STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:3029–3042
Published online 18 August 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3686

A semi-parametric threshold regression analysis of sexually transmitted infections in adolescent women

Zhangsheng Yu[1,*,†], Wanzhu Tu[1,2] and Mei-Ling Ting Lee[3]

[1]*Division of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, U.S.A.*
[2]*Regenstrief Institute, Inc., Indianapolis, IN, U.S.A.*
[3]*Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD, U.S.A.*

SUMMARY

Time-to-event analysis of sexually transmitted infection data is often complicated by the existence of nonproportional hazards and nonlinear independent variable effects. Methods without the proportional hazards assumption, such as threshold regression models, have been successfully used in many applications. This paper seeks to extend the existing threshold regression models to accommodate the nonlinear independent variable effects. Specifically, we incorporated penalized and regression splines to the threshold regression models for added modeling flexibility. Cross validation methods were used for the selection of the number of knots and for the determination of smoothing parameters. Variance estimates were proposed for inference purposes. Simulation results showed that the proposed methods were able to achieve nonparametric function and parametric coefficient estimates that are close to their true values. Simulation also demonstrated satisfactory performance of variance estimates. Using the proposed methods, we analyzed time from sexual debut to the first infection with *Chlamydia trachomatis* infection in a group of young women. Analysis shows that the lifetime number of sexual partners has a nonlinear effect on the risk of *C. trachomatis* infection and the infection risks were differential by ethnicity and age of sexual debut. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: first hitting time model; penalized spline; regression spline; cross validation; survival analysis; nonproportional hazard

1. INTRODUCTION

Infections with *Chlamydia trachomatis*, *Neisseria gonorrhoeae*, and *Trichomonas vaginalis* are among the most prevalent sexually transmitted infections (STIs) in the United States [1, 2]. Delayed treatment of these infections often carries significant morbidity for young women, including pelvic inflammatory disease, ectopic pregnancy, tubal infertility, pre-term birth, and increased

*Correspondence to: Zhangsheng Yu, Division of Biostatistics, 410 W 10th St., Indianapolis, IN 46254, U.S.A.
†E-mail: yuz@iupui.edu

Contract/grant sponsor: National Institutes of Health; contract/grant numbers: RO1 HD042404, RO1 OH008649

Received 9 March 2009
Accepted 1 July 2009

susceptibility to human immunodeficiency virus (HIV) [3–5]. Although screening adolescent women for selected STIs is endorsed by clinical practice guidelines [6, 7], there are no evidence-based recommendations about the beginning age and frequency of screenings [8]. However, young women are at a risk of STI acquisition once they become sexually active, it is important to quantify the time between sexual debut and first STI and to identify behavioral markers that could be used to assess STI risk in the screening.

Methodologically, traditional time-to-event models with proportional hazards assumption such as those proposed by Cox represent a significant restriction in the analysis of adolescent STI data [9]. For example, adolescent women are rarely in stable monogamous relationships, and they usually see partners in their own age groups, it is conceptually difficult to assume the partner effect to be proportional over time since male partners of different ages typically represent different levels of STI risk. Additionally, the accommodation of potential nonlinear covariate effects requires an additive structure such as those proposed by Gray [10].

An alternative approach that has gained much popularity in recent years is the first-hitting-time (FHT)-based threshold regression models. Instead of the proportional hazards assumption used by the Cox regression model, FHT-based threshold regression stipulates that an event of interest occurs when a latent health status process passes a certain threshold [11–14]. Within this framework, several authors have explored threshold regression models with parametric link functions for covariates [15–17]. To enhance the modeling flexibility, this paper extends the existing threshold regression models to incorporate semi-parametric components for the accommodation of potentially nonlinear covariate effects [18].

The proposed semi-parametric threshold regression model is to a large extent motivated by the desire to better understand the effects of certain sexual behavioral markers, such as the number of partners on the timing of initial STI following sexual debut in adolescent women. In this application, potentially nonlinear covariate effects and nonproportional hazards have made it difficult to use the traditional Cox regression models.

Briefly, the general approach of the research is as follows: In the context of the application, we consider a young woman's STI risk as a Wiener process, where the time between sexual debut and first STI is modeled with an inverse Gaussian (IG) distribution [19]. Under this formulation, the drift parameter μ of the latent Wiener process is linked to a semi-parametric function of covariates. Events of STI occur when the latent infection risk process passes the threshold. In Section 2, we express the main parameter of interest, μ, as the sum of a parametric (i.e. linear) component and a nonparametric component with an unspecified functional form. For the nonparametric component, estimation procedures for regression spline and penalized spline are developed parallel to the published works on semi-parametric methods in the generalized linear and proportional hazard models [10, 20, 21]. In Section 2.4, we propose a variance estimate for inferences. The operating characteristics of the proposed model are then evaluated in Section 3 through a simulation study. Data analysis is presented in Section 4.

2. MODELS AND ESTIMATION PROCEDURE

We model censored survival times with an IG distribution with mean $1/\mu$ and variance ν/μ^3

$$f(y;\mu,\nu)=(2\pi y^3\nu)^{-1/2}\exp\left\{-\frac{(1-\mu y)^2}{2\nu y}\right\},\quad 0<y<\infty \tag{1}$$

DOI: 10.1002/sim

where μ, as the reciprocal of the mean survival time, is the failure rate or rate of infection acquisition in the current application, and v is commonly referred as the volatility parameter.

To link covariate effects with the failure rate μ, Whitmore proposed a fully parametric regression model $\mu = X^{\mathrm{T}}\gamma$, with regression coefficients $\gamma^{\mathrm{T}} = (\gamma_1, \ldots, \gamma_p)$ [15]. Here, we extend Whitmore's model by adding $\theta(Z)$, an unspecified smooth function of covariate Z, to the regression model $\mu = \theta(Z) + X^{\mathrm{T}}\gamma$. Under this formulation, the model is able to accommodate potentially nonlinear effect of covariate Z without specification of a pre-determined functional form, in addition to the linear effects of covariate vector X, which adds substantial flexibility to the modeling structure.

We use the following notation throughout the paper: We write the observation data as the ith subject as $(Y_i, Z_i, X_i, \delta_i)$, where Y_i denotes the event time subject to right censoring, Z_i is the covariate with possible nonlinear effects in an unspecified form, and X_i is the covariate vector with parametric coefficients. We write the indicator $\delta_i = 1$ if Y_i is an observed event time, and $\delta_i = 0$ if Y_i is a right censored observation.

For an observed event time $Y_i = y_i$, the log-likelihood function is

$$l_i(\theta, \gamma, v) = \log\{f(y_i; \theta, \gamma, v)\} = -\frac{1}{2}\log(2\pi v) - \frac{3}{2}\log(y_i) - \frac{1 - 2\{\theta(z_i) + x_i^{\mathrm{T}}\gamma\}y_i + \{\theta(z_i) + x_i^{\mathrm{T}}\gamma\}^2 y_i^2}{2v y_i}$$

where $f(y_i; \theta, \gamma, v)$ is the density function of an IG distribution with regression parameters θ, γ.

For a censored observation Y_i, the log-likelihood contribution is

$$l_i^c(\theta, \gamma, v) = \log\{P(Y_i \geqslant y_i)\} = \log\left\{\int_{y_i}^{\infty} f(s; \theta, \gamma, v)\,\mathrm{d}s\right\}$$

The combined log-likelihood function of the observed data can be written as

$$l(\theta, \gamma, v) = \sum_{i=1}^{n}\{\delta_i l_i + (1 - \delta_i) l_i^c\}$$

Model parameters are then estimated based on the estimating equations derived from this log-likelihood function. Under this general form, with some derivation similar to Whitmore [15], the estimating equations for parameters can be written as

$$S(p) = \sum_{i=1}^{n}\left\{\delta_i \frac{\partial \log f(y_i; \theta, \gamma, v)}{\partial p} + (1 - \delta_i)\int_{y_i}^{\infty} \frac{\partial \log f(s; \theta, \gamma, v)}{\partial p}\,\mathrm{d}s\right\} \tag{2}$$

where p represents either one of the θ, γ, v.

2.1. The regression spline approach

We first explore the modeling of the nonlinear covariate effect regression splines. Specifically, we model the nonparametric function $\theta(z)$ as a linear combination of a set of basis functions $B_j(z)$, $j = 1, 2, \ldots, m$, that is, $\theta(z) = \sum_{j=1}^{m} \beta_j B_j(z)$. Herein, we use cubic B-spline basis functions. Details on the generation of the cubic B-spline basis can be found in DeBoor [22]. The primary motivations of our adoption of cubic B-spline basis are computational efficiency and easy implementation, which include the generation of banded design matrices and the avoidance of extreme large values.

DOI: 10.1002/sim

Under the new parameterization of the regression splines, the log-likelihood function of event time Y_i can be written as

$$\log\{f(y_i;\beta,\gamma,v)\} = -\frac{1}{2}\log(2\pi v) - \frac{3}{2}\log(y_i) - \frac{1-2\sum_{j=1}^{m}\mu_{ij}(\beta,\gamma)y_i + \sum_{j=1}^{m}\mu_{ij}(\beta,\gamma)^2 y_i^2}{2vy_i} \tag{3}$$

where $\mu_{ij}(\beta,\gamma)=\beta_j B_j(z_i)+x_i^{\mathrm{T}}\gamma/m$. The log-likelihood function is then considered as the log-likelihood of data with covariates $\{B_1(z_i), B_2(z_i), \ldots, B_m(z_i), x_i^{\mathrm{T}}\}^{\mathrm{T}}$ and coefficients $(\beta_1, \beta_2, \ldots, \beta_m, \gamma^{\mathrm{T}})^{\mathrm{T}}$.

To derive the necessary estimating function, we let $F(\cdot)$ be the IG distribution function. We take advantage of the following properties of the conditional expectations [15]:

$$E[Y_i|Y_i\geqslant a] = \frac{F(1/(v^2a);\mu,v)}{\mu\{1-F(a;\mu,v)\}} \tag{4}$$

and

$$E[1/Y_i|Y_i\geqslant a] = v+\mu^2 E[Y_i|Y_i\geqslant a] - \frac{2avf(a;\mu,v)}{1-F(a;\mu,v)} \tag{5}$$

We then obtain the following estimating equations for β_j and γ by plugging (3)–(5) into (2):

$$\mathbf{0} = S(\beta,\gamma) = \frac{1}{v}\sum_{i=1}^{n}\{B_1(z_i),\ldots,B_m(z_i),x_i^{\mathrm{T}}\}^{\mathrm{T}} \times\left\{1-\sum_{j=1}^{m}\mu_{ij}(\beta,\gamma)M_{i,1}\right\} \tag{6}$$

where $M_{i,k}=\delta_i y_i^k+(1-\delta_i)E_*(y_i^k)$ for $k=1,-1$, and $E_*(y_i^k)=E[Y_i^k|Y_i\geqslant y_i,\beta,\gamma]$. Here, $E_*(y_i^k)$ for $k=1,-1$ can be calculated using (4) and (5).

The estimating equation of parameter v is

$$S(v) = -\frac{n}{2v}+\frac{1}{v^2}\sum_{i=1}^{n}\left[M_{i,-1}-2\sum_{j=1}^{m}\mu_{ij}(\beta,\gamma) + \sum_{j=1}^{m}\mu_{ij}(\beta,\gamma)^2 M_{i,1}\right]=0 \tag{7}$$

Therefore, the estimation procedure iterates the following two steps until convergence:

(a) On the basis of current estimates of β, γ, v, compute $E_*(y_i)$ and $E_*(1/y_i)$, for $i=1,2,\ldots,n$.
(b) Substitute the value in Step (a) into $S(\beta,\gamma)$ and $S(v)$ and solve for new values of β, γ, v.

After obtaining the estimates $\widehat{\beta}_j$, $j=1,\ldots,m$, the regression spline estimate can be constructed as $\widehat{\theta}(z)=\sum_{j=1}^{m}\widehat{\beta}_j B_j(z)$.

DOI: 10.1002/sim

A practical issue frequently encountered in conducting regression spline procedure is the selection of the number and locations of the knots. For a nonlinear function without complicated curvature, it usually suffices to use 3–7 knots to capture the basic shape of the curve. Too many knots may result in wiggly curves. In Section 2.3, we propose a cross validation score for the selection of the number of knots.

Given a fixed number of knots ($m-4$, for example), we suggest to place the knots at $100\times k/(m-3)$ per cent percentiles of observed values of covariate z to provide enough observations around each knot for estimation, where $k=1,\ldots,m-3$. Note that using $m-4$ interior knots will result in having m cubic B-spline basis functions. The regression spline differs from a parametric covariate function because the knot number (or parameters number) increases with sample size.

In the present research, we consider the simplest case of nonlinear effect for a single covariate Z. The proposed method, however, could be easily extended to more general models with multiple nonparametric additive functions, in which case, one set of spline basis can be generated for each nonparametric component. The estimation equations (6) can be adjusted accordingly to allow multiple splines.

2.2. The penalized spline approach

Regression spline approach usually works well when the underlying function has less curvature with a small number of knots. The stability of the spline estimates may become an issue when a large number of knots are involved [10]. Thus, the method may not suit situations where nonlinear functions have more complicated shapes. An alternative method is to use penalized splines, which penalizes the roughness of the nonparametric function estimates. In this section we describe a cubic spline approach with penalty on the second derivative of the nonparametric function.

Following the notation introduced in Section 2.1, we write the penalized log-likelihood as follows:

$$pl(\theta,\gamma,v)=\sum_{i=1}^{n}\{\delta_i l_i+(1-\delta_i)l_i^c\}-\frac{\lambda}{2}\int[\theta^{(2)}(z)]^2\,\mathrm{d}z \tag{8}$$

where λ is the smoothing parameter controlling the degree of smoothness of nonparametric function estimates. The penalty term is written as $(\lambda/2)\int[\theta^{(2)}(z)]^2\,\mathrm{d}z=(\lambda/2)\boldsymbol{\beta}^{\mathrm{T}}\int\mathbf{B}^{(2)}(z)\mathbf{B}^{(2)}(z)^{\mathrm{T}}\,\mathrm{d}z\boldsymbol{\beta}=(\lambda/2)\beta^{\mathrm{T}}\mathbf{P}\boldsymbol{\beta}$, where $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_m)^{\mathrm{T}}$ and $\mathbf{B}^{(2)}(z)=\{B_1^{(2)}(z),\ldots,B_m^{(2)}(z)\}^{\mathrm{T}}$. Note that the penalty matrix $\mathbf{P}$ is known and can be constructed using the fda package in R or Matlab in a straightforward fashion. We include Matlab code of generating cubic B-spline basis and the penalty matrix in Appendix A.

The estimating equations for $\boldsymbol{\beta}$, γ are

$$S_p(\beta,\gamma)=\frac{1}{v}\sum_{i=1}^{n}\{B_1(z_i),\ldots,B_m(z_i),x_i^{\mathrm{T}}\}^{\mathrm{T}}\times\left\{1-\sum_{j=1}^{m}\mu_{ij}(\beta,\gamma)M_{i,1}\right\}-(\lambda\boldsymbol{\beta}^{\mathrm{T}}\mathbf{P},\mathbf{0})^{\mathrm{T}}=\mathbf{0} \tag{9}$$

where $\mathbf{0}$ is a zero vector with p components. Estimating equations for v is in the same form as (7). For a given λ, the estimating procedure will be the same as the iterative algorithm for regression spline approach except that $S_p(\beta,\gamma)$ will be used.

DOI: 10.1002/sim

Bias occurs when the penalty term is introduced into the log-likelihood. The degree of bias depends on the λ. We will evaluate the bias using simulation in next section and compare it with the regression spline estimates numerically. The Matlab code for fitting the penalized spline is available upon request. Regression spline-based estimators can be obtain using the same code by setting the penalty term to zero.

2.3. Selecting the smoothing parameter and number of knots

An appealing idea is to have a data-driven method for the knot number determination. Substituting the estimated parameters into $M_{i,1}$, the estimating function (6) can be viewed as a derivative of a weighted least-square function with weight $\widehat{M}_{i,1}$ and response variable $1/\widehat{M}_{i,1}$. Along this vein, one could construct a cross validation score for knot selection

$$\mathrm{CV}(m)=\sum_{i=1}^{n}\widehat{M}_{i,1}\left[\frac{1}{\widehat{M}_{i,1}}-\sum_{j=1}^{m}\widehat{\beta}_{j(-i)}B_j(z_i)-x_i^{\mathrm{T}}\widehat{\gamma}_{(-i)}\right]^2 \tag{10}$$

where $\widehat{\beta}_{j(-i)},\widehat{\gamma}_{(-i)}$ are the estimates when the ith observation is excluded. Calculating $\mathrm{CV}(m)$ using different values of m, one will select the m^* that minimizes $\mathrm{CV}(m)$. After selecting m^*, (m^*-4) knots will be used for regression spline. The computational load associated with the leaving-one-observation-out cross validation approach is relatively large. Similarly, one could consider the alternative approach of leaving-N-observation-out cross validation to reduce the computational load.

The above CV function can be used on penalized spline analysis. Specifically, the cross validation scores for different values of λ will be calculated and the one minimizing CV will be selected as the smoothing parameter for estimation. The performance of the smoothing parameter selection is assessed in the simulation study.

2.4. Inferences

To make inferences on regression spline-based estimates of the covariate functions and coefficients, we use the negative of the inverse Hessian matrix of the log-likelihood for variance estimate $(\widehat{\beta}_1,\widehat{\beta}_2,\ldots,\widehat{\beta}_p,\widehat{\gamma}^{\mathrm{T}})^{\mathrm{T}}$.

First, the derivative of the estimating equation (6) is calculated as

$$-I(\beta,\gamma)=-\frac{1}{v^2}\sum_{i=1}^{n}\left[v(E_*(y_i)+y_i)+(1-\delta_i)\{E_*[y_i^2]-E_*[y_i]^2\}\left(\sum_{j=1}^{m}\beta_j B_j(z_i)+x_i^{\mathrm{T}}\gamma\right)^2\right]$$
$$\times\{B_1(z_i),\ldots,B_m(z_i),x_i\}^{\otimes 2}$$

It then follows that the covariance estimate is given by $V(\beta,\gamma)=I(\beta,\gamma)^{-1}$. We write the variance estimate for regression spline estimate $\widehat{\theta}(z)$ as

$$\{B_1(z),\ldots,B_m(z)\}V_{\beta,\beta}(\beta,\gamma)\{B_1(z),\ldots,B_m(z)\}^{\mathrm{T}}$$

where $V_{\beta,\beta}(\beta,\gamma)$ is the $m\times m$ submatrix of $V(\beta,\gamma)$ corresponding to the β's.

DOI: 10.1002/sim

For the volatility parameter v, we propose to use the inverse Hessian matrix of the corresponding log-likelihood, which is equal to the derivative of (7)

$$-I(v)=\frac{n}{2v^2}-\frac{1}{4v^3}\sum_{i=1}^{n}\{\delta_i\Omega_i+(1-\delta_i)E_*[\Omega_i]\}$$
$$+\frac{1}{4v^4}\sum_{i=1}^{n}(1-\delta_i)\{E_*[\Omega_i^2]-E_*[\Omega_i]^2\}$$

where

$$\Omega_i=\frac{1-2\{\widehat{\theta}(z_i)+x_i^{\mathrm{T}}\widehat{\gamma}\}y_i+\{\widehat{\theta}(z_i)+x_i^{\mathrm{T}}\widehat{\gamma}\}^2y_i^2}{y_i}$$

Hence, $I(v)^{-1}$ is the variance estimate of v. Note that $I(\beta,\gamma)$, $I(v)$ are submatrices of the information matrix of $I(\beta,\gamma,v)$, the negative of second derivative matrix of log-likelihood, which is difficult to obtain. With censoring in our current setting, using the inverse of submatrix $I(\beta,\gamma)$, $I(v)$ may lead to a biased estimate of the variance of proposed estimates. We will use simulation to evaluate the magnitude of the bias.

The variance calculation of penalized spline-based estimates parallels that of the regression splines. The only difference is the calculation of information matrix, which will be

$$I_p(\boldsymbol{\beta},\gamma)=I(\boldsymbol{\beta},\gamma)+\begin{pmatrix}\lambda\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}$$

where the second term in the right-hand side is a $(m+p)\times(m+p)$ dimensional matrix.

3. SIMULATION

In this section we first evaluate the penalized spline-based estimates by simulation. We also compare the performance of penalized spline-based and regression spline-based estimates. Data are generated using the IG density function as described in model (1). The regression function is specified as

$$\mu=\theta(z)+x_1\gamma_1+x_2\gamma_2$$

Two functional forms of $\theta(z)$ with different curvatures will be considered.

We first consider a simple unimodal function $\theta_1(z)=(4/\sqrt{\pi})\exp(-16(2z-1)^2/2)+1$. Herein, covariate z is generated as a uniform random variable over $[0,1]$; covariate x_1 is generated as $x_1=z/3+\mathrm{Uniform}(0,1)\times\frac{2}{3}$, which has a coefficient of correlation about 0.45 with a covariate z. Covariate x_2 is generated as a binary variable taking values 0 and 1 with equal probabilities. We use $\gamma_1=\gamma_2=0.5$ and the true volatility parameter $v=2$ in the simulation. Censoring times are generated from the exponential distribution with rate 4, independent of event times. The maximum follow-up time is set to be 4. In each data set there are about 20 per cent censored observations.

A total of 400 data sets were generated with 600 observations in each set. We fit semi-parametric threshold regression models using both regression spline and penalized spline approaches. For regression spline, cross validation score (10) is used to choose the number of knots. For penalized

DOI: 10.1002/sim

spline, cross validation score (10) is used to choose the number of knots and smoothing parameter simultaneously. Owing to the computation intensity, we only run simulation using seven different values of smoothing parameter, $\lambda \in \{0, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1, 1\}$. The cross validation score shows a parabola shape as a function of knots number and smoothing parameter. Unique minimum locations are obtained for almost all replicates. For the few replicates with multiple local minimums, the global minimizers were chosen for simplicity. For both methods, the knots were placed at $100 \times k/(m-3)$ per cent ($k = 1, 2, \ldots, m-4$) percentiles of observed values of z, where m is the knots number.

Simulation results are shown in panels in the left of Figure 1. Figure 1(a) compares the mean of penalized spline estimates of covariate function $\theta_1(z)$ with the true covariate function. The dotted line (representing the mean of estimates) is close to the solid line (true values) except at places where the curvature is abundant. Figure 1(b) evaluates the proposed variance estimate by comparing the mean of estimated standard error and empirical standard error of $\widehat{\theta}_1(z)$. The mean of the estimated standard errors is very close to the empirical one. Figure 1(c) shows the empirical coverage probabilities of 95 per cent confidence intervals constructed by using the estimated standard error. The coverage probabilities are generally around the nominal value 95 per cent and the mean of probabilities is 97.5 per cent. On the basis of these numerical results, we conclude that the penalized spline estimate for the nonlinear function and corresponding variance estimates work very well. Similar patterns are observed (plots not provided) for the estimators using regression spline. However, the average coverage probability of the regression spline estimate is about 89.0 per cent only. The mean integrated square error (MISE) of regression spline estimate is 0.130, which is 30 per cent larger than that of penalized spline, 0.101. The MISE is define as

$$\text{MISE} = E \int [\widehat{\theta}(z) - \theta(z)]^2 \, \mathrm{d}z$$

The true function θ_1 is known in the simulation. We approximate this integral by summing over 100 equally spaced points in the range of z. The MISE presented is the average of integrated squared errors over the 400 replicates.

The performance of the parametric regression coefficient estimates $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ is also of interest. Table I summarizes the simulation results using both spline approaches. For the penalized spline approach, over 400 replications, the means are $\hat{\gamma}_1 = 0.491$ and $\hat{\gamma}_2 = 0.504$, which are very close to the true values $\gamma_1 = \gamma_2 = 0.5$. The mean of the estimated standard error is 0.311 and 0.180, which is very close to the empirical standard errors 0.307 and 0.177, respectively. The empirical coverage probabilities are 0.952 and 0.947. Therefore, our simulation results show satisfactory performance of the penalized spline approach with respect to regression coefficients and corresponding variance estimation. The mean of the estimate for v over 400 replications is 1.99, which is very close to the true value 2. The mean of the corresponding variance estimate is 0.118, which is smaller than the empirical standard error 0.127. As we noted before, the use of the inverse of a submatrix of the total information matrix may underestimate the variance. We believe that this effect explains why the variance of $\widehat{v}$ was underestimated. Variance of the volatility parameter, however, is not the focus of our interest for this paper. Hence, we will not pursue this issue here. Similar performance of parametric estimate using regression spline approach can be seen in the first row of Table I.

We also evaluate the performance of two spline approaches when the true nonparametric function is more complicated. We use a bimodal function $\theta_2(z) = (6 \times \text{beta}(z, 30, 17) + 4 \times \text{beta}(z, 3, 11))/10$, where beta is the beta density function. The true function can be seen in the right column of Figure 1. Since there are more curvatures, the bias is larger at the second peak; the estimated

DOI: 10.1002/sim

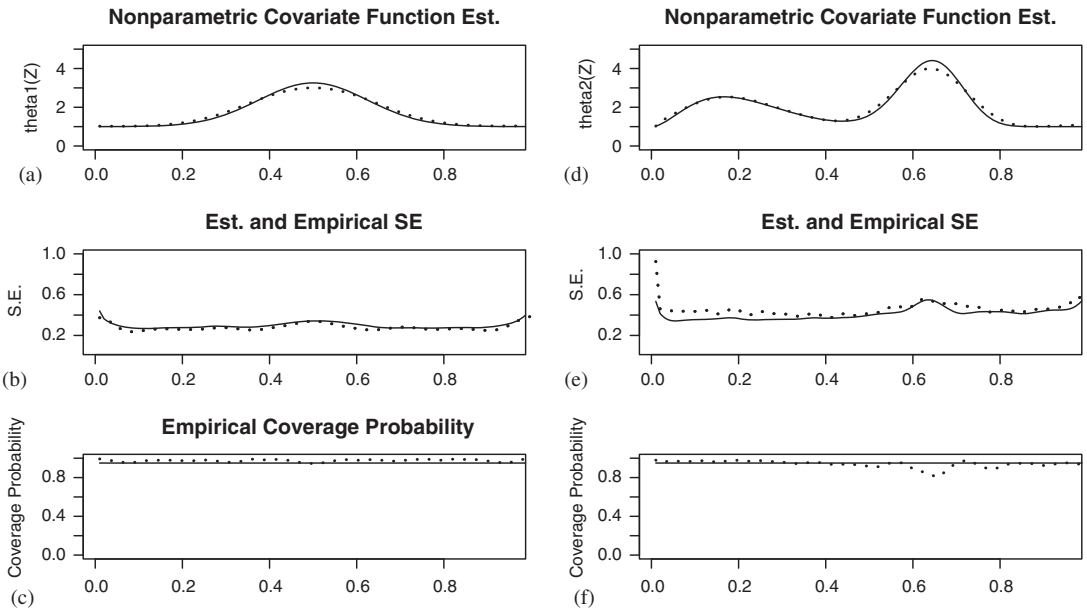

Figure 1. Simulation results for the nonparametric covariate function $\theta_1(z)$ and $\theta_2(z)$: (a) penalized spline estimate of $\widehat{\theta}_1(z)$, dotted, true $\theta_1(z)$, solid; (b) standard errors: estimated, dotted, empirical, solid; (c) empirical coverage probability: the mean coverage probability is 0.975; (d) penalized spline estimate of $\widehat{\theta}_2(z)$, dotted, true $\theta_2(z)$, solid; (e) standard errors of $\theta_2(z)$: estimated, dotted, empirical, solid; and (f) empirical coverage probability of $\theta_2(z)$: the mean coverage probability is 0.940.

Table I. Simulation: parametric coefficient estimates.

			Estimator		Empirical SE		Estimated SE		95 per cent C.I. coverage	
	Parameter	True value	Reg.*	Pen.†	Reg.*	Pen.†	Reg.*	Pen.†	Reg.*	Pen.†
$\theta_1(z)$	γ_1	0.5	0.491	0.491	0.308	0.307	0.308	0.311	95.5	95.2
	γ_2	0.5	0.504	0.504	0.177	0.177	0.177	0.180	94.2	94.7
	v	2	1.98	1.99	0.127	0.127	0.116	0.118	NA	NA
$\theta_2(z)$	γ_1	0.5	0.497	0.499	0.508	0.501	0.490	0.495	93.0	94.7
	γ_2	0.5	0.492	0.494	0.188	0.186	0.180	0.191	94.0	94.5
	v	2	1.98	2.00	0.127	0.125	0.111	0.113	NA	NA

*Results from regression spline-based method.
†Results from penalized spline-based method.

standard errors are smaller than the empirical one too, hence the lower coverage probability at the second peak. This is not surprising due to the smoothing effect. A similar phenomenon can be seen in the literature such as Lin and Zhang [20]. The parametric estimates of penalized spline approach work well in terms of small biases and close to nominal level of coverage probabilities of the confidence intervals. Furthermore, we also found that the cross validation method in penalized

DOI: 10.1002/sim

spline method selects a larger number of knots than regression spline in average. The biases $\widehat{\gamma}_1$, $\widehat{\gamma}_2$ using the penalized spline approach are generally smaller than those of the regression spline approach. The empirical standard error estimates of penalized spline approach are smaller too. The mean interpreted squared error of $\widehat{\theta}_2(z)$ using regression spline is 0.267, which is 40 per cent larger than that of the penalized spline estimate 0.192. These results show that penalized spline approach performs better than regression spline approach in the chosen nonlinear functions. This advantage is clear for a more complicated nonparametric function.

An easier way of fitting a nonlinear curve is to use the polynomial regression. We perform the estimation of $\theta(z)$, γ_1, γ_2 assuming that $\theta(z)$ is a polynomial function. The biases for $\widehat{\gamma}_1$, $\widehat{\gamma}_2$ are substantially larger than the regression spline-based estimates for γ's. The numerical results are omitted due to space limitation. In conclusion, both spline approaches can better identify the functional shape of an unknown nonparametric covaraite function than polynomial functions.

Data sets with different proportions of censored observations were also generated in our simulations. The results show patterns similar to that of Figure 1 and Table I. As expected, the larger the censoring probability, the more biased the estimates are.

4. APPLICATION

In this section we apply the proposed penalized spline threshold regression method to the analysis of epidemiological data generated by a longitudinal study of STI in urban adolescent women. Briefly, 387 women between ages 11 and 17 were enrolled from three adolescent medicine clinics. Upon enrollment, participants were interviewed in person and were tested for infections with *C. trachomatis*, *N. gonorrhoeae* and *T. vaginalis*. Enrolled subjects had quarterly follow-up visits, at which time they were interviewed for sexual behavioral information in the previous quarter. Subjects were retested at their quarterly visits for STI. Infected individuals were treated. To supplement the study data, we reviewed subjects' medical records using an electronic medical record system to identify all STI diagnoses outside of the study venues. Age of sexual debut was ascertained in the enrollment interview. In this analysis, we focused on the effects of age of sexual debut (dichotomized at age 14, the median age of sexual debut in the study cohort), number of unprotected sexual intercourse during the last three months, and life time number of sexual partners reported at enrollment. Because screening recommendations were largely organism specific, we focused on time to the infections with *C. trachomatis* in the current analysis. Among 387 women, about 26.6 per cent are censored by the end of follow up.

Validity of proportional hazards assumption for Cox regression was examined using testing procedures developed by Grambsch and Therneau [23]. The global test of proportional hazards assumption was rejected with a p-value of 0.0134. For specific variables, proportional hazards test for dichotomized age of first intercourse (>14) yielded a p-value of 0.0015, which again led to the rejection of proportional hazard assumption. As discussed in the introduction, we suspected a nonlinear effect for the number of sexual partners because of the potential prophylactic behavioral change in the young women and the differential infection risk of male partners in larger sexual networks. To accommodate such a nonlinear effect, we introduced a nonparametric effect for the lifetime number of sexual partners. Parametric effects were used for the age of first intercourse (>14), ethnicity (African American), and number of unprotected sex events in the past three months (>10).

DOI: 10.1002/sim

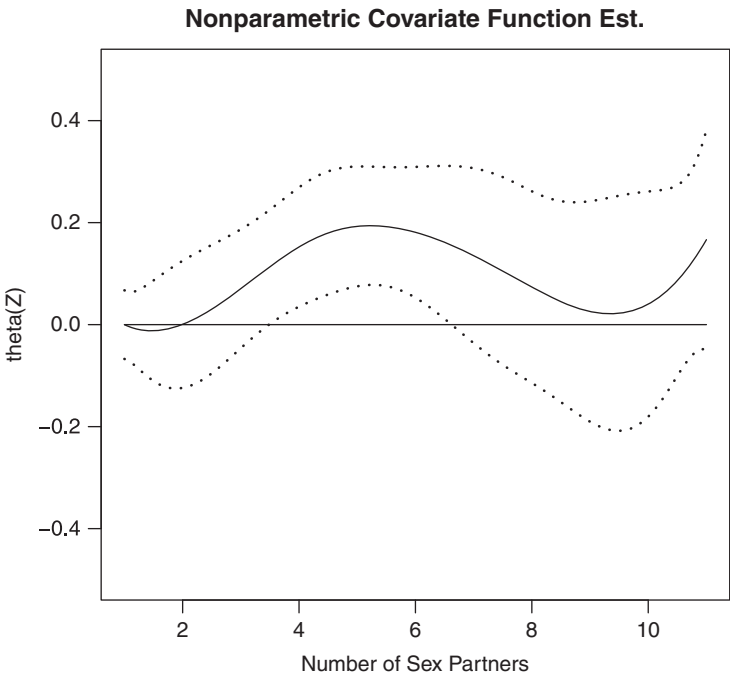

Figure 2. Application: θ (number of sex partners) penalized spline estimator $\widehat{\theta}(z)$, solid, 95 per cent confidence band, dotted.

As described in Section 2.2, the model with three knots and the 0.1 as the smoothing parameter was chosen by minimizing the cross validation score. As demonstrated in Figure 2, infection risk was highest for subjects with four to six partners, which was the range that partner effect on *C. trachomatis* infection became significant. This intriguing observation, though previously not reported in the literature, is perhaps not entirely surprising. For example, one could speculate that adolescent women with relatively fewer (less than four) partners had lower risk because they are usually younger and are seeing younger male partners, who are unlikely sources of infection pathogens, while women who had a larger number of sexual partners (more than seven) are likely to be more mature and more cognizant of the STI risk, and these women may be more cautious in partner selections. Similarly, as demonstrated in Table II, for the number of exposures, we noted that the *C. trachomatis* infection risk for subjects with larger numbers of unprotected sex (more than 10 sex events in the past three months) was higher than those with fewer unprotected sex events although the effect did not reach a level of statistical significance (p-value $= 0.173$). Black adolescents tended to have a higher (0.0902) STI rate than white teens (p-value $= 0.029$). Teens having later sexual debut (after age 14) had slightly higher infection rates than others (p-value $= 0.035$). Again, this seemingly paradoxical observation may reflect the differential STI risk level presented by the male partners. Although the current study is unable to fully disseminate the reasons behind the observations due to the lack of male partner data, the findings nonetheless highlight the complexity of the STI risk in adolescents and the need for a more careful examination of the behavioral markers for STI screening.

DOI: 10.1002/sim

Table II. Parametric coefficient estimates in STI study analysis.

Risk factors	Estimate	p-value
Number of unprotected sex >10	0.0519	0.173
Race (Black)	0.0902	0.029
Age at first sex >14	0.1105	0.035

5. DISCUSSION

Previous studies have demonstrated the values of threshold regression models in applications where proportional hazards assumptions are not justified [14, 15]. This research has further extended the flexibility of current threshold regression methods by incorporating into these models semi-parametric components for the accommodation of potentially nonlinear independent effects. Our research demonstrated the feasibility of adding regression or penalized splines in threshold regression settings. For estimation, we proposed a specific cross validation score for the selection of the number of knots, and for the determination of smoothing parameters. For inference, we proposed a variance estimate. Simulation showed that the proposed method was able to achieve estimates of nonparametric functions and parametric covariate effects that are close to the true values, and confidence intervals based on the variance estimate has a coverage probability close to its nominal level. Penalized spline-based estimates are shown to be superior to the regression spline, especially when the true nonparametric function is complicate. In general, our experience is that the proposed methods are easy to implement. The simulation and the data analysis reported in the current paper were conducted using Matlab, but similar operations could be carried out in other computing platforms. Further extension of the proposed method is possible. For example, a nonWiener process for the modeling of the latent infection risk would further increase the flexibility of the model.

From an application perspective, adding semi-parametric spline components to threshold regression models is not only a logical extension of the threshold regression methods but also represents an expansion of the utility of the methods in epidemiological applications. As we have demonstrated through the analysis of STI data, the proposed models provide a real solution to time-to-event analysis when data exhibit nonproportional hazards and potential nonlinear independent variable effect. Failure to accommodate these features could result in model misspecification and questionable findings. In this sense, the proposed models provide a safeguard against misspecified models in time-to-event analysis. Epidemiologically, the observed nonlinear sexual partner effect indicates that the risk of STI is not a simple function of the number of partners, but also the characteristics of the partners. This finding, in a sense, raises questions about the wisdom of using the number of partners as a risk marker for screening purposes.

APPENDIX A: MATLAB CODE FOR GENERATING CUBIC B-SPLINE BASIS AND THE PENALTY MATRIX*

```
% basis and penalty matrix
% Use the full data set to create the knots set
basis_range = [min(covariate) max(covariate)];
```


DOI: 10.1002/sim

```
breaks = linspace(basis_range(1),basis_range(2),nknot+2);
norder = 4;
nbasis = length(breaks)+ norder −2;
bspline_obj = create_bspline_basis(basis_range,nbasis,norder,breaks);
pen = bsplinepen(bspline_obj);
basisvals = eval_basis(covariate,bspline_obj);
% *: to use this code, one has to install fda package for matlab available
at Prof. Ramsay's website:
% 'http://www.psych.mcgill.ca/misc/fda/software.html'.
```

ACKNOWLEDGEMENTS

The project is supported in part by National Institutes of Health grants RO1 HD042404 and RO1 OH008649.

REFERENCES

1. Centers for Disease Control and Prevention. *Trends in Reportable Sexually Transmitted Diseases in the United States*, U.S. Department of Health and Human Services, CDC, Atlanta, GA, 2007.
2. Weinstock H, Berman S, Cates Jr W. Sexually transmitted diseases among American youth: incidence and prevalence estimates, 2000. *Perspectives on Sexual and Reproductive Health* 2004; **36**(1):6–10.
3. Hillis SD, Owens LM, Marchbanks PA, Amsterdam LF, Mac Kenzie WR. Recurrent chlamydial infections increase the risks of hospitalization for ectopic pregnancy and pelvic inflammatory disease. *American Journal of Obstetrics and Gynecology* 1997; **176**(1 Pt 1):103–107.
4. Paavonen J, Westrom L, Eschenbach D. Pelvic inflammatory disease. In *Sexually Transmitted Diseases* (4th edn), Holmes KKSP, Stamm WE, Piot P *et al.* (eds). McGraw-Hill: New York, NY, 2008; 1017–1050.
5. Van Der Pol B, Kwok C, Pierre-Louis B, Rinaldi A, Salata RA, Chen PL, van de Wijgert J, Mmiro F, Mugerwa R, Chipato T, Morrison CS. Trichomonas vaginalis infection and human immunodeficiency virus acquisition in African women. *Journal of Infectious Diseases* 2008; **197**(4):548–554.
6. Centers for disease control and prevention. Department of health and human services. *Sexually Transmitted Diseases Treatment Guidelines* 2006; **55**(No. RR-11):1–93. MMWR.
7. U.S. Preventive Services Task Force. Screening for chlamydial infection: U.S. preventive services task force recommendation statement. *Annals of Internal Medicine* 2007; **147**(2):128–134.
8. Meyers D, Wolff T, Gregory K, Marion L, Moyer V, Nelson H, Petitti D, Sawaya GF. USPSTF recommendations for STI screening. *American Family Physician* 2008; **77**(6):819–824.
9. Cox DR. Regression models and lifetime table (with Discussion). *Journal of Royal Statistics Society*, *Series B* 1972; **34**:187–220.
10. Gray RJ. Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of American Statistical Association* 1994; **87**:942–951.
11. Aalen OO, Borgan O, Gjessing HK. *Survival and Event History Analysis*: *A Process Point of View* (*Statistics for Biology and Health*). Springer: Berlin, 2008.
12. Aalen OO, Gjessing HK. Understanding the shape of the hazard rate: a process point of view. *Statistics Sciences* 2001; **16**:1–22.
13. Lawless JF. *Statistical Models and Methods for Lifetime Data* (2nd edn). Wiley: New York, 2003.
14. Lee MT, Whitmore GA. Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistics Sciences* 2006; **21**:501–513.
15. Whitmore GA. A regression method for censored inverse-Gaussian data. *Canadian Journal of Statistics* 1983; **11**:305–315.
16. Whitmore GA, Crowder MJ, Lawless JF. Failure inference from a marker process based on bivariate Wiener model. *Lifetime Data Analysis* 1998; **4**:229–251.
17. Lee MT, DeGruttola V, Schoenfeld D. A model for markers and latent health status. *Journal of Royal Statistical Society*, *Series B* 2000; **62**:747–762.

DOI: 10.1002/sim

18. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press: Cambridge, 2003.
19. Chhicara R, Folks JL. *The Inverse Gaussian Distribution Theory*, *Methodology*, *and Applications*. CRC Press: Boca Raton, 1989.
20. Lin X, Zhang D. Inference in generalized additive mixed model by using smoothing splines. *Journal of Royal Statistical Society*, *Series B* 1999; **61**:381–400.
21. Hastie T, Tibshirani R. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* 1990; **46**:1005–1016.
22. Deboor C. *A Practical Guide to Splines*. Springer, GmbH and Co.: Berlin, Heidelberg, 1978.
23. Grambsch P, Therneau T. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; **81**:515–526.

DOI: 10.1002/sim