Computational Statistics and Data Analysis 57 (2013) 491–503

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Variable selection in the additive rate model for recurrent event data

Xiaolin Chen [a], Qihua Wang [a,b,*]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PR China
[b] School of Mathematics and Statistics, Yunnan University, Kunming 650091, PR China

ARTICLE INFO

Article history:
Received 6 October 2011
Received in revised form 14 June 2012
Accepted 20 June 2012
Available online 22 July 2012

Keywords:
Adaptive LASSO
Additive rate model
SCAD
Recurrent event data
Variable selection

ABSTRACT

In this paper, we investigate the variable selection problem for recurrent event data under the additive rate model. According to the explicit estimator of the regression coefficients of the additive rate model, a loss function is constructed. It has a form similar to the ordinary least squares of a linear regression model up to a constant. We develop variable selection procedures by penalizing the loss function with the adaptive L_1 penalty and smoothly clipped absolute derivation penalty, respectively. Under some mild regularity conditions, the oracle properties of both procedures are established. Extensive simulation studies are conducted to examine the performance of our proposed procedures in finite samples. Finally, these methods are applied to the well-known chronic granulomatous disease study.

1. Introduction

Recurrent event data appear frequently in various fields, such as biomedical research, reliability experiments and some other research fields. Examples of recurrent events include repeated opportunistic infections among HIV patients, recurrent seizures in epileptic patients, multiple hospitalizations among organ transplant recipients, tumor metastases, and so on. For the regression analysis of recurrent event data, there exist mainly two classes of methods in the literature: conditional methods and marginal methods. The conditional methods are based on modeling the intensity or hazard functions (Prentice et al., 1981; Anderson and Gill, 1982), while the marginal methods model the mean or rate functions (Lin et al., 2000; Schaubel et al., 2006). Compared with the conditional methods, the marginal methods have fewer assumptions. Furthermore, the mean number of recurrent events is more intuitive than the event intensity. Hence the marginal methods for the recurrent event data have received much attention in recent years.

Variable selection is vital to complex statistical modeling and has been an important topic in regression analysis. As discussed by Breiman (1996), the traditional methods for variable selection such as the best subset selection, AIC, BIC, etc., suffer from several drawbacks such as instability and intensive computation. In the last decade, many new and efficient variable selection methods have been proposed by statisticians. Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) method which minimizes the penalized ordinary least squares with the L_1 penalty function. By a smart analysis, Fan and Li (2001) suggested the smoothly clipped absolute derivation (SCAD) penalty function and proved the oracle property of the regression coefficient estimator from the SCAD penalized ordinary least squares. By modifying the LASSO method, Zou (2006) developed the adaptive LASSO method and established the corresponding oracle property. The adaptive LASSO and SCAD methods are currently the most important and frequently used variable selection methods in various regression analyses.

* Corresponding author at: Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PR China.
E-mail addresses: xlchen@amss.ac.cn (X. Chen), qhwang@amss.ac.cn (Q. Wang).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.06.019

In the regression analysis of failure time data, variable selection and influence diagnostics have been investigated in the literature (Fan and Li, 2002; Zhang and Lu, 2007; Leng and Ma, 2007; Leiva et al., 2007; Martinussen and Scheike, 2009; Li and Gu, 2012). But to the best of our knowledge, there is no literature studying the variable selection problem for recurrent event data except Tong et al. (2009), which proposed the variable selection procedure for the multiplicative rate model based on the SCAD method. As noted by Schaubel et al. (2006), the popularity of the multiplicative rate model derives not only from its utility and wide applicability, but also from convention and the availability of statistical software. The regression coefficients based on the multiplicative rate model reflect relative covariate effects, while those for the additive rate model characterize the absolute covariate effects. Schaubel et al. (2006) also pointed out that the additive rate model may indeed be more appropriate in many applications, especially in the case of continuous covariates. Based on this point, in this paper, we investigate the variable selection problem for recurrent event data under the additive rate model.

The remainder of this article is organized as follows. In Section 2, we introduce the additive rate model and a loss function. The adaptive LASSO and SCAD methods along with their oracle properties and algorithms are stated in Sections 3 and 4 respectively. The performance of the proposed procedures in finite samples is investigated through extensive simulation studies in Section 5 and the chronic granulomatous disease (CGD) study is analyzed to illustrate our procedures in Section 6. The regularity conditions and the proofs of the theorems are given in the Appendix.

2. The additive rate model and a loss function

Let $N^*(t)$ denote the number of recurrent events over time $[0, t]$ and let C be the censoring time of the recurrent events. Denote the $p \times 1$ covariate vector by Z. Define the at risk process to be $Y(t) = I(C \geq t)$. Assume that the end time of the study is τ and that $P(C \geq \tau) > 0$. Subject to right censoring, the observed event process is $N(t) = N^*(t \wedge C)$, i.e., $N(t) = \int_0^t I(C \geq s)dN^*(s)$. For the censoring mechanism, we assume that

$$E\{dN^*(t)|Z, C \geq t\} = E\{dN^*(t)|Z\},$$

which is referred to as independent censoring or noninformative censoring in the literature.

In this article, we consider the following additive rate model:

$$E\{dN^*(t)|Z\} = d\mu_0(t) + \beta_0^T Z dt, \tag{1}$$

where $\mu_0(t)$ is the true baseline mean function and β_0 is the true regression coefficient vector. Let $\{N_i^*(\cdot), C_i, Z_i, i = 1, \ldots, n\}$ be independent and identically distributed samples under model (1). Throughout this paper, for a vector a, $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa^T$.

To estimate β_0, Schaubel et al. (2006) developed the following estimating equation:

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \{Z_i - \bar{Z}(t)\} dM_i(\beta, t) = 0, \tag{2}$$

where

$$\bar{Z}(t) = S^{(1)}(t)/S^{(0)}(t),$$

$$S^{(k)}(t) = n^{-1} \sum_{i=1}^{n} Y_i(t) Z_i^{\otimes k}, \quad k = 0, 1$$

and

$$M_i(\beta, t) = N_i(t) - \int_0^t Y_i(s)\{d\mu_0(s) + \beta^T Z_i ds\}.$$

For simplicity, in what follows, we denote $M_i(\beta_0, t)$ to be $M_i(t)$. It is easy to see that $M_i(t)$ and hence $U(\beta_0)$ are mean-zero processes under model (1) and the assumption of independent censoring. By solving (2), we can obtain the explicit estimator

$$\tilde{\beta} = \left[\sum_{i=1}^{n} \int_0^{\tau} \{Z_i - \bar{Z}(t)\}^{\otimes 2} Y_i(t) dt \right]^{-1} \left[\sum_{j=1}^{n} \int_0^{\tau} \{Z_j - \bar{Z}(t)\} dN_j(t) \right]. \tag{3}$$

Let $A_n = \sum_{i=1}^n \int_0^{\tau} \{Z_i - \bar{Z}(t)\}^{\otimes 2} Y_i(t) dt$ and $U_n = \sum_{i=1}^n \int_0^{\tau} \{Z_i - \bar{Z}(t)\} dN_i(t)$. Then $\tilde{\beta} = A_n^{-1} U_n$, the large sample properties of which were established by Schaubel et al. (2006).

It is easy to see that $\tilde{\beta}$ is the minimizer of the following loss function:

$$l(\beta) = \beta^T A_n \beta - 2\beta^T U_n, \tag{4}$$

i.e., $\tilde{\beta} = \text{argmin}_\beta l(\beta)$. It is noted that (4) has a form similar to the ordinary least squares of a linear regression model up to a constant. In fact, Leng and Ma (2007) and Martinussen and Scheike (2009) used the same idea for the variable selection of the failure time data under the additive risk model (Lin and Ying, 1994). This loss function plays an important role in our variable selection procedures. Based on the loss function, we can implement the variable selection procedures as in the linear regression model. In Sections 3 and 4, we develop variable selection procedures by penalizing the loss function (4) using the adaptive L_1 penalty and smoothly clipped absolute derivation penalty functions, respectively.