Computational Statistics and Data Analysis 71 (2014) 1208–1220

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Integral approximations for computing optimum designs in random effects logistic regression models

C. Tommasi [a], J.M. Rodríguez-Díaz [b], M.T. Santos-Martín [b,*]

[a] *Department of Economics, Business and Statistics, University of Milan, Italy*

[b] *Department of Statistics, University of Salamanca, Spain*

ARTICLE INFO

Article history:
Received 10 March 2011
Received in revised form 23 May 2012
Accepted 28 May 2012
Available online 2 June 2012

Keywords:
Binary regression model
Fisher information matrix
Information matrix
Optimal design of experiments
Influence function

ABSTRACT

In the context of nonlinear models, the analytical expression of the Fisher information matrix is essential to compute optimum designs. The Fisher information matrix of the random effects logistic regression model is proved to be equivalent to the information matrix of the linearized model, which depends on some integrals. Some algebraic approximations for these integrals are proposed, which are consistent with numerical integral approximations but much faster to be evaluated. Therefore, these algebraic integral approximations are very useful from a computational point of view. Locally *D*-, *A*-, *c*-optimum designs and the optimum design to estimate a percentile are computed for the univariate logistic regression model with Gaussian random effects. Since locally optimum designs depend on a chosen nominal value for the parameter vector, a Bayesian *D*-optimum design is also computed. In order to find Bayesian optimum designs it is essential to apply the proposed integral approximations, because the use of numerical approximations makes the computation of these optimum designs very slow.

1. Introduction

The interest in finding optimum designs in the context of regression models with random effects is steadily increasing. See for instance, Mentré et al. (1997), Patan and Bogacka (2007), Schmelter et al. (2007), Graßhoff et al. (2009), Holland-Letz et al. (2011) and Debusho and Haines (2011).

Another setting where optimal designs have been extensively studied is the context of (fixed effects) binary regression models. See Abdelbasit and Plackett (1983), Minkin (1987), Ford et al. (1992), Sitter and Wu (1993), Biedermann et al. (2006) and Sitter and Fainaru (1997), among many others. Recently, Ouwens et al. (2006) have studied optimum designs for logistic models with random intercept. In this paper, different optimum designs are derived for the logistic regression model where not only the intercept but all the coefficients are random.

When the interest is to estimate as precisely as possible the parameters of the model (or some function of them), optimum designs can be computed minimizing some convex criterion function of the Fisher information matrix. Therefore, in order to find such designs, the explicit representation of this matrix is a very valuable tool. The expression of the Fisher information matrix for the random effects logistic regression model is given in Section 2, where the equivalence with the information matrix of the linearized model is also proved. The main contribution, however, is given in Section 3, where some algebraic integral approximations are provided. These approximations are very useful from a computational point of view, because

* Correspondence to: Faculty of Sciences, Plaza de los Caídos, 37008 Salamanca, Spain. Tel.: +34 923294458; fax: +34 923294514.
E-mail addresses: chiara.tommasi@unimi.it (C. Tommasi), juanmrod@usal.es (J.M. Rodríguez-Díaz), maysam@usal.es (M.T. Santos-Martín).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.05.024

the Fisher information matrix of the random effects logistic regression model depends on some integrals which need to be evaluated. In addition, the algorithms to compute optimum designs depend on these integrals as well. If numerical integral approximations are used instead of the proposed algebraic approximations then the algorithms become very slow. In Section 4, the univariate logistic regression model with normally distributed random coefficients is deeply studied. In this context it is shown, through an example and some graphics of the integrals, that the two integral approximations (numerical and algebraic) are consistent, but the algebraic one is faster to be computed. In Section 5, *D*-, *A*- and *c*-optimum designs for this model are derived while Section 6 deals with the important problem of precise estimation of a percentile. Finally, Section 7 describes how the efficiency of the locally *D*-optimum designs changes if the chosen nominal value for the parameter vector differs from the true value. A prior distribution for the parameter vector is also specified and the corresponding Bayesian *D*-optimum design is computed. Differently from the locally optimal designs, the efficiency of the Bayesian *D*-optimum design is always "good" (greater than 97%).

2. The binary regression model and the Fisher information matrix

Let Y be a binary response variable such that $P(Y=1|\beta)=F(\mathbf{x}'\beta)$ where $F(\cdot)$ is a cumulative distribution function (cdf), being two examples the Probit and the Logit models for which $F(\cdot)$ is the Gaussian and Logistic cdf, respectively; "′" denotes transposition, $\mathbf{x}=(1,x_1,\ldots,x_{k-1})'$ is a $k\times 1$ vector of experimental conditions which may be chosen in an experimental domain $\mathcal{X}\subseteq\mathbb{R}^k$ and $\beta=(\beta_0,\beta_1,\ldots,\beta_{k-1})'$ is $k\times 1$ vector of random coefficients. In other words, for each experimental unit there is a vector of unobservable random coefficients $\beta\in\Omega_\beta\subseteq\mathbb{R}^k$ such that $\beta\sim\phi(\beta;\theta)$, where $\phi(\cdot;\cdot)$ is a k-variate probability density function (pdf) which depends on some unknown parameters $\theta\in\Theta\subseteq\mathbb{R}^m$, with $k\le m$. It is well known that if β is a vector of constant unknown coefficients then the Fisher information matrix coincides with information matrix corresponding to the linearized binary regression model. This equivalence holds even when the coefficients are random variables.

In order to have n independent observations, $y_1,\ldots,y_n$, it is assumed that one observation per individual is taken, as in Graßhoff et al. (2009). With this assumption, the log-likelihood function is

$$\log L(\theta)=\sum_{i=1}^{n}\log\int L_c(y_i,\mathbf{x}_i,\beta)\,\phi(\beta;\theta)\,d\beta,$$

where from now on the integration is taken over Ω_β and

$$L_c(y_i,\mathbf{x}_i,\beta)=\left[F(\mathbf{x}_i'\beta)\right]^{y_i}\left[1-F(\mathbf{x}_i'\beta)\right]^{1-y_i}.$$

The Fisher information matrix of an exact design $\xi=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ is the $m\times m$ matrix $\mathcal{I}(\xi;\theta)=\{\mathcal{I}_{rs}(\xi;\theta)\}$, whose (r,s)-item $(r,s=1,\ldots,m)$ is

$$\begin{aligned}\mathcal{I}_{rs}(\xi;\theta)&=E\left[-\frac{\partial^2}{\partial\theta_r\partial\theta_s}\log L(\theta)\right]\\&=\sum_{i=1}^{n}\frac{\int F(\mathbf{x}_i'\beta)\,\frac{\partial}{\partial\theta_r}\phi(\beta;\theta)\,d\beta\cdot\int F(\mathbf{x}_i'\beta_i)\,\frac{\partial}{\partial\theta_s}\phi(\beta;\theta)\,d\beta}{\int F(\mathbf{x}_i'\beta)\,\phi(\beta;\theta)\,d\beta\cdot\left[1-\int F(\mathbf{x}_i'\beta)\,\phi(\beta;\theta)\,d\beta\right]},\end{aligned}\tag{1}$$

where the last equality is obtained after some algebra, assuming the usual regularity conditions on the pdf $\phi(\cdot;\cdot)$.

An alternative expression for the binary regression model is $Y_i=E[Y_i]+\varepsilon_i$, where

$$E[Y_i]=\int F(\mathbf{x}_i'\beta)\,\phi(\beta;\theta)\,d\beta,$$

and ε_i is such that $E[\varepsilon_i]=0$ and

$$\mathrm{Var}[\varepsilon_i]=\mathrm{Var}[Y_i]=\int F(\mathbf{x}_i'\beta)\,\phi(\beta;\theta)\,d\beta\cdot\left[1-\int F(\mathbf{x}_i'\beta)\,\phi(\beta;\theta)\,d\beta\right].$$

If this model is linearized at some nominal values of the parameters, then the information matrix for one observation is given by $\mathbf{g}(\mathbf{x}_i;\theta)\mathbf{g}(\mathbf{x}_i;\theta)'$ where $\mathbf{g}(\mathbf{x}_i;\theta)$ is the $m\times 1$ vector whose j-th item $(j=1,\ldots,m)$ is

$$g_j(\mathbf{x}_i;\theta)=\frac{\int F(\mathbf{x}_i'\beta)\,\frac{\partial}{\partial\theta_j}\phi(\beta;\theta)d\beta}{\sqrt{\int F(\mathbf{x}_i'\beta)\,\phi(\beta;\theta)\,d\beta\cdot\left[1-\int F(\mathbf{x}_i'\beta)\,\phi(\beta;\theta)\,d\beta\right]}}.\tag{2}$$

From (1) it follows that

$$\mathcal{I}(\xi;\theta)=\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i;\theta)\mathbf{g}(\mathbf{x}_i;\theta)'\tag{3}$$