Statistica Neerlandica (2011) Vol. 65, nr. 2, pp. 164–182
doi:10.1111/j.1467-9574.2011.00481.x

Analysis of failure time using threshold regression with semi-parametric varying coefficients

Jialiang Li*

Department of Statistics and Applied Probability, Duke-NUS Graduate Medical School, National University of Singapore

Mei-Ling Ting Lee

Department of Epidemiology and Biostatistics, Biostatistics and Risk Assessment Center, School of Public Health, University of Maryland

Many new statistical models may enjoy better interpretability and numerical stability than traditional models in survival data analysis. Specifically, the threshold regression (TR) technique based on the inverse Gaussian distribution is a useful alternative to the Cox proportional hazards model to analyse lifetime data. In this article we consider a semi-parametric modelling approach for TR and contribute implementational and theoretical details for model fitting and statistical inferences. Extensive simulations are carried out to examine the finite sample performance of the parametric and non-parametric estimates. A real example is analysed to illustrate our methods, along with a careful diagnosis of model assumptions.

Keywords and Phrases: threshold regression, Wiener process, varying coefficients model, inverse Gaussian distribution, bootstrap.

1 Introduction

The analysis of failure time data is a rich area for statistical researchers. One crucial ingredient of a complete modern survival analysis is to link the individual's survival experience with all sorts of prognostic factors, such as treatment assignment, risk factors, personal characteristics and other relevant covariates (chapter 10, Lee, 1992). Many regression models have been proposed to provide insights into the effects of covariates on failure time. Among them the most widely used tool is the Cox proportional hazards model (Cox, 1972). The model offers a convenient interpretation for regression coefficients in terms of changes in log hazards. The assumption of proportional hazards may appear plausible for many previous studies. When this assumption is violated, investigators have proposed alternative model choices

*stalj@nus.edu.sg

and generalizations to relax the assumption. For example, see chapter 7 of Hosmer and Lemeshow (1999).

We consider a novel modelling framework, threshold regression (TR), in this article. The TR method has at its root a stochastic process as a model of an underlying process for health status. The occurrence of the failure event may then be characterized as a first-hitting-time of the process. The model results in an explicit probability distribution for the failure time. The covariates may then be incorporated in the distribution parameters. See Lee and Whitmore (2006) for a detailed review of this method. Unlike the proportional hazards model which usually focuses on hazard ratios and does not offer many insights into underlying determinants of survival, the TR method does not require the proportional hazards assumption and may yield more meaningful analysis for degradation or deterioration of a subject's health.

The implementation of TR is straightforward since one can make use of the wealth of probabilistic properties of the inverse Gaussian distribution. Early application of the method can be found in Whitmore (1983). The idea was further extended in Whitmore, Crowder and Lawless (1998) and Lee, DeGruttola and Schoenfeld (2000) where both authors argued the progression of latent health status directly influences the lifetime distribution. The TR method has attracted abundant attention lately and showed its broad applicability to practical problems. Specifically, Lee, Whitmore and Rosner (2010) considered incorporating time-varying covariates in TR. Pennell, Whitmore and Lee (2009) used Markov chain Monte Carlo to fit a TR model with random effects to address data with non-proportional hazards. To improve the model accuracy, Yu, Tu and Lee (2009) considered additive semi-parametric covariate effects in TR where non-parametric smoothing splines were used to produce the functional effects of some continuous covariates. All these works proved TR may be a useful alternative to the Cox model. Furthermore Lee and Whitmore (2010) discussed the fundamental connection between proportional hazards and TR and showed that proportional hazards regression is for most purposes a special case of TR. In this article we provide more general theoretical results for TR modelled by a state-of-the-art semi-parametric regression formulation including varying coefficients and partial linearity.

Varying coefficients for prognostic factors in survival analysis were considered by several authors recently (Cai *et al.*, 2007; Peng and Fine, 2007; Peng and Huang, 2007; Cai *et al.*, 2008). Allowing the regression coefficients to vary with a time-dependent variable such as age may enhance the interpretability of the model and provide a more correct form of the covariate effects. We use the mature non-parametric technique of local polynomial regression (Fan and Gijbels, 1996) and derive necessary model-specific theoretical results to facilitate inferences.

The article is organized as follows. In section 2 we introduce the semi-parametric TR model and a rough sketch of model estimation. In section 3 we present the detailed computation procedure for the semi-parametric model. We give asymptotic results in section 4. In section 5 we examine the finite sample performance of

parametric and non-parametric estimates by numerical studies. In section 6, we apply our procedure to real data. Model checking for the data analysis will be followed in section 7. We conclude in section 8 with a discussion.

2 Threshold regression model with varying coefficients

We first introduce the probabilistic foundation for the TR method. Let us consider a first-hitting-time (FHT) model which has two basic components: (i) a parent stochastic process and (ii) a boundary set. When the parent process first encounters the boundary set is considered the threshold event (Lee and Whitmore (2006)). Such a probabilistic framework has been widely applied in survival analysis since it has a close resemblance to the lifetime process in human health studies and other practical failure time processes (Eaton and Whitmore, 1977; Lawless, 2003). In fact, by using special constructions, one can obtain the proportional hazards feature from a TR model (Lee and Whitmore, 2010).

A common parent process considered in practice is the Wiener process equipped with a positive initial value and certain mean and variance parameters. This process is appropriate since researchers have found that daily or hourly improvements and decrements in patient health can be accommodated by the bidirectional movements of the Wiener process (Lee and Whitmore, 2006). The boundary is the zero level of the process. It is shown that the time required for the process to reach the zero level for the first time has an inverse Gaussian distribution if the mean parameter is negative so the process tends to drift towards zero (Lancaster, 1972). Under such a general theoretical framework, we consider modelling the survival times with the inverse Gaussian distribution

$$f(y;\mu,v)=(2\pi\pi y^3 v)^{-1/2}\exp\{-(1-\mu y)^2/(2vy)\},\quad y>0, \tag{1}$$

where μ is the reciprocal of the mean survival time and v is the so-called volatility parameter. This distribution can accommodate a wide variety of shapes and is a member of the exponential family. Tweedie (1945) investigated its intriguing mathematical properties and applied it in a clinical trial. He found great success in applying the inverse Gaussian to model the survival distribution for a series of patients treated for cancer while log-normal and Weibull distributions fit such data poorly. Since then many pioneering studies have been followed to extend the application of this distribution in scientific studies. See chapter 1 of Seshadri (1993) for a detailed historical survey.

Previous authors suggest various approaches to incorporate covariates. Specifically, we may model the parameter μ as a regression function of the covariates. Though μ does not have a direct interpretation in terms of the instantaneous failure rate or hazard rate, it does have a direct interpretation in terms of what is happening to the underlying health process. The regression structure for μ dictates how rapidly or slowly the sample path approaches the failure threshold and, thus, in a way, is

even more fundamental because it is a feature of the latent process from which the survival distribution is derived. We therefore also choose to work with the inverse mean parameter in this article.

Many authors considered parametric linear regression models under the inverse Gaussian distribution (Whitmore, 1983; Whitmore *et al.*, 1998; Lee and Whitmore, 2010). Recently Yu *et al.* (2009) developed an additive semi-parametric model which allows partial covariate effects to be functional. However, such a model tends to ignore the interaction effects of covariates and may be limited in some situations. We intend to generalize the previous models to contribute a more accurate modelling form.

We use the following notation in the article. For patient i $(i=1,2,\ldots,n)$, let $Y_i \in \mathbb{R}^+$ be the event time subject to right censoring, $\mathbf{X}_i \in \mathbb{R}^p$, $\mathbf{Z}_i \in \mathbb{R}^q$, and $U_i \in \mathbb{R}$ be covariates, and $\delta_i \in \{0,1\}$ be the indicator of being censored (0) or not (1).

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

$$\begin{aligned} l_i &= \log\{f(y_i;\mu_i,v)\} = -\log(2\pi v)/2 - 3\log(y_i)/2 \\ &\quad -(1-\mu_i y_i)^2/(2vy_i), \end{aligned}$$

where

$$\mu_i = \mathbf{X}_i^T \boldsymbol{\alpha}(U_i) + \mathbf{Z}_i^T \boldsymbol{\beta}, \tag{2}$$

$\boldsymbol{\alpha}(.) = (\alpha_1(.),\ldots,\alpha_p(.))^T$ is a p-dimensional vector of unknown functions, and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_q)^T$ is a q-dimensional vector of unknown parameters. Under Equation (2), the regression coefficients of $\mathbf{X}$ are varying according to U and those of $\mathbf{Z}$ are constant. When all regression coefficients are constant, the model reduces to that of Whitmore (1983). When $\mathbf{X}$ simply involves an intercept, the model reduces to that of Yu *et al.* (2009). Our model thus incorporates all previous regression models. This partially linear varying-coefficient modelling approach has received a lot of research attention in the non-parametric regression literature (Fan and Zhang, 1999, 2000; Zhang and Lee, 2000; Xia, Zhang and Tong 2004; Fan and Huang, 2005). Recently Cai *et al.* (2007, 2008) applied such a non-parametric modelling technique to multivariate survival data under the Cox regression model.

For a censored observation, the log-likelihood contribution is

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

where f in the integral involves the unknown parameters v, $\boldsymbol{\beta}$ and unknown functions $\boldsymbol{\alpha}$ as before. We note that the complementary c.d.f. has a simple form involving the standard normal c.d.f. See Seshadri (1999).

In this article, to estimate the unknown functions, we consider a local polynomial approach. For observed covariates U_i close to the point u, we have

$$\begin{aligned} \boldsymbol{\alpha}(U_i) &\approx \alpha(u) + (U_i-u)\boldsymbol{\alpha}^{(1)}(u) + \cdots + (U_i-u)^k \boldsymbol{\alpha}^{(k)}(u)/k! \\ &= \{\mathbf{u}_i(u) \otimes \mathbf{I}_p\}^T \boldsymbol{\theta}(u), \end{aligned}$$

in which $\mathbf{u}_i(u)=(1,(U_i-u),\ldots,(U_i-u)^k)^T$, $\boldsymbol{\theta}(u)=(\boldsymbol{\alpha}(u)^T,\ldots,\boldsymbol{\alpha}^{(k)}(u)^T/k!)^T$, and the symbol $\otimes$ denotes the Kronecker product.

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

$$L_u=\sum_{i=1}^{n}\{\delta_i l_i+(1-\delta_i)l_i^c\}K_h(U_i-u), \tag{3}$$

where $K_h(.)=K(./h)/h$ for a kernel function K and h is a bandwidth. The kernel reflects that the above model is only applied to the neighborhood of u and weighs down smoothly the contribution of remote data points. We maximize Equation 3 to obtain $\hat{\boldsymbol{\theta}}(u)$ and use the first p entries of $\hat{\boldsymbol{\theta}}(u)$ as the local MLEs $\hat{\boldsymbol{\alpha}}(u)$ of $\boldsymbol{\alpha}(u)$. The computation has to be repeated for a grid of N points for distinct u over the range of U to obtain the complete profile of function values for $\boldsymbol{\alpha}(u)$.

3 Computational issues

The preceding section sketched a roadmap for obtaining the parameter and function estimates. We address some technical details for the implementation in this section.

3.1. Estimating equations for function estimates

Differentiating the local likelihood Equation 3 with respect to the parameters, we obtain a set of $kp+q+1$ estimating equations.

Specifically, estimating equations for $\boldsymbol{\theta}(u)$ and $\boldsymbol{\beta}$ are given by

$$v^{-1}\sum_{i=1}^{n}(\mathbf{X}_i^T\{\mathbf{u}_i(u)\otimes\mathbf{I}_p\}^T,\mathbf{Z}_i^T)^T\times\{1-\mu_i M_{i,1}(\mu_i,v)\}K_h(U_i-u)=\mathbf{0}, \tag{4}$$

where $M_{i,k}(\mu_i,v)=\delta_i y_i^k+(1-\delta_i)E_*(y_i^k;\mu_i,v)$ for $k=-1,1$, and $E_*(y_i^k;\mu_i,v)=E(Y_i^k\mid Y_i\geq y_i;\mu_i,v)$. The two moments of the residual lifetime for inverse Gaussian can be easily evaluated by adopting the following famous formula (Whitmore, 1983),

$$E(Y\mid Y\geq y)=F(1/(\mu^2 y))/[\mu\{1-F(y)\}], \tag{5}$$

$$E(1/Y\mid Y\geq y)=v+\mu^2E(Y\mid Y\geq y)-2yvf(y)/\{1-F(y)\}, \tag{6}$$

where f and F are the density and distribution functions for the inverse Gaussian.

The estimating equation of the volatility parameter v is given by

$$[-n/v+v^{-2}\sum_{i=1}^{n}\{M_{i,-1}(\mu_i,v)-2\mu_i+\mu_i^2M_{i,1}(\mu_i,v)\}]K_h(U_i-u)=0. \tag{7}$$

We may consider an iterative procedure to obtain the parameter estimates. Specifically, given current parameter values, we may estimate $E_*(y^k)$ by using Equations 5 and 6. These values are then plugged into Equations 4 and 7 and we update the parameter values by solving the system of nonlinear equations. Many computer packages such as GAMS, MATLAB, and Mathematica can provide fast algorithms

for solving systems of nonlinear equations by using gradient information. The MATLAB code for all of the computation involved in this article is available from the authors upon request.

We denote the parameter estimates as $\hat{\boldsymbol{\beta}}_u$ and $\hat{v}_u$ to highlight that these estimates are obtained using only local data around u. They are not as efficient as they would be if estimated using the complete data. As we show in section 4, the convergence rate of these estimates is quite slow.

Remark 1. One important operational issue in the above procedure is to determine the bandwidth h. If h is too small, the resulting estimates may exhibit a large variance and be highly irregular. On the other hand, if h is too large, the estimates may not be close to the true functions. Therefore an optimal bandwidth must be sought to balance the variance and bias. We consider leave-one-out cross-validation as the bandwidth selection criterion (Li and Palta, 2009). At each step, we conduct a similar estimation procedure, based on $n-1$ observations excluding the ith datum. This leads to the local log-likelihood l_u^{-i} and the corresponding local MLEs for the function estimates $\hat{\boldsymbol{\alpha}}^{-i}(u)$ and parameter estimates $\hat{\boldsymbol{\beta}}_u^{-i}$ and $\hat{v}_u^{-i}$. We then substitute these estimates into the likelihood for the ith observation L_i and compute its corresponding deviance $D_i(y_i, \hat{\mu}_i) = -2\log L_i$. We select the bandwidth h which corresponds to the minimum mean deviance $D = n^{-1}\sum_{i=1}^{n} D_i$. Zhang and Peng (2010) provided a rigorous justification for the theoretical basis of this method.

3.2. Estimating equations for parameter estimates

The final parameter estimates are obtained via a pseudo profile likelihood approach. The objective function under maximization is the log-likelihood

$$L(\boldsymbol{\beta}, v; \hat{\boldsymbol{\alpha}}) = \sum_{i=1}^{n} \{\delta_i l_i + (1-\delta_i) l_i^c\},$$

where we replace $\boldsymbol{\alpha}$ everywhere in l_i and l_i^c with $\hat{\boldsymbol{\alpha}}$.

In this case, the estimating equations for $\boldsymbol{\beta}$ and v are given by differentiating the above likelihood function,

$$v^{-1} \sum_{i=1}^{n} \mathbf{Z}_i \times \{1 - \hat{\mu}_i M_{i,1}(\hat{\mu}_i, v)\} = \mathbf{0}, \tag{8}$$

$$-n/v + v^{-2} \sum_{i=1}^{n} \{M_{i,-1}(\hat{\mu}_i, v) - 2\hat{\mu}_i + \hat{\mu}_i^2 M_{i,1}(\hat{\mu}_i, v)\} = 0, \tag{9}$$

where $\hat{\mu}_i = \mathbf{X}_i^T \hat{\boldsymbol{\alpha}}(U_i) + \mathbf{Z}_i^T \boldsymbol{\beta}$.

Solving the above equations leads to the final parameter estimates, denoted by $\hat{\boldsymbol{\beta}}$ and $\hat{v}$ without any subscript. We show in section 4 that these global parameter estimates converge at the usual optimal rate $O(n^{-1/2})$.

4 Asymptotic properties of the estimates

In order to make inferences for the parameter and functional estimates, we derive the large sample properties for them in this section. We note that Cai *et al.* (2007, 2008) worked on a local kernel method similar to our approach. However, their procedure was embedded in a Cox proportional hazard structure and therefore closely depended on asymptotic behaviour of the corresponding partial likelihood. Our underlying model form is different and therefore the justification of the asymptotics is also different.

For the non-parametric estimates for functional coefficients, we have the following result.

Theorem 1. *Assume that Conditions (1)–(5) in the Appendix hold. Then as* $n \to \infty$, $h \to 0$ *and* $nh \to \infty$, *we have*

$$(nh)^{1/2}\left(\begin{bmatrix}\hat{\boldsymbol{\alpha}}(u)-\boldsymbol{\alpha}_0(u)\\ \hat{\boldsymbol{\beta}}_u-\boldsymbol{\beta}_0\end{bmatrix}-\text{bias}\right)\to_d N[\mathbf{0},\text{covariance}].$$

The proof of the theorem is given in the Appendix where we present the asymptotic covariance matrix. The bias term is of order $O(h^2)$ and thus is only of theoretical concern. We usually treat such a term as a nuisance in the calculation since omitting it has no practical impact on estimation accuracy.

The above result shows that if we estimate $\boldsymbol{\beta}$ with the local likelihood, the estimator converges to the true parameter at a relatively slower speed $O(\sqrt{nh})$. We show in the next theorem that the estimator based on a profile global likelihood can achieve a better convergence rate.

Theorem 2. *Assume that Conditions (1)–(5) in the Appendix hold. Then as* $n \to \infty$, *we have*

$$n^{1/2}\begin{bmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\\ \hat{v}-v_0\end{bmatrix}\to_d N[\mathbf{0},\text{covariance}].$$

The covariance matrices for the parametric and non-parametric components are usually difficult to estimate with the usual plug-in method where we replace the unknown quantities in the covariance expressions with their sample estimates. A common practice is to employ the bootstrap resampling method to yield inference results such as the $100(1-\alpha)\%$ confidence intervals. specifically, we sample with replacement from the original data and repeatedly estimate parametric and non-parametric components. After we form a bootstrap sample with size B, we may construct confidence intervals for the estimates by using the sample percentiles (Li, Tai and Nott, 2009; Li *et al.*, 2010). One may follow an argument similar to that in Shao and Tu (1995) to support the validity and consistency of the bootstrap method. We exclude further technical details since the argument repeats the proof of the two main theorems and does not offer extra insight.

5 Simulations

In this section we conduct extensive Monte Carlo simulations to examine the finite sample performance of the proposed procedures. The performance of the estimator $\hat{\boldsymbol{\alpha}}(.)$ is assessed by using the square root of average square errors (RASE)

$$\text{RASE} = \{n_{grid}^{-1} \sum_{k=1}^{n_{grid}} \| \hat{\boldsymbol{\alpha}}(u_k) - \boldsymbol{\alpha}(u_k) \|^2\}^{1/2},$$

where $\{u_k\}_{k=1}^{n_{grid}}$ are the grid points at which the functions $\hat{\boldsymbol{\alpha}}(.)$ are evaluated. In our simulation, we employ the Epanechnikov kernel $K(u) = 0.75(1-u^2)_+$ and $n_{grid} = 150$. The performance of the parametric estimator $\hat{\boldsymbol{\beta}}$ is assessed by its estimation bias and variance.

We consider two model specifications:

$$\text{Case I:} \quad \alpha_1(u) = 0.3\exp\{2u-1\} + 15, \alpha_2(u) = 1.2u(1-u), \alpha_3(u) = 1.5 + 5u^3, \boldsymbol{\beta} = (0.5, 0.25)^T$$

$$\text{Case II:} \quad \alpha_1(u) = 0.8\exp\{-2(u-1)\} + 30, \alpha_2(u) = 2\sin^2(2\pi u), \boldsymbol{\beta} = (3, 1.5, 2)^T.$$

For both cases, we generated U from the uniform distribution on [0,1]. In Case I, $\mathbf{X} = (X_1, X_2, X_3)^T$ with $X_1 \equiv 1$, X_2 being equally spaced fixed points on [0,1], and $X_3 \sim N(0,1)$ and $\mathbf{Z}$ follows a bivariate normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{2\times 2}$ with $\sigma_{ij} = (0.5)^{|i-j|}$. In Case II, $\mathbf{X} = (X_1, X_2)^T$ with $X_1 \equiv 1$ and $X_2 \sim N(0,1)$ and $\mathbf{Z}$ follows a trivariate normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{3\times 3}$ with $\sigma_{ij} = (0.5)^{|i-j|}$. The volatility parameter v were chosen to be 0.04 and 11 respectively. The exact survival times were then generated from the inverse Gaussian distribution by adapting the rejection sampling algorithm in Michael, Schucany and Hass (1976). We independently generated censoring times from the exponential distribution and designed the censoring proportion to be roughly 25% of the complete sample in each simulation. Sample sizes of 100 and 200 were considered for both cases.

The estimation results from 1000 simulation runs for parametric components appear in Table 1. The performance of the estimators seem to be rather satisfactory since the biases and standard errors are all quite small. Furthermore, the estimation

Table 1. Estimated parametric components in the simulation study

			n = 100			*n = 200*		
Case	Parameter	True	Bias	SE	MAE	Bias	SE	MAE
I	β_1	0.50	0.032	0.098	0.103	0.028	0.066	0.071
	β_2	0.25	–0.026	0.094	0.097	–0.012	0.065	0.067
	v	0.04	–0.0060	0.0047	0.0075	–0.0013	0.0022	0.0025
II	β_1	3	0.19	2.10	2.18	0.08	1.42	1.44
	β_2	1.5	–0.32	2.48	2.50	–0.09	1.63	1.65
	β_3	2	–0.16	2.15	2.17	0.10	1.35	1.37
	v	11	1.18	5.49	5.61	0.24	3.27	3.30

Bias is the difference between the average of 1000 parameter estimates and the corresponding true parameter. SE is the sample standard deviation of 1000 parameter estimates. MAE is the sample mean absolute error.

will improve as the sample size grows. The improvement for the volatility parameter v is especially remarkable.

It is of interest to assess the impact of estimation of $\boldsymbol{\beta}$ on the estimation of $\boldsymbol{\alpha}(.)$. To this end we also estimated $\boldsymbol{\alpha}(.)$ by using the true value of $\boldsymbol{\beta}$ in the two cases. The RASE of the non-parametric estimates $\hat{\boldsymbol{\alpha}}(.)$ are shown in Figure 1 for the two cases. Generally RASE tends to be smaller when the parameters do not need to be estimated. For both cases with $n = 100$, we also displayed typical non-parametric estimates which are of the median RASE among all 1000 simulations. See Figures 2 and 3.

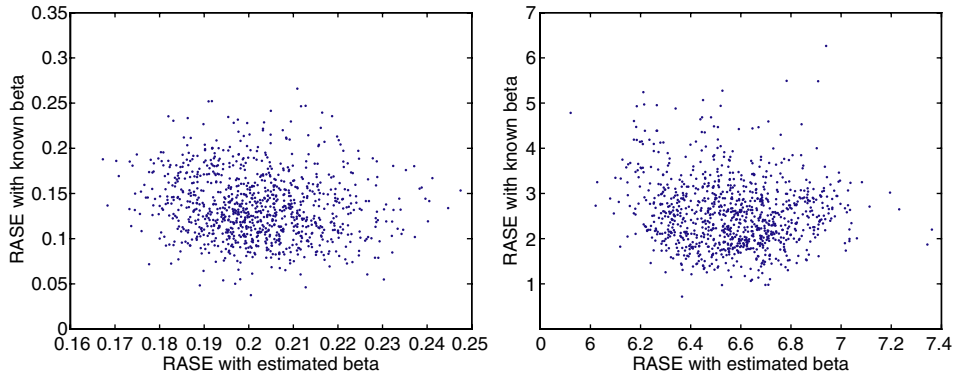

Fig. 1. Plots for RASE in Cases I (left panel) and II (right panel) with sample sizes 100.

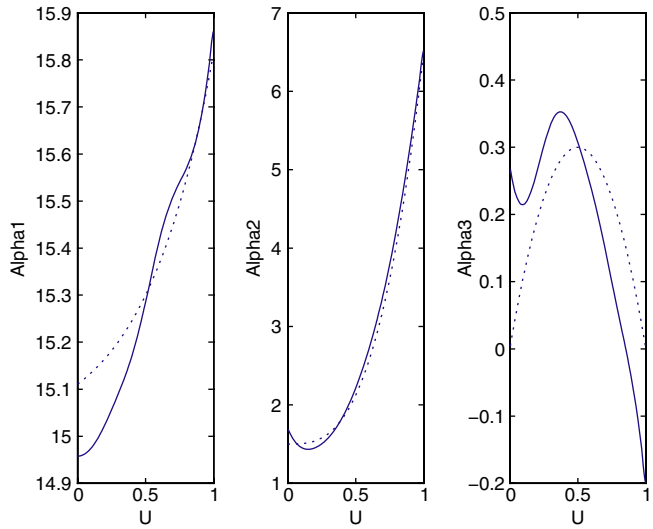

Fig. 2. Estimated varying regression coefficients with median RASE for Case I with sample size 100. Solid lines are the estimates while dotted lines are the true functions. The sample mean integrated square errors for the three functions are 0.406, 0.144 and 0.257 respectively. The bandwidth for the fit shown in the figure is 0.217.

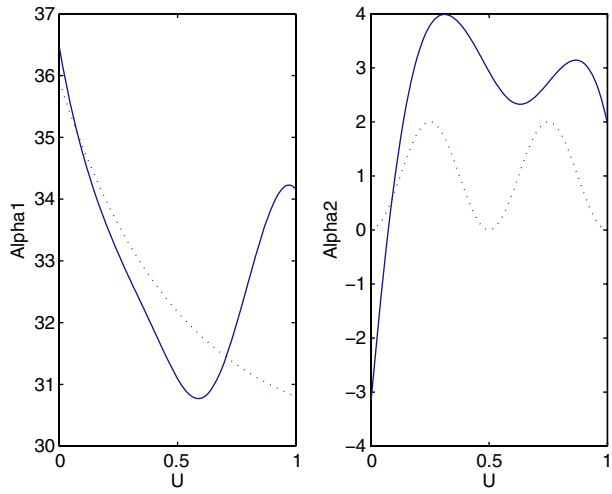

Fig. 3. Estimated varying regression coefficients with median RASE for Case II with sample size 100. Solid lines are the estimates while dotted lines are the true functions. The sample mean integrated square errors for the two functions are 16.04 and 9.59 respectively. The bandwidth for the fit shown in the figure is 0.336.

Table 2. Predictor variables in malignant melanoma dataset

Variable names	Description
ICI	Amount of inflammatory cell infiltrate (ICI)
Ulcerat	Ulceration
Thick	Invasion level or thickness of the tumor
Sex	Sex
Ecells	Indicator of cell type–epithelioid

6 Applications

We illustrate our methods using a malignant melanoma data set that was originally analysed by Drzewiecki and Andersen (1982). The clinical data cover 205 melanoma patients in clinical Stage I of the disease who were treated at the Plastic Surgery Unit in Odense, Denmark from 1964 to 1973. Follow-up was terminated 1 January 1978. Of the 205 patients, 57 had died from malignant melanoma before the closing date of the study. Investigators were interested in studying the effects of important prognostic factors. The predictor variables in this dataset are listed in Table 2.

Previous investigators have noted that the proportional hazards assumption may not hold for this data set (Andersen *et al.*, 1993; Keiding, Andersen and Klein, 1997). We thus choose the TR model to fit the failure time in this data set by the inverse Gaussian distribution. We consider modelling the reciprocal of the mean survival time by Equation 2. For the dichotomous predictors Sex and Ecells, it is sensible to associate their effects with constant regression coefficients. On the other hand, we expect that the effects for ICI, Ulcerat and Thick to be varying with the age of the subjects. It is natural to use age as the index variable U. Consequently in this

analysis, **X** involves the intercept, ICI, Ulcerat and Thick whereas **Z** involves Sex and Ecells.

We carried out the estimation procedure described in section 3. The estimated functional regression coefficients for the intercept, ICI, Ulcerat and Thick, along with 95% confidence intervals, are shown in Figure 4. The estimated coefficients for sex and ecells are reported in Table 3.

The estimated curves for the continuous effects all display non-constant patterns. The functional coefficients corresponding to the intercept start around zero and then decrease gradually. The curve stays below zero in most of the range, indicating a negative association between Age and μ. The functional coefficients corresponding to ICI are flat for Age below 60 and start to climb up rapidly for older subjects. Since the estimated varying coefficients are all positive for ICI, there exists a positive association between ICI and μ. A larger value of ICI might lead to mortality with a larger chance and the ICI effect increases as the subject becomes older. In this dataset there is a portion of subjects (approximately 4%) that are older than 80

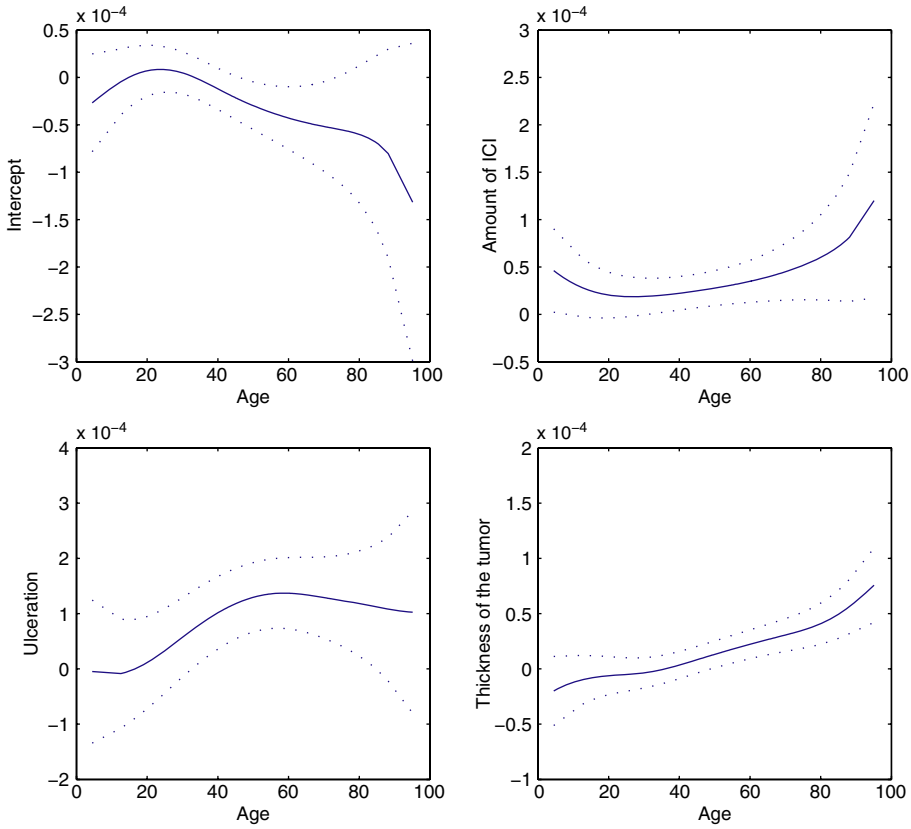

Fig. 4. Estimated varying regression coefficients for the malignant melanoma dataset. Solid lines are the estimates while dotted lines are the 95% bootstrap confidence intervals. The bandwidth for this fit is 29.65.

Table 3. Estimated coefficients for parametric part in malignant melanoma dataset

	Semi-parametric model			Parametric model		
	Coefficients	SE	*P*-value	Coefficients	SE	*P*-value
Sex	3.9×10^{-4}	2.4×10^{-4}	0.0521	2.4×10^{-4}	1.1×10^{-4}	0.0354
Ecells	4.7×10^{-4}	2.1×10^{-4}	0.0126	3.1×10^{-4}	1.2×10^{-4}	0.0122
Intercept	–	–	–	-8.7×10^{-2}	3.7×10^{-4}	<0.0001
Age	–	–	–	8.2×10^{-6}	3.6×10^{-6}	0.0244
ICI	–	–	–	1.3×10^{-4}	7.3×10^{-5}	0.0835
Ulcerat	–	–	–	3.6×10^{-4}	1.2×10^{-4}	0.0019
Thick	–	–	–	3.2×10^{-5}	2.1×10^{-5}	0.1183
v	3.2×10^{-4}	2.6×10^{-5}	<0.0001	5.4×10^{-4}	4.8×10^{-5}	<0.0001

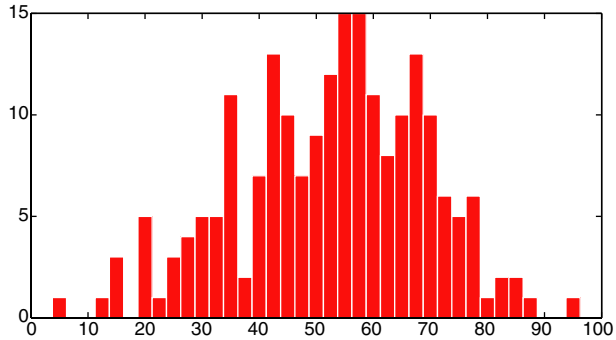

Fig. 5. Distribution of age for malignant melanoma dataset.

(See Figure 5). Therefore constant regression coefficients may not be suitable for this subgroup of samples. The functional coefficients for Ulcerat increase before age 50 in a straight-line pattern and then remain constant afterwards. A larger value of Ulcerat leads to a larger likelihood of mortality. The hazard effect of Ulcerat remains at a fixed level after age 50. The functional coefficients for Thick seem to be increasing with time in a linear pattern and climb a bit more steeply after age 80. A larger value of Thick thus leads to a larger likelihood of mortality. The hazard due to Thick is increasing with age.

The estimated parametric coefficients for the semi-parametric model are provided in Table 3. For the sake of comparison, we also fit a parametric model by treating all the coefficients for Age, ICI, Ulcerat and Thick to be constant and summarize the results in Table 3 as well. Such a simple model specification would require all the curves in Figure 4 to be increasing straightlines. Clearly the nonparametric pattern detected from the semi-parametric model can not be reproduced in a parametric model.

7 Model checking

One should not conclude a regression analysis without a careful check on model assumptions. Unfortunately, there is still a lack of attention in the literature for a

systematic or automatic semi-parametric residual analysis. We provide some relevant discussion in this section, following the results in the preceding section.

We need to check the following: (i) the appropriateness of the selected model form; (ii) the distribution assumption.

We first consider (i). There are many possible alternative model formulations. One particular relevant question is whether we can reduce any functional coefficient to a constant parameter. Specifically, for model Equation 2, we may consider testing the hypothesis:

$$H_0 : \alpha_k(.) = a_{k0}, \quad \text{v.s.} \quad H_1 : \alpha_k(.) \neq a_{k0},$$

for $k \in \{1, 2, \ldots, p\}$. A reasonable approach in semi-parametric analysis is to construct a so-called generalized likelihood ratio (GLR) test. Under the null hypothesis, the functional coefficient $\alpha_k(.)$ is replaced by a constant regression coefficient a_{k0}. We then fit the model and evaluate the maximum log-likelihood $\mathcal{L}_0$. Under the alternative hypothesis, we estimate $\alpha_k(.)$ as a function and obtain the corresponding maximum log-likelihood $\mathcal{L}_1$. The GLR test statistic simply compares the log-likelihoods as follows:

$$\mathcal{T} = \mathrm{r}_K(\mathcal{L}_1 - \mathcal{L}_0),$$

where r_K is a fixed constant which might depends on the kernel function. The asymptotic distribution of $\mathcal{T}$ can be carefully justified to follow an adjusted chi-squared distribution under H_0 (see Eubank, Hart and LaRicca 1993; Eubank and LaRicca, 1993; Fan, (1996), Ramil-Nova and Gonzalez-Manteiga (1998) for rigorous proofs for similar problems). However, it has been empirically observed that the asymptotic null distribution may not suit a finite sample because of the slow convergence rate of non-parametric estimates (Fan, 1996; Fan and Lin, 1998). It is thus advisable to take a numerical approach such as the bootstrap to simulate the null distribution (Fan and Yao, 2005) or an approximate approach to replace the chi-squared distribution with a more accurate distribution (Zhang, 2005). In this article, we choose to conduct a bootstrap resampling procedure to identify the null distribution of $\mathcal{T}$. The P-values from testing the four non-parametric coefficients under the GLR test are all highly significant in this case (all smaller than 0.01). The results indicate that none of these four components should be simplified as a constant coefficient. All are better to be modelled as functions.

To formally check the distribution assumption (ii), we conducted goodness-of-fit tests using the Kolmogorov–Smirnov (KS) test and Cramer–von Mises (CV) test. The two test statistics take the following forms

$$\mathrm{KS} = \sup_t |\, S_*(t) - S_0(t) \,|,$$

$$\mathrm{CV} = \int_{-\infty}^{\infty} (S_*(t) - S_0(t))^2 \mathrm{d}S_0(t),$$

where $S_*(t)$ is the estimated survival function under the semi-parametric model averaged across all covariates at time t in the original data. The reference distribution

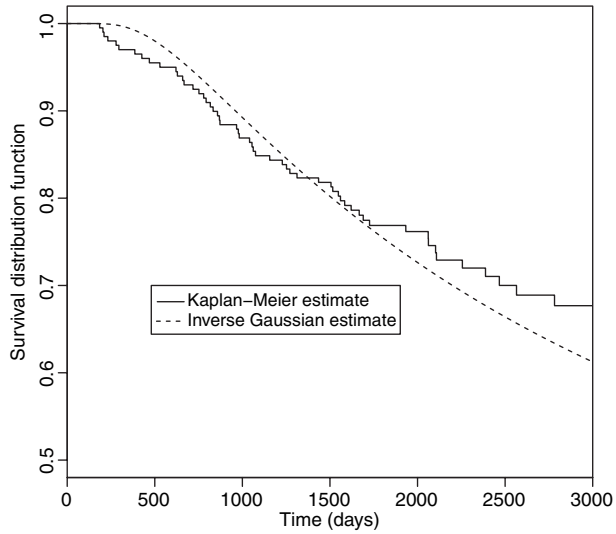

Fig. 6. Comparison of fitted survival distribution function with the Kaplan–Meier non-parametric estimation.

$S_0(t)$ is the Kaplan–Meier estimator which summarizes the survival pattern of the data without any distribution assumption. KS and CV tests for lifetime data appear in many studies (see e.g., D'Agostino and Stephens, 1986). We employed the bootstrap method to simulate data from the null distribution with a bootstrap sample size $B = 500$ and used the bootstrap samples to rebuild the KS and CV tests. The proportions of bootstrap test statistics greater than the observed test statistics are reported as the test P-values. In this case the P-values for the KS and CV tests are not significant (0.2759 and 0.3263), indicating that our assumed inverse Gaussian distribution does not differ significantly from the empirical distribution of the data. Furthermore, we plotted the Kaplan–Meier estimator and the estimated survival function under our semi-parametric model in Figure 6 for a graphical comparison. The two curves agree with each other closely in the observed time range for failure events.

8 Discussion

In the varying coefficients, the role of U acts like an effect modifier in biostatistical research studies (Thompson, 1991; Armitage and Colton, 2005). In this article and many previous publications, the index variable U is treated explicitly as time. We note that in practice the variable U can be any variable potentially affecting the effects of existing covariates and can be either dependent or independent of time (Xia *et al.*, 2004). The estimation of the functional coefficients follows a local kernel construction in this article. In fact the model can be implemented using a penalized

quasi-likelihood estimator with smoothing splines although a different technical justification is needed.

Our analysis of the melanoma dataset shows that some subjects may have a negative estimated μ value. In fact, when $\mu < 0$, the probability density function Equation 1 is improper and the probability $P(Y = \infty) = 1 - \exp\{-2\mu/v\}$ is positive, indicating that the fist-hitting-time may never be observed. The negative sign thus implies a cure rate model for the subject. Therefore our results suggest the development of further refined models involving a cure rate component in the likelihood. For a comprehensive summary of recent literatures on cure rate modelling, see Ma (2009, 2010).

Acknowledgements

This research is supported in part by NMRC/NIG/0054/2009, ARF R-155-000-109-112 and CDC/NIOSH grant OH008649. We thank Professor Zhidong Bai for reading the manuscript and providing valuable suggestions to improve the rigor of the mathematical results. We also acknowledge the helpful comments of Dr. G. A. Whitmore that have greatly improved the presentation of the manuscript.

Appendix

The following technical conditions are needed for the proofs of our theoretical results.

(1) The kernel function $K(.)$ is a bounded symmetric density with compact support.
(2) The density function $p(u)$ of the variable U is of compact support $\mathcal{U}$ and has a bounded second derivative.
(3) The functions $\alpha_j(.)$ have a continuous second derivative for $j = 1, \ldots, p$.
(4) There is an $s > 2$ and some $\epsilon < 2 - s^{-1}$ such that $E \| \mathbf{X} \|^{2s} < \infty$, $E \| \mathbf{Z} \|^{2s} < \infty$, and $n^{2\epsilon - 1} h \to \infty$.
(5) The $p \times p$ matrix $E(\mathbf{X}\mathbf{X}^T \mid U)$ is non-singular for each U in its support. $E(\mathbf{X}\mathbf{X}^T \mid U)$, $E(\mathbf{X}\mathbf{X}^T \mid U)^{-1}$ and $E(\mathbf{X}\mathbf{Z}^T \mid U)$ are all Lipschitz continuous.

Proof of Theorem 1. Let $\xi = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T)$. The left-hand-side of Equation (4) can be regarded as a function $\mathcal{H}_n(\xi)$.

Using Liapounov's central limit theorem and omitting the small order bias term, we have that $\sqrt{nh}\{n^{-1}\mathcal{H}_n(\xi_0)\} \to_d N(0, \Sigma)$. Furthermore, we can show $n^{-1} \| \mathcal{H}_n(\xi) \| = O_p(n^{-1/2} + h^2 + \epsilon_n)$ uniformly for $\xi \in \Xi$ where Ξ is a neighbourhood of ξ_0 such that $\| \xi - \xi_0 \| \leq \epsilon_n$ and $\epsilon_n \to 0$. We note $\| \cdot \|$ is the Euclidean modulus of a matrix.

A Taylor series expansion near the true parameter ξ_0 gives

$$\mathcal{H}_n(\xi) \approx \mathcal{H}_n(\xi_0) + \mathcal{J}_n(\xi - \xi_0),$$

where

$$\mathcal{J}_n = \partial \mathcal{H}_n / \partial \xi = \begin{pmatrix} \mathcal{J}_{11} & \mathcal{J}_{12} \\ \mathcal{J}_{21} & \mathcal{J}_{22} \end{pmatrix},$$

is the derivative of the estimating equation. The four building blocks of $\mathcal{J}_n$ are given by

$$\mathcal{J}_{11} = \frac{\sum_{i=1}^n [\nu M_{i,1} - (1-\delta_i)\{E_*(y_i^2) - E_*(y_i)^2\}\mu_i^2](\mathbf{X}_i^T\{\mathbf{u}_i(u) \otimes \mathbf{I}_p\}^T, \mathbf{Z}_i^T)^{\otimes 2} K_h(U_i - u)}{\nu^2}$$

$$\mathcal{J}_{12} = \mathcal{J}_{21}^T = \frac{\sum_{i=1}^n \mu_i(1-\delta_i)[\{E_*(y_i^2) - E_*(y_i)^2\}\mu_i^2 - E_*(y_i)E_*(y_i^{-1}) + 1](\mathbf{X}_i^T\{\mathbf{u}_i(u) \otimes \mathbf{I}_p\}^T, \mathbf{Z}_i^T)^T K_h(U_i - u)}{2\nu^3}$$

$$\mathcal{J}_{22} = \frac{n}{2\nu^2} - \frac{\sum_{i=1}^n (1-\delta_i)[\{E_*(y_i^2) - E_*(y_i)^2\}\mu_i^4 - 2(E_*(y_i)E_*(y_i^{-1}) - 1)\mu_i^2 + E_*(y_i^{-2}) - E_*(y_i^{-1})^2]K_h(U_i - u)}{4\nu^4},$$

We can show that under the assumed conditions $\mathcal{J}_n = \mathrm{J} + o_p(1)$ and J is positive definite with probability converging to one, evaluated at the true parameter values. This fact enables us to conclude that a unique consistent solution to the estimating equation $\mathcal{H}_n(\xi) = 0$ exists.

By the convexity lemma (Pollard, 1991) and the previous results, we obtain the asymptotic normality of the sequence $\sqrt{nh}(\hat{\xi} - \xi_0)$ whose covariance matrix is $\mathrm{J}^{-1}\Sigma\mathrm{J}^{-1}$. The distribution of $\sqrt{nh}\{(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})^T - (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\}$ is the first p and the $kp+1$th to the $kp+q$th components of the sequence $\sqrt{nh}(\hat{\xi} - \xi_0)$. We thus obtain the claimed asymptotic covariance matrix as

$$\frac{e_0}{p(u)} E\left[\left\{\lambda(\mu\nu)^{-1} + (1-\lambda)\mu^2\nu^{-2}\mathcal{E}_1\right\}(\mathbf{X}^T, \mathbf{Z}^T)^{\otimes 2} \mid U = u\right]^{-1},$$

where $e_0 = \int K^2(u)\mathrm{d}u$, $\mathcal{E}_1 = E_c\{E_*(Y_c)\nu\mu^{-2} - (E_*(Y_c^2) - E_*(Y_c)^2)\}$. The expectation E_c is taken with respect to the observed censored survival time Y_c.

Proof of Theorem 2. By using the assumed conditions and the result from Mack and Silverman (1982), we can show $n^{-1} \| \mathcal{H}_n(\xi) \|$ defined in the previous theorem converges to zero in probability uniformly for any $u \in \mathcal{U}$. It then follows that

$$\sup_{u \in \mathcal{U}} \| \hat{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}_0(u) \| \to_p 0, \tag{10}$$

where again we omit the small order bias terms.

We now consider the profile likelihood function and the resulting estimating Equations 8 and 9. By a standard argument for parametric score equations (Appendix C in Lawless, 2003) and using Equation 10, we can obtain the asymptotic normality of $\sqrt{n}\{(\hat{\boldsymbol{\beta}}, \hat{\nu})^T - (\boldsymbol{\beta}_0^T, \nu_0)^T\}$ with a covariance equal to Ψ^{-1} where

$$\Psi = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix},$$

with the four building blocks given by

$$\Psi_{11} = E\left[\left\{\lambda(\mu v)^{-1} + (1-\lambda)\mu^2 v^{-2}\mathcal{E}_1\right\}\mathbf{Z}\mathbf{Z}^T\right]$$
$$\Psi_{12} = \Psi_{21}^T = (2v^3)^{-1}E\left[\mu(1-\lambda)\left\{\mu^2 v\mathcal{E}_1 - \mathcal{E}_2\right\}\mathbf{Z}\right]$$
$$\Psi_{22} = n/2v^2 - \{\mu^4 v^2\mathcal{E}_1 - 2\mu^2\mathcal{E}_2 + \mathcal{E}_3\}/4v^4,$$

where $\mathcal{E}_1$ is given in the previous theorem, $\mathcal{E}_2 = E_c\{E_*(Y_c)E_*(Y_c^{-1})\} - 1$ and $\mathcal{E}_3 = E_c\{E_*(Y_c^{-2}) - E_*(Y_c^{-1})^2\}$.

References

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (1993), *Statistical models based on counting processes*, Springer-Verlag, Inc., New York, New York.

Armitage, P. and T. Colton (2005), *Encyclopedia of biostatistics*, John Wiley & Sons.

Cai, J., J. Fan H. Zhou and Y. Zhou (2007), Marginal hazard models with varying-coefficients for multivariate failure time data, *The Annals of Statistics* **35**, 324–354.

Cai, J., J. Fan, J. Jiang and H. Zhou (2008), Partially linear hazard regression with varying-coefficients for multivariate survival data, *Journal Royal Statistical Society*, B, **70**, 141–158.

Cox, D. R. (1972), Regression models and life tables (with discussion), *Journal of the Royal Statistical Society B* **34**, 187–220.

D'Agostino, R. B. and M. A. Stephens (eds) (1986), *Goodness-of-fit techniques*, Marcel Dekker, New York, New York.

Drzewiecki, K. T. and P. K. Andersen (1982), Survival with malignant melanoma: a regression analysis of prognostic factors, *Cancer* **49**, 2414–2419.

Eaton, W. W. and G. A. Whitmore (1977), Length of stay as a stochastic process: A general approach and application to hospitalization for schizophrenia, *Journal of Mathematical Sociology* **5**, 273–292.

Eubank, R. L. and V. N. LaRicca (1993), Testing for no effect in non-parametric regression, *Journal of Statistical Planning and Inference* **36**, 1–14.

Eubank, R. L., J. D. Hart and V. N. LaRicca (1993), Testing goodness of fit via non-parametric function estimation techniques, *Communications in Statistics – Theory and Methods* **22**, 3327–3354.

Fan, J. (1996), Test of significance based on wavelet thresholding and Neyman's truncation, *Journal of the American Statistical Association* **91**, 674–688.

Fan, J. and I. Gijbels (1996), *Local polynomial modeling and its applications*, Chapman and Hall, London.

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

Fan, J. and S.-K. Lin (1998), Test of significance when data are curves, *Journal of the American Statistical Association* **93**, 1007–1021.

Fan, J. and Q. Yao (2005), *Nonlinear time series*, Springer, New York.

Fan, J. and W. Zhang (1999), Statistical estimation in varying coefficient models, *Annals of Statistics* **27**, 1491–1518.

Fan, J. and W. Zhang (2000), Simultaneous confidence bands and hypothesis testing in varying-coefficient models, *Scandinavian Journal of Statistics* **27**, 715–731.

Hosmer, D. W. and S. Lemeshow (1999), *Applied survival analysis*, Wiley, New York.

Keiding, N., P. K. Andersen, and J. P. Klein (1997), The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates, *Statistics in Medicine* **16**, 215–224.

Lawless, J. F. (2003), *Statistical models and methods for lifetime data*, 2nd edn, Wiley, New York.

Lee, E. T. (1992), *Statistical methods for survival data analysis*, 2nd edn, Wiley, New York.

Lee, M.-L. T. and G. A. Whitmore (2006), Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary, *Statistical Science*, **21**, 501–513.

Lee, M.-L. T. and G. A. Whitmore (2010), Proportional hazards and threshold regression: their theoretical and practical connections, *Lifetime Data Analysis* **16**, 196–214.

Lee, M.-L. T., V. DeGruttola and D. Schoenfeld (2000), A model for markers and latent health status, *Journal of the Royal Statistical Society: Series B* **62**, 747–762.

Lee, M.-L. T., G. A. Whitmore and B. Rosner (2010), Threshold regression for survival data with time-varying covariates, *Statistics in Medicine* **29**, 896–905.

Li, J. and M. Palta (2009), Bandwidth selection through cross validation for semi-parametric varying-coefficient partially linear models, *Journal of Statistical Computation and Simulation* **79**, 1277–1286.

Li, J., B. C. Tai and D. J. Nott (2009), Confidence interval for the bootstrap P-value and sample size calculation of the bootstrap test, *Journal of Non-parametric Statistics* **21**, 649–661.

Li, J., C. M. Zhang, K. A. Doksum and E. V. Nordheim (2010), Simultaneous confidence intervals for semi-parametric logistic regression and confidence regions for the multi-dimensional effective dose, *Statistica Sinica*, **20**, 637–659.

Ma, S. (2009), Cure model with current status data, *Statistica Sinica*, **19**, 233–249.

Ma, S. (2010), Mixed case interval censored data with a cured subgroup, *Statistica Sinica* **20**, 1165–1181.

Mack, Y. P. and B. W. Silverman (1982), Weak and strong uniform consistency of kernel regression estimates, *Zeitschriftfuer Wahrscheinlichkeitstheorie verw. Gebiete* **61**, 405–415.

Michael, J. R., W. R. Schucany and R. W. Haas (1976), Generating random variates using transformations with multiple roots, *The American Statistician* **30**, 88–90.

Peng, L. and J. P. Fine (2007), Regression modeling of semi-competing risks data, *Biometrics*, **63**, 96–108.

Peng, L. and Y. Huang (2007) Survival analysis with temporal covariate effects, *Biometrika* **94**, 719–733.

Pennell, M. L., G. A. Whitmore and M.-L. T. Lee (2009), Bayesian random effects threshold regression with application to survival data with nonproportional hazards, *Biostatistics*. In press.

Pollard, D. (1991), Asymptotics for least absolute deviations regression estimators, *Econometric Theory* **7**, 186–199.

Ramil-Nova, L. A. and W. Gonzalez-Manteiga (1998), χ^2 Goodness-of-fit tests for polynomial regression, *Communications in Statistics – Theory and Methods*, **27**, 229–258.

Seshadri, V. (1993), *The inverse Gaussian distribution: a case study in exponential families*, Clarendon press, Oxford.

Seshadri, V. (1999), *The inverse Gaussian distribution*, Springer-Verlag, New York.

Shao, J. and D. S. Tu (1995), *The jackknife and bootstrap*, Springer-Verlag, New York.

Thompson, W. D. (1991) Effect modification and the limits of biological inference from epidemiologic data, *Journal of Clinical Epidemiology* **44**, 221–232.

Tweedie, M. C. K. (1945), Inverse Gaussian variate, *Nature* **155**, 453.

Whitmore, G. A. (1983), A regression method for censored inverse-Gaussian data, *Canadian Journal of Statistics* **11**, 305–315.

Whitmore, G. A., M. J. Crowder and J. F. Lawless (1998), Failure inference from a marker process based on a bivariate Wiener model, *Lifetime Data Analysis* **4**, 229–251.

Xia, Y., W. Zhang and H. Tong (2004), Efficient estimation for semivarying-coefficient models, *Biometrika*, **91**, 661–681.

Yu, Z., W. Tu, and M.-L. T. Lee (2009), A semi-parametric threshold regression analysis of sexually transmitted infections in adolescent women, *Statistics in Medicine* **28**, 3029–3042.

Zhang, J.-T. (2005), Approximate and asymptotic distributions of Chi-squared-type mixtures with applications, *Journal of the American Statistical Association* **100**, 273–285.
Zhang, W. and S. Y. Lee (2000). Variable bandwidth selection in varying-coefficient models, *Journal of Multivariate Analysis* **74**, 116–134.
Zhang, W. and H. Peng (2010), Simultaneous confidence band and hypothesis test in generalised varying-coefficient models, *Journal of Multivariate Analysis* **101**, 1656–1680.

Received: June 2010. Revised: September 2010.