Designing experiments for binary data using search algorithms

David Woods and Susan Lewis
Southampton Statistical Sciences Research Institute
School of Mathematics
University of Southampton
Southampton, SO17 1BJ, UK
D.C.Woods@maths.soton.ac.uk; S.M.Lewis@maths.soton.ac.uk

In some experiments in the physical and biological sciences, a binary response is of primary interest and is often described by a generalized linear model. Examples include experiments in food technology and studies in chemistry where the outcome is whether or not a salt is formed in a chemical reaction. For such experiments, designs that are efficient under the assumption of a linear model may be inadequate for the description and prediction of the response. The generation of designs using a search algorithm is addressed for completely randomized designs when a generalized linear model describes the response. A method of assessing the designs is discussed and illustrated by examples.

1. Introduction

Experiments which produce a binary or binomial response arise in a variety of areas such as engineering, chemistry and food technology. Although a generalized linear model (GLM) is often used to describe such a response, methods of designing experiments tailored to these nonlinear models are largely limited to situations when only one or two variables are to be investigated and a simple first order linear predictor is assumed. These methods include sequential approaches, such as Wu (1985), Bayesian methods, for example Chaloner and Larntz (1989), and the application of a maximin criterion, see Sitter (1992) and King and Wong (2000). When several variables are involved in the experiment, classical factorial or response surface designs are sometimes employed, see Myers et al. (2002). These designs are effective when resources allow each treatment to be replicated a large number of times and when the probabilities of success are not close to 0 or 1; in other circumstances such designs may be inefficient.

We address the problem of finding and assessing exact designs for a binary response described by a GLM for experiments that involve several variables. As with other nonlinear models, the design problem is complicated by the dependence of the asymptotic generalized variance of the maximum likelihood estimators of the model parameters on the unknown values of these parameters. An approach is described that uses a search algorithm to generate designs which are robust to the values of the model parameters and other aspects of the model specification. The algorithm uses simulated annealing (Haines, 1987) and is implemented in C++. It can be used to find a single design or to build designs sequentially. A method of assessing competing designs is also indicated. An example is given of a design for a GLM with logit link and the performance of this design is compared with that of a response surface design.

2. Design selection and assessment

Suppose that the treatments in the experiment are allocated at random to the units and a single observation is made on each unit. Let the n observations be held in vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, where $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{Y}) = V$. The link function $g(\cdot)$ of the GLM relates $\boldsymbol{\mu}$ to the linear predictor $\boldsymbol{x}_i'\boldsymbol{\beta}$ through $\mu_i = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta})$, $i = 1, \ldots, n$, where $\boldsymbol{\beta}$ is the vector of p unknown model parameters and $\boldsymbol{x}_i'$ is the ith row of the model matrix X. The form of the diagonal matrix V is determined by the distribution assumed for the response. For binary data, each Y_i has a Bernoulli distribution and $V = \mathrm{diag}\{\mu_i(1-\mu_i)\}$; appropriate link functions are the probit, the complementary log-log and the logit link which is given by

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)\ , \text{ for } i = 1, \ldots, n\,.$$

The asymptotic variance-covariance matrix of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (X'\Delta V \Delta X)^{-1}$ (see, for example, McCullagh and Nelder, 1989, p. 119), where $\Delta = \mathrm{diag}(g'(\mu_i))_{i=1}^n$ and the matrix $\Delta V \Delta$ involves the unknown $\boldsymbol{\beta}$. If a value of $\boldsymbol{\beta}$ is assumed, then the application of a standard optimality criterion within a search algorithm may be used to find a locally optimal design. For example, the use of D-optimality gives a design which maximizes the local objective function

$$\phi(d|\boldsymbol{\beta}) = |M(d, \boldsymbol{\beta})|^{1/p} \text{ for } d \in \mathcal{D}\,,$$

where $\mathcal{D}$ is the set of all possible designs with n runs and $M(d, \boldsymbol{\beta}) = X'\Delta V \Delta X$. This criterion is equivalent to minimizing the asymptotic confidence ellipsoid for the model parameters.

The optimum-in-average criterion of Fedorov and Hackl (1997) overcomes the dependence of design choice on $\boldsymbol{\beta}$ by selecting a design that optimizes the average of a local objective function evaluated over a predefined parameter space; see also Pettersson and Nyquist (2003) for GLM designs. We use this criterion to find exact designs in a continuous design space and note that it can be extended to incorporate uncertainty in the link function and the form of the linear predictor. For full details, see Woods et al. (2004). Uncertainty in the parameter values is incorporated through the choice of parameter space which may strongly influence the resulting designs and should reflect any prior information from the experimenters' knowledge or from pilot experiments.

To implement a search of the design space which is computationally feasible, a surrogate criterion must be formulated that can be evaluated efficiently. We apply the approach of Woods et al. (2004) in choosing a small set $\mathcal{S}$ of parameter vectors that is representative of the specified parameter space. An objective function is then defined over this set as

$$\Phi(d, \mathcal{S}) = \prod_{\boldsymbol{\beta} \in \mathcal{S}} \phi(d|\boldsymbol{\beta})\,.$$

A *compromise design* that maximizes $\Phi(d, \mathcal{S})$ over set $\mathcal{D}$ may be found by search algorithm. The efficiency of a design may be assessed by sampling from the entire parameter space, as illustrated below.

3. Example

An experiment in the food technology industry investigated the joint effects of three variables on a binary response for which a GLM with a logit link was assumed. The experimenters wanted to consider a full second order linear predictor and chose a central composite design (CCD) with 16 runs and axial points at ±1.3. As an alternative, a compromise design of the same size can be found using the same range for each of the variables. A 10-dimensional parameter space was considered which was defined by 10 intervals, one for each model parameter. For two of the variables, an interval $[2, 6]$ was used for the coefficients of the linear terms as these terms were thought likely to have a positive effect on the probability of success. Little was known about each of the remaining eight model parameters and hence intervals $[-2, 2]$ were used. The problem of selecting a representative set $\mathcal{S}$ of parameter vectors from this space is

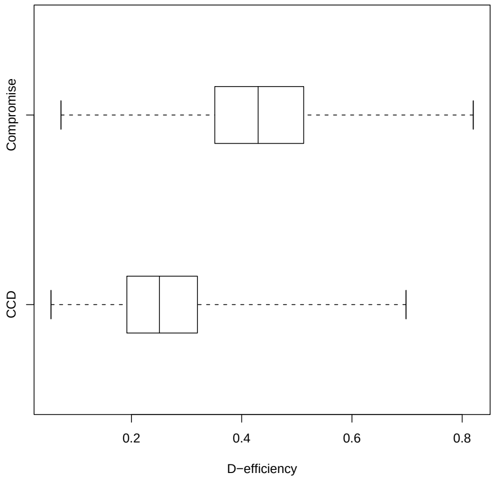

Figure 1: Boxplots of D-efficiencies for a central composite design and a compromise design relative to a locally optimal design for each of a random sample of 10,000 parameter vectors

equivalent to finding a space-filling design in 10 dimensions. This was achieved by latin hypercube sampling and a compromise design was obtained by search algorithm under the criterion of maximization of $\Phi(d, \mathcal{S})$.

The performances of the compromise and CCD designs over the specified parameter space were assessed relative to a locally optimal design d^* found for each of a random sample of 10,000 parameter vectors drawn from the space. This computationally intensive evaluation was carried out using a Beowulf cluster. Figure 1 shows boxplots of the calculated D-efficiencies, $\phi(d|\boldsymbol{\beta})/\phi(d^*|\boldsymbol{\beta})$, for each design and indicates that the compromise design is an improvement over the CCD. For example, the compromise design has probability 0.05 of being less than 25% efficient compared with the locally optimal design; the corresponding probability for the CCD is 0.5. The issue of the non-existence of maximum likelihood estimators (Silvapulle, 1981) should also be considered and leads to the use of several copies of the design, see Woods et al. (2004).

4. Discussion

In order to design an experiment for a binary or binomial response described by a GLM, the dependence of the performance of a design on the unknown model parameters must be overcome. One approach is to use a standard factorial or response surface design but, as illustrated in the above example, these designs may be outperformed by more robust designs tailored to the individual problem using available prior information. The method outlined here requires less computational effort than a corresponding Bayesian or maximin approach and can be applied to find designs for a response described by any GLM.

If the experimental units are to be arranged in blocks, then a blocking factor needs to be incorporated into the linear predictor. For fixed block effects, an extension of the above algorithm is computationally feasible. The inclusion of random block effects leads to a generalized linear mixed model. A challenge for finding designs is then to tailor the design selection to the model estimation procedure. Penalized quasi-likelihood (Breslow and Clayton, 1993) provides a simple method of estimation involving a first-order Taylor series expansion of $g(\boldsymbol{Y})$ about $\boldsymbol{\mu}$. However, this approximation is known to be poor for blocks composed of a small number of units. The incorporation into design search algorithms of recent advances in modeling, such as the use of Monte Carlo methods (McCulloch and Searle, 2001, ch.10), presents even greater computational challenges and is an area for future investigation.

Acknowledgements

Part of this work is joint with J.A.Eccleston (University of Queensland) and K.G.Russell (University of Wollongong). The first author was supported by EPSRC grant GR/R67729.

REFERENCES

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Amer. Statist. Ass.*, **88**, 9–25.

Chaloner, K. and Larntz, K. (1989) Optimal Bayesian design applied to logistic regression experiments. *J. Statist. Plann. Inference*, **21**, 191–208.

Fedorov, V. V. and Hackl, P. (1997) *Model-oriented design of experiments.* New York: Springer.

Haines, L. M. (1987) The application of the annealing algorithm to the construction of exact optimal designs for linear-regression models. *Technometrics*, **29**, 439–447.

King, J. and Wong, W. K. (2000) Minimax D-optimal designs for the logistic model. *Biometrics*, **56**, 1263–1267.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models.* London: Chapman and Hall, 2nd edn.

McCulloch, C. E. and Searle, S. R. (2001) *Generalized, Linear and Mixed Models.* New York: Wiley.

Myers, R. H., Montgomery, D. C. and Vining, G. G. (2002) *Generalized linear models with applications in engineering and the sciences.* New York: Wiley.

Pettersson, H. and Nyquist, H. (2003) Computation of optimum in average designs for experiments with finite design space. *Comm. Statist. Simulation Comput.*, **32**, 205–221.

Silvapulle, M. J. (1981) On the existence of maximum likelihood estimators for the binomial response model. *J. Roy. Statist. Soc. B*, **43**, 310–313.

Sitter, R. R. (1992) Robust designs for binary data. *Biometrics*, **48**, 1145–1155.

Woods, D. C., Lewis, S. M., Eccleston, J. A. and Russell, K. G. (2004) Designs for generalized linear models with several variables and model uncertainty. Submitted for publication.

Wu, C. F. J. (1985) Efficient sequential designs with binary data. *J. Amer. Statist. Soc.*, **80**, 974–984.

RÉSUMÉ

Les réponses binaires sont de grand intérêt dans le cadre de beaucoup d'expriences dans les sciences physiques et biologiques, notamment dans les processus chimiques de formation de sels qui ont lieu lors de la combinaison d'acides et de bases. Dans la conception des telles expriences, l'utilisation de modèles linéaires peut s'avrer inadéquate dans la description et prédiction de la réponse. Nous nous occupons de la génération pratique de designs complètement randomisés à l'aide d'algorithmes de recherche dans les cas où un modèle linéaire généralisé décrit la réponse.