p_i) \log\left(\frac{1 - p_i}{1 - q}\right), \quad i = 1, \ldots, n.$$

For a complete sample of n independent observations, the KL information is defined as

$$KL(\mathbf{p} : \mathbf{q}) = \sum_{i=1}^{n} KL(\boldsymbol{p}_i : \boldsymbol{q}), \tag{2.3}$$

where $\mathbf{p} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n)$ and $\mathbf{q} = (\boldsymbol{p}_0, \ldots, \boldsymbol{p}_0)$. In terms of this measure, the logistic regression model satisfies the following property:

Theorem 2.1. *Consider a binary response variable Y and a set of explanatory variables $x_1, \ldots, x_k$, measured as $(y_i, x_{i1}, \ldots, x_{ik})$ for $i = 1, \ldots, n$ independent observations, with $p_i = P(Y_i = 1)$. In the class of models with those explanatory variables that have fixed value $s_j = \sum_{i=1}^{n} y_i x_{ij}$ for $\sum_{i=1}^{n} p_i x_{ij}$, $j = 1, \ldots, k$, the logistic regression model (2.1) is the closest to the model of constant success probability (2.2) in terms of the Kullback–Leibler information.*

The proof is omitted, because it is a special case of the more general result in Theorem 3.1, provided in Section 3. The given sums s_j in Theorem 2.1 are the sufficient statistics for the $\{\beta_j\}$ parameters of model (2.1), the likelihood equations for which are

$$s_j = \sum_{i=1}^{n} p_i x_{ij}, \quad j = 1, \ldots, k.$$

Other binary regression models do not have such reduced sufficient statistics, but it seems sensible to fix these statistics in finding such a result as a way of keeping constant, in some sense, the information about the effects of explanatory variables.

3. A binary response model based on ϕ-divergence

Using Theorem 2.1 as a starting point and replacing the Kullback–Leibler information by a more general class of divergences that includes KL as a special case, we obtain a generalized class of regression models for a binary response that includes logistic regression as a special case. The family of divergences we use for this purpose is the ϕ-divergence one.

The ϕ-divergence family is a general family of divergence measures, introduced by Csiszàr (1963), and is based on ϕ, a real-valued convex function on $[0, +\infty)$, with $\phi(1) = \phi'(1) = 0$, $0\phi(0/0) = 0$ and $0\phi(x/0) = x\phi_\infty$ with $\phi_\infty = \lim_{x\to\infty}[\phi(x)/x]$ (see Pardo, 2006). In our context, the ϕ-divergence of n binomial distributions from the baseline distribution of a constant probability, based on n independent observations, is defined by

$$D_\phi(\mathbf{p} : \mathbf{q}) = \sum_{i=1}^{n} D_\phi(\boldsymbol{p}_i : \boldsymbol{p}_0) = \sum_{i=1}^{n} \left[q\phi\left(\frac{p_i}{q}\right) + (1 - q)\phi\left(\frac{1 - p_i}{1 - q}\right) \right]. \tag{3.1}$$

Theorem 3.1. *Let ϕ be a twice differentiable and strictly convex function and let $F(x) = \phi'(x)$, for all x. Consider a binary response variable Y and a set of explanatory variables $x_1, \ldots, x_k$, measured as $(y_i, x_{i1}, \ldots, x_{ik})$ for $i = 1, \ldots, n$ independent observations, with $p_i = P(Y_i = 1)$. Then, in the class of models with those explanatory variables that have fixed value*

$s_j = \sum_{i=1}^n y_i x_{ij}$ for $\sum_{i=1}^n p_i x_{ij}$, $j = 1, \ldots, k$, the model

$$F\left(\frac{p_i}{q}\right) - F\left(\frac{1-p_i}{1-q}\right) = \sum_{j=1}^k \beta_j x_{ij}, \tag{3.2}$$

under the constraint that $0 < p_i < 1$, is the closest to the model of constant success probability, $p_i^{(0)} = q$ for all i, in terms of the ϕ-divergence.

Proof. This is a constraint minimization problem, solved by the method of Lagrange multipliers. The function to be minimized is $D_\phi(\mathbf{p} : \mathbf{q})$ subject to the restrictions $\sum_{i=1}^n p_i x_{ij} = \sum_{i=1}^n y_i x_{ij}$ $(j = 1, \ldots, k)$. Let $\tilde{p}_i = 1 - p_i$, and we also add the constraint $p_i + \tilde{p}_i = 1$. Thus, the Lagrange function is

$$L(\mathbf{p}) = D_\phi(\mathbf{p} : \mathbf{q}) + \sum_{i=1}^n c_i(p_i + \tilde{p}_i - 1) + \sum_{j=1}^k b_j \left[\sum_{i=1}^n y_i x_{ij} - \sum_{i=1}^n p_i x_{ij}\right],$$

where $\{c_i\}$ and $\{b_j\}$ are Lagrange multipliers. Setting $\partial L(\mathbf{p})/\partial p_i = 0$, we obtain

$$\phi'\left(\frac{p_i}{q}\right) + c_i - \sum_{j=1}^k b_j x_{ij} = 0, \quad i = 1, \ldots, n.$$

For $\alpha_i = -c_i$ and $\beta_j = b_j$, and since $F = \phi'$, we conclude that

$$F\left(\frac{p_i}{q}\right) = \alpha_i + \sum_{j=1}^k \beta_j x_{ij}, \quad i = 1, \ldots, n. \tag{3.3}$$

To solve (3.3) with respect to p_i, we require the existence of F^{-1}. This is ensured by the strict monotonicity of F, due to $F'(x) = \phi''(x) > 0$ for all x, because ϕ is strictly convex. Thus (3.3) leads to the expression

$$p_i = q \cdot F^{-1}\left(\alpha_i + \sum_{j=1}^k \beta_j x_{ij}\right), \quad i = 1, \ldots, n. \tag{3.4}$$

Analogously, by $\partial L(\mathbf{p})/\partial \tilde{p}_i = 0$ we obtain

$$1 - p_i = \tilde{p}_i = (1-q)F^{-1}(\alpha_i), \quad i = 1, \ldots, n, \tag{3.5}$$

and the fact that (3.4) and (3.5) must add to 1 for all i yields the constraints

$$qF^{-1}\left(\alpha_i + \sum_{j=1}^k \beta_j x_{ij}\right) + (1-q)F^{-1}(\alpha_i) = 1, \quad i = 1, \ldots, n, \tag{3.6}$$

which the parameters of our model must satisfy. Also, L has a minimum at p_i, since the Hessian matrix is positive definite ($\phi'' > 0$). Now, from (3.3) and (3.5), we have the result (3.2) that

$$F\left(\frac{p_i}{q}\right) - F\left(\frac{1-p_i}{1-q}\right) = \sum_{j=1}^k \beta_j x_{ij}. \quad \square$$

4. Parameter interpretation and characteristic special cases

The parameter β_j in model (3.2) reflects departures from the model of constant success probability due to the jth explanatory variable. The case $\beta_1 = \cdots = \beta_k = 0$ is equivalent to the model of constant success probability (2.2). As in ordinary logistic regression, interchanging the binary response categories results in the coefficient of x_{ij} changing from β_j to $\tilde{\beta}_j = -\beta_j$.

To interpret an individual β_j, we shall focus on two observations that differ only for explanatory variable x_j. Now, from (3.2), for two observations i and i' differing only in x_j, we have

$$F\left(\frac{p_i}{q}\right) - F\left(\frac{1-p_i}{1-q}\right) - F\left(\frac{p_{i'}}{q}\right) + F\left(\frac{1-p_{i'}}{1-q}\right) = \beta_j(x_{ij} - x_{i'j}). \tag{4.1}$$

In the case of KL divergence, $F(\frac{p_i}{q})$ and $F(\frac{1-p_i}{1-q})$ reduce to the log odds $\log\frac{p_i}{q}$ and $\log\frac{1-p_i}{1-q}$, respectively, as seen below. Furthermore, for $x_{ij} - x_{i'j} = 1$, relation (4.1) reduces to the well-known result for logistic regression that β_j equals the log odds ratio. Thus, the difference on the left-hand side of Eq. (4.1) can be regarded as a generalized log odds comparison, scaled through F.

We next highlight some members of the general class (3.2):

1. For $\phi(x) = x\log(x) - x + 1, x > 0, F^{-1}(y) = e^y$ and the ϕ-divergence simplifies to the Kullback–Leibler divergence. In this case restriction (3.6) leads to $\exp(\alpha_i) = \left[1 - q + q\exp(\sum_{j=1}^k \beta_j x_{ij})\right]^{-1}$, and model (3.2) becomes

$$p_i = \frac{q\exp\left(\sum_{j=1}^k \beta_j x_{ij}\right)}{1 - q + q\exp\left(\sum_{j=1}^k \beta_j x_{ij}\right)}, \quad i = 1, \dots, n, \tag{4.2}$$

which is simply the standard logistic regression model (2.1) with $\beta_0 = \log[q/(1-q)]$. In terms of odds, model (4.2) becomes

$$\frac{p_i}{1-p_i} = \frac{q}{1-q}\exp\left(\sum_{j=1}^k \beta_j x_{ij}\right), \quad i = 1, \dots, n. \tag{4.3}$$

Expression (4.3) reveals the "departure from constant success probability" nature of the logistic regression model.

2. For $\phi(x) = \frac{1}{2}(x-1)^2$, the ϕ-divergence is the Pearsonian distance $X^2(\mathbf{p} : \mathbf{q}) = \sum_{i=1}^n X^2(\boldsymbol{p}_i : \boldsymbol{p}_0) = \sum_{i=1}^n \frac{(p_i - q)^2}{q(1-q)}$ and constraints (3.6) give $\alpha_i = -q\sum_{j=1}^k \beta_j x_{ij}$. Thus, model (3.2) simplifies to a *linear probability model*,

$$p_i = q\left[1 + (1-q)\sum_{j=1}^k \beta_j x_{ij}\right], \quad i = 1, \dots, n. \tag{4.4}$$

This model is also the linear transformation model of Aranda-Ordaz (1981). In this case, the positivity of $\{p_i\}$ requires $\sum_{j=1}^k \beta_j x_{ij} > -\frac{1}{1-q}$, while $p_i < 1$ implies the restriction $\sum_{j=1}^k \beta_j x_{ij} < \frac{1}{q}$, for all i.

3. When D_ϕ is the power divergence measure (Read and Cressie, 1988), based on

$$D_\phi(\boldsymbol{p}_i : \boldsymbol{p}_0) = [\lambda(\lambda+1)]^{-1}\left[p_i\left(\frac{p_i}{q}\right)^\lambda + (1-p_i)\left(\frac{1-p_i}{1-q}\right)^\lambda - 1\right],$$

then ϕ depends on a real-valued parameter λ and equals $\phi_\lambda(x) = \frac{1}{\lambda(\lambda+1)}[x^{\lambda+1} - x - \lambda(x-1)], x > 0$ (Pardo, 2006). Model (3.2) then becomes

$$\left(\frac{p_i}{q}\right)^\lambda - \left(\frac{1-p_i}{1-q}\right)^\lambda = \lambda\sum_{j=1}^k \beta_j x_{ij}, \quad i = 1, \dots, n, \tag{4.5}$$

while the constraints (3.6) take the form

$$q\left[1 + \lambda\left(\alpha_i + \sum_{j=1}^k \beta_j x_{ij}\right)\right]^{1/\lambda} + (1-q)\,[1 + \lambda\alpha_i]^{1/\lambda} = 1, \quad i = 1, \dots, n.$$

Incorporating these constraints, (4.5) leads to

$$p_i = q\left[1 + \lambda\left(\alpha_i + \sum_{j=1}^k \beta_j x_{ij}\right)\right]^{1/\lambda}, \quad i = 1, \dots, n, \tag{4.6}$$

In terms of odds, the expression for model (4.6) is

$$\frac{p_i}{1-p_i} = \frac{q}{1-q}\left(1 + \lambda\frac{\sum_{j=1}^k \beta_j x_{ij}}{1 + \lambda\alpha_i}\right)^{1/\lambda}, \quad i = 1, \dots, n.$$

In order to ensure that $p_i < 1$ we need $\alpha_i < \frac{q^{-\lambda} - 1}{\lambda} - \sum_{j=1}^k \beta_j x_{ij}$. In case λ is even, we additionally need $\alpha_i > -\frac{1}{\lambda} - \sum_{j=1}^k \beta_j x_{ij}$. These constraints complicate the fitting of model (4.6) and suggest that the usefulness of the model for general λ is quite limited. Unless $\lambda = 1/2$, there is no closed-form expression for $\{\alpha_i\}$. When $\lambda = 0$ or $\lambda = -1$,

Table 1
Change in clinical condition by degree of infiltration.
Source: Cochran (1954).

	Degree of infiltration		Proportion	Model fit		
Clinical change	High	Low	High	Power divergence ($\lambda = 1.673$)	Logistic ($\lambda = 0$)	Linear ($\lambda = 1$)
Worse	1	11	0.083	0.083	0.140	0.112
Stationary	13	53	0.197	0.198	0.194	0.190
Slight improvement	16	42	0.276	0.278	0.262	0.270
Moderate improvement	15	27	0.357	0.345	0.344	0.348
Marked improvement	7	11	0.389	0.405	0.436	0.428

$\phi_0(x) = \lim_{\lambda\to 0}[\phi_\lambda(x)]$ and $\phi_{-1}(x) = \lim_{\lambda\to -1}[\phi_\lambda(x)]$. Model (4.6) reduces to model (4.4) when $\lambda = 1$ and to (4.2) when $\lambda \to 0$.

5. Model fitting and example

In some cases, maximum likelihood estimates can be obtained with ordinary software for maximizing functions by supplying the likelihood function to be maximized. For the models introduced in this article, the complicating factor is the constraints to keep probabilities in the (0, 1) interval. When the model fits the data well, we can ignore the constraints in the model-fitting process.

For example, consider the power divergence class of models, when at setting i of the explanatory variables we observe the binomial variate y_i based on n_i trials. Then, one can express the model in the form

$$p_i = \left[\beta_0^* + \sum_{j=1}^{k} \beta_j^* x_{ij}\right]^{1/\lambda}, \tag{5.1}$$

where $\beta_0^* = q^\lambda$ and $\beta_j^* = \lambda q^\lambda \beta_j$. We then maximize the usual binomial likelihood function, $\Pi_i[p_i^{y_i}(1-p_i)^{n_i-y_i}]$, treating λ as fixed or as a parameter to estimate.

For the example below, we did this using PROC NLMIXED in SAS. This requires reasonable initial estimates. With a single explanatory variable, one simple way to get these is to find sample proportions $\hat{p}_i$ and plot $\hat{p}_i^\lambda$ against x_i for various λ to suggest a version that linearizes the relationship, and then regress $\hat{p}_i^\lambda$ against x_i for that power to find initial estimates of β_0^* and β_1^*. Such a procedure will not be adequate when the model fit violates the constraints for probabilities, but the model is not sensible in any case for such situations.

We illustrate using Table 1, from Cochran (1954), used there to illustrate a trend test for a binary response with quantitative predictor. The table refers to an experiment on the use of sulfones and streptomycin drugs in the treatment of leprosy. The degree of infiltration at the start of the experiment measures a type of skin damage. The response is the change in the overall clinical condition of the patient after 48 weeks of treatment. We use response scores (0, 1, 2, 3, 4). The question of interest is a comparison of the mean change for the two infiltration levels. Cochran noted that such an analysis is equivalent to a trend test treating the binary variable as the response. That test is sensitive to linearity between clinical change and the proportion of cases with high infiltration.

The logistic model with a linear trend using equally spaced clinical change scores fits well, having deviance 0.63 with $df = 3$. The linear probability model, which is the power divergence model with $\lambda = 1$, also fits well, with deviance 0.26 with $df = 3$. Fitting the power divergence model with λ as a parameter, we get $\hat{\beta}_0^* = 0.0154$, $\hat{\beta}_1^* = 0.0512$, and $\hat{\lambda} = 1.673$, with deviance 0.05, while df reduces to 2. Table 1 also shows the sample proportions and the fitted proportions for the three models. The more complex model has the advantage of a precise fit, as is illustrated in Fig. 1, but at the cost that the interpretation of parameter estimates is not as simple.

Although we show this example to illustrate a model from the generalized class presented in this article, we do not feel that such models have broad scope for applications, because of the constraint issues and the lack of simple interpretation of parameters compared to standard models. We feel that the results in this article are mainly of some theoretical interest for providing a property for logistic regression and related models. However, in concrete situations where the standard models fail, they may be helpful for identifying a more appropriate scale.

6. Generalizations

In this section, we consider some generalizations of the results in previous sections.

For a given set of explanatory variables, Theorem 2.1 found that the logistic regression model with those predictors was closest (in terms of the Kullback–Leibler information) to the null model of constant success probability, subject to the constraint that the sufficient statistics for the logistic model equal their expected values. What if instead we constrain only a subset of the sufficient statistics to equal their expected values, corresponding to a subset of the explanatory variables? Then, at least for the Kullback–Leibler case, the binary response model with the full set of predictors that is closest to the

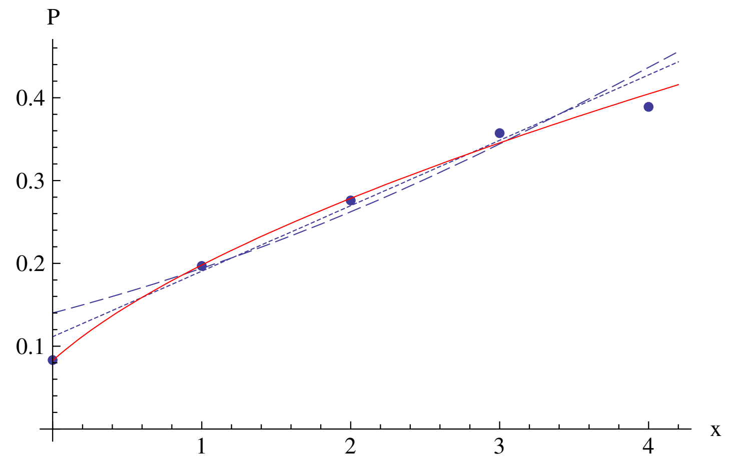

Fig. 1. Sample proportions (dots) and estimated probabilities using logistic regression (dashed curve), linear probability model (fine dashed line), and power divergence model with $\lambda = 1.673$ (solid curve).

null logistic regression model having the remaining explanatory variables as predictors is the logistic regression model. This result, stated as the following theorem, has a proof following the same lines as the proof of Theorem 3.1.

Theorem 6.1. *Consider the logistic regression model, denoted by M_0, containing the first k_0 explanatory variables of the set $x_1, \dots, x_k$. Then, in the class of binary regression models with fixed value $s_j = \sum_{i=1}^n y_i x_{ij}$ for $\sum_{i=1}^n p_i x_{ij}$, $j = k_0 + 1, \dots, k$, the logistic regression model (2.1) is the closest to the model M_0 in terms of the Kullback–Leibler information.*

The analogous result does not hold for the general model (3.2) in terms of an arbitrary ϕ-divergence measure. The form of the closest model then is

$$p_i = qF^{-1}\left(\alpha_i + \sum_{j=1}^{k_0} \beta_j x_{ij}\right) F^{-1}\left(\alpha_i + \sum_{j=k_0+1}^{k} \beta_j x_{ij}\right)$$

and it cannot be further simplified. Note that if we allow in this last model different ϕ-functions for different blocks of explanatory variables then we are led to a generalized model in which blocks of explanatory variables enter with different scalings.

Just as we generalized binary logistic regression to a family of binary regression models, in a similar manner multinomial logistic regression models generalize. For the baseline-category logit model for a categorical response with c categories, let

$$p_{ih} = P(Y_i = h) = \frac{\exp\left(\beta_{h0} + \sum_{j=1}^{k} \beta_{hj} x_{ij}\right)}{1 + \sum_{s=1}^{c-1} \exp\left(\beta_{s0} + \sum_{j=1}^{k} \beta_{sj} x_{ij}\right)}, \quad h = 1, \dots, c-1.$$

Let y_{ih} denote an indicator variable that equals 1 if $Y_i = h$. Then, in terms of KL information, under the constraint that the sufficient statistics $\sum_i x_{ij} y_{ih} = \sum_i x_{ij} p_{ih}$, for $j = 1, \dots, k$ and for $h = 1, \dots, c-1$, this model is the closest to the model of constant probabilities, by which $p_{ih} = q_h$ for all i and h. More generally, in terms of the ϕ-divergence, the closest model to that of constant probabilities is

$$F\left(\frac{p_{ih}}{q_h}\right) - F\left(\frac{p_{ic}}{q_c}\right) = \sum_j \beta_{hj} x_{ij}, \quad h = 1, \dots, c-1,$$

where $F(x) = \phi'(x)$. Incorporating the probability constraints, the model for a particular outcome probability is alternatively expressed as

$$p_{ih} = q_h F^{-1}\left(a_i + \sum_j \beta_{hj} x_{ij}\right), \quad h = 1, \dots, c-1,$$

where the parameters a_i ($i = 1, \dots, n$) are determined by

$$\sum_{h=1}^{c-1} q_h F^{-1}\left(a_i + \sum_j \beta_{hj} x_{ij}\right) + q_c F^{-1}(a_i) = 1.$$

The proof is also a straightforward generalization of the proof for Theorem 3.1, using Lagrange multipliers.

The result also applies to useful special cases of this multinomial model. For example, the stereotype model (Anderson, 1984) replaces the linear predictor $\sum_{j=1}^{k}\beta_{hj}x_{ij}$ by the multiplicative form $\phi_h\sum_{j=1}^{k}\beta_j x_{ij}$, for category scores $\{\phi_h\}$ that are themselves parameters and have a constraint such as $\phi_c = 0$. For given category scores $\{\phi_h\}$ and for given scaled correlations $\sum_{i=1}^{n}\sum_{h=1}^{c-1}p_{ih}\phi_h x_{ij}$ between the category scores and each explanatory variable, $j = 1,\ldots,k$, the stereotype model is the closest to the model of constant probabilities in terms of the Kullback–Leibler information. In terms of ϕ-divergence, the closest model has the generalized stereotype form

$$F\left(\frac{p_{ih}}{q_h}\right) - F\left(\frac{p_{ic}}{q_c}\right) = \phi_h\sum_j \beta_j x_{ij} \quad h = 1,\ldots,c-1.$$

In fact, the results for binary regression generalize to discrete generalized linear models. For the Kullback–Leibler divergence, under the constraint that $\sum_i(y_i x_{ij} - \mu_i x_{ij}) = 0$ for all j, when there are no additional constraints the loglinear model is the closest model to the null model of a constant mean, $\mu_i = \mu_0$. This follows because the Lagrangian function with Lagrangian multipliers $\{b_j\}$ is

$$L = \sum_i \mu_i\log(\mu_i/\mu_0) + \sum_{j=1}^{k} b_j\left[\sum_{i=1}^{n}(y_i x_{ij} - \mu_i x_{ij})\right],$$

which results in

$$\partial L/\partial\mu_i = \log(\mu_i/\mu_0) + 1 - \sum_{j=1}^{k} b_j x_{ij} = 0$$

and the model $\mu_i = \mu_0\exp\left(\sum_j b_j x_{ij} - 1\right)$. In particular, the constraint equations $\sum_i y_i x_{ij} = \sum_i \mu_i x_{ij}$ for all j are the likelihood equations for the generalized linear model using the canonical link function. As an important special case, the Poisson loglinear model is the closest model to the model of a constant mean in the class of models with these constraint equations. For the more general ϕ-divergence with $F = \phi'$, under the same constraints, the closest model has the form

$$\mu_i = \mu_0 F^{-1}\left(\sum_j b_j x_{ij}\right). \tag{6.1}$$

Analogous results occur in the literature for association models and correlation models for two-way contingency tables (Goodman, 1985), in terms of distance from the independence model. Gilula et al. (1988) showed that association models are closest to independence in terms of the Kullback–Leibler measure, while correlation models (which, like model (4.4), are linear in the probability) are closest in terms of Pearsonian distance. A general class of association models based on the ϕ-divergence has been introduced by Kateri and Papaioannou (1995). In comparing association and correlation models, Goodman (1985, p. 32) pointed out that the parametric scores in correlation models must satisfy certain constraints to ensure that cell probabilities lie in the (0, 1) interval, but such constraints were not needed for the corresponding scores in association models. This is also the situation in our context, since (4.4) and (4.6) require constraints whereas (4.2) does not.

In the case of square contingency tables, similar results have been proved for the quasi-symmetry model (Kateri and Papaioannou, 1997) and the ordinal quasi-symmetry model (Kateri and Agresti, 2007).

Acknowledgements

We are grateful to the Associate Editor and a referee for several comments that helped to improve the paper.

References

Anderson, J.A., 1984. Regression and ordered categorical variables. Journal of the Royal Statistical Society B 46, 1–30.

Aranda-Ordaz, F.J., 1981. On two families of transformations to additivity for binary response data. Biometrika 68, 357–364.

Cochran, W.G., 1954. Some methods of strengthening the common χ^2 tests. Biometrics 10, 417–451.

Csiszàr, I., 1963. Eine informationstheoretische Ungleichung und ihre Anwendungen auf den Beweis der Ergozitat von Markoffschen Ketten. A Mayar Tudomanyos Academia Mathematikai Kutato Intezelent Kozlemezyri 8, 85–108.

Gilula, Z., Krieger, A.M., Ritov, Y., 1988. Ordinal association in contingency tables: Some interpretive aspects. Journal of the American Statistical Association 83, 540–545.

Goodman, L.A., 1985. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models and asymmetry models for contingency tables with or without missing entries. The Annals of Statistics 13, 10–69.

Kateri, M., Papaioannou, T., 1995. f-divergence association models. International Journal of Mathematical and Statistical Science 3, 179–203.

Kateri, M., Papaioannou, T., 1997. Asymmetry models for contingency tables. Journal of the American Statistical Association 92, 1124–1131.

Kateri, M., Agresti, A., 2007. A class of ordinal quasi symmetry models for square contingency tables. Statistics & Probability Letters 77, 598–603.

Pardo, L., 2006. Statistical Inference Based on Divergence Measures. Chapman & Hall.

Read, T.R.C., Cressie, N.A.C., 1988. Goodness-of-fit Statistics for Discrete Multivariate Data. Springer-Verlag, New York.

Stukel, T., 1988. Generalised logistic models. Journal of the American Statistical Association 83, 426–431.