Journal of Statistical Planning and Inference 149 (2014) 140–151

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

New optimal design criteria for regression models with asymmetric errors

Lucy L. Gao, Julie Zhou *

Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada V8W 3R4

ARTICLE INFO

Article history:
Received 3 September 2013
Received in revised form
6 January 2014
Accepted 19 January 2014
Available online 23 January 2014

Keywords:
A-optimal design
D-optimal design
Non-normal errors
Polynomial regression
Robust design
Second-order least squares estimator
Scale invariance
Shift invariance
Symmetric design
Trigonometric regression

ABSTRACT

Optimal regression designs are usually constructed by minimizing some scalar functions of the covariance matrix of the ordinary least squares estimator. However, when the error distribution is not symmetric, the second-order least squares estimator is more efficient than the ordinary least squares estimator. Thus we propose new design criteria to construct optimal regression designs based on the second-order least squares estimator. Transformation invariance and symmetry properties of the new criteria are investigated, and sufficient conditions are derived to check for these properties of D-optimal designs. The results can be applied to both linear and nonlinear regression models. Several examples are given for polynomial, trigonometric and exponential regression models, and new designs are obtained.

1. Introduction

Consider a linear regression model

$$y_i = \boldsymbol{z}^\top(\mathbf{x}_i)\boldsymbol{\theta} + \varepsilon_i, \quad i = 1, \cdots, n, \tag{1}$$

where y_i is the ith response observed at $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})^\top$ of p independent variables $x_1, \ldots, x_p$, $\boldsymbol{z}(\mathbf{x}) \in R^q$ is a vector of known functions of $\mathbf{x}, \boldsymbol{\theta} \in R^q$ is a vector of unknown regression parameters, and the errors ε_i's are independent with mean 0 and variance σ^2. Optimal regression designs aim to choose optimal design points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ such that we can get the most information about the unknown regression parameters or the regression response $\boldsymbol{z}^\top(\mathbf{x})\boldsymbol{\theta}$. Optimal regression designs have been studied extensively in the literature and many optimal design criteria have been proposed and studied; see, for example, Fedorov (1972) and Pukelsheim (1993).

The ordinary least squares estimator (OLSE) is usually used to estimate $\boldsymbol{\theta}$ since it is the best linear unbiased estimator. The OLSE is given by

$$\hat{\boldsymbol{\theta}}_{OLS} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{y},$$

* Corresponding author. Tel.: +1 250 721 7470.
E-mail address: jzhou@uvic.ca (J. Zhou).

0378-3758/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2014.01.005

where $\mathbf{y}=(y_1,\ldots,y_n)^\top$, and the ith row of matrix $\mathbf{Z}$ is $\boldsymbol{z}^\top(\mathbf{x}_i)$. The covariance matrix of $\hat{\boldsymbol{\theta}}_{OLS}$ is $Cov(\hat{\boldsymbol{\theta}}_{OLS})=\sigma^2(\mathbf{Z}^\top\mathbf{Z})^{-1}$. The commonly used A-optimal, D-optimal and E-optimal designs minimize the trace, the determinant and the largest eigenvalue of $Cov(\hat{\boldsymbol{\theta}}_{OLS})$ respectively. Optimal regression designs have been derived for various regression models and design spaces, and they perform well under the ideal model assumptions for (1):

(i) the response function is correct,
(ii) the errors ε_i's are i.i.d. $N(0,\sigma^2)$.

However, these assumptions are never true in practice.

When there are possible violations of the model assumptions, robust regression designs against small departures have been explored. Many robust designs have been constructed to deal with small departures in the response function, in the equal variance of the errors, in the correlation structure of the errors, and combinations of those departures. For example, Huber (1981), Li and Notz (1982), Notz (1989) and Wiens (1992) investigated robust designs against small departures in the response function. Fang and Wiens (2000) studied robust designs against small departures in the response function and in the equal variance of the errors. Chen et al. (2008) worked on minimax designs for polynomial models with various types of heteroscedastic errors, and Wiens and Zhou (1999) and Zhou (2001) worked on minimax designs robust against small departures in the response function and in the correlation structure of the errors. However very little work has been done to study optimal/robust designs when there is a small departure from the assumption of normal distribution. Thus, in this paper we will explore optimal regression designs when the error distribution is asymmetric.

If the errors ε_i's have non-normal distributions, there are more efficient estimators than the OLSE. When the exact distribution of the error term is known, the maximum likelihood estimator can be applied to estimate $\boldsymbol{\theta}$. However the exact distribution of the error term is often unknown in practice. In this situation, the second-order least squares estimator (SLSE) proposed in Wang and Leblanc (2008) is more efficient than the OLSE if the third moment of the error is nonzero, i.e., the error distribution is asymmetric. The SLSE does not depend on the exact distribution of the error term, so we propose to use the SLSE to construct optimal regression designs for asymmetric errors in this paper. In particular, for a given model we construct the optimal distribution of $\mathbf{x}$, and the design points $\mathbf{x}_1,\ldots,\mathbf{x}_n$ are selected randomly from the optimal distribution. A-optimal and D-optimal designs are defined based on the SLSE, various properties including scale and shift invariance and symmetry of D-optimal designs are explored, and sufficient conditions are derived to check for these properties. In addition, optimal designs based on the SLSE and OLSE are compared. All results can be applied to both linear and nonlinear regression models.

The paper is organized as follows. Section 2 reviews the SLSE and its asymptotic distribution. In Section 3, optimal designs based on the SLSE are defined and compared with optimal designs based on the OLSE. In Section 4, transformation invariance of D-optimal designs is studied, and a sufficient condition is obtained to check for transformation invariance. Section 5 investigates the symmetry of D-optimal designs. Section 6 presents examples of D-optimal designs. Concluding remarks are in Section 7. All proofs are given in the Appendix.

2. Second-order least squares estimator

The SLSE was first introduced for nonlinear models with measurement errors in Wang (2003, 2004). Since then it has been studied and extended for regression models without measurement errors, for example, Wang and Leblanc (2008) for nonlinear models, Abarin and Wang (2009) for censored regression models, and Chen et al. (2012) for robust SLSE for linear models. The SLSE is asymptotically normally distributed and is more efficient than the OLSE if the third moment of the errors is nonzero. Here we review the SLSE and its related results in Wang and Leblanc (2008) for the following model:

$$y_i=g(\mathbf{x}_i;\boldsymbol{\theta})+\varepsilon_i,\quad i=1,\ldots,n, \tag{2}$$

where $g(\mathbf{x}_i;\boldsymbol{\theta})$ can be a linear or nonlinear function of $\boldsymbol{\theta}\in R^q$, and the errors ε_i's are independent and have mean $E(\varepsilon|\mathbf{x})=0$ and $E(\varepsilon^2|\mathbf{x})=\sigma^2$.

Denote the parameter vector as $\boldsymbol{\gamma}=(\boldsymbol{\theta}^\top,\sigma^2)^\top$. We assume that $\mathbf{y}$ and $\boldsymbol{\epsilon}$ have finite fourth moments, where $\mathbf{y}=(y_1,\ldots,y_n)^\top$ and $\boldsymbol{\epsilon}=(\varepsilon_1,\ldots,\varepsilon_n)^\top$. The SLSE $\hat{\boldsymbol{\gamma}}_{SLS}$ for $\boldsymbol{\gamma}$ is defined as the measurable function that minimizes

$$Q_n(\boldsymbol{\gamma})=\sum_{i=1}^n \boldsymbol{\rho}_i^\top(\boldsymbol{\gamma})\boldsymbol{W}_i\boldsymbol{\rho}_i(\boldsymbol{\gamma}),$$

where $\boldsymbol{\rho}_i(\boldsymbol{\gamma})=(y_i-g(\mathbf{x}_i;\boldsymbol{\theta}),y_i^2-g^2(\mathbf{x}_i;\boldsymbol{\theta})-\sigma^2)^\top$, and $\boldsymbol{W}_i=\boldsymbol{W}(\mathbf{x}_i)$ is a 2×2 positive semidefinite matrix which may depend on $\mathbf{x}_i$. Notice that parameters $\boldsymbol{\theta}$ and σ^2 are estimated together in $\hat{\boldsymbol{\gamma}}_{SLS}$, and $\boldsymbol{W}_i$ can be any positive semidefinite matrix.

Suppose the true parameter value of model (2) is $\boldsymbol{\gamma}_0=(\boldsymbol{\theta}_0^\top,\sigma_0^2)^\top$, and design points $\mathbf{x}_1,\ldots,\mathbf{x}_n$ are randomly selected from a distribution ξ of $\mathbf{x}$. Under some regularity conditions, the asymptotic covariance matrix of the SLSE is shown to be

$$V(\hat{\boldsymbol{\gamma}}_{SLS})=\lim_{n\to+\infty}Cov(\sqrt{n}\hat{\boldsymbol{\gamma}}_{SLS})=\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1},$$