Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 51 (2006) 801 – 808

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Meta-heuristic algorithms for parameter estimation of semi-parametric linear regression models

Guoqing Zheng[a,*], Pingjian Zhang[b]

[a] *Department of Mathematics, South China Agricultural University, Guangzhou 510640, China*
[b] *School of Software Engineering, South China University of Technology, Guangzhou 510641, China*

Received 11 May 2005; accepted 5 September 2005
Available online 4 October 2005

Abstract

Consider the semi-parametric linear regression model $Y=\beta' X+\varepsilon$, where ε has an unknown distribution F_0. The semi-parametric MLE $\tilde{\beta}$ of β under this set-up is called the generalized semi-parametric MLE(GSMLE). Although the GSML estimation of the linear regression model is statistically appealing, it has never been attempted due to difficulties with obtaining the GSML estimates of β and F until recent work on linear regression for complete data and for right-censored data by Yu and Wong [2003a. Asymptotic properties of the generalized semi-parametric MLE in linear regression. Statistica Sinica 13, 311–326; 2003b. Semi-parametric MLE in simple linear regression analysis with interval-censored data. Commun. Statist.—Simulation Comput. 32, 147–164; 2003c. The semi-parametric MLE in linear regression with right censored data. J. Statist. Comput. Simul. 73, 833–848]. However, after obtaining all candidates, their algorithm simply does an exhaustive search to find the GSML estimators. In this paper, it is shown that Yu and Wong's algorithm leads to the so-called dimension disaster. Based on their idea, a simulated annealing algorithm for finding semi-parametric MLE is proposed along with techniques to reduce computations. Experimental results show that the new algorithm runs much faster for multiple linear regression models while keeping the nice features of Yu and Wong's original one.

Keywords: Generalized likelihood; Simulated annealing algorithm; Semi-parametric models; Running cost

1. Introduction

Linear regression is one of the most powerful tools among statistical methods. Its applications cover almost every field, from economics, engineering, physics, life and biological sciences to social sciences. In this paper, we consider the following multiple linear regression:

$$Y=\beta' X+\varepsilon, \tag{1.1}$$

where β is the unknown regression coefficient of dimension m, ε has an unknown cdf F_0. Model (1.1) is a semi-parametric problem since no further assumption is made upon (β, ε).

For the linear regression model (1.1), there exist several estimators for β in the literature: the least square estimator (LSE), the Theil–Sen estimator (Theil, 1950; Sen, 1968), various M-estimators (Huber, 1964; Ritov, 1990), the adaptive

[*] Corresponding author. Tel.: +86 2033366286; fax: +86 2039380218.
E-mail address: guoqingzh@yahoo.com (G. Zheng).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.09.001

estimators (Bickel, 1982), the L-estimators and the R-estimators (Montegomery and Peck, 1992). While enjoying the optimal statistical properties, LSE is not efficient unless F_0 is a normal distribution. Besides LSE, the Theil–Sen estimator, the L-estimators and the R-estimators are based on order statistics like the median of regression slopes, offering computationally simple alternatives. The Theil–Sen estimator has good small-sample efficiency, particularly when the error term is heteroscedastic (Wilcox, 1998). While having desirable computational and statistical properties, R-estimators (and also L-estimators) cannot reject outliers properly. It is observed (Hampel et al., 1986) that R-estimators either "reject" always or never in the sense of giving zero influence.

The maximum likelihood method provides an alternate estimation. Recently, Yu and Wong (2003a, 2003b, 2003c) have studied the GSML estimators under the semi-parametric setting. To begin with, define the generalized maximum likelihood function of observations $T_1, \ldots, T_n$ to be (Kiefer and Wolfowitz, 1956)

$$L = \prod_{i=1}^{n} f(T_i), \quad f(t) = F(t) - F(t-) \text{ and } F \in \mathscr{F} \tag{1.2}$$

where $T_i = Y_i - \beta' X_i$ and $\mathscr{F} = \{F : F \text{ is an increasing function}, \ F(-\infty) = 0 \text{ and } F(\infty) = 1\}$. The MLE of the cdf, also called the generalized MLE, is the function F that maximizes L over $\mathscr{F}$. It is well known that the generalized MLE is the empirical distribution function. Moreover, the semi-parametric MLE of (F_0, β), denoted by $\left(\tilde{F}_0, \tilde{\beta}\right)$, is a pair of (F, β) that maximizes

$$L(F, \beta) = \prod_{i=1}^{n} f(T_i) \quad \text{over all } (F, \beta) \in \mathscr{F} \times \mathscr{R}^m$$
$$\text{where } f(t) = F(t) - F(t-). \tag{1.3}$$

For fixed β, the likelihood function L in (1.3) is maximized by the empirical cdf

$$\tilde{F}_\beta(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{Y_i - \beta' X_i \leqslant t\right\}, \tag{1.4}$$

where $\mathbf{1}(\cdot)$ is the indicator function. Then, any $\tilde{\beta}$ that maximizes $l(\beta) \equiv L\left(\tilde{F}_\beta, \beta\right)$ is a semi-parametric MLE of β and the corresponding $\tilde{F}_\beta$ is the semi-parametric MLE of the cdf F_0. A non-iterative algorithm for finding all semi-parametric MLE is given in Yu and Wong (2003a, 2003c) and some asymptotic properties of the semi-parametric MLE are discussed in Yu and Wong (2003b).

Note that the GSMLE is akin to a class of efficient M-estimators studied in Zhang and Li (1996). Their M-estimators are zero-crossing points of a function $\Phi(\beta)$, which has the form

$$\Phi = \sum_{i=1}^{n} \phi\left(Y_i - \bar{Y} - \beta'\left(X_i - \bar{X}\right)\right)\left(X_i - \bar{X}\right), \quad \phi(t) = \frac{\partial}{\partial t} \ln f(t), \tag{1.5}$$

where f is a pdf. In the M-estimator approach, f is replaced by some estimator $\hat{f}$. Since $(\partial/\partial t)\tilde{F}_\beta(t) = 0$ a.e., $\tilde{\beta}$ is trivially an M-estimator with $\hat{f} = \tilde{F}_\beta$. Thus, the M-estimator approach provides no information to the GSMLE $\tilde{F}_\beta$. To find the GSML estimators, some exhaustive search algorithms and their variants are presented in Yu and Wong (2003a, c), which are computationally intensive especially for high-dimensional models. Note that this is essentially an optimization problem featured with a non-smooth objective function, many local extremum values and huge state space even for moderate dimensions, we shall adopt the simulated annealing algorithm (SA) to find the GSML estimators. For an introduction to SA, see Arts and Korst (1990), van Laarhoven and Arts (1987), for a survey on applications of SA, see Kang et al. (1998), van Laarhoven and Arts (1987).

The rest of the paper is organized as follows: in Section 2, a review of Yu and Wong's Algorithm for the semi-parametric MLE problem is given, and the analysis of its running cost is carried out. An introduction to SA and a new algorithm based on SA to find semi-parametric MLE are given in Section 3. Applications and discussions of the new algorithm are contained in Section 4. Comparisons with other algorithms are made in Section 5 and some final remarks and conclusions are drawn in Section 6.