Statistica Sinica **19** (2009), 1319-1335

LOCAL POLYNOMIAL MODELLING FOR VARYING-COEFFICIENT INFORMATIVE SURVIVAL MODELS

Wenyang Zhang, Yan Sun, Jin-Ting Zhang and Duolao Wang

University of Bath, Shanghai University of Finance and Economics, National University of Singapore and London School of Hygiene and Tropical Medicine

Abstract: A proportional hazard function together with partial likelihood estimation is the most common approach to the analysis of censored data. However, partial likelihood estimation is established on the grounds that the censoring is non-informative. The partial likelihood approach enjoys many good properties when the censoring is indeed non-informative. However, in reality, censoring can be informative. One pays a price in the efficiency of the estimator if partial likelihood estimation is used when the censoring is indeed informative. This problem is particularly acute in the nonparametric case. When censoring is informative, to make use of the information provided by the censoring times, it is better to take the local complete likelihood approach. Motivated by the data set about the first birth interval in Bangladesh, we propose here a varying-coefficient proportional hazard function to fit informatively censored data. We take the complete likelihood approach coupled with local linear modelling to estimate the functional coefficients involved in the model. Asymptotic properties of the proposed estimator are established, that show the proposed estimator is indeed more efficient than the maximum local partial likelihood estimator. A simulation study was conducted to demonstrate how much the proposed estimator improves the efficiency of the maximum local partial likelihood estimator when sample size is finite. In reality, we do not know whether censoring is informative or not, and a cross-validation based criterion is proposed to check whether the censoring is informative or not. Finally, the proposed varying-coefficient proportional hazard function, together with the proposed estimation method, is used to analyse the first birth interval in Bangladesh, leading to some interesting findings.

Key words and phrases: Informative censoring, local complete likelihood estimation, local linear modelling, maximum local partial likelihood estimation, proportional hazard function, varying-coefficient models.

1. Introduction

1.1. Preamble

Censored data appear frequently in medical science, finance, economics, sociology, and so on. A common approach to analyzing the censored data is maximum partial likelihood estimation under the proportional hazard function assumption, see Cox (1972). Although there are some other approaches such as the unbiased transformation-based regression approach, see Fan and Gijbels (1994), the proportional hazard function together with partial likelihood estimation is probably the most influential. For parametric settings, as far as the estimation is concerned, the partial likelihood estimation approach has been well-established, see Cox and Oakes (1984, Chap.8). Parametric modelling is a traditional modelling approach. It has many advantages when the model is correctly specified, but misspecification can lead to large bias.

Fan, Gijbels and King (1997) investigated the maximum local partial likelihood estimation for the standard nonparametric setting in which only one covariate was included. Their nonparametric setting does not easily extend to the multiple nonparametric setting with more covariates, due to the 'curse of dimensionality'. Fan and Li (2006) have a very interesting essay discussing statistical challenges with high dimensionality. Here we have to assume that the underlying model has a specific structure, which leads to semiparametric modelling.

Semiparametric modelling is an appealing approach. There have been many semiparametric models, together with estimation methods, proposed recently. Examples include additive modelling (Breiman and Friedman (1985) and Hastie and Tibshirani (1990, Chap.4)), low-dimensional interaction modelling (Friedman (1991); Stone, Hansen, Kooperberg and Truong (1997)), multiple-index models (Xia, Tong, Li and Zhu (2002)), varying-coefficient models (Hastie and Tibshirani (1993)); Cai, Fan and Li (2000); Fan and Huang (2005)), and so on.

1.2. Modelling strategy and motivation

Different models explore different aspects of high-dimensional data and incorporate different prior knowledge into modelling and approximation. Which model should be used depends on what the data we analyse is like and what aspect we want to explore. The data for our study come from the Bangladesh Demographic and Health Survey (BDHS) of 1996-97 (Mitra, Al-Sabir, Cross and Jamil (1997)), a cross-sectional nationally representative survey of ever-married women aged between 10 and 49. What we are interested in is how several factors, commonly found to be associated with the fertility behaviour, affect the first birth interval which is defined as the duration in months between marriage and the first birth. Some women in the sample had not had a birth by the time of the survey and are therefore right-censored. We use y to denote the first birth interval, and X to denote the vector of all factors concerned. We start with the standard proportional hazard function, and assume the conditional hazard function of y given X is

$$h_y(t|X) = h_{0,y}(t) g_1(X^{\mathrm{T}}\mathbf{b}_1).$$

We also take the censoring time into account, and use C to denote the censoring time. We assume the censoring time also depends on the covariate X, and the conditional hazard function of C given X is

$$h_c(t|X) = h_{0,c}(t)g_2(X^{\mathrm{T}}\mathbf{b}_2).$$

The coefficient $\mathbf{b}_1$ can be interpreted as the impact of the factors and the model assumes that the impact of the factors is constant. This assumption is not plausible as Bangladesh has experienced many changes so the impact of the factors should vary with time. Further, to explore how the impact varies with time is important, because the dynamic patterns of the impact may reveal how society changes with time. We are thus led to the varying-coefficient proportional hazard function

$$h_y(t|U,X) = h_{0,y}(t)g_1\left(X^{\mathrm{T}}\mathbf{b}_1(U)\right),$$

where $h_y(t|U,X)$ is the conditional hazard function of y given X and time U. We also assume the impact of the factors on the conditional hazard function of censoring time vary with time. This leads to the conditional hazard function of C given X and U of

$$h_c(t|U,X) = h_{0,c}(t)g_2\left(X^{\mathrm{T}}\mathbf{b}_2(U)\right).$$

Partial likelihood estimation coupled with local linear modelling can be used to construct estimators of $\mathbf{b}_1(U)$ and $\mathbf{b}_2(U)$.

Partial likelihood estimation is established on the grounds that the censoring is non-informative. The partial likelihood estimation enjoys many good properties when the censoring is indeed non-informative, see Wong (1986). However, the censoring can be informative, which means the censoring mechanism may involve some parameters which appear in the survival mechanism, and are to be estimated. Please notice that informative censoring is different than dependent censoring. Dependent censoring means the survival time is dependent on censoring time; Cheng, Hall and Yang (2007) is an interesting paper about nonparametric inference under dependent truncation. Informative censoring means the censoring time carries some information about the parameters of interest; mathematically, the distribution of censoring time involves the parameters of interest. Intuitively, the partial likelihood estimation would pay a price in efficiency of the estimator obtained, as it does not make use of the information provided by the censoring time. This problem is particularly acute in the nonparametric and semiparametric settings as we make fewer assumptions as in the parametric setting, and should be careful to avoid losing any information. To make use of the information provided by survival and censoring times, we take a complete likelihood approach coupled with local linear modelling. We show that the proposed estimator is more efficient than the maximum local partial likelihood estimator when censoring is indeed informative.

For our case, that censoring is informative means that some components of $\mathbf{b}_1(U)$ are the same as some components of $\mathbf{b}_2(U)$. Without loss of generality, we assume the first p_1 components of $\mathbf{b}_1(U)$ are the same as the first p_1 components of $\mathbf{b}_2(U)$. This leads to

$$\begin{aligned} h_y(t|U,X) &= h_{0,y}(t)g_1(X_{01}^{\mathrm{T}}\boldsymbol{\beta}_0(U) + X_{02}^{\mathrm{T}}\boldsymbol{\beta}_1(U)), \\ h_c(t|U,X) &= h_{0,c}(t)g_2(X_{01}^{\mathrm{T}}\boldsymbol{\beta}_0(U) + X_{02}^{\mathrm{T}}\boldsymbol{\beta}_2(U)), \end{aligned} \tag{1.1}$$

where $(X_{01}^{\mathrm{T}}, X_{02}^{\mathrm{T}})^{\mathrm{T}} = X$, $h_{0,y}(t)$ and $h_{0,c}(t)$ are unknown baseline functions, $g_1(\cdot)$ and $g_2(\cdot)$ are known link functions, $\boldsymbol{\beta}_j(\cdot)$, $j = 0, 1, 2$, are unknown functional coefficients to be estimated. Although (1.1) is motivated by the data about the first birth interval in Bangladesh, apparently it can be used to analyse the survival data from a wider range of scientific areas. The goal of this paper is to introduce a methodology to deal with informatively censored data; from now on, y represents survival time rather than first birth interval only, C is censoring time, X and U are covariates where $X = (x_1, \ldots, x_p)^{\mathrm{T}}$ is a p dimensional vector, and U is a scalar.

The estimation method, together with the associated asymptotic theory developed in this paper, straightforwardly applies to the case where the censoring and survival time depend on different covariates by changing notations.

Generally, we do not know whether censoring is informative or not. We would pay a price in bias if the proposed informative proportional hazard function is used when the censoring is indeed non-informative. Further, we need to detect which unknown functions in the survival mechanism also appear in the censoring mechanism when the censoring is indeed informative. All these questions are essentially model selection questions. There are many criteria proposed to select an optimal model. The most commonly used are AIC, BIC and cross-validation (Stone (1977)). Fan and Li (2002) proposed a promising model selection approach based on smoothly clipped absolute deviation penalty. We rely on cross-validation to select the optimal model.

The paper is organised as follows. We begin in Section 2 with a description of the proposed estimation procedure for the unknown functional coefficients. In Section 3, we discuss the model selection criterion based on cross-validation. The asymptotic properties of the proposed estimator are presented in Section 4. A simulation study is conducted in Section 5 to demonstrate how much the proposed estimator improves on maximum local partial likelihood estimator when censoring is informative and sample size is finite. Finally, in Section 6, we explore how several factors which are commonly found to be associated with fertility behaviour affect the length of first birth intervals in Bangladesh, and how the impacts vary with time based on the proposed model and estimation procedure.

2. Estimation Procedure

We assume the (U_i, X_i, y_i), $i = 1, \ldots, n$, are samplied from (U, X, y), independent and identically distributed. The y_i are right-censored and the censoring times, C_i, are samplied from C, and assumed to be independent and identically distributed. The distribution of C depends on (U, X). The conditional hazard functions of y and C given (U, X) satisfy (1.1). The observed data are $(t_i, U_i, X_i, \delta_i)$, $i = 1, \ldots, n$, where $t_i = \min(y_i, C_i)$, $\delta_i = I(y_i < C_i)$, $t = \min(y, C)$, and $\delta = I(y < C)$.

We assume y_i is independent of C_i given (U_i, X_i). The conditional likelihood function of (t_i, δ_i), $i = 1, \ldots, n$, given (U_i, X_i), $i = 1, \ldots, n$, is

$$L = \prod_{i=1}^{n} \left\{ \int_{t_i}^{\infty} f(u, t_i | U_i, X_i) du \right\}^{\delta_i} \left\{ \int_{t_i}^{\infty} f(t_i, v | U_i, X_i) dv \right\}^{1-\delta_i},$$

where $f(u, v|U, X)$ is the conditional joint density function of (C, y) given (U, X). Let

$$X_i = (X_{i1}^{\mathrm{T}}, X_{i2}^{\mathrm{T}})^{\mathrm{T}}, \quad \Lambda_c(t) = \int_0^t h_{0,c}(u) du, \quad \Lambda_y(t) = \int_0^t h_{0,y}(u) du,$$

$$g_{1i} = g_1\Big(X_{i1}^{\mathrm{T}} \boldsymbol{\beta}_0(U_i) + X_{i2}^{\mathrm{T}} \boldsymbol{\beta}_1(U_i)\Big), \quad g_{2i} = g_2\Big(X_{i1}^{\mathrm{T}} \boldsymbol{\beta}_0(U_i) + X_{i2}^{\mathrm{T}} \boldsymbol{\beta}_2(U_i)\Big).$$

By simple calculation, we have the log-likelihood function

$$\begin{aligned}\mathcal{L} = \sum_{i=1}^{n} \Big[& - g_{2i} \Lambda_c(t_i) - g_{1i} \Lambda_y(t_i) + (1 - \delta_i) \Big\{ \log h_{0,c}(t_i) + \log g_{2i} \Big\} \\ & + \delta_i \Big\{ \log h_{0,y}(t_i) + \log g_{1i} \Big\} \Big]. \end{aligned} \tag{2.1}$$

For any given u, by Taylor's expansion, we have $\boldsymbol{\beta}_j(U_i) \approx \mathbf{a}_j + \mathbf{b}_j(U_i - u)$, $j = 0, 1, 2$, when U_i is in a small neighbourhood of u. This leads to the local log-likelihood function

$$\begin{aligned}\ell_0 = \sum_{i=1}^{n} \Big[& - g_{2ih}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_2, \mathbf{b}_2) \Lambda_c(t_i) - g_{1ih}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_1, \mathbf{b}_1) \Lambda_y(t_i) \\ & + (1 - \delta_i) \Big\{ \log h_{0,c}(t_i) + \log g_{2ih}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_2, \mathbf{b}_2) \Big\} \\ & + \delta_i \Big\{ \log h_{0,y}(t_i) + \log g_{1ih}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_1, \mathbf{b}_1) \Big\} \Big] K_h(U_i - u), \end{aligned} \tag{2.2}$$

where $K_h(\cdot) = K(\cdot/h)/h$, h is the bandwidth, $K(\cdot)$ is the kernel function,

$$g_{1ih}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_1, \mathbf{b}_1) = g_1\Big(X_{i1}^{\mathrm{T}}\mathbf{a}_0 + X_{i1}^{\mathrm{T}}\mathbf{b}_0(U_i - u) + X_{i2}^{\mathrm{T}}\mathbf{a}_1 + X_{i2}^{\mathrm{T}}\mathbf{b}_1(U_i - u)\Big),$$

$$g_{2ih}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_2, \mathbf{b}_2) = g_2\Big(X_{i1}^{\mathrm{T}}\mathbf{a}_0 + X_{i1}^{\mathrm{T}}\mathbf{b}_0(U_i - u) + X_{i2}^{\mathrm{T}}\mathbf{a}_2 + X_{i2}^{\mathrm{T}}\mathbf{b}_2(U_i - u)\Big).$$

The least informative modelling approach, see Wong (1986), is employed to deal with $\Lambda_y(t)$ and $\Lambda_c(t)$, i.e., $\Lambda_y(t)$ has a possible jump λ_{yj} only at the observed failure time t_j and $\Lambda_c(t)$ has a possible jump λ_{cl} only at the observed censoring time t_l. This means

$$\begin{aligned} \Lambda_y(t_j) &= \sum_{i=1}^n \delta_i I(t_i \le t_j)\lambda_{yi}, & h_{0,y}(t_j) &= \delta_j \lambda_{yj}, \\ \Lambda_c(t_j) &= \sum_{i=1}^n (1-\delta_i) I(t_i \le t_j)\lambda_{ci}, & h_{0,c}(t_j) &= (1-\delta_j)\lambda_{cj}, \end{aligned} \tag{2.3}$$

which, together with (2.2), leads to

$$\begin{aligned} \ell_1 = \sum_{i=1}^n \Bigg[& - g_{2ih}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_2, \mathbf{b}_2) \sum_{j=1}^n (1-\delta_j) I(t_j \le t_i)\lambda_{cj} \\ & - g_{1ih}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_1, \mathbf{b}_1) \sum_{j=1}^n \delta_j I(t_j \le t_i)\lambda_{yj} \\ & + (1-\delta_i)\Big\{ \log \lambda_{ci} + \log g_{2ih}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_2, \mathbf{b}_2)\Big\} \\ & + \delta_i \Big\{ \log \lambda_{yi} + \log g_{1ih}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_1, \mathbf{b}_1)\Big\}\Bigg] K_h(U_i - u), \end{aligned} \tag{2.4}$$

which is the objective function used for estimation. Let $R = \{i : \delta_i = 1, 1 \le i \le n\}$, and $R_c = \{i : \delta_i = 0, 1 \le i \le n\}$. Maximising ℓ_1 with respect to λ_{yi} and λ_{cj}, $i \in R$, $j \in R_c$, we get the maximizer

$$\hat{\lambda}_{yi} = K_h(U_i - u)\Bigg[\sum_{j=1}^n I(t_j \ge t_i) g_{1jh}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_1, \mathbf{b}_1) K_h(U_j - u)\Bigg]^{-1}, \quad i \in R,$$

$$\hat{\lambda}_{ci} = K_h(U_i - u)\Bigg[\sum_{j=1}^n I(t_j \ge t_i) g_{2jh}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_2, \mathbf{b}_2) K_h(U_j - u)\Bigg]^{-1}, \quad i \in R_c.$$

Substituting $\hat{\lambda}_{yi}$ and $\hat{\lambda}_{cj}$ for λ_{yi} and λ_{cj} respectively in (2.4), $i \in R$, $j \in R_c$, we have

$$\ell_2 = \sum_{i=1}^n K_h(U_i - u)\Bigg\{(1-\delta_i)\log \frac{g_{2ih}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_2, \mathbf{b}_2)}{\sum_{j=1}^n I(t_j \ge t_i) g_{2jh}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_2, \mathbf{b}_2) K_h(U_j - u)}$$

$$+\delta_i \log \frac{g_{1ih}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_1, \mathbf{b}_1)}{\sum_{j=1}^n I(t_j \ge t_i) g_{1jh}(\mathbf{a}_0, \mathbf{b}_0, \mathbf{a}_1, \mathbf{b}_1) K_h(U_j - u)} \Bigg\} + D, \tag{2.5}$$

where D is independent of $\mathbf{a}_i$, $\mathbf{b}_i$, $i = 0, 1, 2$. Maximize ℓ_2 with respect to $\mathbf{a}_i$, $\mathbf{b}_i$, $i = 0, 1, 2$, to get the maximizer $(\hat{\mathbf{a}}_i, \hat{\mathbf{b}}_i)$, $i = 0, 1, 2$. The estimator of $\boldsymbol{\beta}_i(u)$ is taken to be $\hat{\mathbf{a}}_i$. As there is no closed form for the maximizer of ℓ_2, the Newton-Raphson algorithm is used to maximize ℓ_2.

From (2.5), it is easy to see that $(\hat{\mathbf{a}}_1, \hat{\mathbf{b}}_1)$, the estimator of the functional coefficient that does not appear in the censoring mechanism, is the same as the maximiser of

$$\sum_{i=1}^n K(U_i - u)\delta_i \log \frac{g_{1ih}(\mathbf{a}_1, \mathbf{b}_1)}{\sum_{j=1}^n I(t_j \ge t_i) g_{1jh}(\mathbf{a}_1, \mathbf{b}_1) K_h(U_j - u)}, \tag{2.6}$$

which is the local partial likelihood function based on the observed survival times. So, for the functional coefficient that does not appear in the censoring mechanism, there is no difference between its estimators obtained by either the proposed local complete likelihood approach or the local partial likelihood approach based on the observed survival times.

When the censoring is non-informative, $\mathbf{a}_0$ and $\mathbf{b}_0$ would disappear from ℓ_2, and to maximise ℓ_2 would be equivalent to maximising (2.6) and

$$\sum_{i=1}^n K(U_i - u)(1 - \delta_i) \log \frac{g_{2ih}(\mathbf{a}_2, \mathbf{b}_2)}{\sum_{j=1}^n I(t_j \ge t_i) g_{2jh}(\mathbf{a}_2, \mathbf{b}_2) K_h(U_j - u)},$$

which is the local partial likelihood function based on the observed censoring times. So, the proposed estimator would be exactly the same as maximum local partial likelihood estimator when censoring is non-informative.

3. Model Selection

In this section, we devise a cross-validation-based criterion to assess whether the censoring is informative or not, and find the optimal model.

Because model selection is in the global sense, the criterion for model selection cannot be based on a local likelihood function such as ℓ_2 at (2.5). We have to start from the original log conditional joint density function of (t_i, δ_i) given (U_i, X_i),

$$\mathcal{D}_i = (1-\delta_i)\Big\{\log h_{0,c}(t_i) + \log g_{2i}\Big\} + \delta_i\Big\{\log h_{0,y}(t_i) + \log g_{1i}\Big\} - g_{2i}\Lambda_c(t_i) - g_{1i}\Lambda_y(t_i).$$

Because $\Lambda_c(\cdot)$ and $\Lambda_y(\cdot)$ are unknown nuisance functions, we use least informative modelling, i.e. (2.3), to model them. Substituting (2.3) for $\Lambda_c(\cdot)$, $\Lambda_y(\cdot)$, $h_{0,c}(\cdot)$, and $h_{0,y}(\cdot)$ in (2.1), then maximising $\mathcal{L}$ with respect to λ_{yi} and λ_{cj}, $i \in R$, $j \in R_c$, we have

$$\check{\lambda}_{yi} = \left(\sum_{j=1}^{n} I(t_j \geq t_i) g_{1j} \right)^{-1}, \ i \in R, \quad \check{\lambda}_{ci} = \left(\sum_{j=1}^{n} I(t_j \geq t_i) g_{2j} \right)^{-1}, \ i \in R_c.$$

Substituting (2.3) for $\Lambda_c(\cdot)$, $\Lambda_y(\cdot)$, $h_{0,c}(\cdot)$, and $h_{0,y}(\cdot)$ in $\mathcal{D}_i$, then replacing λ_{cj} and λ_{yj} by $\check{\lambda}_{cj}$ and $\check{\lambda}_{yj}$ respectively, we have

$$\check{\mathcal{D}}_i = (1-\delta_i) \log \frac{g_{2i}}{\sum_{j=1}^{n} I(t_j \geq t_i) g_{2j}} + \delta_i \log \frac{g_{1i}}{\sum_{j=1}^{n} I(t_j \geq t_i) g_{1j}} - D_{0i},$$

where

$$\begin{aligned} D_{0i} = g_{1i} & \sum_{j=1}^{n} \delta_j I(t_j \leq t_i) \left(\sum_{k=1}^{n} I(t_k \geq t_j) g_{1k} \right)^{-1} \\ & + g_{2i} \sum_{j=1}^{n} (1-\delta_j) I(t_j \leq t_i) \left(\sum_{k=1}^{n} I(t_k \geq t_j) g_{2k} \right)^{-1}. \end{aligned}$$

Because $\sum_{i=1}^{n} D_{0i} = n$, D_{0i} does not contribute anything to the cross-validation and can be dropped from $\check{\mathcal{D}}_i$, which leads to

$$F_i = (1-\delta_i) \log \frac{g_{2i}}{\sum_{j=1}^{n} I(t_j \geq t_i) g_{2j}} + \delta_i \log \frac{g_{1i}}{\sum_{j=1}^{n} I(t_j \geq t_i) g_{1j}}.$$

F_i is the quantity on which the proposed cross-validation is based.

For each i, $i = 1, \ldots, n$, delete the ith observation and, based on the other $n-1$ observations, use the estimation procedure proposed above to compute the estimators of $\boldsymbol{\beta}_j(\cdot)$, $j = 0$, 1, 2. Denote them by $\hat{\boldsymbol{\beta}}_j^{\setminus i}(\cdot)$, $j = 0, 1, 2$. Compute

$$\begin{aligned} \hat{g}_{1j} &= g_1 \left(X_{j1}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_0^{\setminus i}(U_j) + X_{j2}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_1^{\setminus i}(U_j) \right), \\ \hat{g}_{2j} &= g_2 \left(X_{j1}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_0^{\setminus i}(U_j) + X_{j2}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_2^{\setminus i}(U_j) \right), \quad j = 1, \ldots, n. \end{aligned}$$

Let $\hat{F}_i$ be F_i with g_{1j} and g_{2j} being replaced, respectively, by $\hat{g}_{1j}$ and $\hat{g}_{2j}$, $j = 1, \ldots, n$. The cross-validation sum is

$$\mathrm{CV} = n^{-1} \sum_{i=1}^{n} \hat{F}_i,$$

and the chosen model is the one maximising CV.

The proposed CV can also be used to select bandwidth. However, it is difficult to accomplish model selection and bandwidth selection simultaneously.

Smaller bandwidth would have the same effect as fewer coefficients involved in the censoring mechanism: both would lead to smaller bias and larger variance. As our primary interest lies in the coefficients related to the survival time, we assume censoring is noninformative and ignore the first term in F_i when constructing the CV to select bandwidth. The selected bandwidth is the one maximizing

$$n^{-1}\sum_{i=1}^{n}\delta_i \log\frac{\hat{g}_{1i}}{\sum_{j=1}^{n} I(t_j \geq t_i)\hat{g}_{1j}}.$$

As we assume the censoring is noninformative when selecting bandwidth, one may think the selected bandwidth would tend to larger. It is true that assuming censoring to be noninformative would favor larger bandwidth, but, on the other hand, cross-validation as a criterion favors smaller bandwidths. So, intuitively the selected bandwidth should be all right. Our simulation results show this bandwidth selection does work well.

4. Asymptotic Properties

For clearer notation, we suppress u when there is no ambiguity. We use I_k to denote a $k\times k$ identity matrix, $\mathbf{0}_{k\times l}$ to be a $k\times l$ matrix with all entries being zero, $q = 2p-p_1$, and H to be a diagonal matrix with size $2q$. The first q elements on the diagonal of H are 1, and the last q elements are h. For any function $G(\cdot)$, we write $G'(\cdot)$ and $G''(\cdot)$ for its first and second derivative, respectively. Let

$$\begin{aligned}
&\boldsymbol{\eta}(u) = \Big(\boldsymbol{\beta}_0(u)^{\mathrm{T}}, \boldsymbol{\beta}_1(u)^{\mathrm{T}}, \boldsymbol{\beta}_2(u)^{\mathrm{T}}\Big)^{\mathrm{T}}, \qquad \hat{\boldsymbol{\eta}}(u) = \Big(\hat{\boldsymbol{\beta}}_0(u)^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_1(u)^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_2(u)^{\mathrm{T}}\Big)^{\mathrm{T}},\\
&\hat{\boldsymbol{\eta}}'(u) = \Big(\hat{\boldsymbol{\beta}}'_0(u)^{\mathrm{T}}, \hat{\boldsymbol{\beta}}'_1(u)^{\mathrm{T}}, \hat{\boldsymbol{\beta}}'_2(u)^{\mathrm{T}}\Big)^{\mathrm{T}}, \quad \boldsymbol{\xi}(u) = \Big(\boldsymbol{\eta}(u)^{\mathrm{T}}, \boldsymbol{\eta}'(u)^{\mathrm{T}}\Big)^{\mathrm{T}},\\
&\hat{\boldsymbol{\xi}}(u) = \Big(\hat{\boldsymbol{\eta}}(u)^{\mathrm{T}}, \hat{\boldsymbol{\eta}}'(u)^{\mathrm{T}}\Big)^{\mathrm{T}},
\end{aligned}$$

$$\mu_i = \int s^i K(s)ds, \quad \nu_i = \int s^i K^2(s)ds, \quad P(v,x,u) = P(t \geq v|X = x, U = u),$$

$f(u)$ be the density of U and, for $l,i,j = 1,2$,

$$\begin{aligned}
&\theta_{0l}(u) = (\boldsymbol{\beta}_0(u)^{\mathrm{T}}, \boldsymbol{\beta}_l(u)^{\mathrm{T}})^{\mathrm{T}}, \alpha^*_{l0}(v,u) = f(u)E\Big[P(v,X,u)g_l(\theta_{0l}(u)^{\mathrm{T}}X)|U = u\Big],\\
&\alpha^*_{lj}(v,u) = f(u)E\Big[P(v,X,u)g_l'(\theta_{0l}(u)^{\mathrm{T}}X)X_{0j}|U = u\Big],\\
&\rho^*_{lij}(v,u) = f(u)E\Big\{\frac{P(v,X,u)[g_l'(\theta_{0l}(u)^{\mathrm{T}}X)]^2}{g_l(\theta_{0l}(u)^{\mathrm{T}}X)X_{0i}X_{0j}^{\mathrm{T}}}\Big|U = u\Big\},\\
&\rho^*_{1ij}(u) = \int_0^\tau \rho^*_{1ij}(v,u)d\Lambda_y(v), \quad \rho^*_{2ij}(u) = \int_0^\tau \rho^*_{2ij}(v,u)d\Lambda_c(v),
\end{aligned}$$

$$\Gamma^*_{1ij}(u) = \rho^*_{1ij}(u) - \int_0^\tau \alpha^*_{1i}(v,u)\alpha^*_{1j}(v,u)^{\mathrm{T}}\alpha^*_{10}(v,u)^{-1}h_{0,y}(v)dv,$$
$$\Gamma^*_{2ij}(u) = \rho^*_{2ij}(u) - \int_0^\tau \alpha^*_{2i}(v,u)\alpha^*_{2j}(v,u)^{\mathrm{T}}\alpha^*_{20}(v,u)^{-1}h_{0,c}(v)dv,$$
$$\Delta(u) = \begin{pmatrix} \Gamma^*_{111}(u)+\Gamma^*_{211}(u) & \Gamma^*_{112}(u) & \Gamma^*_{212}(u) \\ \Gamma^*_{121}(u) & \Gamma^*_{122}(u) & \mathbf{0}_{(p-p_1)\times(p-p_1)} \\ \Gamma^*_{221}(u) & \mathbf{0}_{(p-p_1)\times(p-p_1)} & \Gamma^*_{222}(u) \end{pmatrix},$$

where we use τ to denote a fixed number, and $\Omega(u)$ is the same as $\Delta(u)$ but with $\Gamma^*_{lij}(u)$ replaced by $\rho^*_{lij}(u)$. $\mathbf{e}^{\mathrm{T}}_{2q\times q} = (I_q, \mathbf{0}_{q\times q})$.

Theorem. *Under the conditions in the Appendix* (*at* `http://www.stat.sinica.edu.tw/statistica`), *we have*

$$\sqrt{nh}\Big\{H(\hat{\boldsymbol{\xi}}(u)-\boldsymbol{\xi}(u)) - \frac{1}{2}h^2\mu_2\mathbf{e}_{2q\times q}\boldsymbol{\eta}''(u)\Big\} \xrightarrow{D} \mathbf{N}(\mathbf{0}_{2q\times 1}, \Sigma(u)),$$

where

$$\Sigma(u) = \begin{pmatrix} \Delta(u)^{-1}\nu_0 & \mathbf{0}_{q\times q} \\ \mathbf{0}_{q\times q} & \Omega(u)^{-1}\mu_2^{-2}\nu_2 \end{pmatrix}.$$

Corollary. *Under the conditions of the Theorem, we have*

$$\sqrt{nh}\Big(\hat{\boldsymbol{\eta}}(u)-\boldsymbol{\eta}(u) - \frac{1}{2}h^2\mu_2\boldsymbol{\eta}''(u)\Big) \xrightarrow{D} \mathbf{N}(\mathbf{0}_{q\times 1}, \Delta(u)^{-1}\nu_0).$$

From the Corollary, it is easy to see that the variance of the proposed estimator $\hat{\boldsymbol{\beta}}_0(u)$ of $\boldsymbol{\beta}_0(u)$ is

$$\Big\{\Gamma^*_{111}(u)+\Gamma^*_{211}(u)-\Gamma^*_{112}(u)\Gamma^*_{122}(u)^{-1}\Gamma^*_{121}(u)\Big\}^{-1}(nh)^{-1}(1+o(1)).$$

In Section 2, we have seen that $\hat{\boldsymbol{\beta}}_1(u)$ is the same as the maximum local partial likelihood estimator of $\boldsymbol{\beta}_1(u)$ based on the observed survival times. This implies that the variance of the maximum local partial likelihood estimator of $\boldsymbol{\beta}_1(u)$ is the same as the variance of $\hat{\boldsymbol{\beta}}_1(u)$,

$$\Big\{\Gamma^*_{122}(u)-\Gamma^*_{121}(u)\Gamma^*_{111}(u)^{-1}\Gamma^*_{112}(u)\Big\}^{-1}(nh)^{-1}(1+o(1)).$$

Because the positions of $\boldsymbol{\beta}_1(u)$ and $\boldsymbol{\beta}_0(u)$ in $h_y(t|U,X)$ in (1.1) are in equilibrium, the variance of the maximum local partial likelihood estimator of $\boldsymbol{\beta}_0(u)$ should be

$$\Big\{\Gamma^*_{111}(u)-\Gamma^*_{112}(u)\Gamma^*_{122}(u)^{-1}\Gamma^*_{121}(u)\Big\}^{-1}(nh)^{-1}(1+o(1)).$$

Similarly, the bias of the maximum local partial likelihood estimator of $\boldsymbol{\beta}_0(u)$ should be $2^{-1}h^2\mu_2\boldsymbol{\beta}_0''(u)$. Now, it is clear the proposed estimator of $\boldsymbol{\beta}_0(u)$ shares the same bias with the maximum local partial likelihood estimator of $\boldsymbol{\beta}_0(u)$, but has smaller variance by the amount

$$\Big[\Big\{\Gamma^*_{111}(u) - \Gamma^*_{112}(u)\Gamma^*_{122}(u)^{-1}\Gamma^*_{121}(u)\Big\}^{-1} \\ -\Big\{\Gamma^*_{111}(u) + \Gamma^*_{211}(u) - \Gamma^*_{112}(u)\Gamma^*_{122}(u)^{-1}\Gamma^*_{121}(u)\Big\}^{-1}\Big](nh)^{-1}(1+o(1)),$$

which is of the same order as the variance itself. So, the proposed estimator is more efficient and the improvement in efficiency is significant.

5. Simulation Study

In this section, we are going to use a simulated example to assess how well the proposed local complete likelihood based estimation method works. We compare the proposed estimator with the maximum local partial likelihood estimator, and demonstrate that the proposed estimator is indeed more efficient when sample size is finite.

Example. In (1.1), take $n = 1,000$, $p_1 = 1$, $p = 2$, $g_1(\cdot) = g_2(\cdot) = \exp(\cdot)$, $h_{0,c}(t) = h_{0,y}(t) = 4t^3$, $\boldsymbol{\beta}_0(u) = \sin(2\pi u)$, $\boldsymbol{\beta}_1(u) = 2\sin(\pi u)$, and $\boldsymbol{\beta}_2(u) = \cos(\pi u)$. The covariates (U_i, X_i), $i = 1, \ldots, n$, are generated from the uniform distribution.

For any unknown function $h(t)$, if $\hat{h}(t)$ is an estimator of $h(t)$, the mean integrated squared error (MISE) of $\hat{h}(t)$ is

$$\text{MISE} = E\Big\{\int \Big(h(t) - \hat{h}(t)\Big)^2 dt\Big\}.$$

We performed 100 simulations, the average censoring rate across the 100 simulations was 36%. In each simulation, the proposed local complete likelihood estimation and local partial likelihood estimation were used, respectively, to estimate the unknown functional coefficients. The kernel function in the estimation procedure was taken to be the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$. The accuracy of the obtained estimators was assessed by their MISEs.

To compare the proposed local complete likelihood based estimator with the maximum local partial likelihood estimator, for different bandwidths, we computed the MISEs of the estimators of $\boldsymbol{\beta}_i(\cdot)$, $i = 0, 1, 2$, obtained by either the proposed local complete likelihood estimation or local partial likelihood estimation. The results are presented in Figure 1. It can be seen that the MISE of the estimator of $\boldsymbol{\beta}_0(\cdot)$ obtained by the proposed local complete likelihood estimation

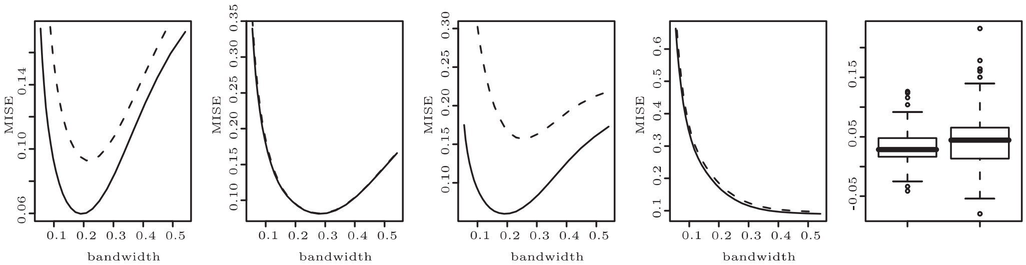

Figure 1. From left to right, one was the MISEs of the estimators of $\boldsymbol{\beta}_0(\cdot)$, of $\boldsymbol{\beta}_1(\cdot)$, of $\boldsymbol{\beta}_0(\cdot)$, and of $\boldsymbol{\beta}_2(\cdot)$. The solid lines are the MISEs of the estimators obtained by the proposed local complete likelihood estimation; the dashed lines in the first two figures are the MISEs of the maximum local partial likelihood estimators based on the observed survival times; the dashed lines in the last two figures are the MISEs of the maximum local partial likelihood estimators based on the observed censoring times. In the last figure, the boxplot of $cv - cv_1$ is on the right, that of $cv - cv_2$ is on the left.

is well below the MISE of the maximum local partial likelihood estimator of $\boldsymbol{\beta}_0(\cdot)$ for any bandwidth. For $\boldsymbol{\beta}_1(\cdot)$ and $\boldsymbol{\beta}_2(\cdot)$, there is no difference between the local complete likelihood estimation and local partial likelihood estimation.

To examine how well the proposed bandwidth selection serves the estimation of functional coefficients, we set the proposed local complete likelihood estimation with bandwidth selected by the proposed bandwidth selection to estimate $\boldsymbol{\beta}_0(\cdot)$ and $\boldsymbol{\beta}_1(\cdot)$ and computed the MISEs of the estimators. The resulting MISEs were 0.061 for $\boldsymbol{\beta}_0(\cdot)$, and 0.107 for $\boldsymbol{\beta}_1(\cdot)$, which are quite small. In fact, from the first two figures in Figure 1, we can see that 0.061 and 0.107 are almost the minimums of the MISEs of $\boldsymbol{\beta}_0(\cdot)$ and $\boldsymbol{\beta}_1(\cdot)$, respectively. This indicates the proposed bandwidth selection serves the estimation well.

Finally, to examine how well the proposed cross-validation criterion works, we set the sample size to be 300. We denote the CVs of the true model, where $X_{01} = x_1$, by cv, the model assuming censoring is noninformative by cv_1, and the misspecified model, the model mistakenly assuming $X_{01} = x_2$, by cv_2. We performed 100 simulations and computed $cv - cv_1$ and $cv - cv_2$ for each simulation. The boxplots of $cv - cv_1$ and $cv - cv_2$ are shown in Figure 1. From Figure 1, we can see the proposed cross-validation works well on model selection.

6. Analysis of First Birth Intervals in Bangladesh

Family planning and health programs in developing countries usually advocate delayed childbearing and increased birth spacing, with the aims of controlling fertility levels and improving maternal and child health. Indicators of the success of such programs include the timing of the first birth (the first birth interval) and the length of time between subsequent births. Of particular interest are changing impacts of important factors, over time, on the length of birth intervals.

We present an analysis of the first birth interval in Bangladesh, using data from a nationally representative survey of ever-married women of reproductive age.

The proposed informative varying-coefficient proportional hazard function allows us to explore changes in the impacts of several factors on women's first birth interval over time. Education level and region are commonly found to be associated with fertility behaviour, for example.

Bangladesh has seen a dramatic decrease in fertility. This fertility decline is generally attributed to a successful national family planning program (see, for example, Cleland, Phillips, Amin and Kamal (1994)). The decrease in fertility is commensurate with an increase in the age at first marriage; a nationally representative survey of women in 1996−97 (Mitra et al. (1997)) found that the median age at marriage was 13.3 years among respondents aged 45−49 at the time of survey, compared to 15.3 years for respondents aged 20−24. There has been a slower increase in the age at first birth; the median age at first birth was 16.9 years among women aged 45−59, and 18.4 years among the younger cohort. These trends in age at marriage and age at first birth imply that the length of the first birth interval (measured from the age at marriage, since pre-marital fertility is rare in Bangladesh) has become shorter over time, which suggests that the fertility decline is due to increased birth spacing rather than delayed childbearing. In this paper, we examine the trend in the duration of first birth intervals, as well as the impact of background characteristics.

The data for our study come from the Bangladesh Demographic and Health Survey (BDHS) of 1996−97 (Mitra et al. (1997)), a cross-sectional nationally representative survey of ever-married women aged between 10 and 49. The analysis is based on a sample of 8189 women.

The dependent variable, y_i, is the duration in months between marriage and the first birth for the ith woman. A small number of women (0.6% of the total sample size) reported a pre-marital birth, and they are excluded from the analysis. When a woman was asked for the date of her first marriage in the BDHS, the intention was to collect the age at which she started to live with her husband. However, it is likely that some older women reported the age at which they were formally married which, in Bangladesh, can take place at a very young age and some time before puberty (Mitra et al. (1997)). For this reason, we calculate the first birth interval assuming a minimum *effective* age at marriage of twelve years. The youngest age at first birth in the sample was twelve years and this is assumed to be the youngest age at which a woman can reasonably be at risk of giving birth. 11.53 % of women in the sample had not had a birth by the time of the survey and are therefore right-censored.

Table 1. The CVs of The Models

Model	$\{U, x_1\}$	$\{U, x_2\}$	$\{U, x_3\}$	$\{U, x_4\}$	$\{U, x_5\}$
CV	-7.5454	-7.5636	-7.5228	-7.5513	-7.5640

$\{U, x_i\}$ *stands for the model with the coefficient of* x_i *and the trend not appearing in the mechanism of censoring.*

We consider several covariates which are commonly found to be associated with fertility behaviour in Bangladesh. The selected categorical covariates include the woman's religion (Muslim or other) (x_1), type of region of residence (urban or rural) (x_2), the woman's level of education (categorised as none, primary +) (x_4), and the husband's level of education (primary - or secondary +) (x_5). The selected continuous covariates include the woman's age at first marriage in years (x_3), and year of marriage (U).

The proposed cross-validation criterion was used to assess whether the censoring is informative or not. Ideally we would compute the CVs for all models, but this appears computationally impossible as there are 64 possible models. So, we appealed to the backward elimination method to reduce the computational burden. We started with the full model, the one with coefficients of all variables appearing in the mechanism of censoring, and computed its CV. The CV of the full model was -7.850. The CVs of the models in the final step in the backward elimination procedure are presented in Table 1.

From Table 1, it is clear that the censoring is informative, and the impacts of all selected variables except the year of marriage and the woman's age at first marriage appear in the mechanism of censoring. The proposed informative varying-coefficient model, with $X_{01} = (x_1, x_2, x_4, x_5)^{\mathrm{T}}$ and $X_{02} = (1, x_3)^{\mathrm{T}}$, was used to fit the data. The proposed local complete likelihood estimation was employed to estimate the functional coefficients. The kernel function for the local linear modelling in the estimation procedure was again the Epanechnikov kernel, and the bandwidth was selected to be 10% of the range of year of marriage by the proposed cross-validation based bandwidth selection procedure. The results obtained are presented in Figure 2.

Figure 2 shows that the hazard of a first birth increases with year of marriage. Thus, the younger marriage cohorts have shorter birth intervals. This trend is expected since the increase in the age at marriage during this period has been accompanied by relatively little change in the age at first birth. Cultural norms regarding the age to childbearing have been slower to change than norms about the age at marriage. The figure also shows a clearly nonlinear dynamic pattern with a periodic feature.

From Figure 2, it can be seen that Muslim women have shorter first birth interval than other religions, however, the difference decreases sharply from 1960

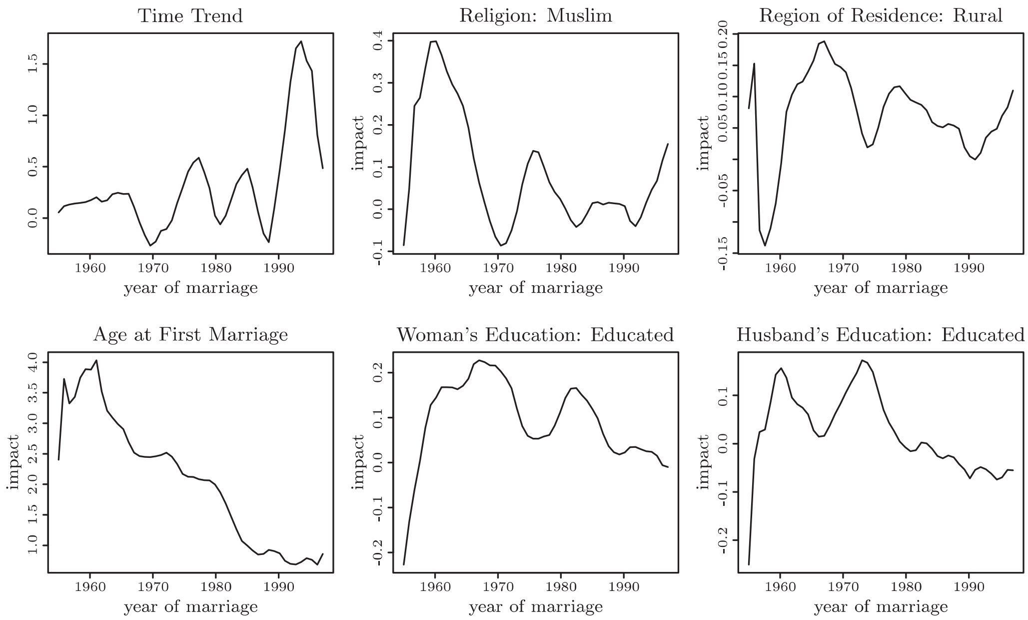

Figure 2. The solid lines are the impacts of the factors concerned. For example, the middle one on the upper panel is the impact of being Muslim.

to 1970, then it slightly increases, then decreases again. After 1980, the difference becomes very small. After 1995, the difference seems larger again. However, the post-1995 pattern should be interpreted with caution since there are relatively few women in the sample who married after 1995. Urban women have longer first birth intervals than rural women. Figure 2 also shows that the difference is varying with time.

From Figure 2, we can see the impact of age at first marriage on the hazard of a first birth is always positive. This can be interpreted as women who married at very young ages are likely to delay having their first child. We can also see the impact is decreasing with time. This pattern is as expected. The age at first marriage in Bangladesh is increasing and the age at first birth does not change very much, so the impact of age at first marriage is getting small.

From Figure 2 it can be seen that, compared to women with no education, the hazard of a first birth is higher (i.e. the duration of the first birth interval is lower) for women educated to primary level or beyond. The longer birth intervals among women with no education may be partly explained by the higher frequency with which these women report their age at formal marriage rather than their age at cohabitation. Calculating the duration to the first birth from an origin of age twelve for these women may have artificially inflated the length of their birth intervals. Figure 2 also shows that the impact of the husband's education on the length of the first birth interval is positive before 1980, and negative thereafter.

Acknowledgement

Yan Sun's research was supported by National Science Foundation of China (grant 10801093). The authors would like to thank an associate editor and the referees for helpful comments.

References

Anderson, P. K., Borgan, O, Gill. R. D. and Keiding, N. (1993). Statistical Models Based on Counting Process. Springer- Verlag, New York.

Anderson, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10**, 1100-1120.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.*, **80**, 580–619.

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying coefficient models. *J. Amer. Statist. Assoc.* **95**, 888-902.

Cheng, M.-Y., Hall, P. and Yang, Y.-J. (2007). Nonparametric inference under dependent truncation. ACTA Scientiarum Mathematicarum (Szeged).

Cleland, J., Phillips, J. F., Amin, S. and Kamal, G. M. (1994). *The Determinants of Reproductive Change in Bangladesh: Success in a Challenging Environment.* The World Bank, Washington D.C.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall, London.

Fan, J. and Gijbels, I. (1994). Censored regression: nonparametric techniques and their applications. *J. Amer. Statist. Assoc.* **89**, 560-570.

Fan, J., Gijbels, I. and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *Ann. Statist.* **25**, 1661-1690.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031-1057.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74-99.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians* (Edited by M. Sanz-Sole, J. Soria, J.L. Varona and J. Verdera), Vol. III, 595-622.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19**, 1-141.

Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hastie, T. J. and Tibshirani, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55**, 757-796.

Mitra, S. N., Al-Sabir, A., Cross, A. R. and Jamil, K. (1997). Bangladesh and Demographic Health Survey 1996-1997. Dhaka and Calverton, MD: National Institute of Population Research and Training (NIPORT), Mitra and Associates, and Macro International Inc.

Stone, C. J., Hansen, M., Kooperberg, C. and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist.* **25**, 1371-1470.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika* **64**, 29-35.

Wong, W. (1986). Theory of partial likelihood. *Ann. Statist.* **14**, 88-123.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space (with Discussion). *J. Roy. Statist. Soc. Ser. B* **64**, 363-410.

Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK.

E-mail: W.Zhang@bath.ac.uk

School of Economics, Shanghai University of Finance and Economics, 777 Guoding Rd., Shanghai, 200433, China.

E-mail: yansun2002cn@yahoo.com.cn

Department of Statistics and Applied probability, University of Singapore, Singapore 119077.

E-mail: zhangjt@stat.nus.edu.sg

Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WCIE 7HT, UK.

E-mail: Duolao.Wang@lshtm.ac.uk

(Received August 2007; accepted May 2008)