Statistica Sinica **19** (2009), 721-730

D-OPTIMAL DESIGNS FOR POISSON REGRESSION MODELS

K. G. Russell[1], D. C. Woods[2], S. M. Lewis[2] and J. A. Eccleston[3]

[1] *University of Wollongong,* [2] *University of Southampton and* [3] *University of Queensland*

Abstract: We consider the problem of finding an optimal design under a Poisson regression model with a log link, any number of independent variables, and an additive linear predictor. Local D-optimality of a class of designs is established through use of a canonical form of the problem and a general equivalence theorem. The results are applied in conjunction with clustering techniques to obtain a fast method of finding designs that are robust to wide ranges of model parameter values. The methods are illustrated through examples.

Key words and phrases: Clustering, locally optimal design, log-linear models, robust design.

1. Introduction and Notation

We consider experiments in which the ith observation has a Poisson distribution with rate λ_i dependent on p independent variables through the log-linear model

$$\ln(\lambda_i) = \eta_i = f(x_i)^{\mathrm{T}}\beta = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\,, \quad i = 1, \ldots, n\,, \tag{1.1}$$

where $x_i = (x_{1i}, \ldots, x_{pi})^{\mathrm{T}}$, $f(x_i) = (1, x_i^{\mathrm{T}})^{\mathrm{T}}$, $\beta_0, \ldots, \beta_p$ are unknown constants, and $\beta_j \neq 0$ for $j > 0$ (see McCullagh and Nelder (1989, Chap. 6)). Our aim is to find a design for an experiment that enables efficient estimation of $\beta = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$ in the sense of minimizing the volume of the $100(1-\alpha)\%$ confidence ellipsoid for β, that is, a D-optimal design. A complication is that, in common with all non-linear models, the optimal design depends on the unknown values of the model parameters. Locally optimal designs can be found by assuming particular values for the parameters that can be updated in a sequence of experiments (see Atkinson, Donev and Tobias (2007, Chap. 17)). Alternative ways of overcoming parameter dependence are through Bayesian design (Chaloner and Larntz (1989) and Firth and Hinde (1997)), maximin criteria (Sitter (1992) and Biedermann, Dette and Pepelyshev (2006)), and compromise or

parameter-robust design (Woods, Lewis, Eccleston and Russell (2006) and Dror and Steinberg (2006)).

Little guidance is available on how to design a multivariable experiment for Poisson regression. For single variable toxicology experiments, Minkin (1993) found locally optimal designs for estimation of the slope parameter in terms of an "effective dose", and compared the performance of the optimal designs with designs having different numbers of equally-spaced support points. For models with one or two variables, Wang, Myers, Smith and Ye (2006) investigated the dependence of locally D-optimal designs on functions of the parameter values, and Wang, Smith and Ye (2006) developed sequential designs. For a single variable, Ford, Torsney and Wu (1992) used a transformation of the design space to a canonical form, together with geometric arguments, to find locally optimal designs for a class of nonlinear models including Poisson regression.

2. Locally D-Optimal Designs

An approximate design $\xi \in \Xi$ in design space $\mathcal{X}$ with finite support is represented as

$$\xi = \left\{ \begin{array}{c} x_1\ x_2 \ldots x_s \\ \nu_1\ \nu_2 \ldots \nu_s \end{array} \right\},$$

where $x_i \in \mathcal{X}$, $\mathcal{X}$ is a compact subset of $\mathbb{R}^p$, $0 < \nu_i \leq 1$ and $\sum_{i=1}^{s} \nu_i = 1$. Under (1.1), the information matrix for ξ is

$$\begin{aligned} M(\xi, \beta) &= \sum_{i=1}^{s} \nu_i w(x_i) f(x_i) f(x_i)^{\mathrm{T}} \\ &= X^{\mathrm{T}} W X\,, \end{aligned}$$

where $w(x_i) = \lambda_i = \exp(\eta_i)$, $X = (f(x_1), \ldots, f(x_s))^{\mathrm{T}}$ and $W = \operatorname{diag}\{\nu_i w(x_i)\}_{i=1}^{s}$.

We want to find a D-optimal design ξ^* satisfying

$$|M(\xi^*, \beta)| = \max_{\xi \in \Xi} |M(\xi, \beta)|\,.$$

The efficiency of a design ξ is then measured relative to ξ^* as

$$\left\{ \frac{|M(\xi, \beta)|}{|M(\xi^*, \beta)|} \right\}^{\frac{1}{p+1}}. \tag{2.1}$$

In order to suppress the dependence of this design problem on β, we follow Ford et al. (1992) and apply a linear transformation to $f(x_i)$ to obtain $f(z_i) = Bf(x_i)\,, \quad i = 1, \ldots, s$, where $z_i = (z_{1i}, \ldots, z_{pi})^{\mathrm{T}} \in \mathcal{Z}$,

$$B = \begin{pmatrix} B_{11} & 0 \\ 0 & B_{22} \end{pmatrix}, \qquad B_{11} = \begin{pmatrix} 1 & 0 \\ \beta_0 & \beta_1 \end{pmatrix},$$

$B_{22} = \operatorname{diag}\{\beta_2, \ldots, \beta_p\}$ and $\beta_j \neq 0$ $(j = 1, \ldots, p)$. It follows from (1.1) that $\eta_i = \left(B^{-1} f(z_i)\right)^{\mathrm{T}} \beta = \sum_{j=1}^{p} z_{ji}$. Let $\psi \in \Psi$ be a design measure over the induced design space $\mathcal{Z}$. Then

$$\psi = \left\{ \begin{matrix} z_1 \, z_2 \ldots z_s \\ \nu_1 \, \nu_2 \ldots \nu_s \end{matrix} \right\}.$$

As the D-optimal design criterion is invariant to a linear transformation of the design space (see, for example, Pukelsheim (1993, Chap. 6)), it is sufficient to find a locally optimal design over Ψ; see also Ford et al. (1992).

We consider a design region $\mathcal{X}$ of rectangular shape. In this common situation, the above transformation maps $\mathcal{X}$ to another rectangular region. Sitter and Torsney (1995) gave an alternative transformation, which need not be shape preserving, and described induced design regions of various shapes.

Let e_j be the j*th* column vector of the $p \times p$ identity matrix, $j = 1, \ldots, p$.

Lemma. *With* $\eta_i = \left(B^{-1} f(z_i)\right)^{\mathrm{T}} \beta = \sum_{j=1}^{p} z_{ji}$, *where* $a_j \leq z_{ji} \leq b_j$, *for* a_j, b_j *constants, and* $b_j - a_j \geq 2$ $(j = 1, \ldots, p)$, *a D-optimal design for the canonical first-order Poisson regression model is given by*

$$\psi^* = \left\{ \begin{matrix} z_1^* & z_2^* & \cdots & z_{p+1}^* \\ \frac{1}{p+1} & \frac{1}{p+1} & \cdots & \frac{1}{p+1} \end{matrix} \right\},$$

where $z_j^* = b - 2e_j$, $j = 1, \ldots, p$, *and* $z_{p+1}^* = b$ *for* $b = (b_1, \ldots, b_p)^{\mathrm{T}}$.

The proof is outlined in the Appendix. Clearly, the lemma can be extended to regions $\mathcal{Z}$ where some or all of the components are unbounded from below.

Example 1. Wang, Myers, Smith and Ye (2006) reported D-optimal designs for (1.1) with $p = 1$ and $p = 2$, and support points defined in terms of $q_i = \lambda_i / \lambda_c$, where $\lambda_c = \exp(\beta_0)$ and $\lambda_i = \exp(\eta_i)$. In their context of toxicity studies, where $x_{ji} \geq 0$ and $\beta_j < 0$ $(i = 1, \ldots, s;\ j = 1, 2)$, the canonical variables satisfy $z_{1i} \leq \beta_0$, $z_{2i} \leq 0$. The D-optimal support points are $\{z_1 : \beta_0 - 2, \beta_0\}$ for $p = 1$, and $\{(z_1, z_2) : (\beta_0 - 2, 0), (\beta_0, -2), (\beta_0, 0)\}$ for $p = 2$. At these support points, $q_1 = \exp(-2)$, $q_2 = 1$ for $p = 1$, and $q_1 = q_2 = \exp(-2)$ and $q_3 = 1$ for $p = 2$, matching the results from Wang, Myers, Smith and Ye (2006).

An optimal design in $\mathcal{X}$ space for finitely bounded variables follows directly from the Lemma by application of the inverse transformation to obtain $f(x_i) = B^{-1} f(z_i)$, $i = 1, \ldots, p + 1$.

Theorem. *A D-optimal design for Poisson regression, with* $\eta_i = f(x_i)^{\mathrm{T}} \beta$, $l_j \leq x_{ji} \leq u_j$ *and* $|\beta_j (u_j - l_j)| \geq 2$ $(j = 1, \ldots, p)$, *has the* $p + 1$ *equally weighted*

support points:

$$x_j^* = c - \Big(\frac{2}{\beta_j}\Big)e_j,\ j = 1, \ldots, p,$$
$$x_{p+1}^* = c,$$

for $c = (c_1, \ldots, c_p)^{\mathrm{T}}$, *where* $c_j = u_j$ *if* $\beta_j > 0$ *and* $c_j = l_j$ *if* $\beta_j < 0$.

Remark 1. The D-optimal design does not depend on the value of β_0 and is invariant to permutation of the variable labels.

Remark 2. In practice, the requirement $|\beta_j(u_j - l_j)| \geq 2$ is not overly restrictive. For example, the use of the standardized design space $[-1, 1]^p$ requires $|\beta_j| \geq 1$, $j = 1, \ldots, p$.

Example 2. For $p = 2$, suppose that $\beta = (1, -2, 3)^{\mathrm{T}}$, and $x_{1i} \in [0, 10]$, $x_{2i} \in [0, 12]$ $(i = 1, \ldots, s)$. Then the design with equally weighted support points $\{(1, 12), (0, 34/3), (0, 12)\}$ is D-optimal.

Remark 3. When $b_j - a_j < 2$ for some $j = 1, \ldots, p$, it can be shown that the D-optimal *saturated* design for the canonical model has equally weighted support points $z_i = b + \max(-2, a_i - b_i)e_i$ $(i = 1, \ldots, p)$, and $z_{p+1} = b$. Ford et al. (1992) proved that this design is D-optimal for $p = 1$ over Ψ. In general, the D-optimal design over Ψ is not saturated when $b_j - a_j < 2$ for some j. For example, if $\beta = (-0.91, 0.04, -0.69)^{\mathrm{T}}$ and $x_{ji} \in [-1, 1]$ $(i = 1, \ldots, s;\ j = 1, 2)$, then $z_1 \in [-0.95, -0.87]$, $z_2 \in [-0.69, 0.69]$, and the D-optimal design in $\mathcal{X}$ has support points $(-1, -1)$, $(-1, 1)$, $(1, -1)$, and $(1, 1)$, with respective weights $\nu_1 = 0.311$, $\nu_2 = 0.163$, $\nu_3 = 0.313$, and $\nu_4 = 0.213$.

3. Robust Design

Often experimenters have little information about parameter values prior to observing the data. Woods et al. (2006) found *compromise* designs for GLMs which are robust to wide ranges of parameter values. Dror and Steinberg (2006) approximated these methods by using a K-means clustering algorithm (see, for example, Hastie, Tibshirani and Friedman (2001, Chap. 14)). They applied the algorithm to the design points of a large number of locally optimal designs found by computer search for ranges of parameter values. The cluster means were then used as equally weighted support points of a *cluster design*. The use of the theorem in Section 2 allows the efficient computation of a cluster design, ξ^c, by removing the need to perform computer searches, as follows.

Algorithm

1. *Define a* p*-dimensional parameter space* $\mathcal{B}$ *for* $\tilde{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, *and a design space* $\mathcal{X}$ *that satisfies the conditions in the theorem for all* $\tilde{\beta} \in \mathcal{B}$.

2. *Sample N vectors $\tilde{\beta}_k$ $(k = 1, \ldots, N)$ from $\mathcal{B}$ using uniform quasi-random numbers, generated as a Sobol sequence; see* Gentle (2003, Chap. 3).
3. *For each $\tilde{\beta}_k$, apply the theorem to construct a locally optimal design ξ_j^*.*
4. *Apply a clustering algorithm to the total of $N(p+1)$ design points (see below).*
5. *Use each cluster mean as an equally weighted support point of ξ^c.*

Example 3. Suppose that $p = 2$, $x_{1i}, x_{2i} \in [-1, 1]$ $(i = 1, \ldots, p+1)$, and $\mathcal{B} = [1, 1+\alpha] \times [-1-\alpha, -1]$ with $\alpha > 0$. Figure 1 (a) shows the ensemble of design points from the locally D-optimal designs obtained from the theorem using $N = 1,000$ values of $\tilde{\beta}$ for each of $\alpha = 1, 5, 20$. All the designs include the point $(1, 1)$. The figure shows that the spread of the remaining values for each of x_{1i} and x_{2i} increases with α, reflecting increasing uncertainty in the value of $\tilde{\beta}$.

Figure 1 (a) suggests that the natural clusters in the design points are not well-described by spheres of equal volume, and hence a more flexible clustering algorithm than K-means may be advantageous. We compared K-means with the model-based clustering algorithm of Fraley and Raftery (2002) that assumes that points arise from a mixture of normal distributions with differing means and possibly differing covariance matrices. The EM algorithm may be used to estimate the conditional probability of each data point belonging to each cluster, and the parameters for each component distribution. We also investigated the number of support points (clusters) that should be selected. To allow estimation of the variance components in the model-based clustering, the locally optimal design points were "jittered" slightly through the addition of a small amount of uniform random noise.

Example 3 cont. For each of the K-means and model-based techniques, a cluster design was formed with $s = 3, \ldots, 22$ support points for each of $\alpha = 1, 5, 20$. The efficiency (2.1) of each design was calculated for each value of $\tilde{\beta}_k$ $(k = 1, \ldots, N)$. Figure 1 (b) shows how the median and minimum efficiencies vary with s and α. The use of model-based clustering frequently results in higher median and minimum efficiencies than K-means. This is often achieved with fewer support points, as for $\alpha = 5$ where the design with highest efficiency is found from model-based clustering and has the three support points (1.0,-1.0), (0.3,-1.0), (1.0,-0.3). As α increases, the parameter space $\mathcal{B}$ increases in volume and hence efficiencies are lower. The most efficient design for $\alpha = 20$ has more support points than that for smaller α. Generally, K-means designs require more support points than designs from model-based clustering, see Figure 1 (b). For fixed numbers of support points, designs from the two methods may differ substantially, see Figure 2.

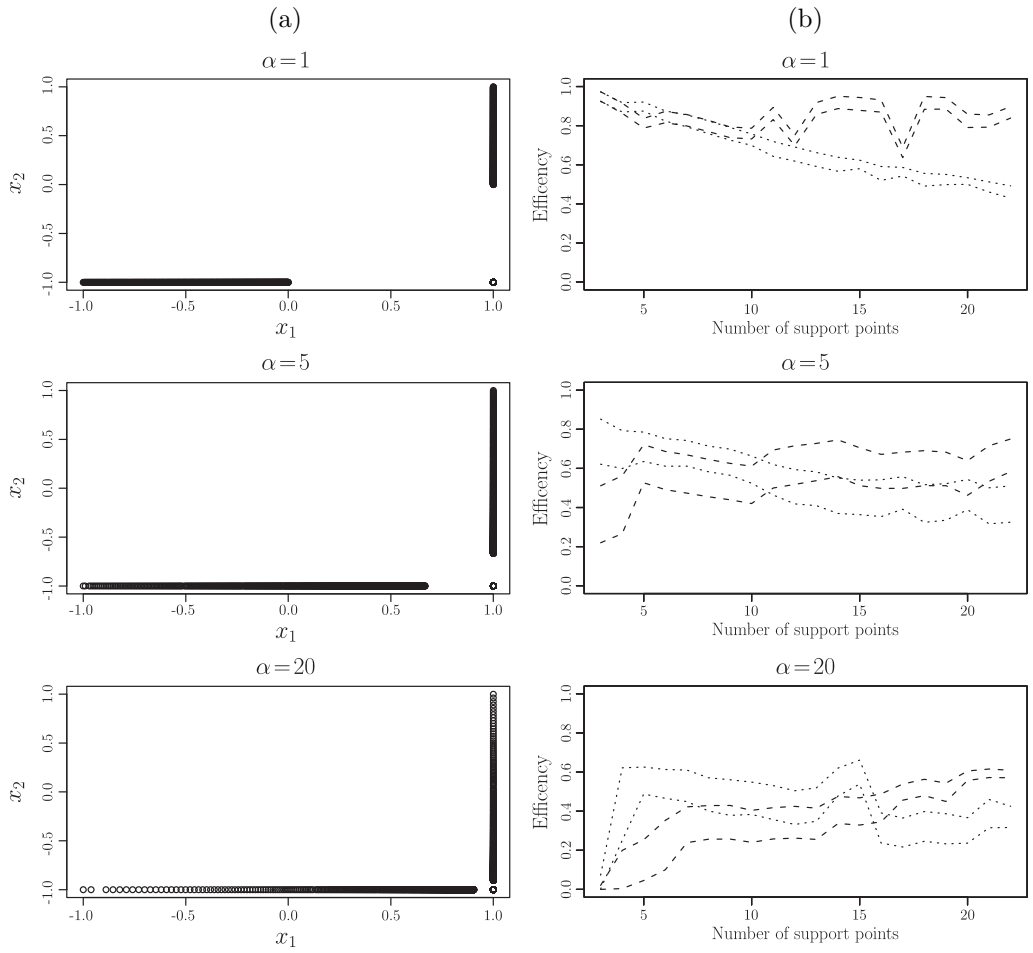

Figure 1. Example 3 for $\alpha = 1, 5, 20$: (a) design points ($\circ$), often overlapping, from 1,000 locally optimal designs for each parameter space; (b) median and minimum efficiencies for $3, \ldots, 22$ support points for cluster designs with K-means (dashed) and model-based (dotted).

This approach to finding designs is particularly useful for experiments with large numbers of variables, as in screening experiments. To avoid the need to evaluate designs with many different numbers of support points, as in Example 3, standard metrics from the unsupervised learning literature for the selection of the number of clusters can be employed, such as the Bayesian Information Criterion (BIC); see Fraley and Raftery (2002).

Example 4. For $p = 10$ variables and a first-order Poisson regression model

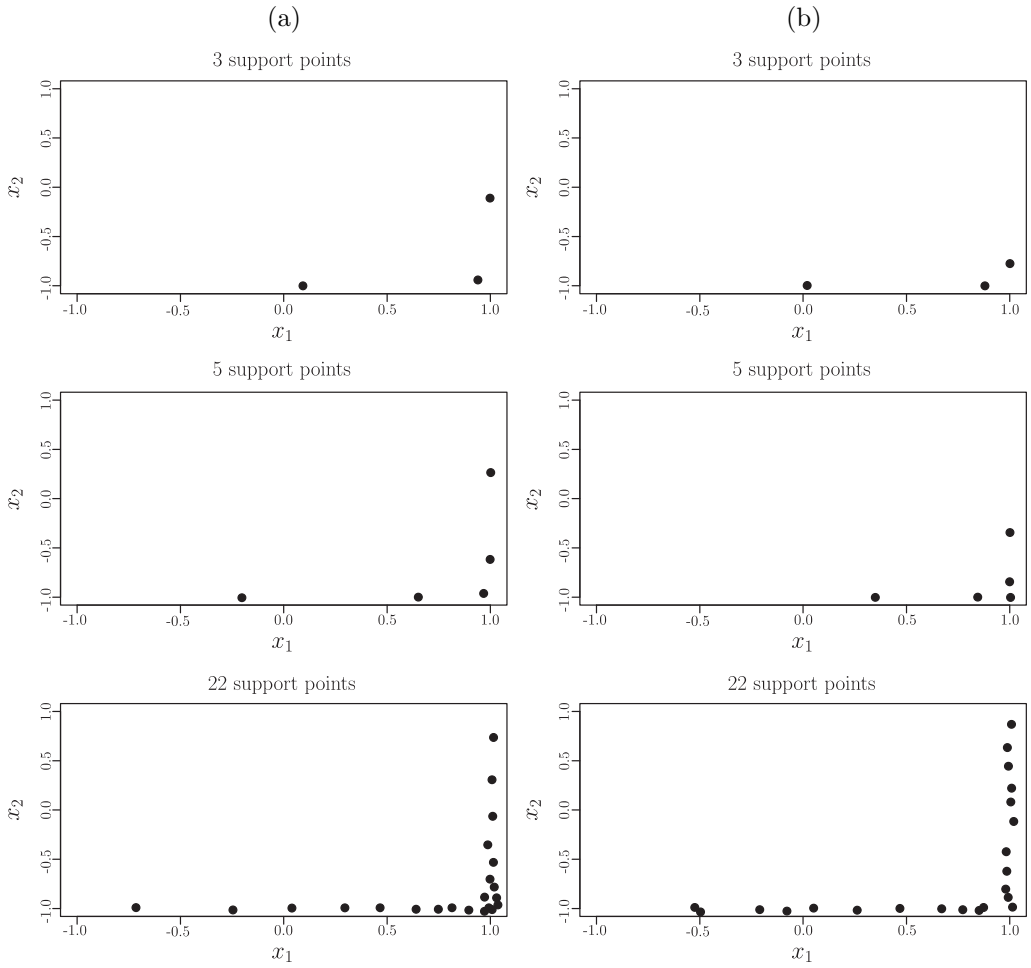

Figure 2. Cluster designs for Example 3 for $\alpha = 20$ and $3, 5, 22$ support points: (a) K-means; (b) model-based clustering.

with log link, suppose a robust design is required across the parameter space

$$\beta_j \in \begin{cases} [1, 1+\alpha] & \text{for } j = 1, 3, 5, 7, 9\,, \\ [-1-\alpha, -1] & \text{for } j = 2, 4, 6, 8, 10\,. \end{cases} \tag{3.1}$$

Cluster designs were found using model-based clustering for (3.1) with $\alpha = 1, 2, 3$, and $N = 1,000$. For each robust design, the number of support points (clusters), chosen using BIC, was found to be 21. Table 1 gives the median and minimum efficiencies across $\mathcal{B}$ and shows that, for each value of α, the cluster design performs well across the parameter space. As in Example 3, the median and minimum efficiencies decrease as α increases, but good performance is main-

Table 1. Median and minimum efficiencies for the model-based cluster designs in Example 4, with parameter spaces defined through (3.1).

	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$
median eff.	0.936	0.877	0.748
minimum eff.	0.895	0.803	0.633

tained even for $\alpha = 3$. The use of the BIC statistics allows an informed choice of the number of support points without the need to evaluate more than one design.

4. Conclusions

The results presented in this paper allow the analytic construction of D-optimal designs for first-order Poisson regression with a log-link, and demonstrate their use in the fast construction of designs robust to parameter values. First-order models are commonly used in practical data analysis, and are particularly appropriate for the analysis of data from experiments early in scientific investigations. Hence the methods from this paper are particularly important for screening experiments that may involve a large number of variables. The use of the Theorem enables much larger problems to be tackled than computer search currently allows.

Acknowledgement

This work was supported by grants from the Engineering and Physical Sciences Research Council, UK, and the Australian Research Council. We are grateful to Dr. Nema Dean (University of Glasgow) for helpful advice on model-based clustering, and to the referees for their thorough and insightful comments.

Appendix. Proof of the Lemma

The result is proved using a general equivalence theorem (Atkinson, Donev and Tobias (2007, p.122)). Let $M(\psi^*)$ be the information matrix for the design ψ^* defined in the lemma. We show that the standardised variance of the predicted response at point z_0, $d(z_0, \psi^*) = w(z_0) f(z_0)^{\mathrm{T}} M^{-1}(\psi^*) f(z_0)$, satisfies $d(z_0, \psi^*) \leq p + 1$ for all $z_0 \in \mathcal{Z}$, and $d(z_0, \psi^*) = p + 1$ at each support point z_i^* of ψ^*.

Some algebra shows that symmetric matrix $M^{-1}(\psi^*)$ has (i, j)th entry

$$\frac{p+1}{4\exp(\sum_{i=1}^{p} b_i)} m^{ij} \,,$$

where $m^{11} = (2 - \sum_{i=1}^{p} b_i)^2 + e^2 \sum_{i=1}^{p} b_i^2$, $m^{1j} = 2 - \sum_{i=1}^{p} b_i - e^2 b_{j-1}$, $m^{jj} = 1 + e^2$ $(2 \leq j \leq p+1)$, $m^{ij} = 1$ $(2 \leq i < j \leq p+1)$, and $e = \exp(1)$. Further,

$$d(z_0, \psi^*) = \frac{p+1}{4} \exp\left(\sum_{i=1}^{p} z_{i0} - \sum_{i=1}^{p} b_i \right) g(z_0) \,,$$

where

$$g(z_0) = m^{11} + 2\sum_{i=2}^{p+1} m^{1i} z_{(i-1)0} + 2\sum_{i=2}^{p}\sum_{j=i+1}^{p+1} m^{ij} z_{(i-1)0} z_{(j-1)0} + \sum_{i=2}^{p+1} m^{ii} z_{(i-1)0}^2 \,.$$

It is easy to show that $d(z_0, \psi^*) = p+1$ for $z_0 = z_i^*$, $i = 1, \ldots, p+1$. The Karush-Kuhn-Tucker theorem (see, for example, Borwein and Lewis (2000, pp.160-161)) provides the following necessary conditions for the constrained maximization of $d(z_0, \psi^*)$ to be achieved:

$$\frac{\partial d(z_0, \psi^*)}{\partial z_{j0}} - \mu_j = 0 \,, \tag{A.1}$$

subject to

$$\mu_j (b_j - z_{j0}) = 0 \,, \quad \mu_j \geq 0 \,, \quad z_{j0} \leq b_j \,, \tag{A.2}$$

for $j = 1, \ldots, p$, where the μ_j are Lagrange multipliers.

If $\mu_j > 0$ for every j, then (A.1) and (A.2) imply that $z_{j0} = b_j$ and $d(z_0, \psi^*) = p+1$. If $\mu_j = 0$ for at least one j, without loss of generality set $\mu_1 = \ldots = \mu_r = 0$ $(1 \leq r \leq p)$. Then to satisfy (A.2), $z_{(r+1)0} = b_{r+1}, \ldots, z_{p0} = b_p$. From (A.1), after some algebra, $z_{j0} = b_j - t$, $j = 1, \ldots, r$, where $t = 2/r$ or $t = 4/(r + e^2)$. For each solution $\partial d / \partial z_{j0}^* \geq 0$ $(j = r+1, \ldots, p)$, and hence from (A.1), $\mu_j \geq 0$, satisfying (A.2). As $d(z_0, \psi^*) = t(p+1)/2$ if $t = 2/r$, and $d(z_0, \psi^*) = (p+1)\ \exp(-rt)$ if $t = 4/(r+e^2)$, the maximum value of $d(z_0, \psi^*)$ over $\mathcal{Z}$ is $p+1$ when $r = 1$ and $z_{10} = b_1 - 2$. From the general equivalence theorem mentioned, ψ^* is D-optimal.

References

Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS*. Oxford University Press, Oxford.

Biedermann, S., Dette, H. and Pepelyshev, A. (2006). Some robust design strategies for percentile estimation in binary response models. *Canad. J. Statist.* **34**, 603-622.

Borwein, J. M. and Lewis, A. S. (2000). *Convex Analysis and Nonlinear Optimization.* Springer, New York.

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

Dror, H. A. and Steinberg, D. M. (2006). Robust experimental design for multivariate generalized linear models. *Technometrics* **48**, 520-529.

Firth, D. and Hinde, J. P. (1997). Parameter neutral optimum design for non-linear models. *J. Roy. Statist. Soc. Ser. B* **59**, 799-811.

Ford, I., Torsney, B. and Wu, C. F. J. (1992). The use of a canonical form in the construction of locally optimal designs for nonlinear problems. *J. Roy. Statist. Soc. Ser. B* **54**, 569-583.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *J. Amer. Statist. Assoc.* **97**, 611-631.

Gentle, J. E. (2003). *Random Number Generation and Monte Carlo Methods*, 2nd edition. Springer, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edition. Chapman and Hall, New York.

Minkin, S. (1993). Experimental design for clonogenic assays in chemotherapy. *J. Amer. Statist. Assoc.* **88**, 410-420.

Pukelsheim, F. (1993). *Optimal Design of Experiments*. Wiley, New York.

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

Sitter, R. R. and Torsney, B. (1995). D-optimal designs for generalized linear models. In *MODA 4 – Advances in Model-Oriented Data Analysis* (Edited by C. P. Kitsos and W. G. Müller), 87-102. Physica, Heidelberg.

Wang, Y., Myers, R. H., Smith, E. P. and Ye, K. (2006). D-optimal designs for Poisson regression models. *J. Statist. Plann. Inference* **136**, 2831–2845.

Wang, Y., Smith, E. P. and Ye, K. (2006). Sequential designs for a Poisson regression model. *J. Statist. Plann. Inference* **136**, 3187-3202.

Woods, D. C., Lewis, S. M., Eccleston, J. A. and Russell, K. G. (2006). Designs for generalized linear models with several variables and model uncertainty. *Technometrics* **48**, 284-292.

School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia.

E-mail: kgr@uow.edu.au

Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK.

E-mail: D.Woods@soton.ac.uk

Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK.

E-mail: S.M.Lewis@soton.ac.uk

Department of Mathematics, School of Physical Sciences, University of Queensland, Brisbane QLD 4072, Australia.

E-mail: jae@maths.uq.edu.au

(Received October 2007; accepted January 2008)