Journal of Multivariate Analysis 141 (2015) 67–80

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Semiparametric linear transformation model with differential measurement error and validation sampling

Xuan Wang, Qihua Wang*

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 15 July 2014
Available online 19 June 2015

AMS 2000 subject classifications:
62N01
62N02
62P10

Keywords:
Right-censored data
Transformation model
Estimating equation

ABSTRACT

For the semiparametric linear transformation model with covariate measurement error and validation sampling, we propose an estimation method to estimate the covariate coefficient. The method updates the validation set based estimator to get a more efficient estimator using the data information available on the whole cohort. It can be used to deal with both differential and nondifferential measurement errors. Consistency and asymptotic normality are established for the proposed estimator and a closed form formula is derived for the limiting variance–covariance matrix. Simulation studies and a real data analysis are used to illustrate the performances of the proposed method.

1. Introduction

In survival analysis, failure time data usually include covariate information. Sometimes, it is expensive or time cost to collect complete covariate information from the entire cohort. Such true covariate information is then collected only on a sample from the full study cohort. However, proxy covariate information is usually available on every subject in the full cohort. This data can be viewed as survival data with covariates subject to measurement error.

Let T be the failure time, $\mathbf{Z}$ be the true p-dimensional covariate vector. Due to measurement error, $\mathbf{Z}$ is measured only on a validation subset whereas the proxy covariate $\mathbf{Z}^*$ is obtained for all subjects. If the conditional distribution of T given $(\mathbf{Z}, \mathbf{Z}^*)$ is the same as that of T given $\mathbf{Z}$, we say $\mathbf{Z}^*$ has nondifferential measurement error. Otherwise, $\mathbf{Z}^*$ has differential measurement error. There is a large number of researchers on models with covariate measurement error and validation data for uncensored data, see [1,8,10,12], etc.

The effects of covariates or risk factors on the failure time can be studied by fitting a failure time model to the survival data. The most popular model is the Cox proportional hazards model. Besides the Cox model, there are other alternative models. Here we consider semiparametric linear transformation model due to its generality. The semiparametric linear transformation model, which is referred to as linear transformation model in the rest of this paper, assumes

$$H(T) = -\beta'\mathbf{Z} + \varepsilon, \tag{1}$$

where H is an unknown increasing function, ε is a random variable with a known distribution which is independent of $\mathbf{Z}$ and β is an unknown p-dimensional covariate coefficient of interest. Model (1) includes the proportional hazards model and the proportional odds model as special cases with ε following the extreme value distribution and the standard logistic distribution, respectively.

* Corresponding author.
E-mail address: qhwang@amss.ac.cn (Q. Wang).

http://dx.doi.org/10.1016/j.jmva.2015.05.017

For the Cox model with covariate measurement error and validation sampling, Wang et al. [11] proposed a regression calibration method. The corresponding estimator is in general inconsistent when the covariate measurement error is considerable or the measurement error is nonnormally distributed. For estimating approaches under the linear transformation model without measurement error, see [5,3]. Chen and Wang [4] considered linear transformation model with covariate measurement error for right censored data. However, their method needs to assume parametric models for both the paired differences of the measurement errors and paired differences of the true covariates. As a method based on that of [5], it suffers from the restriction that the censoring variable must be independent of the covariates and violation of this assumption would cause serious bias as shown by the simulation results of Sinha and Ma [9]. Sinha and Ma [9] proposed another method for linear transformation model with covariate measurement error based on [3]. However, they made a parametric model assumption regarding the possibly unobserved true covariate given the other always observed true covariates, while a misspecification of this assumption will lead to a biased estimator, as commented by the authors.

In this paper, we propose an estimating method (without any parametric model assumption) that leads to consistent estimator under the semiparametric linear transformation model with covariate measurement error and validation sampling. Unlike all the above mentioned approaches that require the measurement error to be nondifferential, the proposed method is feasible when the measurement error is differential. We first obtain an estimator based on the validation set using true covariate by the method of [3]. The proposed estimator is then defined by updating the validation set based estimator with the proxy covariate information and hence is more efficient than the validation set based estimator.

The remainder of this paper is as follows. In Section 2, we describe the proposed method. Theoretical properties for the proposed estimator are also derived. We evaluate the finite sample performances of the proposed method and apply the method to the analysis of a real data set in Section 3. A discussion is provided in Section 4. Finally, the proofs of theorems are given in the Appendix.

2. Estimating method

We consider the case where the validation sample is a simple random sample from the cohort study. Let V denote the index set of individuals in the sampled validation set, N denote the sample size of the whole cohort, n that of the validation set and $\rho = n/N$ the validation ratio. For subjects in the validation set, we observe $(\tilde{T}, \delta, \mathbf{Z}, \mathbf{Z}^*)$, and for subjects outside the validation set, we observe $(\tilde{T}, \delta, \mathbf{Z}^*)$, where $\tilde{T} = T \wedge C$, $\delta = I(T \le C)$ and C is the censoring variable.

We can obtain a consistent estimator of β based on the validation set using the method of [3]. Let $N(t) = I(\tilde{T} \le t, \delta = 1)$, $Y(t) = I(\tilde{T} \ge t)$ and $M(t) = N(t) - \int_0^t Y(s)\, \mathrm{d}\Lambda\{\beta_0'\mathbf{Z} + H_0(s)\}$, where $(\beta_0, H_0(\cdot))$ is the true value of $(\beta, H(\cdot))$ and $\Lambda(\cdot)$ is the cumulative hazard function of ε. We assume that the hazard function of ε, denoted by $\lambda(t)$, is strictly positive and $\dot{\lambda}(t)$ is bounded and continuously differentiable on $(-\infty, K)$, where K is any finite number. Based on the mean zero martingale process $M(t)$, we can construct estimating equations

$$\sum_{i\in V}\Big[\mathrm{d}N_i(t) - Y_i(t)\,\mathrm{d}\Lambda\{\beta'\mathbf{Z}_i + H(t)\}\Big] = 0, \quad (0 \le t \le \tau), \tag{2}$$

$$\sum_{i\in V}\int_0^\tau \mathbf{Z}_i\Big[\mathrm{d}N_i(t) - Y_i(t)\,\mathrm{d}\Lambda\{\beta'\mathbf{Z}_i + H(t)\}\Big] = 0, \tag{3}$$

where $H(\cdot)$ is a nondecreasing function satisfying $H(0) = -\infty$ and $\tau = inf\{t : Pr(\tilde{T} > t) = 0\}$. We denote the solution of (2) and (3) by $(\hat{\beta}, \hat{H}(\cdot))$.

The estimator $\hat{\beta}$ uses only the validation set information. To exploit the proxy covariate information, we update the validation set based estimator to get a more efficient estimator using the data information available on the whole cohort. Denote $L(\cdot)$ to be an unknown increasing function and γ an unknown p-dimensional vector. Let $(\hat{\gamma}, \hat{L}(\cdot))$ be the solution to the following estimating equations

$$\sum_{i\in V}\Big[\mathrm{d}N_i(t) - Y_i(t)\,\mathrm{d}\Lambda\{\gamma'\mathbf{Z}_i^* + L(t)\}\Big] = 0, \quad (0 \le t \le \tau), \tag{4}$$

$$\sum_{i\in V}\int_0^\tau \mathbf{Z}_i^*\Big[\mathrm{d}N_i(t) - Y_i(t)\,\mathrm{d}\Lambda\{\gamma'\mathbf{Z}_i^* + L(t)\}\Big] = 0. \tag{5}$$

Fix γ and solve (4) for $L(\cdot)$, the solution of $L(\cdot)$, denoted by $\hat{L}(t, \gamma)$, is a uniquely defined increasing step function with jumps only at observed failure times $t_1, t_2, \ldots, t_K$. Let

$$U^*(\gamma) = \sum_{i\in V}\int_0^\tau \mathbf{Z}_i^*\Big[\mathrm{d}N_i(t) - Y_i(t)\mathrm{d}\Lambda\{\gamma'\mathbf{Z}_i^* + \hat{L}(t,\gamma)\}\Big]. \tag{6}$$

Then $\hat{\gamma}$ which solves $U^*(\gamma) = 0$ is a consistent estimator of γ_0, and correspondingly, $\hat{L}(t, \hat{\gamma})$, also denoted as $\hat{L}(t)$, estimates $L_0(t)$ consistently, where $(\gamma_0, L_0(\cdot))$ is the unique solution to

$$E\Big[\mathrm{d}N(t) - Y(t)\,\mathrm{d}\Lambda\{\gamma'\mathbf{Z}^* + L(t)\}\Big] = 0, \quad (0 \le t \le \tau),$$