Journal of Statistical Planning and Inference 178 (2016) 128–139

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Optimal designs based on the maximum quasi-likelihood estimator

Gang Shen [a], Seung Won Hyun [a,*], Weng Kee Wong [b]

[a] Department of Statistics, North Dakota State University, Fargo, ND 58102, USA

[b] Department of Biostatistics, University of California, Los Angeles, CA 90095, USA

ARTICLE INFO

Article history:
Received 28 March 2016
Received in revised form 6 July 2016
Accepted 7 July 2016
Available online 15 July 2016

MSC:
00-01
99-00

Keywords:
Approximate design
Design efficiency
Dose-finding study
Equivalence theorem
Heteroscedasticity
Maximum likelihood estimator

ABSTRACT

We use optimal design theory and construct locally optimal designs based on the maximum quasi-likelihood estimator (MqLE), which is derived under less stringent conditions than those required for the MLE method. We show that the proposed locally optimal designs are asymptotically as efficient as those based on the MLE when the error distribution is from an exponential family, and they perform just as well or better than optimal designs based on any other asymptotically linear unbiased estimators such as the least square estimator (LSE). In addition, we show current algorithms for finding optimal designs can be directly used to find optimal designs based on the MqLE. As an illustrative application, we construct a variety of locally optimal designs based on the MqLE for the 4-parameter logistic (4PL) model and study their robustness properties to misspecifications in the model using asymptotic relative efficiency. The results suggest that optimal designs based on the MqLE can be easily generated and they are quite robust to mis-specification in the probability distribution of the responses.

1. Introduction

Model assumptions are frequently made for drawing statistical inference but they may not be tenable in practice. The optimal design constructed under wrong model assumptions can be very inefficient. When the model is nonlinear, a further complication is that nominal values for the model parameters are required before the optimal designs can be implemented. It is thus important to find a design that is robust to various forms of mis-specification in the statistical model.

Different approaches have been proposed to find designs robust to various departures from the model assumptions. They typically concern mis-specifications in the error distribution, the response mean function, and the form of heteroscedasticity in the model. For instance, Burridge and Sebastiani (1994) and Dette and Wong (1999) proposed optimal designs when the variance is a function of the mean response, and Chen et al. (2008) proposed optimal minimax designs for a heteroscedastic polynomial model. However the bulk of the work focused on a single violation from the model assumptions. Some examples are Läuter (1974), Stigler (1971), Lee (1987, 1988), Studden (1982), Song and Wong (1998a,b) and Dette and Wong (1996, 1999); the first six considered model mis-specification in mean function and the last three concerned mis-specification in the heteroscedasticity structure.

Much of the research in optimal design assumes that the design criterion is formulated in terms of the Fisher Information matrix. This matrix can only be computed under the full specification of the probability distribution for the response. In

* Corresponding author.
E-mail address: Seung.W.Hyun@NDSU.edu (S.W. Hyun).

http://dx.doi.org/10.1016/j.jspi.2016.07.002
0378-3758/

contrast, the quasi-likelihood method introduced by Weddernburn (1974) assumes only a functional relationship between the mean and the variance of the response. This is a tenable assumption in practical situations, see for example, Box and Hill (1974), Bickel (1978) and, Jobson and Fuller (1980). This suggests that designs based on the quasi-likelihood method are likely to be more robust to model assumptions than methods that require a full parametric probability model.

There is very little work on optimal designs based on the MqLE for nonlinear regression models despite numerous successful applications of the quasi-likelihood method for generalized linear models (McCullagh and Nelder, 1989). A clear exception is work by Niaparast (2009), and Niaparast and Schwabe (2013), who derived optimal designs based on the MqLE for Poisson regression models. The first paper focused on the theoretical construction of D-optimal designs for the Poisson models with random intercepts, and the second paper extended the work to a general mixed effects Poisson model with random coefficients and so the paper deals with random slopes rather than random intercepts. However, their work concerns only finding locally D-optimal designs for the case when the response follows a Poisson distribution.

We consider a more general situation and develop theory to find various types of optimal designs based on the MqLE where the response may be a discrete or a continuous random variable. We also show such optimal designs can be generated from algorithms currently used to find optimal designs based on the Fisher Information matrix. We list advantages of such designs compared to other designs based on other efficient estimators such as the MLE and the LSE when the sample size is large and study the robustness properties of D-optimal designs to mis-specified nominal parameter values. We focus on locally optimal designs (Chernoff, 1953) where nominal values for the model parameters are required before they can be implemented. Bayesian optimal designs can also be directly constructed from our methodology but will not be discussed here. One of the key results is that we show here that asymptotically, locally optimal designs based on the MqLE also have the same property as those constructed based on the MLEs but without the normality distributional assumption. When responses are from a member of the class of exponential family distributions, we prove that the optimal designs based on the MqLE are asymptotically just as efficient as the optimal designs based on the MLE. Such designs are also asymptotically more efficient than the designs based on asymptotic best linear unbiased estimators, such as LSE. When responses are not from a member of the exponential family of distributions, we show optimal designs based on the MqLE also perform well based on an asymptotic relative efficiency measure.

Section 2 reviews preliminaries and discusses the asymptotic optimality of MqLEs. In Section 3, we review approximate designs and equivalence theorems for checking optimality of an approximate design. We then develop the theory and an algorithm to find a variety of optimal designs based on the MqLE. As an illustrative application, we construct locally optimal designs for dose–response studies using the widely used 4-parameter logistic model (4PL-model). We construct optimal designs for estimating models and meaningful function of the model parameters in the 4PL-model and compare its performance with those based on MLEs and a uniform design with 10 points when there is mis-specification in (i) the nominal values of the model parameters, (ii) the heteroscedasticity structure, and (iii) the error distributions. Section 4 concludes with a summary of our work and recommendations.

2. Preliminaries

Suppose we have a dose–response study and we have resources to take n observations at d distinct doses $x_1, \ldots, x_d$ with n_i replicates at x_i, $i = 1, \ldots, d$ and $\sum_{i=1}^{d} n_i = n$. Let y be the response at x with $E(y) = \mu$ and $Var(y) = v(\mu)$ where $\mu = \mu(\boldsymbol{\theta}; x)$ and μ is a known function of an unknown r-dimensional parameter vector $\boldsymbol{\theta}$ and dose x. For simplicity, we write $\mu(\boldsymbol{\theta}; x)$ as $\mu(\boldsymbol{\theta})$ or simply μ if there is no confusion. Our stipulation $Var(y) = v(\mu)$ covers the situation of Niaparast (2009) but not the situation of Niaparast and Schwabe (2013) since the latter considered $Var(y) = \mu + \mu^2 c(x)$ and $c(x)$ is not a function of μ. Another difference is that the two papers assumed the n_i replications at x_i are not independent and we do here as in most dose–response studies. Assuming that $\inf_\mu v(\mu) > 0$, the log quasi-likelihood of y (with dispersion ϕ) is defined as

$$q(\mu; y) \triangleq \int_{-\infty}^{\mu} \frac{y - t}{\phi v(t)} dt.$$

Let $\dot{q} \triangleq \frac{\partial}{\partial \boldsymbol{\theta}} q(\mu; y)$ (quasi-score) and $\ddot{q} \triangleq \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} q(\mu; y)$. Weddernburn (1974) showed that $q(\mu; y)$ behaves like a likelihood function in the sense that

$$E(\dot{q}) = 0,\ E(\dot{q}\dot{q}') = -E(\ddot{q})$$

and $-E(\ddot{q}) = \frac{1}{\phi v(\mu)} \dot{\mu}(\boldsymbol{\theta}) \dot{\mu}'(\boldsymbol{\theta})$ (quasi-information), where $\dot{\mu}(\boldsymbol{\theta}) \triangleq \frac{\partial \mu}{\partial \boldsymbol{\theta}}$. Given $y_1, y_2, \ldots, y_n$, the maximum quasi-likelihood estimate (MqLE) of $\boldsymbol{\theta}$ maximizes the joint quasi-likelihood $\sum_{i=1}^{n} q(\mu_i; y_i)$.

More generally, let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be the vector of observations at the n doses and let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ be its mean vector. Without loss of generality, let the first d elements of $\boldsymbol{y}$ correspond to observations from the d distinct doses $x_1, \ldots, x_d$ and let $\boldsymbol{\mu}_\star = (\mu(\boldsymbol{\theta}; x_1), \ldots, \mu(\boldsymbol{\theta}; x_d))'$ be the mean vector of the first d observations from the d distinct doses. Clearly $\boldsymbol{\mu}_\star = C\boldsymbol{\mu}$, where C is the first d rows of the n-dimensional identity matrix I_n.