Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 3894–3903
journal of statistical planning and inference
www.elsevier.com/locate/jspi

M-estimation in regression models for censored data

Zhezhen Jin*

Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 W 168th Street, New York, NY 10032, USA

Received 30 December 2005; received in revised form 7 September 2006; accepted 6 April 2007
Available online 4 May 2007

Abstract

In this paper, we study M-estimators of regression parameters in semiparametric linear models for censored data. A class of consistent and asymptotically normal M-estimators is constructed. A resampling method is developed for the estimation of the asymptotic covariance matrix of the estimators.

Keywords: Accelerated failure time model; M-estimator; Rank estimator; Resampling; Semiparametric model; Survival data; Variance estimation

1. Introduction

A way to study the covariate effects on the response is to use the semiparametric linear regression model in which the response is sum of a linear combination of covariates and error term whose distribution is completely unspecified.

For uncensored data, it is well known that the use of M-estimation methods for the model can overcome some of the robustness limitations of the least squares approach (Huber, 1981; Hampel et al., 1986). For right censored data, Ritov (1990), Zhou (1992) and Lai and Ying (1994) developed M-estimators for the model. The approach of Zhou (1992) is based on a weighted data approach similar to that in Koul et al. (1981), which requires that the censoring time is independent of covariates. The approaches of Ritov (1990) and Lai and Ying (1994) are extensions of the estimator of Buckley and James (1979) to general M-estimators. Under certain conditions, both Ritov (1990) and Lai and Ying (1994) rigorously proved that their estimating equation has a solution that is $\sqrt{n}$-consistent and asymptotically normal. But these results do not guarantee consistency of all solutions. They were not able to show that the estimating equation has a unique solution, nor were they able to devise an algorithm that could be used in finding a consistent estimator. Also, the variance of resulting estimator is difficult to estimate well because it involves the unknown density function of error term and its derivative.

In this note, we develop a new M-estimation procedure for the semiparametric linear model with right censored data. The method is easy to implement and yields a class of estimators which are consistent and asymptotically normal. We also develop a resampling method for the estimation of the limiting covariance matrix.

In the next section, the semiparametric linear regression model is introduced and the new M-estimation method is presented. Simulation studies are presented in Section 3 and some remark is given in Section 4. Proofs are outlined in Appendix A.

* Tel.: +1 212 305 9404; fax: +1 212 305 9408.
E-mail addresses: zj7@columbia.edu, zjin@biostat.columbia.edu (Z. Jin).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.04.008

2. Model and *M*-estimation

Let (Y_i, X_i, δ_i), $i=1, \ldots, n$, be independent and identically distributed observed data, where X_i is a $p \times 1$ covariate, $Y_i = \min\{T_i, C_i\}$ with T_i being failure time and C_i being the censoring time, $\delta_i = 1\{T_i \leqslant C_i\}$ is the failure indicator. Throughout the paper, it is assumed that T_i and C_i are independent conditional on X_i.

2.1. Accelerated failure time model

The semiparametric linear regression model is of form

$$T_i = X_i^{\mathrm{T}}\beta_0 + \varepsilon_i, \tag{1}$$

where β_0 is the unknown true $p \times 1$ parameter of interest and ε_i $(i = 1, \ldots, n)$ are unobservable independent random errors with a common but completely unspecified distribution function $F(\cdot)$. (Thus, the mean of ε is not necessarily 0). In survival analysis in which the response is the logarithm transformation of the survival time, the model (1) is the accelerated failure time (AFT) model.

2.2. M-estimation

Ritov (1990), Lai and Ying (1994) considered the following estimating equation $\Psi(\beta; s)=0$ with $s(\cdot)$ being a known function and

$$\Psi(\beta; s) = \sum_{i=1}^{n}(X_i - \bar{X})\left\{\delta_i s(Y_i - X_i^{\mathrm{T}}\beta) + (1-\delta_i)\frac{\int_{Y_i - X_i^{\mathrm{T}}\beta}^{\infty} s(u)\,\mathrm{d}\hat{F}_{n,\beta}(u)}{1-\hat{F}_{n,\beta}(Y_i - X_i^{\mathrm{T}}\beta)}\right\}, \tag{2}$$

where $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$, and $\hat{F}_{n,b}(t)$ is the Kaplan–Meier estimator based on data $\{(Y_i - X_i^{\mathrm{T}}b, \delta_i), i = 1, \ldots, n\}$. Specifically,

$$\hat{F}_{n,b}(t) = 1 - \prod_{i:e_i(b)<t}\left[1 - \frac{\delta_i}{\sum_{j=1}^{n} 1\{e_j(b) \geqslant e_i(b)\}}\right], \tag{3}$$

where $e_i(b) = Y_i - X_i^{\mathrm{T}}b$, $i = 1, \ldots, n$.

Notice that, if $s(u) = u$, then (2) becomes the Buckley–James estimating function based on least squares principle (Buckley and James, 1979); if $s(u) = |u|^{r-1}$ with $r > 1$, then (2) becomes estimating function of L_r regression; if $s(u) = \mathrm{sign}(u)$, then (2) becomes estimating function of least absolute deviation; and (2) becomes the maximum likelihood estimating function when $s(u) = -\dot{f}(u)/f(u)$, where $f(u)$ is the density function of ε and $\dot{f}(u)$ is its derivative.

Recently, Jin et al. (2006a) studied the least squares approach with $s(u) = u$ for the model (2) along the line of Buckley and James (1979). Below we extend their approach for general function $s(u)$. Specifically, with an initial estimator $\hat{\beta}_{(0)}$, which is $\sqrt{n}$ consistent, obtain a new estimator $\hat{\beta}_{(1)}$ by the one-step Newton–Raphson approximation:

$$\hat{\beta}_{(1)} = \hat{\beta}_{(0)} - \left[\left.\frac{\partial \mathrm{U}(\beta; \hat{\beta}_{(0)}, s)}{\partial \beta}\right|_{\beta=\hat{\beta}_{(0)}}\right]^{-1} \mathrm{U}(\hat{\beta}_{(0)}; \hat{\beta}_{(0)}, s), \tag{4}$$

where

$$\mathrm{U}(\beta; b, s) = \sum_{i=1}^{n}(X_i - \bar{X})\left\{s(e_i(\beta)) + (1-\delta_i)\left[\frac{\int_{e_i(b)}^{\infty} s(u)\,\mathrm{d}\hat{F}_{n,b}(u)}{1-\hat{F}_{n,b}(e_i(b))} - s(e_i(b))\right]\right\}. \tag{5}$$