Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Estimation for semiparametric varying coefficient models with different smoothing variables under random right censoring

Seong J. Yang

Department of Statistics, Hankuk University of Foreign Studies, Republic of Korea

ARTICLE INFO

Article history:
Received 2 November 2017
Accepted 4 December 2017
Available online xxxx

AMS 2000 subject classifications:
primary 62G08
secondary 62N01

Keywords:
Profile least squares
Local linear smooth backfitting
Mean-preserving transformation
Random right censoring
Semiparametric varying coefficient model

ABSTRACT

In this paper we study a semiparametric varying coefficient model when the response is subject to random right censoring. The model gives an easy interpretation due to its direct connectivity to the classical linear model and is very flexible since nonparametric functions which accommodates various nonlinear interaction effects between covariates are admitted in the model. We propose estimators for this model using mean-preserving transformation and establish their asymptotic properties. The estimation procedure is based on the profiling and the smooth backfitting techniques. A simulation study is presented to show the reliability of the proposed estimators and an automatic bandwidth selector is given in a data-driven way.

1. Introduction and model

Semiparametric regression models have attracted many researchers since they compromise between flexibility and simplicity which are main advantages of parametric and nonparametric models, respectively. Among them, the semiparametric varying coefficient model has been extensively studied in recent years due to its direct connectivity to the classical linear model. The model assumes that $E(Y|\mathbf{X}, \mathbf{Z}, \mathbf{U}) = X_1\beta_1 + \cdots + X_p\beta_p + Z_1\alpha_1(U_1) + \cdots + Z_d\alpha_d(U_d)$ for some response Y and a set of covariates $\mathbf{X} = (X_1, \ldots, X_p)^\top$, $\mathbf{Z} = (Z_1, \ldots, Z_d)^\top$, $\mathbf{U} = (U_1, \ldots, U_d)^\top$, where β_j and α_j are unknown parameters of interest. This model can be viewed as an extension of the classical linear model in the sense that some of regression coefficients are allowed to depend on another variables, and is a semiparametric model since both parametric and nonparametric parts are included in the model. It can give more flexibility into the model while maintaining a simple structure of the linear model which promises easy interpretation. By introducing linear predictors (X_j and Z_j) in the model, the curse of dimensionality can be avoided. Examples of literature include Ahmad, Leelahanon, and Li (2005), Fan and Huang (2005), Long, Ouyang, and Shang (2013), Xia, Zhang, and Tong (2004) and Zhang, Lee, and Song (2002) among others. However, most of studies focused on the case where all smoothing variables are the same, that is, $U_1 = \cdots = U_d$. Recently, Yang and Park (2014) studied a more general model, assuming different coefficients have different smoothing variables, and Yang and Lee (2014) further extended it to the case that the distribution of the response obeys the exponential family. This generalization makes the model much more flexible since various non-linear interaction effects can be dealt with.

On the other hand, it is often encountered that a response is only partially observed. Most frequently, the response can be censored to the right, for example in medical research, due to the termination of studies or subjects' leaving studies by

E-mail address: sjyang@hufs.ac.kr.

https://doi.org/10.1016/j.jkss.2017.12.001
1226-3192/

Please cite this article in press as: Yang, S.J., Estimation for semiparametric varying coefficient models with different smoothing variables under random right censoring. Journal of the Korean Statistical Society (2017), https://doi.org/10.1016/j.jkss.2017.12.001.

various reasons. This is called random right censoring. The objective of this paper is to study a very flexible semiparametric model which takes the form of:

$$E(\phi(Y)|\mathbf{X}, \mathbf{Z}, \mathbf{U}) = \mathbf{X}^\top \boldsymbol{\beta} + \mathbf{Z}^\top \boldsymbol{\alpha}(\mathbf{U}) \tag{1}$$

where Y is subject to random right censoring, $\mathbf{X} = (X_1, \ldots, X_p)^\top$, $\mathbf{Z} = (Z_1, \ldots, Z_d)^\top$, $\mathbf{U} = (U_1, \ldots, U_d)^\top$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and $\boldsymbol{\alpha}(\mathbf{u}) = (\alpha_1(u_1), \ldots, \alpha_d(u_d))^\top$. Here, $\phi(\cdot)$ is a known function, which is introduced to include various functions according to one's interest, not to deal with random right censoring. For example, it may be the identity function or $I(\cdot \le t)$, which produces the conditional mean or the conditional probability function given covariates, respectively. When the response is censored, we have incomplete information on data so that well-known regression techniques are no longer directly applicable. To address this problem, data transformation techniques have been introduced. Buckley and James (1979), Koul, Susarla, and Van Ryzin (1981) and Leugans (1987) proposed different types of transformations in the classical linear model. The main idea behind them is to transform the data in a way that the conditional mean of the response given covariates is preserved. In other words, we have the same target function after transformation, which enables us to use some known regression techniques as long as the transformed data is observable. In nonparametric regression models, Fan and Gijbels (1994) proposed an estimation method with a more general transformation and a dependent censored data case was considered later in El Ghouch and Van Keilegom (2008).

Recent works have shown some advances in the semiparametric varying coefficient model with randomly censored data. Bravo (2014) studied the estimation of the parametric and nonparametric components based on profile least squares and established their asymptotic normalities. This work is an extension of Fan and Huang (2005) to censored data. Recently, Lin and Zhou (2016) proposed estimators for this model with right-censored and length-biased data. However, their works focused on the case that all coefficient functions have the same univariate smoothing variables, which makes estimation relatively simpler. A univariate smoothing technique (such as local constant or local linear expansion to the unknown coefficient functions) is enough for their estimation. A more flexible varying coefficient model without the parametric part in censored data was studied in Yang, El Ghouch, and Van Keilegom (2014). Since different coefficient functions depend on different smoothing variables in the model considered therein, they employed the smooth backfitting method that is known to successfully circumvent the curse of dimensionality in multivariate smoothing.

This paper is an extension of Yang et al. (2014) to a semiparametric model. In many applications, the primary interest lies on the estimation of the parametric part. We first propose an estimator for $\boldsymbol{\beta}$ based on profiling to remove the unknown coefficient functions in the estimation procedure and prove its asymptotic normality. The estimation and asymptotic properties for the nonparametric part will be also treated and a simple smoothing parameter selection method is discussed based on the theoretical results.

The rest of the paper is organized as follows. In Section 2 we present the estimation method of the parametric part based on profiling and local linear smooth backfitting. To deal with censoring, a simple data transformation is used. Section 3 introduces estimators for the nonparametric part and their theoretical properties. Section 4 is devoted to simulation studies to see how our estimators perform. Finally, a bandwidth selector is proposed in Section 5.

2. Estimation of parametric part

Let $\mathbf{W} = (\mathbf{X}^\top, \mathbf{Z}^\top, \mathbf{U}^\top)^\top$ be a set of covariates and let Y and C be a response and a censoring variable, respectively. We assume that a random sample $(\mathbf{W}_i, Y_i),\ i = 1, 2, \ldots, n$, of $(\mathbf{W}, Y)$ obeys the model (1). In the presence of random right censoring, we observe a random sample (T_i, δ_i) of a pair (T, δ) instead of observing Y_i. Here, T is the observed time which is the minimum of Y and C, and δ indicates whether the censoring occurs or not. Formally, they are defined as

$$T = \min\{Y, C\},\ \ \delta = I(Y \le C)$$

In this case, a direct estimation of the model is not possible. To solve this problem, we consider the following mean-preserving transformation

$$Y_G = \frac{\delta\phi(T)}{1 - G(T-)}$$

proposed in Koul et al. (1981) where G is the distribution function of the censoring variable. It is known that the conditional mean of the response given covariates is preserved with this transformation under some assumptions given below. That is,

$$E(Y_G|\mathbf{W} = \mathbf{w}) = E(\phi(Y)|\mathbf{W} = \mathbf{w}). \tag{2}$$

To give some intuition why the above transformation ensures Eq. (2), this transformation sets Y_G to zero when the response is censored ($\delta = 0$) and inflates $\phi(Y)$ by $1/(1 - G(Y-))$ when the response is observed ($\delta = 1$). Note that the conditional mean of δ given $\mathbf{W}$ and Y is $1 - G(Y-)$ under the assumptions given below. Thus, $1 - G(T-)$ in the denominator of the transformation is a kind of compensation for δ. We would like note that Y_G depends on the choice of ϕ, i.e., $Y_G = Y_G(\phi)$, but we omit the dependence in the expression throughout the paper for notational convenience. We would also like to note that Eq. (2) holds for any ϕ.

Then, regarding Y_G as a new response, one can estimate the unknown coefficients by using conventional regression techniques as if there were no censoring. Note that Y_G is observable as long as we assume that G is known.