Journal of Statistical Planning and Inference 179 (2016) 22–35

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Model fitting and optimal design for a class of binary response models

Subir Ghosh [a,*], Hans Nyquist [b]

[a] University of California, Riverside, USA
[b] Stockholm University, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 12 January 2015
Received in revised form 17 April 2016
Accepted 6 July 2016
Available online 20 July 2016

Keywords:
Binary response
Efficiency comparison
Estimating equations
Maximum likelihood
Model discrimination
Odds ratio
Optimum design

ABSTRACT

A class of binary response models is considered for describing the data on a response variable having two possible outcomes and q explanatory variables when the odds ratios on the response are a linear function of the explanatory variables. The models provide the closed form solutions of the maximum likelihood estimating equations for the parameter estimation under a Bernoulli setup. A data example is presented to demonstrate the better goodness of fit of a model within this class in comparison with the logit, probit, and complimentary log–log models. The design conditions are given and locally optimal designs are presented for some special cases under the $\mathscr{D}$-, $\mathscr{A}$-, and $\mathscr{E}$-, optimality criterion functions. Two designs, one efficient for identifying one model and other efficient for identifying another model, are then compared for their discrimination abilities between two models even before the data collection.

1. Introduction

The study of dependence between response and explanatory variables is a major area of investigation in statistics with its complexities and challenges. Controlled experiments are performed to collect data on the variables and the models are assumed for the analysis of the data. Developing efficient methods of data collection and analysis are fundamental statistical endeavors. An important situation arises when the response variable is binary in having two possible outcomes. For example, the responses may be alive or dead in an experiment exploring the toxicity of a pesticide (Shi and Renton, 2013), accept or reject a bid in a contingent valuation experiment (Lim et al., 2014), or correct or wrong answer in an achievement test (Finch and Cassady, 2014). Here "success" and "failure" are used as generic representations for the two categories. In this paper a particular class of models is considered where the odds of the binary response variable are linearly dependent on the explanatory variables. This class is related to many practically useful and important models. The exact solution of maximum likelihood estimation of the parameters for a Bernoulli model is presented. Design issues are investigated and locally optimal designs are presented in some special cases. Model discrimination issues are also discussed with an illustrative example. Ghosh and Dutta (2013) worked out the issues for some continuous response models but in here the issues are addressed for the binary response models.

Consider a binary response variable Y with two realized values 1 (success) and 0 (failure). One of the most used models for studying the dependence of Y on q explanatory variables $X_1, \ldots, X_q$, is the logit model (McCullagh and Nelder, 1989;

* Correspondence to: Department of Statistics, University of California, Riverside, CA 92521-0138, USA.
E-mail address: subir.ghosh@ucr.edu (S. Ghosh).

http://dx.doi.org/10.1016/j.jspi.2016.07.001
0378-3758/

Fackle-Fornius and Nyquist, 2009; Ghosh and Banerjee, 2010). The log-odds for the logit model is a linear combination of the explanatory variables

$$\log_e\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right)=\alpha+\boldsymbol{\beta}'\mathbf{x}$$

where $\mathbf{x}=(x_1,\ldots,x_q)'$, and $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_q)'$. However, as is noted by e.g. Thomas (1981), many plausible models may not be adequately represented by a logit function. An alternative model, used by Thomas (1981) for analyzing the dependence of lung cancer incidence on exposure of asbestos, utilizes the odds being a linear function of the explanatory variables

$$\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}=\alpha+\boldsymbol{\beta}'\mathbf{x}, \tag{1}$$

see also Holford (2002, Ch. 6) and VanderWeele and Vansteelandt (2011). Models in this class become

$$P(Y=1\mid X_1=x_1,\ldots,X_q=x_q)=\pi(\mathbf{x})=\frac{\alpha+\boldsymbol{\beta}'\mathbf{x}}{1+\alpha+\boldsymbol{\beta}'\mathbf{x}}, \tag{2}$$

where $(\alpha,\boldsymbol{\beta}')\neq(0,\mathbf{0}')$ and $(\alpha+\boldsymbol{\beta}'\mathbf{x})\geq 0$. Clearly, $\pi(\mathbf{x})\leq 1$ and the "=" holds approximately for all practical considerations when $(\alpha+\boldsymbol{\beta}'\mathbf{x})$ becomes very large. Moreover, $\pi(\mathbf{x})>0$ when $(\alpha+\boldsymbol{\beta}'\mathbf{x})>0$, $\pi(\mathbf{x})=\frac{1}{2}$ when $(\alpha+\boldsymbol{\beta}'\mathbf{x})=1$, and $\pi(\mathbf{x})=0$ when $(\alpha+\boldsymbol{\beta}'\mathbf{x})=0$. The conditional expectation of Y given $(X_1=x_1,\ldots,X_q=x_q)$ is $E(Y\mid X_1=x_1,\ldots,X_q=x_q)=\pi(\mathbf{x})$. In the language of the generalized linear models (GLM), McCullagh and Nelder (1989), the linear predictor θ and link function g are

$$\theta(\mathbf{x})=\alpha+\boldsymbol{\beta}'\mathbf{x},\quad \pi(\mathbf{x})=\frac{\theta(\mathbf{x})}{1+\theta(\mathbf{x})},\qquad g(\pi(\mathbf{x}))=\theta(\mathbf{x})=\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}. \tag{3}$$

Fig. 1 displays the dependence of $\pi(\mathbf{x})$ on $\theta(\mathbf{x})$. The first graph in Fig. 1 portrays the plot of $\pi(\mathbf{x})$ for $0\leq\theta(\mathbf{x})\leq 5$ and the second graph for $0\leq\theta(\mathbf{x})\leq 1000$.

The data on $Y, X_1,\ldots,$ and X_q collected from an experiment are represented by $(y_{ij}, x_{i1},\ldots,x_{iq})$, $j=1,\ldots,n_i$, $i=1,\ldots,k$. At the ith design point $\mathbf{x}^{(\mathbf{i})'}=(x_{i1},\ldots,x_{iq})$, the replicated observations are y_{ij}, $j=1,\ldots,n_i$. Denote $\mathbf{x_s}=(x_{1s},\ldots,x_{ks})'$, $s=1,\ldots,q$, and the $(k\times q)$ matrix $\mathbf{D}=(\mathbf{x}^{(\mathbf{1})},\ldots,\mathbf{x}^{(\mathbf{k})})'=(\mathbf{x_1},\ldots,\mathbf{x_q})$ representing the design whose k rows are the design points. Given the data, the ML estimation of the parameters $\alpha,\beta_1,\ldots,\beta_q$ is considered first. The admissibility of estimated parameters in satisfying the conditions $\theta_i>0$, $i=1,\ldots,k$ is then discussed. At the design stage, the locally optimum designs are presented in the framework of Robbins–Monro–Chernoff (Robbins and Monro, 1951; Chernoff, 1953) by determining the optimum values of $\theta_1,\ldots,$ and θ_k in some special cases. Determining the optimum values of $\theta_1,\ldots,$ and θ_k is used by many authors for logit, probit, and complementary log–log models (see, Atkinson et al., 2007, Section 22.4, pp. 398–410) including a few references to the original work (Abdelbasit and Plackett, 1983; Ford et al., 1992; Khan and Yazdi, 1988; Mathew and Sinha, 2001; Minkin, 1987; Sitter and Wu, 1993). For the logit model, which is different from the model in (1), the celebrated $\mathscr{D}$-optimal equally replicated symmetric design for $k=2$ and $q=1$ is given by $\theta_1=-1.5434$ and $\theta_2=1.5434$ (Abdelbasit and Plackett, 1983; Minkin, 1987; Atkinson et al., 2007). Again, the negative value of θ_1 in here is not in violation of the condition $\theta_i>0$ for the model in (1).

The first goal in this paper is to fit the model in (1) to data collected from an experiment. The second goal is to determine the optimal choice of the design $\mathbf{D}$ for a given value of $(n_1,\ldots,n_k)$. The third goal is to determine the optimal choice of $(n_1,\ldots,n_k)$ for a given design $\mathbf{D}$. The fourth goal is to determine the optimal choices of $\mathbf{D}$ and $(n_1,\ldots,n_k)$. The fifth goal is to consider the model discrimination situation where several possible models are chosen for fitting to the data.

The model (1) is a special case of a general class of models

$$\pi(\theta)=1-(c+(t-1)\theta)^{1/(1-t)},\quad t\neq 1, \tag{4}$$

when $c=1+\alpha$ and $t=2$. The models in (4) are characterized by the assumption that the rate of increase in $\pi(\mathbf{x})$ with respect to θ is proportional to the tth power of $(1-\pi(\mathbf{x}))$ or equivalently

$$\frac{d}{d\theta}\pi(\theta)=c(\gamma_2-\pi(\theta))^t, \tag{5}$$

for a proportionality constant c (>1) and $\gamma_2=2$. This is to say that the larger the probability for observing a success, the smaller is the derivative of $\pi(\theta)$. This differential equation is a special case of the Bernoulli equation (Zwillinger, 1997) and can easily be solved.

When $t=1$, Eq. (5) gives

$$\pi(\mathbf{x})=1-e^{\alpha-\boldsymbol{\beta}'\mathbf{x}},$$

where $\pi(\mathbf{x})$ resembles the distribution function of an exponential distribution and the constant α is interpreted as the log of the probability for no response when $\mathbf{x}=\mathbf{0}$, i.e. $\alpha=\ln(1-\pi(\mathbf{0}))$. This is the so called log binomial model and has been extensively used in epidemiology for estimating risks (Skov et al., 1998; McNutt et al., 2003). It can also be noted that