Journal of Multivariate Analysis 122 (2013) 20–34

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Quantile regression analysis of case-cohort data

Ming Zheng, Ziqiang Zhao, Wen Yu *

Department of Statistics, School of Management, Fudan University, Shanghai 200433, PR China

ARTICLE INFO

Article history:
Received 15 October 2012
Available online 29 July 2013

AMS subject classifications:
62N01
62N02

Keywords:
Case-cohort design
Counting process
Estimating equation
Random weighting
Simple random sampling
Uniform consistency
Weak convergence

ABSTRACT

Case-cohort designs provide a cost effective way to conduct epidemiological follow-up studies in which event times are the outcome variables. This paper develops a quantile regression approach to the analysis of case-cohort data. Quantile regression is a highly useful tool to delineate relationships between the outcome variable and covariates. Unbiased functional estimating equations are constructed, resulting in asymptotically unbiased estimators. Efficient algorithms based on minimizing L_1-type convex functions are given. Uniform consistency and weak convergence of the resulting estimators are established. Error estimation and confidence intervals are obtained by applying a specially designed resampling procedure for case-cohort data. Simulation studies are conducted to assess the performance of the proposed method. An example is also provided for illustration.

1. Introduction

The case-cohort design proposed by Prentice [22] provides a cost effective way of conducting large epidemiological cohort studies when event time is the outcome variable. In a typical epidemiological cohort study, subjects are sampled from a population and their disease trajectories are followed. Although only a small fraction of the subjects will develop the disease of interest, the usual regression analysis requires the exposure histories are ascertained for all subjects. In the case-cohort design, however, the covariate information is collected only for the diseased subjects (i.e., cases) and a randomly sampled subcohort of censored observations. For rare diseases, the size of the subcohort can be much smaller than that of the entire cohort, resulting in substantial savings.

Standard analyses of cohort studies with event time outcomes are usually conducted using the Cox regression model [2]. For the case-cohort design, Prentice [22] proposed a pseudo-likelihood approach to analyze case-cohort data under the Cox regression model. Further developments were made by Self and Prentice [23], Lin and Ying [16], Chen and Lo [5], Chen [3], Kulich and Lin [15], etc. Kulich and Lin [14] developed the case-cohort analysis under the additive hazard model, another important type of hazard-based regression model. Chen [4] discussed the analysis of case-cohort data using the proportional odds model. More recently, Lu and Tsiatis [17]and Chen and Zucker [6] analyzed case-cohort data under the class of linear transformation models, which includes the Cox model and the proportional odds model as special cases.

An important alternative to the Cox model in survival analysis is the accelerated failure time (AFT) model; cf. [7]. This model relates the mean of the logarithm of the event time linearly to the covariates. Due to its connection to the classical linear regression and its ease of interpretation of the regression effect, it is of interest to develop parallel approaches to handling case-cohort data. The case-cohort regression analysis under the AFT model was studied by Kong and Cai [13].

* Corresponding author.
E-mail address: wenyu@fudan.edu.cn (W. Yu).

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.07.004

As a significant extension of classical linear regression, quantile regression has received much attention since the influential work of Koenker and Bassett [12]. A linear quantile regression model links the conditional quantiles of the outcome to the covariates linearly. Specifically, let Y be the outcome variable and $\tilde{Z}$ the corresponding $(p-1)\times 1$ covariate vector. Let $Z = (1, \tilde{Z}^\top)^\top$. For any $\tau \in [0, 1)$, the conditional quantile of Y given Z is defined as $Q_Y(\tau|Z) = \sup\{t : \mathrm{P}(Y \leqslant t|Z) \leqslant \tau\}$. A linear quantile regression model assumes that for each $\tau \in (0, 1)$,

$$Q_Y(\tau|Z) = Z^\top \beta(\tau), \tag{1}$$

where $\beta(\tau)$, the parameter vector representing the relationship of covariates to the τ-th quantile of Y, may depend upon τ and is right continuous. Model (1) describes the relationship of the covariates with the conditional distribution of Y rather than the conditional mean only, as in the AFT model. The estimates of the regression parameters can be obtained efficiently using a linear programming algorithm. Resampling techniques can be adopted to make inferences about the regression parameters, cf., [10]. A nice review of quantile regression models can be found in [11].

When the outcome variable is an event time, estimation and inference for the regression parameters of quantile models become complicated due to the censoring. For cohort data with complete covariate data, much effort has devoted to the development of estimation procedures for the regression parameters. Early work includes Powell [19,20]'s modified least absolute deviation method with always observable censoring times. After that, Ying et al. [25] proposed a semiparametric estimation procedure under the restrictive assumption that censoring time is independent of event time and covariates. Under the conditionally independent censoring assumption, Portnoy [21] developed a recursively reweighed estimating procedure. More recently, Peng and Huang [18] novelly developed a series of martingale-type estimating functions assuming conditionally independent censoring. The resulting estimating equations can be solved by minimizing a sequence of L_1-type convex functions. They also systematically studied the large sample properties of the proposed estimators and constructed corresponding inference procedures.

In this paper, we develop a quantile regression approach for the analysis of case-cohort data. Unbiased estimating functions that account for the missing covariates in case-cohort designs are constructed. The subcohort can be drawn either by simple random sampling with fixed subcohort sizes or by independent Bernoulli sampling with arbitrary selection probabilities. The proposed estimating equations for the regression parameters can be solved by minimizing L_1-type convex functions. We establish the large sample properties, including consistency and asymptotic normality, of the resulting estimators. However, the corresponding limiting variances depend on unknown density functions which may not be estimated well nonparametrically with case-cohort data. In order to obtain the complete inference procedures, we develop a new random weighting approach for case-cohort data to estimate the standard errors.

The remainder of the paper is organized as follows. In Section 2, we introduce notation, describe the specification of the model, and construct the estimating functions. The algorithm for solving the resulting estimating equations is then developed and the new resampling technique is described. In Section 3, the finite sample performance of the proposed estimating approach is assessed by a series of simulation studies. A real example is used to illustrate the proposed method in Section 4. Section 5 concludes. All technical details are summarized in the Appendix.

2. Main results

In this section, we first introduce necessary notation and describe the quantile regression model. Then we develop unbiased estimating functions that account for case-cohort data. The algorithm for obtaining the estimators is designed. Finally, we develop the resampling procedures for standard error estimation and inferences.

2.1. Notation and model specification

We will use T to denote the event time and $\tilde{Z}$ the corresponding $(p-1)\times 1$ covariate vector. Let C be the censoring time, $X = T \wedge C$ and $\Delta = I\{T \leqslant C\}$, where $\wedge$ is the minimum operator and $I\{\cdot\}$ represents the indicator function. We assume that C is independent of T given covariates $Z = (1, \tilde{Z}^\top)^\top$.

For the cohort study with complete covariates observations, the data are assumed to be n independently and identically distributed (i.i.d.) replicates of (X, Δ, Z) and are denoted by $\{(X_i, \Delta_i, Z_i), i = 1, \ldots, n\}$. Define $F_T(t|Z) = \mathrm{P}(T \leqslant t|Z)$ and $\Lambda_T(t|Z) = -\log\{1 - F_T(t|Z)\}$ to be the conditional distribution function and conditional cumulative hazard function of T given Z, respectively. Let $N(t) = \Delta I\{X \leqslant t\}$ and $M(t) = N(t) - \Lambda_T(t \wedge X|Z)$. Let $N_i(t)$ and $M_i(t)$, $i = 1, \ldots, n$, be the sample analogs of $N(t)$ and $M(t)$. It is not difficult to see that $M_i(t)$ is a martingale process with an appropriate σ-filtration.

Under the case-cohort designs, covariates are available only on the cases, i.e., those subjects with $\Delta_i = 1$, and on the subcohort. In this section we assume that the subcohort is drawn from the entire cohort by simple random sampling scheme with fixed size denoted by $\tilde{n}$. Let ξ_i be a binary variable. It takes values 1 and 0, indicating whether or not the i-th subject in the original cohort is selected into the subcohort. Under the simple random sampling, $(\xi_1, \ldots, \xi_n)$ is uniformly distributed on $\{(d_1, \ldots, d_n) \in \{0, 1\}^n : \sum_{i=1}^n d_i = \tilde{n}\}$. Moreover, the subcohort indicators ξ_i's are independent of the data.

A linear quantile regression model is specified to link the event time T and the covariates Z. Since the event time is always non-negative, we take the logarithm of the event time to be the outcome variable, i.e., $Y = \log T$. We still use $Q_Y(\tau|Z)$ to denote the conditional quantile of Y given Z. Model (1) is assumed for Y and Z, i.e., $Q_Y(\tau|Z) = Z^\top \beta(\tau)$, or equivalently, $Q_T(\tau|Z) = \exp\{Z^\top \beta(\tau)\}$.