Optimal Bayesian Design Applied
to Logistic Regression Experiments

^M
Kathryn Chaloner and Kinley Larntz
School of Statistics
University of Minnesota
Technical Report No. 483
December 1986

UNIVERSITY OF MINNESOTA

SCHOOL OF STATISTICS

Optimal Bayesian Design Applied
to Logistic Regression Experiments

Kathryn Chaloner and Kinley Larntz
School of Statistics
University of Minnesota
Technical Report No. 483
December 1986

A traditional way to design a binary response experiment is to design the experiment to be most efficient for a best guess of the parameter values. A design which is optimal for a best guess however may not be efficient for parameter values close to that best guess. We propose designs which formally account for the prior uncertainty in the parameter values. A design for a situation where the best guess has substantial uncertainty attached to it is very different from a design for a situation where approximate values of the parameters are known. We derive a general theory for concave design criteria for non-linear models and then apply the theory to logistic regression. Two numerical algorithms are described, an algorithm using the Nelder and Mead version of the simplex algorithm, which requires that the number of design points be specified, and an algorithm like the Wynn-Federov algorithm for linear design. The Wynn-Federov algorithm is much the slower of the two. Designs found by the Nelder-Mead algorithm are examined for a range of prior distributions and a range of criteria. The theoretical results are used to verify that the designs are indeed optimal. A general finding is that as the uncertainty in the prior distribution increases so does the number of support points in the optimal design.

KEY WORDS: Algorithm; A-optimality; Binary data; D-optimality; Equivalence theorem; Nelder-Mead algorithm; Optimization; Simplex method.

Kathryn Chaloner is Assistant Professor, Department of Applied Statistics, University of Minnesota, St Paul MN 55108. Kinley Larntz is Director and Professor, Program in Statistics, Washington State University, Pullman WA 99164-6212. The research of Kathryn Chaloner was supported by a single quarter leave award from the University of Minnesota.

1. Introduction

Designing an experiment for a binary response is important in the biological sciences for bioassay experiments and in the engineering sciences for reliability testing. We will focus specifically on a logistic regression experiment.

A logistic regression experiment is just one example of a non-linear design problem. The non-linear problem is intrinsically much harder than the linear problem as the efficiency of a design depends on the unknown parameters. A common approach is to design an experiment to be efficient for a best guess of the parameter values. This approach leads to what are termed "locally optimal" designs, the term being introduced by Chernoff(1953). A natural generalization of this approach is to use a prior distribution on the unknown parameters rather than a single guess and it is this approach we use here and justify using Bayesian arguments.

Non-Bayesian design for logistic regression is discussed at length in Finney(1978) and in Abdelbasit and Plackett(1983). Various Bayesian approaches are given in Tsutakawa(1972,1980), Zacks(1977), Freeman(1970) and Owen(1975). There is a large literature on sequential design for binary response models, see for example Wu(1985), but we will not consider the sequential approach in this paper. In many experimental situations sequential design is impractical or too expensive, (see Anbar(1984)).

In this paper we first present a unifying theory of optimal Bayesian design for non-linear problems. This theory parallels the theory for linear problems as presented in Silvey(1980) and Fedorov(1972). The theory is seen to be valid, under regularity conditions, for general concave optimality criteria. The general theory developed in section 2 discusses an equivalence theorem of

Whittle(1973) and shows how the theorem applies to Bayesian design problems. The theorem also provides a method of verifying whether or not a particular design is optimal, or close to optimal. In section 3 we discuss the logistic regression model and derive expressions for the directional derivatives needed to use the theorem. Readers interested in the specific results on how to design a logistic regression experiment may wish to omit these sections and go straight to section 4 where numerical algorithms are discussed.

Two concave criteria will be used. The first criterion corresponds to maximizing the average over the prior distribution of the log of the determinant of the information matrix. This maximizes, approximately, the expected increase in Shannon information provided by the experiment as suggested by Lindley(1956) and given a decision theoretic basis by Bernardo(1979). It was used by Läuter(1972,1974) for mixtures of linear models. This criterion is equivalent to D-optimality in linear design. The second criterion corresponds to minimizing the approximate expected posterior variance of the quantities of interest. This criterion requires the experimenter to specify what is to be estimated or what is to be predicted and the relative importance of these predictions. A weighted trace of the information matrix is then averaged over the prior distribution and minimized. In linear problems this criterion is called A-optimality.

In section 4 we discuss two algorithm developed for finding these designs. The first algorithm is an implementation of the version of the simplex algorithm developed by Nelder and Mead(1965). This algorithm finds the best design for a fixed number of design points. The second algorithm is suggested by the theory of section 2 and is an algorithm similar to that of Wynn(1970) and Fedorov(1972). The first algorithm is much faster than the second and probably the most useful in practice. The second algorithm is useful for verifying that a

design obtained from the first algorithm is indeed optimal, or almost so, and that the number of design points is sufficient.

In section 5 an experiment to investigate these designs for a range of prior distributions is described and the results discussed. The most noticeable result is that as the uncertainty in the prior distribution increases so does the number of support points required. In section 6 the designs from section 5 are compared and contrasted to the corresponding locally optimal designs. The locally optimal designs can be very inefficient in terms of the Bayesian criteria. In addition the locally D-optimal design takes observations on only two design points and does not allow for any test of the adequacy of the model. We also examine whether designs that maximize the average log determinant of the information matrix are generally efficient in terms of minimizing average variances. In section 7 we examine a few designs in more detail.

Tsutakawa(1972,1980) and Abdelbasit and Plackett(1983) restrict designs to take observations at equally spaced design points with equal numbers of observations at each point. No such restrictions are made here.

2. The theory of optimal Bayesian design

2.1 The criteria

To introduce the notation, suppose that an experiment is to be designed by choosing n values of the design variable x from an experimental region X . At each of the n values of x_i an independent observation y_i will be available. Let the data be denoted $y^T = (y_1, \dots, y_n)$ and the unknown parameters $\theta^T = (\theta_1, \dots, \theta_k)$. The contribution to the likelihood for a single observation y_i made at x is $p(y_i | \theta, x)$.

For most models using exact posterior distributions is intractable and asymptotic arguments must be used. Under easily satisfied assumptions the posterior distribution of θ is approximately a multivariate normal distribution with mean the maximum likelihood estimate, $\hat{\theta}$, and variance covariance matrix the inverse of the observed Fisher information matrix $I(\hat{\theta})$. Recall that this is the matrix of second derivatives of the log likelihood function evaluated at the maximum likelihood estimate of θ . The estimated expected Fisher information matrix could also be used but is generally less accurate. These approximations are given, for example, in Berger(1985) page 224.

By expanding the definition of a design to include any probability measure η on X we define the normalized matrix $I(\hat{\theta}, \eta)$ by

$$[I(\hat{\theta}, \eta)]_{ij} = - \int \left[\frac{\delta^2}{\delta \theta_i \delta \theta_j} \log(p(y|\theta, x)) \right]_{\theta=\hat{\theta}} \eta(dx) .$$

The posterior distribution of θ using a design measure η is therefore approximately multivariate normal $N_{\mathbf{k}}(\hat{\theta}, (nI(\hat{\theta}, \eta))^{-1})$.

Preposterior expected losses can therefore be approximated using the prior distribution of θ for the predictive distribution of $\hat{\theta}$. The first criterion we consider is to choose the design measure η maximizing $\phi_1(\eta)$ where

$$\phi_1(\eta) = E_{\theta} \log \det I(\theta, \eta) .$$

For a design measure η for which $I(\theta, \eta)$ is singular for θ values of non-zero prior probability we define $\phi_1(\eta)$ to be $-\infty$.

The second criterion is $\phi_2(\eta)$, which requires that the parameters to be estimated are specified and possibly weighted. The criterion is

$$\phi_2(\eta) = - E_{\theta}(\text{tr } B(\theta) I(\theta, \eta)^{-1}) ,$$

where $B(\theta)$ is a symmetric k by k matrix. If linear combinations of the θ_i are of interest then $B(\theta)$ does not depend on θ and is a matrix of known weights. If non-linear combinations of the θ_i 's are of interest then $B(\theta)$ has entries which are functions of θ .

We now define H to be the set of all probability measures on X and view the design problem as finding a measure in H that maximizes $\phi(\eta)$. We will only consider concave functions ϕ on H and if we assume H is compact there will be an optimal design measure in H . Both ϕ_1 and ϕ_2 are concave on H for logistic regression experiments and many other design problems.

2.2 General equivalence theorem

In many linear design problems the information matrix does not depend on θ and there is a compact set of possible information matrices to optimize over. The set is a finite dimensional subset of Euclidean space. The general equivalence theorems of optimal design are usually stated in terms of this set and directional derivatives defined on this set. This is the approach in Silvey(1980). For the non-linear problem we need to optimize over the set of measures H directly and define directional derivatives on H . Whittle(1973) uses this approach and it is this version of the general equivalence theorem we need to invoke.

For two measures η_1 and η_2 in H the derivative at η_1 in the direction of η_2 is denoted by $F(\eta_1, \eta_2)$ and is defined, when the limit exists, by

$$F(\eta_1, \eta_2) = \lim_{\epsilon \downarrow 0} \epsilon^{-1} (\phi((1-\epsilon)\eta_1 + \epsilon\eta_2) - \phi(\eta_1)) .$$

The extreme points of H are the measures which put point mass at a single point x in X . We denote such a measure η_x . If $F(\eta_1, \eta_2)$ is linear in η_2 then ϕ is said to be differentiable and

$$F(\eta_1, \eta_2) = \int F(\eta_1, \eta_x) \eta_2(dx) .$$

We further define $d(\eta, x)$ to represent $F(\eta, \eta_x)$.

Theorem 1 of Whittle(1973) is usually used in the context of linear design but we use it here for non-linear design. This theorem forms the basis for a numerical algorithm used later and is reproduced below in our notation.

Theorem (Whittle(1973))

a) If ϕ is concave, then a ϕ -optimal design η_0 can be equivalently characterized by any of the three conditions

- i) η_0 maximized $\phi(\eta)$
- ii) η_0 minimizes $\sup_{x \in X} d(\eta, x)$
- iii) $\sup d(\eta_0, x) = 0$.

b) The point (η_0, η_0) is a saddle point of F in that

$$F(\eta_0, \eta_1) \leq 0 = F(\eta_0, \eta_0) \leq F(\eta_2, \eta_0) \quad \text{for all } \eta_1, \eta_2 \in H .$$

c) If ϕ is differentiable, then the support of η_0 is contained in the set of x for which $d(\eta_0, x) = 0$ almost everywhere in η_0 measure.

The proof is given by Whittle and requires only that H is compact and that the derivatives are continuous in x .

The algorithms and results of linear optimal design theory therefore apply. The only results from linear theory not available are those which require the

existence of a finite dimensional set of information matrices. We cannot therefore guarantee that there is an optimal design on at most $k(k+1)/2$ support points as we can in linear problems. Furthermore we cannot even guarantee that there is an optimal design on a finite number of support points.

2.3 The criteria ϕ_1 and ϕ_2

The two criteria ϕ_1 and ϕ_2 are now examined and their derivatives derived.

Lemma 1

If $\log \det I(\theta, \eta)$ and $-\text{tr } B(\theta) I(\theta, \eta)^{-1}$ are both concave functions over H almost everywhere in θ , for the prior measure on θ , then ϕ_1 and ϕ_2 are concave functions on H .

This lemma can be proved easily by using the concavity for each θ and then taking expectation over θ . As expectation is a linear operator the concavity is preserved. For many non-linear design problems, including logistic regression, we have concavity for each θ .

Lemma 2

For measures η such that $I(\theta, \eta)$ is non-singular almost everywhere in the prior measure on θ we have

i) for ϕ_1

$$F(\eta_1, \eta_2) = E_{\theta} \text{tr } I(\theta, \eta_2) I(\theta, \eta_1)^{-1} - k.$$

ii) for ϕ_2

$$F(\eta_1, \eta_2) = E_{\theta} \text{tr } B(\theta) I(\theta, \eta_1)^{-1} I(\theta, \eta_2) I(\theta, \eta_1)^{-1} + \phi_2(\eta_1).$$

The proof of this lemma is straightforward, paralleling results for linear theory (see, for example, Silvey(1980) pages 21 and 48). As expectation is a linear operator the expressions are the expectations of those if θ were known. We will specifically exclude from consideration situations where the optimal

design η_0 is singular for θ values of positive prior measure. This condition is not particularly restrictive for ϕ_1 -optimality which implicitly assumes that all the parameters are of interest. This condition does however exclude some cases of ϕ_2 -optimality where there are nuisance parameters but it does not exclude logistic regression experiments where there are only two parameters.

We now proceed to use these general results for designing logistic regression experiments.

3. The logistic regression design problem

The logistic regression model is such that, for an observation at a value x of the explanatory variable, a Bernoulli response y_i is observed. There are two parameters $\theta^T = (\mu, \beta)$ and the probability of success is

$$p(x, \theta) = 1/(1 + \exp(-\beta(x-\mu)))$$

The model is a generalized linear model with the logit transformation of the expected response being linear in x . The parameter μ is the value of x at which the probability of success is .5 and β is the slope in the logit scale. The parameters μ and β can be thought of as representing the location and scale of an underlying tolerance distribution.

We will assume that X is closed and bounded and therefore H will be compact. We also suppose that the support of the prior measure is bounded. This requirement insures non-singularity of $I(\theta, \eta)$ for all measures η except for those corresponding to point mass at a particular x . These conditions are sufficient for the results of Section 2 to be applied to logistic regression.

For a design measure, η , on X putting n_i weight at x_i , $i=1, \dots, k$, $\sum n_i=1$, define

$$w_i = p(x_i, \theta)(1-p(x_i, \theta))$$

$$t = \sum_{i=1}^k n_i w_i$$

$$\bar{x} = t^{-1} \sum_{i=1}^k n_i w_i x_i$$

$$s = \sum_{i=1}^k n_i w_i (x_i - \bar{x})^2.$$

Note that w_i , t , $\bar{x}$ and s all depend on θ , but this dependence has been dropped to simplify the notation. The information matrix can be written as

$$I(\theta, \eta) = \begin{pmatrix} \beta^2 t & -\beta t(\bar{x} - \mu) \\ -\beta t(\bar{x} - \mu) & s + t(\bar{x} - \mu)^2 \end{pmatrix}.$$

The inverse of the information matrix is therefore

$$I(\theta, \eta)^{-1} = \beta^{-2} \begin{pmatrix} 1/t + (\bar{x} - \mu)^2/s & \beta(\bar{x} - \mu)/s \\ \beta(\bar{x} - \mu)/s & \beta^2/s \end{pmatrix}.$$

The ϕ_1 -criterion and several versions of ϕ_2 -optimality will now be examined and expressions will be found for the criteria and their derivatives.

3.1 ϕ_1 -optimality

For ϕ_1 -optimality, maximizing the average log determinant, the expressions simplify to give

$$\phi_1(\eta) = E_{\theta}[\log(\beta^2 ts)]$$

and if we define

$$w(x, \theta) = p(x, \theta)(1 - p(x, \theta))$$

we have

$$d(\eta, x) = E_{\theta} \{w(x, \theta)[t^{-1} + (\bar{x} - x)^2 s^{-1}]\} - 2 \quad .$$

For any design η , we see that if $|x| \rightarrow \infty$ then $w(x, \theta) \rightarrow 0$ at an exponential rate, and therefore $d(\eta, x) \rightarrow -2$. In Section 2 it was shown that at points of support of the optimal design this derivative must be exactly zero. For sufficiently wide X , the design points of the optimal design will therefore not be at the boundary of X . Rather if η_0 is the optimal design the support points will be at the roots of the function $d(\eta_0, x)$.

3.2 ϕ_2 -optimality in general

The weighted trace criterion of ϕ_2 -optimality is perhaps more appealing than ϕ_1 -optimality as it corresponds to the familiar squared error loss function. It requires that the experimenter carefully specify exactly what is to be estimated or predicted.

The specification of what is to be estimated provides the matrix $B(\theta)$. If, for example, the only parameter of interest is μ , then $B(\theta)$ can be written as cc^T with $c=(1,0)^T$. If both μ and β are of equal interest then $B(\theta)$ is the identity matrix.

It is often of interest, especially in bioassay experiments, to estimate the value of x at which the probability of success, $p(x, \theta)$, is a particular value. Suppose that we wish to estimate x_0 where $\text{logit}(p(x_0, \theta)) = \gamma$, then $x_0 = \mu + \gamma\beta^{-1}$ which is a non-linear function of θ . Standard asymptotic arguments give $B(\theta) = c(\theta)c(\theta)^T$ where $c(\theta)$ is a vector of derivatives of the non-linear function. Specifically, $c(\theta)^T = (1, -\gamma\beta^{-2})$. If several percentile response points are of interest then a distribution could be put on γ and $B(\theta)$ averaged over this distribution. This is a standard way of using the weighted trace criterion in a Bayesian framework (see, for example, Chaloner(1984), for this approach in linear models).

We implement the design algorithm for three choices of $B(\theta)$. The first choice is where μ is the only parameter of interest. The second is where the only parameter of interest is the value x such that $p(x, \theta) = .95$. The third is when we are interested in estimating the value of x at which $\text{logit}(p(x, \theta)) = \gamma$ and γ has a uniform distribution over $[-1, 1]$. This range on γ gives probabilities between .23 and .87. These three choices are by no means exhaustive of the many which might occur in real experimental situations but they represent a range of criteria which might be appropriate in particular experiments.

These three examples of ϕ_2 -optimality will be examined and optimal designs found for a range of prior distributions. The criterion functions and derivatives are now derived for these examples of ϕ_2 -optimality. Recall that the derivative involves the weighted trace of the matrix $I(\theta, \eta)^{-1} I(\theta, \eta_x) I(\theta, \eta)^{-1}$. Expressions for the entries of this matrix are given in the Appendix.

3.3 ϕ_2 -optimality for the estimation of μ

When interest is in the estimation of μ alone then the (1,1) entry of $B(\theta)$ is 1 and all other entries are zero. For any design η , with more than one support point,

$$\phi_2(\eta) = -E_{\theta}\{\beta^{-2}[\tau^{-1} + (\bar{x} - \mu)^2 s^{-1}]\}$$

and

$$d(\eta, x) = E_{\theta}\{w(x, \theta)(\beta s t)^{-2}[t(\bar{x} - x)(\bar{x} - \mu) + s]^2\} + \phi_2(\eta) .$$

Estimating μ alone is of interest in some biological applications, the parameter μ is often referred to as the LD50 as it is the 50th percentile of the underlying logistic tolerance distribution.

3.4 ϕ_2 -optimality for any or several percentile response points

Following the earlier discussion in 3.2, to estimate $\mu + \gamma\beta^{-1}$ we have

$$\phi_2(\eta) = -E_{\theta}\{\beta^{-2}[\tau^{-1} + (\gamma - \beta(\bar{x} - \mu))^2\beta^{-2}s^{-1}]\}$$

and

$$d(\eta, x) = E_{\theta}\{w(x, \theta)(\beta^2 s t)^{-2}[t(\bar{x} - x)(\beta(\bar{x} - \mu) - \gamma) + \beta s]^2\} \\ + \phi_2(\eta) .$$

Substituting $\gamma = 0$ gives the expressions in 3.3 for the special case of estimating μ alone. Substituting $\gamma=2.944$ gives the situation where we want to estimate the value of x at which $p(x,\theta)=-.95$. This value of x is referred to as the LD95.

If a distribution is put over γ to represent interest in several percentile response points then define

$$B(\theta) = \begin{pmatrix} 1 & -E(\gamma)/\beta^2 \\ -E(\gamma)/\beta^2 & E(\gamma^2)/\beta^4 \end{pmatrix} .$$

The expectation is over the distribution on γ . An example where γ has a uniform distribution on $[-1,1]$ will later be used. In this case $E(\gamma)=0$ and $E(\gamma^2)=1/3$ and the criterion and derivative become

$$\begin{aligned} \phi_2(\eta) &= -E_{\theta} \{ \beta^{-2} [t^{-1} + (\bar{x} - \mu)^2 s^{-1} + (3\beta^2 s)^{-1}] \} \\ d(\eta, x) &= E_{\theta} \{ w(x, \theta) (\beta^2 s t)^{-2} [\beta^2 (t(\bar{x} - x)(\bar{x} - \mu) + s)^2 + t^2 (\bar{x} - x)^2 3^{-1}] \} \\ &\quad + \phi_2(\eta) . \end{aligned}$$

A uniform $[-1,1]$ distribution on γ represents interest in calibrating the central part of the logistic response curve.

4. Two algorithms for finding optimal designs

We have used two basic algorithms to find optimal designs. The most effective seems to be to use the Nelder and Mead(1965) version of the simplex algorithm for optimizing a function of several variables. This method is a numerical optimization method which does not require derivatives. This method works well, especially if the number of dimensions is less than about 10. It

does require that the number of design points is specified. The algorithm also requires that there are no constraints on the variables and so the variables must be transformed to satisfy this. By evaluating the derivative $d(\eta, x)$ over possible values of x it is possible to verify whether or not a design obtained by this Nelder-Mead algorithm is indeed optimal, or very close to optimal and whether or not a sufficient number of design points have been used.

The second algorithm parallels the Wynn-Fedorov algorithm developed by Wynn(1970,1972) and Fedorov(1972). A sequence of measures $\{\eta_i\}$ is generated by

$$\eta_{i+1} = (1 - \alpha_i) \eta_i + \alpha_i \eta_{x(i+1)}$$

where $x(i+1)$ is the value of x maximizing $d(\eta_i, x)$. Therefore at each stage of the algorithm first $x(i+1)$ is chosen and then α_i is chosen. This algorithm is straightforward to program.

The choice of the region X is not very important for either algorithm if the region is large enough to give optimal design points in the interior not on the boundary. If a design η is optimal for a particular X and the derivative $d(\eta, x)$ is negative outside of X the value of ϕ cannot be improved upon by expanding the X region.

The implementation of these two algorithms will now be described. Both implementations used the same numerical integration routine. The numerical integration routine was the routine adapt from the core math library from the National Bureau of Standards (cmlib), which was edited to use double precision arithmetic. All calculations were performed on a Vax 11/750.

4.2 The Nelder-Mead algorithm

The fortran subroutine of O'Neill(1971) was used. The algorithm is for unconstrained minimization and the problem must therefore be transformed.

Suppose that we wish to find the optimal design on k points. Let the design points be x_i constrained to lie in $X=[a,b]$ and the design weights be n_i which are constrained to sum to 1. The transformed problem is an unconstrained problem in $2k-1$ variables denoted by $z_1, \dots, z_{2k-1}$. Define

$$\epsilon = 10^{-6}, \quad c = (a+b)/2, \quad \text{and } d = \pi/(a+b+2\epsilon) .$$

Then the transformations used were:

$$z_i = \tan (d(x_i - c)) \quad \text{for } i = 1, 2, \dots, k$$

$$z_{k+i} = \log (n_i/n_k) \quad \text{for } i = 1, 2, \dots, k-1 .$$

After this transformation using the algorithm is straightforward.

The output of the algorithm can be examined and the values possibly rounded. For example if the optimal design is in fact on less than k points then one or both of two things may happen. The algorithm may produce two or more x_i 's very close together or some of the optimal n_i 's may be extremely small.

The starting designs used in the implementation were usually uniform on k equally spaced points.

4.3 The Wynn-Fedorov algorithm

This algorithm was implemented for some examples but the results of using it are not reported here as the Nelder-Mead algorithm was so much more efficient. We describe the algorithm however as it is an effective, if slow, method of finding a design. It is extremely easy to program.

This algorithm was implemented by dividing X into a number of points and at each stage of the algorithm evaluating the derivative at each point in order to

choose $x(i+1)$. There are several options for the choice of step lengths, α_i . Wynn's approach is to use $1/i$ and Fedorov's approach is to choose the α_i which maximizes the increase in ϕ (for D-optimality a closed form expression for this α_i can be found). Fedorov's algorithm gives therefore a monotonically increasing criterion function whereas Wynn's step lengths do not. A combination of these two was actually implemented by starting with an α_i of $1/i$ and if this choice of α_i gave an increase in the criterion then it was used, otherwise α_i was decreased by one half until it did give an increase (up to a maximum of 20 decreases). In this way we obtain a monotone sequence. The algorithm was stopped when the maximum derivative was less than .001. Setting the convergence criterion at a smaller value of the maximum derivative was impractical as the algorithm could be very slow and take many hours of CPU time, especially for prior distributions which represent considerable uncertainty.

This implementation could be improved upon by using a more efficient way of choosing $x(i+1)$ and perhaps a different way of choosing α_i . The Nelder-Mead algorithm was much faster for the examples we examined.

4.4 Symmetry

Both of these algorithms were modified to be much more efficient for situations where the optimal design measure is symmetric. Symmetry occurs for example when the prior distribution for μ is independent of that for β and the prior distribution for μ is symmetric around some value, μ_0 say, at the center of the interval X . In addition the criterion must be invariant to which of the Bernoulli responses is labelled "success" and which "failure"; the ϕ_1 criterion is always so.

Specifically, for any design η let η^* be the design which is the reflection of η around μ_0 and suppose that ϕ is such that $\phi(\eta) = \phi(\eta^*)$. Then by concavity the

symmetric design $(\eta(x) + \eta^*(x))/2$ is at least as good as η . Therefore for every asymmetric design η there is a symmetric design which is at least as good.

The symmetric version of the algorithms are implemented by setting μ_0 to zero and $X = [-a, a]$. The design points and the design weights are both symmetric around zero.

The symmetric version of the Nelder-Mead algorithm is a optimization over only $k-1$ variables. If k is odd, one design point is fixed at zero and the other design points are at x and $-x$ for $(k-1)/2$ values of x . The design weight at x is identical to that at $-x$. If k is even there are $k/2$ values of x and $-x$ and $k/2$ weights subject to one constraint. The reduction of dimensions by a factor of 2 is very important, especially when the optimal design is on about 10 points.

In the Wynn-Fedorov algorithm the number of derivatives that need to be evaluated at each stage is reduced by one half. In this formulation the set of measures H is taken to be the set of symmetric measures on X and the extreme points η_x taken to be measures with mass .5 at x and .5 at $-x$.

As with the non-symmetric versions of the algorithm judicious rounding can simplify the designs obtained from implementing these algorithms. The symmetric versions are clearly much faster than the asymmetric versions when they are appropriate.

5.A numerical investigation

A numerical investigation using the Nelder-Mead algorithm will now be described. Designs were found for a range of prior distributions and a range of criteria. It was assumed that μ and β were independent in the prior distribution and each had a uniform distribution on an interval. Three different intervals were chosen for each parameter and all nine combinations were used to generate

nine different prior distributions in a 3×3 factorial structure. The intervals for μ were all centered at zero and were $[-.1, .1]$, $[-.3, .3]$ and $[-1.0, 1.0]$. The intervals for β were $[6.9, 7.1]$, $[6.0, 8.0]$ and $[4.0, 10.0]$ and so they were all centered at 7. These intervals were motivated by the example in Brown(1966) and represent a range of prior distributions from quite precise to quite vague. Even the vague prior distribution, however, recognizes that the sign of β is known and a range is known for μ . This knowledge would be reasonable to assume in many real experimental situations.

For each of the nine prior distributions optimal designs were found, using the Nelder-Mead algorithm, for each value of k between 2 and 11, (recall that k is the number of support points). Four criteria were used, ϕ_1 -optimality and three versions of ϕ_2 -optimality which were: first estimating the LD50 ($\gamma=0$), second estimating LD95 ($\gamma=2.944$) and third estimating the values of x where the logit of the probability of response is γ and γ has a uniform distribution over $[-1, 1]$. For ϕ_1 -optimality and for the ϕ_2 -optimality criteria other than estimating LD95 the optimal design is symmetric and therefore for these three criteria the symmetric version of the algorithm was used.

The design space X was chosen to be wide enough so that the optimal design points did not occur on the boundary. The starting designs were uniformly spaced design points with uniform weights.

The results of this experiment are summarized in the next few sections.

5.1 ϕ_1 -optimality

Table 1 highlights some of the results for maximizing the average log determinant. The table gives the maximum value of the determinant after rounding to 4 decimal places on a maximum of 11 support points. The table also gives the minimum number of support points which gave this value after rounding.

Examination of Table 1 reveals that as our uncertainty increases so does the minimum number of support points required to attain the optimal value. For μ uniform on $[-.1, .1]$ optimal designs for all three prior distributions on β are on only $k=2$ points. For μ uniform on $[-.3, .3]$ the optimal design is on $k=3$ points and for μ uniform on $[-1.0, 1.0]$ more than 7 points are needed. The designs appear to be much more sensitive to the prior distribution on μ than on β .

These designs are displayed graphically in figure 1 as probability mass functions on $[-1, 1]$.

5.2 ϕ_2 -optimality for the estimation of μ

Table 2 gives the corresponding results for minimizing the expected variance of μ . Again, as for ϕ_1 -optimality, the more uncertainty there is in the prior distribution the more support points are needed. Figure 2 is a graphical display of the designs and we can clearly see the number of support points increasing as the uncertainty increases.

5.3 ϕ_2 -optimality for the estimation of LD95

In engineering problems it may well be an extreme percentile, such as the point where the probability of response is .95, that is of interest. We therefore examine the ϕ_2 -optimal designs for this problem. Unlike the two previous examples the optimal design is not necessarily symmetric about the prior mean for μ .

Table 3 is a table derived using the same method as for tables 1 and 2. These designs are illustrated in figure 3. In figure 3 the designs are illustrated as measures on the interval $[-1.2, 1.2]$ but are otherwise on the same scale as the other figures. Note that the central plot in the figure corresponding to μ uniform on $[-.3, .3]$ and β uniform on $[6, 8]$, is supposed to be a 4 point design

but is graphically indistinguishable from a 3 point design. Although the design found has mass .55 at .426 and mass .20 at .425, we have drawn the design with mass .75 at .4265. We examine this design again in section 7 and we will see that a plot of the derivative function indicates that the optimal design should probably have 3 support points.

The designs in table 3 are interesting. They are clearly not symmetric and can be quite different depending on the prior distribution. We noted that the criterion function to be optimized in this case does not appear to be as well behaved as for the other criteria studied and could well be causing irregular behavior in the optimization algorithm.

5.4 ϕ_2 -optimality for the average percentile response point

As described earlier we used a criterion designed to minimize the average variance of $\mu + \gamma\beta^{-1}$ where γ has a uniform distribution over $[-1,1]$. That is we are interested in calibrating the central part of the response curve. The results are summarized in table 4 and figure 4, using the same method of rounding to 4 decimal places. The designs are symmetric around zero and all are within the interval $[-1,1]$. Again we see that as prior uncertainty increases so does the number of support points.

5.5 The locally optimal design

In these examples the best guess for (μ, β) would be $(0, 7)$. The locally ϕ_1 -optimal design is to put weight .5 at each of $+.2205$ and $-.2205$. These are the points at which the probability $p(x, \theta)$, for these values of the parameters θ , are .176 and .824.

5.6 General conclusions

In general as the uncertainty in the prior distribution increases so does the number of support points. This result is in accord with the results of

Tsutakawa(1972, 1980) and Abdelbasit and Plackett(1983) who restricted their designs to be equally spaced and have equal weights. The designs we found make no such restriction and indeed the designs, in general, are neither equally spaced nor have equal weight on all design points.

A feature common to all of the four design criteria was that as the uncertainty in the prior distribution increased so did the length of time the Nelder-Mead algorithm took to converge. In Section 7 designs for two prior distributions will be examined in more detail and we will see that as the uncertainty in the prior distribution increases the derivative function $d(\eta, x)$ of a design close to optimal can be very flat compared to the derivative function for a very informative prior distribution. This is consistent with the criterion value being fairly constant in the region close to the optimal value.

6. Is the ϕ_1 -optimal design really "all purpose"?

A claim often made for D-optimality is that it is an all purpose criterion and using this criterion to design experiments should lead to generally informative experiments. A natural question to ask is whether this is, in fact, true. Table 5, therefore, gives the efficiency of the locally D-optimal design, as described in section 5.5, in terms of the three ϕ_2 -optimality criteria and Table 6 the efficiencies of the ϕ_1 -optimal designs, described in section 5.1. The entries are the value of the ϕ_2 -criterion divided by the value at the optimal designs as given in Tables 2, 3 and 4.

Under all three criteria the locally optimal design performs abismally when the prior distribution on μ is uniform on $[-1,1]$. The ϕ_1 -optimal design, alternatively, takes account of the uncertainty in the parameters. The ϕ_1 -optimal design is, however, inefficient for estimating the LD95.

In summary, for these prior distributions, the ϕ_1 -optimality criterion does seem to be reasonably efficient in terms of the two symmetric ϕ_2 -optimality criteria. If the purpose of the experiment is clearly specified, however, in terms of ϕ_2 -optimality then ϕ_2 -optimality should be used.

7. Two examples

In order to further understand these designs we take two particular prior distributions and study the optimal designs in detail. We take the prior distribution where β is between 6 and 8 and two choices of the prior distribution on μ , namely μ is between $-.3$ and $.3$ and between -1 and 1 . These two prior distributions represent a moderate amount of prior information and quite vague prior information. These two examples illustrate the difference caused by changing the prior distribution on μ .

First consider the prior distribution representing a moderate amount of information, μ uniform on $[-.3, .3]$ and, independently, β uniform on $[6, 8]$. That the ϕ_1 -optimal design in table 7 is indeed optimal can be verified by plotting the derivative function, $d(\eta, x)$. The plot is given in figure 5.1 and it is clear that the function is negative almost everywhere and has exactly three roots at the design points. Plots are also given of the corresponding derivatives for the criteria for the other three optimal designs in 5.2, 5.3 and 5.4. Again, the derivative functions are not exactly everywhere negative, but almost so. For ϕ_1 -optimality and for estimating the LD95 the roots are clearly seen on the plot and correspond almost exactly to the support points. Recall that the LD95 optimal design was on four points with two very close together. The derivative plot indicates that only 3 points should be required. The other two criteria, ϕ_2 -optimality for the LD50 and for the average of percentile response points

give very similar derivative functions. The functions do have maxima at, approximately, the design points but they are hard to discern from the plot as the function is so flat.

In contrast, now consider the prior distribution, with vaguer prior information, where μ is uniform on $[-1,1]$ and β is again uniform on $[6,8]$. The designs found in section 5 for the four different criteria are given in table 8. Figure 6 gives the corresponding plot of the derivative function $d(\eta, x)$, again for values of x between -1.2 and 1.2 . In comparison to figure 5 we see that the derivative is much flatter in all four cases. The flatness of the function makes it difficult to see the roots. As the function is so flat changing the design a little can change the roots considerably, but change the criterion value very little. Nevertheless, the Nelder-Mead algorithm has clearly succeeded in finding a design which is close to optimal under each of the four criteria.

8. Conclusion

The Nelder-Mead algorithm is efficient in finding good design which are close to optimal using these Bayesian criteria. The results of section 2 are useful in understanding the problem and in verifying that the designs obtained from the Nelder-Mead algorithm are indeed close to optimal and a sufficient number of support points are being used.

The prior distributions used in this paper, uniform and independent prior distributions for μ and β , clearly do not always apply. The methodology we have presented, however, applies to any situation where bounds are available on the parameter space. We believe that we have demonstrated that this methodology is preferable to designing for a single best guess of the parameter values. We also believe that this methodology provides a framework for the careful examination

of any experimental procedure in which the data analysis will include fitting a logistic regression model.

These design are easy to derive. Even if an experimenter can not use one of these design exactly they provide a useful benchmark with which to calibrate other designs. They promise to be quite useful as the number of support points is generally greater than the number of parameters and therefore the designs enable the appropriateness of the model to be examined after the experiment is performed. This methodology is clearly effective for the important problem of experimental design for logistic regression and may well be useful in other non-linear design problems.

APPENDIX

Denote the matrix A by

$$A = I(\theta, \eta)^{-1} I(\theta, \eta_x) I(\theta, \eta)^{-1}$$

then the derivative for the ϕ_2 -criterion is

$$d(\eta, x) = E_\theta \operatorname{tr} B(\theta) A + \phi_2(\eta) .$$

The entries of the matrix A can be written as the following:

$$A_{11} = w(x, \theta) (\beta s t)^{-2} [t(\bar{x} - x)(\bar{x} - \mu) + s]^2$$

$$A_{12} = w(x, \theta) (\beta t s^2)^{-1} (\bar{x} - x) [t(\bar{x} - x)(\bar{x} - \mu) + s]$$

$$A_{21} = A_{12}$$

$$A_{22} = w(x, \theta) (\bar{x} - x)^2 s^{-2} .$$

These expressions simplify the algebra required to find the derivatives for different choices of $B(\theta)$.

REFERENCES

- Abdelbasit, K.M. and Plackett, R.L. (1983). Experimental design for binary data. Jour. Amer. Statist. Assoc. 78, 90-98.
- Anbar, D. (1984). Stochastic approximation methods and their use in bioassay and phase I clinical trials. Commun. Statist. - Theor. Meth., 13(19), 2451-2467.
- Berger, J.O. (1985). Statistical Decision Theory and Bayesian Analysis (second edition), Springer-Verlag, New York.
- Bernardo, J.M. (1977). Expected information as expected utility. Ann. Statist. 7, 686-690.
- Brown, B.W., Jr., (1966). Planning a quantal assay of potency. Biometrics, 22, 322-329.
- Chaloner, K. (1984). Optimal Bayesian experimental design for linear models. Ann. Statist. 12, 283-300.
- Chernoff, H. (1953). Locally optimal designs for estimating parameters. Ann. Math. Statist. 24, 586-602.
- Fedorov, V.V. (1972). The Theory of Optimum Experiments. Transl. and edited W.J. Studden and E.M. Klimko, Academic Press, New York.
- Finney, D.J. (1978) Statistical Methods in Biological Assay, (third edition), Macmillan, New York.
- Freeman, P.R. (1970). Optimal Bayesian sequential estimation of the median effective dose. Biometrika, 57, 79-89.
- Läuter, E. (1974) Experimental design in a class of models. Math. Oper. u. Statist., 5, 379-396.
- Läuter, E. (1976). Optimal multipurpose designs for regression models. Math. Oper. u. Statist., 7, 51-68.
- Lindley, D.V. (1956). On a measure of information provided by an experiment. Ann. Math. Statist. 27, 986-1005.
- Nelder, J.A. and Mead, R. (1965). A simplex method for function minimization. Computer J., 7, 308-313.
- O'Neill, R. (1971). Function minimization using a simplex procedure, Algorithm AS47. Applied Statistics, 20, 338-345.
- Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Jour. Amer. Statist. Assoc., 70, 351-356.

- Silvey, S.D. (1980). Optimal Design. Chapman and Hall, London and New York.
- Tsutakawa, R.K. (1972). Design of an experiment for bioassay. Jour. Amer. Statist. Assoc., 67, 584-590.
- Tsutakawa R.K. (1980). Selection of dose levels for estimating a percentage point of a logistic quantal response curve. Applied Statistics, 29, 25-33.
- Whittle, P. (1973). Some general points in the theory of optimal experimental design. Jour. R. Statist. Soc., Ser. B, 35, 123-130.
- Wynn, H.P. (1970). The sequential generation of D-optimal experimental designs. Ann. Math. Statist., 41, 1655-1664.
- Wynn, H.P. (1972). Results in the theory and construction of D-optimum experimental designs. Jour. R. Statist. Soc., Ser. B, 34, 133-147.
- Wu, C.F.J. (1985). Efficient sequential designs with binary data. Jour. Amer. Statist. Assoc., 80, 974-984.
- Zacks, S. (1977) Problems and approaches in design of experiments for estimation and testing in non-linear problems. In: Multivariate Analysis IV, ed. P.R. Krishnaiah, North-Holland, Amsterdam, 209-223.

		μ		
		$[-.1, .1]$	$[-.3, .3]$	$[-1.0, 1.0]$
	$[6.9, 7.1]$	-3.0407 (2)	-3.3729 (3)	-4.4804 (7)
β	$[6.0, 8.0]$	-3.0519 (2)	-3.3786 (3)	-4.5783 (7)
	$[4.0, 10.0]$	-3.1461 (2)	-3.4297 (3)	-4.5610 (8)

TABLE 1: Maximum value of ϕ_1 -optimality criterion for 9 different prior distributions. The numbers in parenthesis are the minimum number of design points required to obtain this value to 4 decimal places using the Nelder-Mead algorithm.

		μ		
		$[-.1, .1]$	$[-.3, .3]$	$[-1.0, 1.0]$
	$[6.9, 7.1]$	.1101 (2)	.1585 (3)	.3523 (6)
β	$[6.0, 8.0]$	.1121 (2)	.1611 (3)	.3561 (6)
	$[4.0, 10.0]$	.1321 (2)	.1863 (3)	.3915 (7)

TABLE 2: Minimum average variance for estimating μ for 9 different prior distributions. The numbers in parenthesis are the minimum number of design points required to obtain this value to 4 decimal places using the Nelder-Mead algorithm.

		μ		
		$[-.1, .1]$	$[-.3, .3]$	$[-1.0, 1.0]$
	[6.9, 7.1]	.4454 (2)	.6086 (8)	1.2652 (8)
β	[6.0, 8.0]	.4623 (2)	.6253 (4)	1.2806 (8)
	[4.0, 10.0]	.6339 (2)	.7944 (4)	1.4301 (10)

TABLE 3: Minimum average variance for estimating LD95 for 9 different prior distributions. The numbers in parenthesis are the minimum number of design points required to obtain this value to 4 decimal places using the Nelder-Mead algorithm.

		μ		
		$[-.1, .1]$	$[-.3, .3]$	$[-1.0, 1.0]$
	[6.9, 7.1]	.1452 (2)	.1873 (3)	.3959 (6)
β	[6.0, 8.0]	.1487 (2)	.1906 (3)	.4002 (7)
	[4.0, 10.0]	.1853 (2)	.2244 (3)	.4413 (7)

TABLE 4: Minimum average variance for estimating the points where the logit response is between -1 and 1 for 9 different prior distributions. The numbers in parenthesis are the minimum number of design points required to obtain this value to 4 decimal places using the symmetric version of the Nelder-Mead algorithm.

		μ			
		$[-.1, .1]$	$[-.3, .3]$	$[-1.0, 1.0]$	
β	[6.9, 7.1]	i)	1.27	1.02	211.
		ii)	1.53	1.65	81.3
		iii)	1.10	1.03	190.
	[6.0, 8.0]	i)	1.26	1.02	233.
		ii)	1.53	1.66	88.5
		iii)	1.10	1.04	210.
	[4.0, 10.0]	i)	1.21	1.03	480.
		ii)	1.62	1.75	165.
		iii)	1.04	1.06	430.

TABLE 5: The efficiency of the locally D-optimal design in terms of the three versions of ϕ_2 -optimality. The entries in the table are the value of the average variance under the locally D-optimal design divided by the minimum variance as given in tables 2, 3, and 4 for the three ϕ_2 criteria.

The criteria are

- i) minimizing average variance of LD50
- ii) minimizing average variance of LD95
- iii) minimizing average variance of $\mu + \gamma\beta^{-1}$ for γ uniformly distributed on $[-1, 1]$.

		μ			
		$[-.1, .1]$	$[-.3, .3]$	$[-1.0, 1.0]$	
β	[6.9, 7.1]	i)	1.28	1.02	1.07
		ii)	1.52	1.41	1.19
		iii)	1.12	1.00	1.06
	[6.0, 8.0]	i)	1.28	1.02	1.15
		ii)	1.52	1.40	1.39
		iii)	1.11	1.00	1.15
	[4.0, 10.0]	i)	1.21	1.01	2.14
		ii)	1.62	1.43	2.41
		iii)	1.04	1.00	1.06

TABLE 6: The efficiency of the ϕ_1 -optimal design in terms of the three versions of ϕ_2 -optimality. The entries in the table are the value of the average variance under the ϕ_1 -optimal design divided by the minimum variance as given in tables 2, 3 and 4 for the three ϕ_2 criteria.

The criteria are

- i) minimizing average variance of LD50
- ii) minimizing average variance of LD95
- iii) minimizing average variance of $\mu + \gamma\beta^{-1}$ for γ uniformly distributed on $[-1, 1]$.

ϕ_1 -optimal	x	-.3083	0.0	.3083	
	$\eta(x)$	.3683	.2634	.3683	
ϕ_2 -optimal for estimating LD50	x	-.2536	0.0	.2536	
	$\eta(x)$	.4535	.0930	.4535	
ϕ_2 -optimal for estimating LD95	x	-.3594	.0231	.4263	.4254
	$\eta(x)$	.0283	.2188	.5504	.2025
ϕ_2 -optimal for γ -uniform on [-1,1]	x	-.2806	0.0	.2806	
	$\eta(x)$	.4047	.1906	.4047	

TABLE 7: The optimal designs giving the optimal criterion values to 4 decimal places as given in tables 1, 2, 3 and 4 as calculated by the Nelder-Mead algorithm for a prior distribution where μ is uniform on $[-.3, .3]$ and independent of β which is uniform on $[6, 8]$.

ϕ_1 -optimal	x	.9526	-.5660	-.2672	0.0
	$\eta(x)$	.1179	.1563	.1534	.1447
	x	.9526	.5660	.2672	
	$\eta(x)$	.1179	.1563	.1534	
ϕ_2 -optimal for estimating LD50	x	-.9436	-.5345	-.1752	
	$\eta(x)$	.1812	.1627	.1560	
	x	.9436	.5345	.1752	
	$\eta(x)$	.1812	.1627	.1560	
ϕ_2 -optimal for estimating LD95	x	-.7794	-.5294	-.3251	-.0700
	$\eta(x)$	.0327	.0859	.1343	.1761
	x	.2421	.5871	.5917	1.0408
	$\eta(x)$	.1700	.0876	.0850	.2284
ϕ_2 -optimal for γ uniform on [-1,1]	x	-.9608	-.5700	-.2594	0.0
	$\eta(x)$	.1691	.1511	.1246	.1103
	x	.9608	.5700	.2594	
	$\eta(x)$	.1691	.1511	.1246	

TABLE 8: The optimal designs giving the optimal criterion values to 4 decimal places as given in tables 1, 2, 3 and 4 as calculated by the Nelder-Mead algorithm for a prior distribution where μ is uniform on $[-1,1]$ and independent of β which is uniform on $[6,8]$.

Figure 1: ϕ_1 -optimal designs from Table 1 plotted as probability mass functions on $[-1, 1]$.

Figure 2: ϕ_2 -optimal designs from Table 2 for estimating the LD50, μ , plotted as probability mass functions on $[-1, 1]$.

Figure 3: ϕ_2 -optimal designs from Table 3 for estimating the LD95 plotted as probability mass functions on $[-1.2, 1.2]$.

μ $[-.1, .1]$ $[-.3, .3]$ $[-1, 1]$ $[6.9, 7.1]$ β $[6, 8]$ $[4, 10]$

Figure 4: ϕ_2 -optimal designs from Table 4 for estimating the average percentile response plotted as probability mass functions on $[-1, 1]$.

Figure 5.1: ϕ_1 -optimality.

Figure 5.2: ϕ_2 -optimality for LD50.

Figure 5.3: ϕ_2 -optimality for LD95.

Figure 5.4: ϕ_2 -optimality for average percentile.

Figure 5: Derivative functions $d(\eta, x)$ plotted against x for each of the four optimal designs found for the prior distribution with μ uniform on $[-.3, .3]$ and β uniform on $[6, 8]$.