Optimal Design and Lack of Fit in Nonlinear Regression Models

Timothy E. O'Brien

ABSTRACT: This paper points out that so-called optimal designs for nonlinear regression models are often limited when the assumed model function is not known with complete certainty and argues that robust designs - near optimal designs but with extra support points - can be used to also test for lack of fit of the model function. A simple robust design strategy - which has been implemented with a popular software package - is also presented and illustrated.

KEYWORDS: D-optimality; Model mis-specification ; Nonlinear models; Robust designs

1 Introduction

Researchers often find that nonlinear regression models are more applicable for modelling various biological or chemical processes than are linear ones since they tend to fit the data well and the models and model parameters are more scientifically meaningful. These researchers are often in a position of obtaining optimal or near-optimal designs for a given nonlinear model. A common shortcoming of most optimal designs for nonlinear models used in practical settings, though, is that these designs often have only p support points where p is the number of model parameters. Such designs may present no problem when the model function is assumed to be known with complete certainty, but researchers typically desire designs which are near-optimal for the assumed model but which contain "extra" design points that can be used to test for model adequacy. This paper introduces and illustrates such a "robust" design procedure which also has been implemented using a popular software package.

In G.U.H Seeber, Francis, B.J. , Hatzinger R & Steckel-Berger, G
(editors) 1995, Statistical Modelling
New York : Springer pp201-206.

2 Optimal Design Theory

The design problem for the nonlinear model

$$y_i = \eta\left(\mathbf{x}_i, \theta\right) + \varepsilon_i \qquad i = 1, ..., n \tag{1}$$

typically involves choosing an n-point design, ξ, to estimate some function of the p-dimensional parameter vector, θ, with high efficiency. Here ξ can be written as

$$\xi = \left\{ \begin{array}{c} \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \\ \omega_1, \omega_2, ..., \omega_n \end{array} \right\}$$

where the design vectors (or points), $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$, are elements of the designs space Δ and are not necessarily distinct, and the associated weights, $\omega_1, \omega_2, ..., \omega_n$, are non-negative real numbers that sum to one. Alternatively, ξ can be expressed in terms of its m (m $\leq$ n) distinct support points s_1, s_2, ..., s_m, and their associated weights, λ_1, λ_2, ..., λ_m.

When the residuals in (1) are uncorrelated Gaussian random variables with zero mean and constant variance (without loss of generality taken to equal one), the Fisher information matrix is given by $\mathbf{M}(\xi,\theta) = \mathbf{V}^T \Omega \mathbf{V}$, where $\mathbf{V}$ is the n×p Jacobian of η and $\Omega = \mathrm{diag}\{\omega_1, \omega_2, ..., \omega_n\}$, and the corresponding variance function of η for the design ξ is given by

$$d(\mathbf{x}, \xi, \theta) = \frac{\partial \eta(\mathbf{x})}{\partial \theta^T} \mathrm{M}^{-1}(\xi, \theta) \frac{\partial \eta(\mathbf{x})}{\partial \theta} \tag{2}$$

where $\partial\eta(\mathrm{x}) / \partial\theta$ is of dimension p×1 and a generalized inverse is used whenever M is singular.

Optimal designs typically minimize some convex function of $\mathbf{M}^{-1}$. For example, designs which minimize the determinant $\left|\mathbf{M}^{-1}(\xi, \theta^o)\right|$ are called locally D-optimal, and those that minimize the maximum (over all $\mathbf{x} \in \Delta$) of the variance function in (2) are called locally G-optimal; the term "locally" is used here to emphasize that the design is based on an initial estimate of the parameter vector, θ^o. Further, the General Equivalence Theorem of Kiefer and Wolfowitz (1960) establishes the equivalence between locally D- and G-optimal designs; a corollary to this theorem states that the variance function in (2) evaluated using a D-optimal design achieves its maximum value at the support points of this design.

To illustrate, consider the two-parameter intermediate product model function (IP2),

$$\eta\left(x, \theta_1, \theta_2\right) = \frac{\theta_1}{\theta_1 - \theta_2} \left(e^{-\theta_2 x} - e^{-\theta_1 x}\right),$$

and the initial parameter estimates $\theta_1^o = 0.70$ and $\theta_2^o = 0.20$. In this case the locally D-optimal design, ξ_D, associates the weight $\lambda = \frac{1}{2}$ with each of the support points $s_1 = 1.229$ and $s_2 = 6.858$. D-optimality of this design can be established by noting that the corresponding variance function, $d(x,\xi_D,\theta^o)$, reaches its maximum value at x = 1.229 and x = 6.858. An algorithm is

presented in O'Brien (1995) which uses the SAS$^{(R)}$ software package and the results of the General Equivalence Theorem to obtain and verify locally D-optimal designs.
For the previous example, note that the number of support points of the D-optimal design is equal to the number of parameters (2). Thus, regardless of the final sample size chosen, half of the observations are to be taken at x = 1.229 and the other half at x = 6.858. Although this design is "optimal" for estimating the two parameters of the IP2 model function, it provides no opportunity to check the validity of the assumed model function. The design strategy introduced below is suggested for situations where robust designs, or designs with "extra" support points, are desired.

3 Lack of Fit in Regression Models

When replicates are taken at at least one of a design's support points, lack of fit of the assumed model function (to the means model) can be tested using the F-statistic

$$F_{m-p,n-m} = \frac{SSLF\,/\,(m-p)}{SSPE\,/\,(n-m)}$$

where SSLF and SSPE are the lack-of-fit and pure-error sums of squares, respectively, m is the number of support points, n is the final sample size and p is the number of model parameters. This statistic may be used to test lack of fit for nonlinear models when intrinsic curvature is negligible, and can be adjusted using the methods given in Hamilton and Wiens (1987) when it is not. Obviously, if the number of support points of a given design is equal to the number of model parameters, no test for inadequacy of the assumed model can be made using this test.
For a given final sample size (n) and model function with p parameters, we are interested in determining (in some sense) the optimal number of support points to maximize the power of this lack-of-fit test. One such measure is to choose m to minimize the quantile F statistic, $F_{\alpha,m-p,n-m}$, for, say, $\alpha = 0.05$. Interestingly, our preliminary research has shown that, at least when n > 2p, this "optimal" m is approximately $\frac{n+2p}{3}$. For example, if a sample size of n = 20 is used to fit the IP2 model function (where p = 2), then indeed $F_{.05,m-p,n-m}$ is minimized for m = 8. The main point here is that to test for lack of fit of the assumed model function, we usually wish to choose designs with more than p support points.

4 A Robust Design Strategy

Strategies to obtain robust designs, or designs with "extra" support points, include Bayesian D-optimality discussed in Chaloner and Larntz (1989), Qoptimality introduced in O'Brien (1992) and the nesting design approach given in O'Brien (1994). As none of methods are intended to obtain designs to test for general departures of the assumed model function, we recommend the following strategy be used in these situations.

4.1 The Algorithm

- Find the (locally) D-optimal design, ξ_D. Often this will have only p support points (see Gaffke, 1987), and the following assumes that this is indeed the case.
- Choose the number "de" between 0 and 1 (typically around .90), and find all (t) values of x such that

$$d(x, \xi_D, \theta^o) = p\left[\left(\frac{p+1}{p}\,de\right)^p - 1\right] \tag{3}$$

- Choose as a final design r_1 replicates of the p support points of ξ_D and r_2 replicates of the t points obtained from the previous step.

4.2 Justification

Our motivation for suggesting the above algorithm is the following. Let ξ_x represent a one-point design putting all weight at the support point x. Then the design $\xi_N = \frac{p}{p+1}\xi_D + \frac{1}{p+1}\xi_x$ associates the weight $\frac{1}{p+1}$ with each of s_1, s_2, ...,s_p (the support points of ξ_D) and x. One measure of the "distance" between ξ_D and ξ_N is the D-efficiency (Atkinson and Donev, 1992)

$$de = \left[\frac{|M(\xi_N)|}{|M(\xi_D)|}\right]^{1/p}$$

which, in the current case, is equal to $\frac{p}{p+1}\left[1 + d(x, \xi_D, \theta^o)\,/\,p\right]^{1/p}$. Solving this expression for $d(x, \xi_D, \theta^o)$ in terms of "de" yields the expression in (3).

4.3 Implementation

Programs to obtain and verify D-optimal designs for nonlinear regression models and to graph the corresponding variance function using the SAS$^{(R)}$ software package are given in O'Brien (1994). These programs have also been adapted to obtain the t values of x which satisfy (3) for a given D-efficiency, "de". Also, these programs (available from the author) have also been extended to incorporate Bayesian D-optimality, subset D-optimality, and heteroskedastic error structures.

4.4 An Example

By way of illustration, consider again the IP2 model function with initial parameter estimates $\theta_1^o = 0.70$ and $\theta_2^o = 0.20$. Taking de = 0.90, we include those values of x such that the corresponding variance function equals 1.645; this yields the t = 4 values x = 0.761, 1.909, 4.890, and 9.366. The final design would then consist of r_1 replicates of the D-optimal support points 1.229 and 6.858 and r_2 replicates of the check points 0.761, 1.909, 4.890, and 9.366. This design has the advantage of having "extra" support points to test for model mis-specification, yet in being "near optimal" in the sense of having a reasonably high final efficiency.

4.5 Final Efficiencies

Denote

$$\xi_x = \left\{ \begin{array}{c} x_1, x_2, ..., x_t \\ \frac{1}{t}, \frac{1}{t}, ..., \frac{1}{t} \end{array} \right\}$$

where x_1, x_2, ...,x_t are the t points that satisfy (3) for given values of "de" and p, and note that $M(\xi_x) = \frac{1}{t}V_x^T V_x$, where V_x is the t × p Jacobian matrix evaluated at x_1, x_2, ...,x_t. Then the final D-efficiency of the design with r_1 replicates of the D-optimal support points and r_2 replicates of the support points of ξ_x is given by

$$de_F = \frac{(r_1 p)^{1-\frac{t}{p}}}{r_1 p + r_2 t} \left| r_1 p \mathbf{I}_t + r_2 \mathbf{D}(\mathbf{x}, \xi_D, \theta^o) \right|^{1/p}, \tag{4}$$

where $\mathbf{D}(\mathbf{x}, \xi_D, \theta^o) = \mathbf{V}_x \mathbf{M}^{-1}(\xi_D) \mathbf{V}_x^T$ is the corresponding *variance-covariance matrix function* (c.f., equation (2)). Note that for the IP2 model function and the six-point design comprising one replicate of ξ_D and one replicate of ξ_x, the final D-efficiency is 88%, and this number can be increased (up to 100%) by increasing the number of observations chosen at the support points of ξ_D.

5 Discussion

The above robust design strategy has been used for sixteen data sets and model functions used in practical settings including the four-parameter log-logistic model used by Vølund (1978) to model a process with one explanatory variable and a five-parameter growth model used by Gerig, *et al.* (1989) to detect synergy of two chemicals. In all settings, near-optimal designs were easily obtained with the SAS software package, a package commonly used by practitioners. The key to this design procedure is its simplicity for obtaining near-optimal designs with extra support points for situations where the model function is not known with complete certainty.

References

Atkinson, A.C. and Donev, A.N. (1992). *Optimum Experimental Designs.* Oxford: Clarendon Press.

Chaloner, K. and Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments. *J. Stat. Plann. Inf.*, **21**, 191-208.

Gaffke, N. (1987). On D-optimality of exact linear regression designs with minimum support. *J. Stat. Plann. Inf.*, **15**, 189-204.

Gerig, T.M., Blum, U., Meier, K. (1989). Statistical analysis of the joint inhibitory action of similar compounds. *J. Chem. Ecol.*, **15**, 2403-2412.

Hamilton, D., Wiens, D. (1987). Correction factors for F ratios in nonlinear regression. *Biometrika*, **74**, 423-5.

Kiefer, J., Wolfowitz, J. (1960). The equivalence of two extremum problems. *Can. J. Math.*, **12**, 363-6.

O'Brien, T.E. (1992). A note on quadratic designs for nonlinear regression models. *Biometrika,* **79**, 847-9.

O'Brien, T.E. (1994). A new robust design strategy for sigmoidal models based on model nesting. In Dutter, R. and Grossmann, W., eds., *Proceedings in Computational Statistics: Compstat, 1994,* Heidelberg: Physica-Verlag, 97-102.

O'Brien, T.E. (1995). Obtaining and verifying optimal designs for nonlinear regression models using SAS software. To appear in *Proceedings of SUGI 20.*

Vølund, A. (1978). Application of the four-parameter logistic model to bioassay: comparison with slope ratio and parallel line models. *Biometrics,* **34**, 357-65.