PARTIALLY LINEAR ADDITIVE HAZARDS REGRESSION FOR CLUSTERED AND RIGHT CENSORED DATA

Wei Chen
School of Zhangjiagang, Jiangsu University of Science and Technology

Fengling Ren
School of Computer Science and Engineering, Xinjiang University of Finance and Economics

https://doi.org/10.5109/4844359

出版情報：Bulletin of informatics and cybernetics. 54 (4), pp.1-14, 2022. 統計科学研究会
バージョン：
権利関係：

PARTIALLY LINEAR ADDITIVE HAZARDS REGRESSION FOR CLUSTERED AND RIGHT CENSORED DATA

by

Wei Chen and **Fengling Ren**

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.54, No. 4

FUKUOKA, JAPAN
2022

Bulletin of Informatics and Cybernetics, Vol. 54, No. 4, 2022

PARTIALLY LINEAR ADDITIVE HAZARDS REGRESSION FOR CLUSTERED AND RIGHT CENSORED DATA

By

Wei Chen[*] and **Fengling Ren**[†]

Abstract

For analyzing clustered survival data, a flexible partially linear additive hazards model is proposed. To accommodate the nonlinear effects, the unknown regression function is approximated by B-splines. All regression coefficients are estimated through a system of pseudo-score functions. Under certain conditions, the proposed estimators are shown to be asymptotically normal, where a consistent estimator of the covariance matrix is given. Simulation studies are also conducted to evaluate the finite sample performance of the proposed method, which is illustrated using a real data set from an AIDS clinical trial.

Key Words and Phrases: B-spline, Clustered data, Marginal model, Partially linear additive hazards, Right censored data.

1. Introduction

In survival analysis, the Cox model might be the most popular model to analyze survival data and has been extensively studied in various contexts, where the risk ratios are of interest. In contrast to it, the additive hazards (AH) model focuses on modelling the risk difference and could be more plausible and reasonable than the former in many applications, see Buckley (1984) and Aalen *et al.* (2008). One such case is the epidemiological study in Kulich and Lin (2000). Particularly, for univariate failure time data, Lin and Ying (1994) proposed the following AH model

$$\lambda(t|\mathbf{Z}) = \lambda_0(t) + \boldsymbol{\beta}^\top \mathbf{Z}, \tag{1}$$

where $\boldsymbol{\beta}$ is the regression coefficient vector corresponding to a p-dimensional covariates $\mathbf{Z}$ and $\lambda_0(t)$ is an unknown baseline hazard function. The model (1) above has attracted much attention and been studied in various contexts. For more details, see references listed in Afzal *et al.* (2017).

It is worth noting that most of the works mentioned above concern the univariate survival data, i.e., the failure times are mutually independent, or assume that the covariates affect the hazard rate linearly. However, multivariate failure time data often arise in practical applications due to natural or artificial clustering, and the linear assumption

[*] School of Zhangjiagang, Jiangsu University of Science and Technology, Zhangjiagang, 215600, P.R. China. Email: chenweixiyang@sina.com.cn

[†] School of Computer Science and Engineering, Xinjiang University of Finance and Economics, Urumqi, 830012, P.R. China.

may not always be appropriate. For example, during follow-up, each patient may experience several different events or recurrent events, or the same type of disease may affect clustered organs of the same subject (Yin (2007)). As the Colon Cancer Study in Moertel *et al.* (1990), age at study entry may effect nonlinearly on the times to recurrence of cancer and death. Additionally, in an AIDS clinical trial study, as one of the motivations of this paper, CD4 count is often considered to be a marker for antiretroviral treatment response and HIV disease progression. When modelling this relationship, researchers usually tend to introduce a specific or unspecified nonlinear function to describe it, e.g., Guo and Carlin (2004) and Mandal *et al.* (2019).

For clustered data, including but not limited to the recent papers listed below, many models and inference approaches have been developed. When there does not exist censoring, Cheng *et al.* (2014) considered efficient estimation of the parameters in a generalized partially linear additive models for longitudinal/clustered data. Geraci (2019) developed methods for the modelling and estimation of nonlinear conditional quantile functions when data are clustered within two-level nested designs. Wang *et al.* (2020) studied the weighted quantile average estimation technique for the parameter in additive partially linear models with missing covariates. When the data exists right censoring, Yin and Cai (2004) derived a class of estimation methods and asymptotic properties for the marginal additive hazards model. Yin (2007) proposed a class of graphical and numerical methods to assess the overall fitting adequacy of the marginal additive hazards model. Li and Yin (2009) proposed a generalized method of moments approach to the accelerated failure time model with correlated survival data, which was also studied by Johnson and Strawderman (2009), and furthermore discussed by Fu *et al.* (2021). Zeng and Cai (2010) proposed a class of additive transformation risk models and developed an estimating equation approach. Eriksson *et al.* (2014) compared the marginal approach and the conditional approach in the context of a Cox regression analysis, where they treated within-cluster correlation as if it was introduced by unobserved cluster level covariates. Zhang and Kwun (2014) discussed a flexible individual frailty model, where the multivariate exponential distributed frailties are introduced. Pan *et al.* (2015) developed the estimating equations for inferring the regression parameters in the AH model with random effects. Geerdens *et al.* (2018) suggested a local likelihood approacha to infer a parametric conditional copula whose parameter depends on a cluster-level covariate in a functional way.

On the other hand, to handle the nonlinear effects, some semi-parametric or nonparametric approaches have been proposed in the literature, such as the transformation model (Mandal *et al.* (2019)) and various versions of partially linear models. Explicitly speaking, based on the partially linear proportional hazards model, Liu *et al.* (2016) proposed a new penalised pseudo-partial likelihood method to select important covariates for multivariate failure time data. Afzal *et al.* (2017) considered partly linear AH model for left-truncated and right-censored data, and recently Afzal *et al.* (2021) proposed a hierarchical bi-level variable selection approach for right censored data in the linear part of this model, where the covariates are naturally grouped. Song *et al.* (2019) considered a partially time-varying coefficient proportional hazards model, where corrected score and conditional score approaches are employed to accommodate potential measurement error. Engebretsen and Glad (2020) used the monotone splines lasso and proposed two methods for fitting a partially linear monotone model. Zou *et al.* (2020) studied the quantile regression estimation and variable selection for the partially linear single-index models with censoring indicators missing at random. More details can be available from Cheng *et al.* (2014), Geraci (2019), Wang *et al.* (2020) and the references therein.

However, to the best of our knowledge, there is no result available in the literature for the partially linear additive hazards (PLAH) regression model with clustered survival data, which assumes that the marginally conditional hazard function depends on some covariate variables in linear relationship but is nonlinearly related to other covariates. In the following, we will consider this situation and present an estimating equation method for regression parameters in the linear part and the nonparametric part. Our method has several desirable features. First, by approximating the nonlinear function with B-splines, the proposed estimators can be obtained in a closed form. Thus our method avoids the "curse of dimensionality" problem. Second, the implementation of the proposed approach is comfortable via some existing softwares. Third, the simulation studies indicate the performance is satisfactory from the viewpoint of bias, coverage probability and average estimated error induced by approximation.

The remainder of the paper is organized as follows. The model formulation is presented in Section 2, where a system of estimating equations for making inference are proposed and the asymptotic normality of the proposed estimators is established under some regular conditions. In Section 3, simulation studies are carried out to evaluate the proposed approach under various scenarios. In Section 4, a HIV data set was analyzed as illustration. Some concluding remarks are given in Section 5. Additional simulation results are contained in the Supplementary Materials.

2. Estimation method

Consider a study consisting of n independent clusters. Let T_{ik} and C_{ik} denote the failure time and censoring time of the k-th subject in the i-th cluster, $i = 1, ..., n$, and $k = 1, ..., K$, respectively. Given the p-dimensional covariates vector $\mathbf{Z}_{ik}$ and univariate continuous covariate W_{ik}, the conditional hazard function of T_{ik} is assumed to have the following form

$$\lambda_{T_{ik}}(t|\mathbf{Z}_{ik}, W_{ik}) = \lambda_0(t) + \boldsymbol{\beta}^\top \mathbf{Z}_{ik} + \varphi(W_{ik}), \tag{2}$$

where $\boldsymbol{\beta}$ is a p-dimensional regression coefficient vector, $\lambda_0(t)$ is the unknown baseline hazard function, and $\varphi(.)$ is the nonlinear regression function, which is smooth but unspecified. Therefore, the observations consist of $(\tilde{T}_{ik} = min(T_{ik}, C_{ik}), \Delta_{ik} = I(T_{ik} \leq C_{ik}), \mathbf{Z}_{ik}, W_{ik})$ with $I(.)$ being the indicator function, $i = 1, ..., n$, and $k = 1, ..., K$. We assume that subjects in the same cluster are exchangeable.

Motivated by the work of Yin (2007), for each (i, k), we define the counting process $N_{ik}(t) = I(\tilde{T}_{ik} \leq t, \Delta_{ik} = 1)$, and the at-risk process $Y_{ik}(t) = I(\tilde{T}_{ik} \geq t)$. Denote

$$M_{ik}(t) = N_{ik}(t) - \int_0^t Y_{ik}(u)[d\Lambda_0(u) + \boldsymbol{\beta}^\top \mathbf{Z}_{ik} + \varphi(W_{ik})du],$$

which is a local square integrable martingale. If $\boldsymbol{\beta}$ and $\varphi(.)$ are known, then the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ can be estimated by

$$\hat{\Lambda}_0(t) = \sum_{i=1}^n \sum_{k=1}^K \int_0^t \frac{dN_{ik}(u) - Y_{ik}(u)[\boldsymbol{\beta}^\top \mathbf{Z}_{ik} + \varphi(W_{ik})]du}{\sum_{i=1}^n \sum_{k=1}^K Y_{ik}(u)}.$$

Due to the fact that the nonparametric component $\varphi(.)$ is a totally unspecified function, the direct application of the estimating function is not practicable. Note that using splines to model unknown function is very common in statistics. Here we propose to approximate $\varphi(.)$ with a linear combination of B-spline basis functions as follows,

$$\varphi(w) \approx \sum_{j=1}^{J} \alpha_j \tilde{B}_j(w),$$

where $(\tilde{B}_1(w), ..., \tilde{B}_J(w))^\top$ is a vector of normalized B-spline basis functions of order l with q internal knots lying in the support of W, and $J = q + l - 1$. For more details, see Schumaker (1981). To ensure the identifiability of $\varphi(.)$, we assume the requirement $E[\varphi(W)] = 0$ holds. Therefore, we consider the centering version

$$B_j(w) = \tilde{B}_j(w) - \sum_{i=1}^{n}\sum_{k=1}^{K} \tilde{B}_j(W_{ik})/(n \times K).$$

Then the final expression of $\varphi(.)$ used in the later estimation method is

$$\varphi(w) \approx \varphi_n(w) = \sum_{j=1}^{J} \alpha_j B_j(w) = \boldsymbol{\alpha}^\top \mathbf{B}(w),$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_J)^\top$ and $\mathbf{B}(w) = (B_1(w), ..., B_J(w))^\top$. Substituting φ_n for φ in (2), one obtain the proposed spline model

$$\begin{aligned} \lambda_{T_{ik}}(t|\mathbf{Z}_{ik}, W_{ik}) &= \lambda_0(t) + \boldsymbol{\beta}^\top \mathbf{Z}_{ik} + \boldsymbol{\alpha}^\top \mathbf{B}(W_{ik}) \\ &= \lambda_0(t) + \boldsymbol{\gamma}^\top \mathbf{X}_{ik} \end{aligned} \tag{3}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$, $\mathbf{X}_{ik} = (\mathbf{Z}_{ik}^\top, \mathbf{B}(W_{ik})^\top)^\top$. Following the idea of Lin and Ying (1994) and under the assumptions mentioned above, we propose to estimate $\boldsymbol{\gamma}$ by the solution, denoted by $\hat{\boldsymbol{\gamma}}_n$, to the following estimating function

$$\mathbf{U}(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \int_0^\tau \mathbf{X}_{ik} d\tilde{M}_{ik}(t),$$

where $\tilde{M}_{ik}(t)$ is obtained by replacing the Λ_0 and φ in $M_{ik}(t)$ by $\hat{\Lambda}_0$ and φ_n, respectively, τ is the end time of a study. After some algebra, we have

$$\begin{aligned} \mathbf{U}(\boldsymbol{\gamma}) &= \textstyle\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau [\mathbf{X}_{ik} - \bar{\mathbf{X}}(u)][dN_{ik}(u) - Y_{ik}(u)\boldsymbol{\gamma}^\top \mathbf{X}_{ik} du] \\ &= \textstyle\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau [\mathbf{X}_{ik} - \bar{\mathbf{X}}(u)] dN_{ik}(u) \\ &\quad - \left\{ \textstyle\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau Y_{ik}(u)[\mathbf{X}_{ik} - \bar{\mathbf{X}}(u)]^{\bigotimes 2} du \right\} \gamma, \end{aligned} \tag{4}$$

where $\mathbf{a}^{\bigotimes 2} = \mathbf{a}\mathbf{a}^\top$ for a column vector $\mathbf{a}$, and

$$\bar{\mathbf{X}}(u) = \frac{\sum_{i=1}^n \sum_{k=1}^K Y_{ik}(u)\mathbf{X}_{ik}}{\sum_{i=1}^n \sum_{k=1}^K Y_{ik}(u)}.$$

Furthermore, we can obtain an analytic closed form of the resulting estimator

$$\hat{\gamma}_n = \left\{ \sum_{i=1}^{n}\sum_{k=1}^{K} \int_0^\tau Y_{ik}(u)[\mathbf{X}_{ik} - \bar{\mathbf{X}}(u)]^{\bigotimes 2} du \right\}^{-1} \left\{ \sum_{i=1}^{n}\sum_{k=1}^{K} \int_0^\tau [\mathbf{X}_{ik} - \bar{\mathbf{X}}(u)] dN_{ik}(u) \right\}. \tag{5}$$

To derive the asymptotic normality of the proposed estimators, some notations are needed. Suppose that

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} Y_{ik}(u)\mathbf{X}_{ik}^{r}$$

uniformly converges to $\pi_r(u)$ for $u \in [0,\tau]$, $r = 0, 1$. Define

$$A_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau Y_{ik}(u)[\mathbf{X}_{ik} - \bar{\mathbf{X}}(u)]^{\bigotimes 2} du$$

and

$$\Sigma = E\left\{\left[\sum_{k=1}^{K}\int_0^\tau Y_{1k}(u)[\mathbf{X}_{1k} - \frac{\pi_1(u)}{\pi_0(u)}]dM_{1k}(u)\right]^{\bigotimes 2}\right\},$$

and assume that A_n converges in probability to a nonsingular deterministic matrix A. Under the above assumptions, and suppose the nonlinear regression function φ is indeed a B-spline function, i.e., the function φ_n, similar to the proof of Theorem 1 in Yin and Cai (2004), one can show that

$$n^{1/2}(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0) \to N(0, A^{-1}\Sigma A^{-1}), \tag{6}$$

in distribution as $n \to \infty$, where $\boldsymbol{\gamma}_0 = (\boldsymbol{\beta}_0^\top, \boldsymbol{\alpha}_0^\top)^\top$ is the true value. A consistent estimator of the covariance matrix can be obtained by substituting A and Σ by A_n and

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}\int_0^\tau [\mathbf{X}_{ik} - \bar{\mathbf{X}}(u)]d\hat{M}_{ik}(u)\right]^{\bigotimes 2},$$

where $\hat{M}_{ik}(u)$ is obtained from $\tilde{M}_{ik}(u)$ with some unspecified quantities therein replaced by their estimators. Based on these results, one can construct the 95% confidence intervals of β_0 and α_0, thus the point-wise confidence interval of $\varphi(w_0)$ at a fixed value w_0 can be obtained.

Remark 1. How to calculate the integral in equation (5) is critical. Combining the definitions of $Y_{ik}(u)$ and $N_{ik}(u)$, the integral in the second term on the right hand of the equality (5) becomes a summarization of integrand at every jump $\tilde{T}_{ik}$. The difficulty mainly arises from the first term. But note that $Y_{ik}(u)$ is a piecewise constant function, whose points of division are the ordered different observation times $\tilde{T}_{ik}$, we can exchange the integral and the summarization, and divide the integral interval into several subintervals. At this time, the integrand in each subinterval becomes a constant. Thus the computation can be conducted. More details can be found in the section 3.1 of page 6 in Anders and Scheike (2012), where the R package ahaz is introduced and is used to finish our simulation studies.

Remark 2. In model (2), the covariates are presumed to be time-independent. In fact, this requirement can be relaxed and easily extended to the case, where some time-dependent covariates are observed, in line with the strategy in Yin and Cai (2004).

Remark 3. To conclude the asymptotic normality of the proposed estimators, we impose the nonlinear regression function φ is a B-spline function, where the knots and order are assumed to be known. The primary purpose is to avoid the complex derivative of the limiting property, and the same skill is also adopted in other areas, such as Afzal *et al.* (2017,2021). Even if this requirement does not hold strictly, the following simulation studies demonstrate that the approximation error can be ignorable.

Remark 4. Selection of the optimal number and locations of interior knots and the order of spline often is consider to be important for bringing superior numerical performance when applying the spline method. As argued by Lu and McMahan (2018), The order controls the overall smoothness of the spline estimator, whereas the knot set specification controls model flexibility. Too many (few) knots lead to over (under) fitting. In this paper, quadratic B-splines was used and the interior knots with fixed number 3 were placed at the equally-spaced quantiles of the observations. Although one can select optimal number of knots through some model selection criterions, as done in the subsection 2.3 in Lu and McMahan (2018), we find from our experience that if the observed covariate W_{ik} is not sparse and skewed seriously, the affection induced by different choices of number of interior knots is not serious, and can be omitted, which is also displayed in the following Table 3.

Remark 5. The computation of the proposed method can be easily implemented by the R packages *Splines* and *ahaz*, where the function *bs* and *ahaz* (Anders and Scheike (2012)) are used in this paper.

3. Simulation studies

In this section, we conducted simulation studies to evaluate the finite-sample performance of the proposed estimation procedure in different settings. The simulation set-up is partly adapted from Afzal *et al.* (2017), Johnson and Strawderman (2009) and Yin (2007). In all simulation settings, we consider the situation of $K = 2$, i.e., the cluster size was two, and the marginally conditional hazard functions of the failure times T_{ik} given the covariates $(\mathbf{Z}_{ik}, W_{ik})$ were

$$\lambda_{T_{ik}}(t|\mathbf{Z}_{ik}, W_{ik}) = \lambda_0(t) + \boldsymbol{\beta}_0^\top \mathbf{Z}_{ik} + \varphi_0(W_{ik}), \tag{7}$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02})^\top$, $\mathbf{Z}_{ik} = (Z_{1ik}, Z_{2ik})^\top$, $\lambda_0(t)$ is the baseline hazard function, and $\varphi_0(.)$ is the nonlinear regression function. Specifically, for each cluster, we generated the two failure times under the Clayton copula function from the bivariate distribution function

$$F(t_1, t_2) = [F_1(t_1) + F_2(t_2) - 1]^{-1/\theta},$$

where $F_k(t_k) = \exp(-\Lambda_0(t_k) - \boldsymbol{\beta}_0^\top \mathbf{Z}_{ik} t_k - \varphi_0(W_{ik}) t_k)$, $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$, and $\theta(> 0)$ is the correlation parameter, which takes differen values to yield the Kendall's τ. All covariates and censoring times involved are generated independently. The true regression coefficients $\boldsymbol{\beta}_0$ and function $\varphi_0(.)$ are determined explicitly in the following Examples 1-2, respectively.

In these simulations, quadratic B-splines were used to approximate the regression function, where the interior knots with fixed number 3 were placed at the equally-spaced quantiles of the observations of W_{ik}. For each study design, $N = 1000$ independent data sets based on the above-mentioned model with sample sizes $n = 100, 200, 400, 800$ were generated and analyzed using the proposed approach. We calculated the estimated biases (BIAS), the sample standard deviations (SD), the average of the estimated standard errors (SE) and the 95% empirical coverage probabilities (CP) for the regression coefficient estimators. In addition, for the nonlinear part, we recorded the averaged integrated absolute bias (IABIAS) for a estimator $\hat{\varphi}_j(.)$, $(j = 1, ..., N)$, defined by

$$\text{IABIAS} = \frac{1}{N \times ngrid} \sum_{j=1}^{N} \sum_{i=1}^{ngrid} |\hat{\varphi}_j(w_i) - \varphi(w_i)|,$$

where $ngrid$ is the number of grid points w_i between 3 and 9 with step length 0.1. We also plotted the 95% point-wise confidence intervals and the average estimated curves of φ_0.

Example 3.1. Based on the above model (7), we specified the variables and parameters as follows. $\lambda_0(t) = 5$, $(\beta_{01}, \beta_{02})^\top = (0.3, 0.5)^\top$, covariates $(Z_{1ik}, Z_{2ik})^\top$ are generated from $Z_{1ik} \sim Bernoulli(0.5)$, $Z_{2ik} \sim Unif(-1, 1)$, and $\varphi_0(W) = 0.3((W - 6)^2 - 3)$ with $W \sim Unif(3, 9)$. The censoring time C_{ik} follows a uniform distribution $Unif(0, a)$ with the constant a chosen to obtain average right censoring rates of about 20% and 50%, respectively. The correlation parameter θ takes values 0, 0.5 and 3 to yield the Kendall's $\tau = 0, 0.2, 0.6$, respectively.

Table 1 presents the results on estimations of regression parameters, and the corresponding estimated curves are shown in Fig. 1. It can be seen that the proposed estimators of $(\beta_{01}, \beta_{02})^\top$ seem to be unbiased, the coverage probabilities almost reach the nominal level 0.95 irrespective to the sample size and censoring rate, and the sample standard deviations and the average of the estimated standard errors are in agreement, which demonstrates the proposed variance estimates are reasonable. At the same correlation and right censoring rate, as sample size increases, the proposed estimators tend to have smaller BIAS, SD, SE and IABIAS. The average estimated curves almost overlapped the true curve on the entire range, and the space between the upper and lower 95% point-wise confidence intervals become narrower with an increase of sample size. In addition, more figures can be available from the author upon request.

Example 3.2. In this example, we specified the variables and parameters as follows. $\lambda_0(t) = t$, $(\beta_{01}, \beta_{02})^\top = (0.5, 1)^\top$, covariates $(Z_{1ik}, Z_{2ik})^\top$ are generated from $Z_{1ik} \sim$ *Bernoulli* (0.5), $Z_{2ik} \sim |N(0, 0.4)|$, and $\varphi_0(W) = \sin(\pi(W/3-1))$ with $W \sim Unif(3, 9)$. The censoring time C_{ik} also follows a uniform distribution $Unif(0, a)$ with the constant a chosen to obtain average right censoring rates of about 20% and 52%, respectively. Here the correlation parameter θ still takes values 0, 0.5 and 3 to yield the Kendall's $\tau = 0, 0.2, 0.6$, respectively.

Table 2 presents the results on estimations of regression parameters, and the corresponding estimated curves are shown in Fig. 2. One can see that the results are similar to those presented in the Example 1. It is worthwhile to point out that compared with that in Example 1, the IABIAS is smaller and the estimated curves are more close to the true one and have a narrower 95% point-wise confidence intervals under each scenario. In addition, more figures can be available from the author upon request.

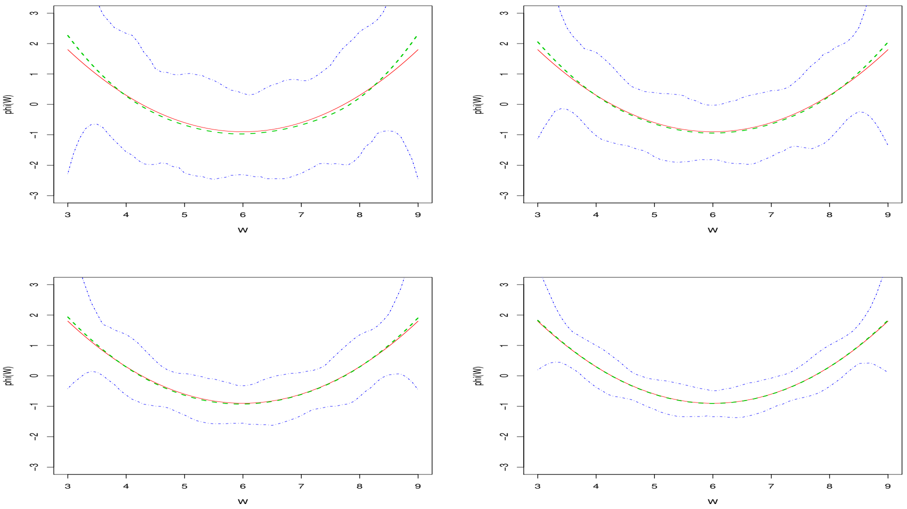

Figure 1: The solid line (red) is the true curve. The dashed lines (green) are average estimated curves for $n = 100, 200, 400, 800$ with 1000 duplications for Example 3.1, under (θ, CR)=(0.5, 20%) by row, respectively. The dash-dotted lines (blue) are the 95% point-wise confidence intervals

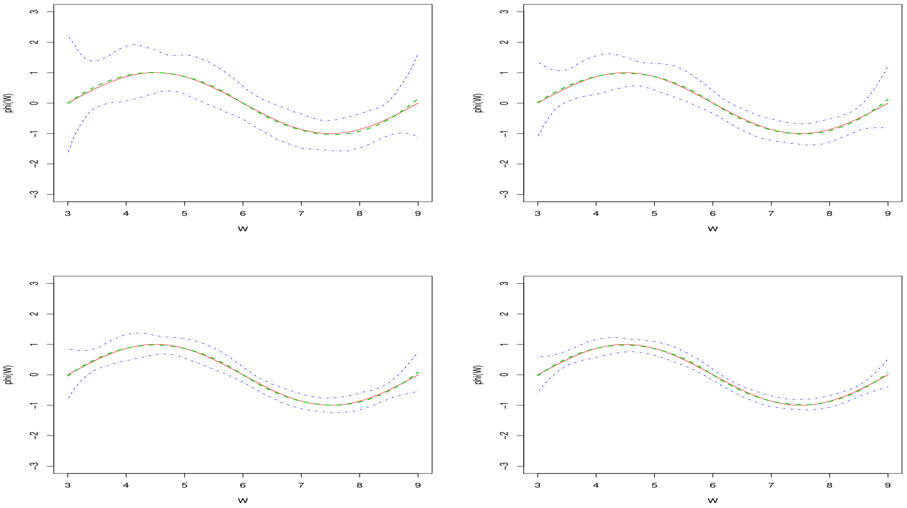

Figure 2: The solid line (red) is the true curve. The dashed lines (green) are average estimated curves for $n = 100, 200, 400, 800$ with 1000 duplications for Example 3.2, under (θ, CR)=(0.5, 20%) by row, respectively. The dash-dotted lines (blue) are the 95% point-wise confidence intervals

Table 1: Simulation results of $(\beta_{01}, \beta_{02}) = (0.3, 0.5)$ and $\varphi(.)$ in Example 3.1

			β_{01}				β_{02}				$\varphi(.)$
n	θ	CR	BIAS	SE	SD	CP	BIAS	SE	SD	CP	IABIAS
50	0	20%	0.029	0.861	0.892	0.943	0.013	0.752	0.748	0.949	0.861
		50%	0.071	1.077	1.061	0.959	-0.003	0.936	0.961	0.951	1.013
	0.5	20%	0.011	0.852	0.848	0.949	0.015	0.741	0.747	0.952	0.812
		50%	-0.009	1.077	1.093	0.949	0.021	0.939	0.920	0.959	1.017
	3	20%	-0.023	0.859	0.868	0.950	-0.006	0.748	0.748	0.958	0.818
		50%	0.059	1.072	1.060	0.956	0.029	0.935	0.945	0.953	1.014
200	0	20%	0.038	0.584	0.581	0.960	0.001	0.508	0.521	0.948	0.550
		50%	-0.030	0.735	0.764	0.942	0.012	0.635	0.646	0.939	0.681
	0.5	20%	0.029	0.584	0.586	0.945	0.018	0.507	0.512	0.953	0.560
		50%	0.010	0.739	0.735	0.951	-0.011	0.640	0.633	0.955	0.677
	3	20%	0.008	0.588	0.596	0.950	-0.000	0.510	0.519	0.953	0.559
		50%	-0.027	0.736	0.740	0.950	0.006	0.638	0.619	0.961	0.686
400	0	20%	-0.014	0.405	0.401	0.954	0.006	0.351	0.347	0.958	0.386
		50%	-0.020	0.514	0.523	0.951	-0.005	0.446	0.451	0.950	0.480
	0.5	20%	0.008	0.406	0.422	0.947	0.009	0.353	0.359	0.952	0.386
		50%	-0.007	0.514	0.524	0.951	0.003	0.445	0.449	0.962	0.470
	3	20%	0.017	0.406	0.412	0.949	0.021	0.352	0.357	0.940	0.377
		50%	0.015	0.514	0.500	0.959	0.014	0.445	0.449	0.962	0.478
800	0	20%	-0.009	0.284	0.279	0.950	-0.011	0.246	0.247	0.953	0.273
		50%	0.008	0.360	0.358	0.952	0.000	0.312	0.306	0.953	0.330
	0.5	20%	0.019	0.284	0.274	0.958	-0.014	0.246	0.241	0.959	0.267
		50%	0.009	0.360	0.359	0.951	-0.009	0.312	0.304	0.957	0.330
	3	20%	0.001	0.285	0.271	0.961	0.000	0.247	0.254	0.939	0.263
		50%	-0.008	0.359	0.361	0.950	0.015	0.311	0.303	0.951	0.330

Notes: BIAS: the estimated biases; SD: sample standard deviation; SE: average standard error estimate; CP: the 95% empirical coverage probability; IABIAS: the averaged integrated absolute bias; n: sample size; CR: the right censoring rate.

4. Real data analysis

In this section, we illustrate the proposed estimation method by analyzing a real data set. For technical details on modelling specifications, we follow the choices used in the simulation studies in Section 3.

The data set comes from a recent clinical trial, which was described by Guo and Carlin (2004), and the primary objective is to compare the efficacy and safety of two antiretroviral drugs, i.e. didanosine (ddI) or zalcitabine (ddC), in treating patients who had failed or were intolerant of zidovudine (AZT) therapy. In this study, a total of 467 HIV-infected patients were enrolled and randomly assigned to receive either ddI or ddC. CD4 counts were recorded at study entry and again at 2, 6, 12, and 18-month visits, and the times to death were also recorded. Therefore, one patient is treated as a cluster. And the data set can be available in JM package in statistical software R. In this paper, we include five covariates as main effects in our analysis: CD4 counts, observation time at which the CD4 cells count was recorded (obstime), drug (ddI = 1, ddC = 0), gender (male = 1, female = 0), PrevOI (previous opportunistic infection (AIDS diagnosis) at study entry = 1, no AIDS diagnosis = 0), and AZT (AZT failure = 1, AZT intolerance = 0). Among of them, only CD4 and obstime are subject-specific covariates, and others are cluster-level covariates. Let T_i denote the times to death of the ith patient, we

Table 2: Simulation results of $(\beta_{01}, \beta_{02}) = (0.5, 1)$ and $\varphi(.)$ in Example 3.2

			β_{01}				β_{02}				$\varphi(.)$
n	θ	CR	BIAS	SE	SD	CP	BIAS	SE	SD	CP	IABIAS
100	0	20%	0.020	0.298	0.295	0.960	0.053	0.678	0.678	0.960	0.299
		50%	0.001	0.352	0.353	0.958	0.075	0.801	0.808	0.944	0.333
	0.5	20%	0.023	0.296	0.284	0.967	0.095	0.678	0.694	0.944	0.287
		50%	0.008	0.351	0.349	0.959	0.028	0.796	0.818	0.948	0.328
	3	20%	0.018	0.298	0.300	0.954	0.090	0.685	0.713	0.941	0.289
		50%	0.014	0.353	0.358	0.952	0.086	0.802	0.807	0.954	0.334
200	0	20%	0.011	0.204	0.206	0.954	0.041	0.465	0.452	0.958	0.203
		50%	0.016	0.241	0.244	0.947	0.032	0.547	0.529	0.967	0.225
	0.5	20%	0.010	0.204	0.200	0.958	0.040	0.463	0.467	0.948	0.196
		50%	0.011	0.241	0.240	0.949	0.012	0.547	0.534	0.948	0.226
	3	20%	0.008	0.204	0.204	0.949	0.035	0.463	0.465	0.945	0.199
		50%	0.015	0.241	0.244	0.948	0.042	0.545	0.551	0.951	0.224
400	0	20%	0.005	0.142	0.138	0.949	0.031	0.322	0.323	0.953	0.139
		50%	0.004	0.168	0.163	0.955	-0.007	0.379	0.376	0.948	0.156
	0.5	20%	0.003	0.142	0.141	0.960	0.006	0.322	0.332	0.949	0.136
		50%	0.003	0.168	0.172	0.943	0.015	0.381	0.383	0.951	0.158
	3	20%	-0.004	0.142	0.143	0.952	0.021	0.323	0.321	0.955	0.136
		50%	0.001	0.168	0.166	0.948	0.020	0.380	0.387	0.956	0.156
800	0	20%	-0.001	0.100	0.099	0.954	0.018	0.226	0.229	0.947	0.097
		50%	0.005	0.118	0.121	0.959	0.009	0.266	0.256	0.950	0.110
	0.5	20%	-0.001	0.099	0.104	0.936	0.007	0.226	0.231	0.939	0.095
		50%	0.007	0.118	0.120	0.946	-0.008	0.267	0.254	0.961	0.111
	3	20%	0.002	0.099	0.099	0.953	0.004	0.225	0.223	0.948	0.097
		50%	0.005	0.118	0.116	0.954	0.007	0.267	0.267	0.936	0.109

Notes: BIAS: the estimated biases; SD: sample standard deviation; SE: average standard error estimate; CP: the 95% empirical coverage probability; IABIAS: the averaged integrated absolute bias; n: sample size; CR: the right censoring rate.

analyze the data by fitting the following PLAH model

$$\begin{aligned}\lambda_{T_i}(t|\mathbf{Z}_{ik}, CD4_{ik}) = \lambda_0(t) + \beta_1 obstime_{ik} + \beta_2 drug_i + \beta_3 gender_i \\ + \beta_4 prevOI_i + \beta_5 AZT_i + \varphi(CD4_{ik}),\end{aligned}$$

where $\mathbf{Z}_{ik} = (obstime_{ik}, drug_i, gender_i, prevOI_i, AZT_i)^\top$.

Table 3 and Fig 3 present the estimation results in terms of the regression parameters in the cases that the degree is taken to be 2 and the number of interior knots is taken to be 3, 5, and 8. It is clear that the proposed method produces very close estimates in these cases, which indicate that the method is robust to the choice of number of knots. one can see that the covariate effects of both drug and prevOI are statistically significant at level 0.05, which means that the ddI group has a higher risk than the ddC group, thus the ddC group has a little better survival, and patients who had a negative AIDS diagnosis at study entry have better average survival rates than those who had a positive diagnosis. These conclusions also were found in Section 3.2 of Guo and Carlin (2004), where they used the variable Stratum, which is the same as the covariate prevOI denoted here. In addition, the covariate obstime seems to be also a significant risk factor and behaves similar to those obtained by Fu *et al.* (2021). Fig. 3 indicates that the risk caused by CD4 count decreases as the number of CD4 increases before it reaches around 15. Not surprisingly, CD4 count often is seen as a protective biomarker for preventing progression to AIDS. It is worthwhile to point out that the effect of CD4 seems to not

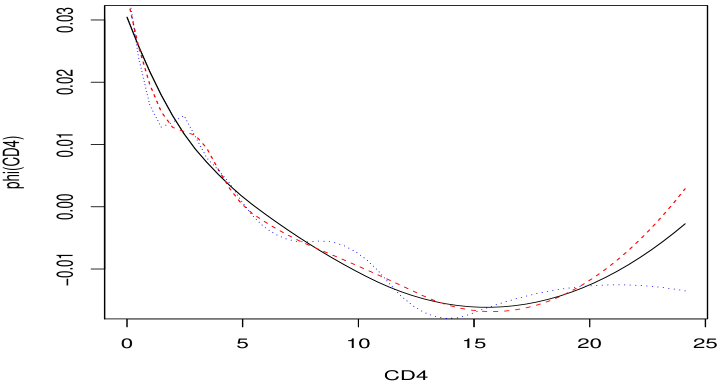

Figure 3: The estimated curves of $\varphi(.)$ with the number of interior knots being equal to 3 (gray solid line), 5 (red dashed line), 8 (blue dotted line), respectively when analyzing the HIV data.

be linear, and the log transformation and square root transformation of CD4 are used for HIV data analysis by Mandal *et al.* (2019) and Guo and Carlin (2004), respectively.

As pointed out by one reviewer, when the number of interior knots is set to be 3 and 5, the estimated function seems to have a increasing trend after CD4 taking value 20. In contrast to it, the performance under 8 knots is fairly flat. we compute the maximum of CD4 in the sample and quantilea at different levels τ, the results are listed in the Table 4. From it, we can see that the maximum of CD4 in the sample is far away from the other values and the number of observations around the maximum is almost null. Therefor, when the number of knots is small, it will bring relatively large separation between the knots and cause the underfitting in the right tail.

Table 3: Esimation results of regression coefficients under the PHAH model for the HIV data

		$q=3$			$q=5$			$q=8$	
Parameter	Est	SE	p-value	Est	SE	p-value	Est	SE	p-value
obstime (β_1)	-0.0023	0.0001	0.0000	-0.0023	0.0001	0.0000	-0.0023	0.0001	0.0000
drug (β_2)	0.0044	0.0021	0.0339	0.0044	0.0021	0.0339	0.0046	0.0021	0.0295
gender (β_3)	-0.0028	0.0039	0.4634	-0.0030	0.0039	0.4467	-0.0035	0.0040	0.3745
prrvOI (β_4)	0.0117	0.0026	0.0000	0.0116	0.0026	0.0000	0.0117	0.0026	0.0000
AZT (β_5)	0.0022	0.0032	0.4814	0.0023	0.0032	0.4610	0.0024	0.0032	0.4550

Table 4: Summary of the variable CD4 in the HIV data

Min	Q1	Q2	Mean	Q3	Max	$\tau=0.9$	$\tau=0.95$	$\tau=0.99$	$\tau=0.999$
0	3.162	5.477	7.023	10.440	24.125	15.205	16.763	19.234	19.847

5. Conclusion

In this paper, we developed a well-defined estimation method for partially linear additive hazards model to analyze clustered failure time data. B-splines are used to approximate the unknown regression function, which significantly reduces the number of unknown parameters while maintaining adequate modelling flexibility. Simulation studies show that the proposed method estimates all regression parameters accurately and efficiently. Specially, the estimation method is robust to the choice of number of knots, and can been easily implemented using some existing softwares.

Technically, the proposed method can be extended to the scenario where multiple covariates have nonlinear effects on the conditional hazard function, i.e.,

$$\lambda_{T_{ik}}(t|\mathbf{Z}_{ik}, W_{ik}) = \lambda_0(t) + \beta_0^\top \mathbf{Z}_{ik} + \sum_{j=1}^{J} \varphi_j(W_{ijk}),$$

where $\varphi_j(.)$ are the nonlinear regression functions for the covariate W_{ijk}. However, some cautions are needed for handing the knots, where the dimension of the covariates induced by introducing of spline basis functions may be high. In addition, the method presented here can adapt to other types of survival data, such as current status data.

Topics for future work also include development of goodness-of-fit test and model diagnostics. For example, how to select the covariates who have nonlinear effects in the model is important and how to determine the optimal model in terms of determination of number and locations of knots. Besides, When the working independence assumption does not hold, such as with the informative cluster size, it is necessary to develop new inference methodology.

Acknowledgement

The author is deeply grateful to the referee and the editor for their helpful comments and suggestions. The research was supported by the Young Natural Science Research Program of Jiangsu University of Science and Technology.

References

Aalen, O., Borgan, O. and Gjessing, H. (2008). *Survival and event history analysis: A process point of view*, Springer.

Afzal, A. R., Dong, C. and Lu X. (2017). Estimation of partly linear additive hazards model with left-truncated and right-censored data. *Statistical Modelling*, **6**, 423-448.

Afzal, A. R., Yang, J. and Lu, X. (2021). Variable selection in partially linear additive hazards model with grouped covariates and a diverging number of parameters. *Computational Statistics*, **36**, 829-855.

Anders, G. R. and ScheikeT. H. (2012). Coordinate descent methods for the penalized semiparametric additive hazards model. *Journal of Statistical Software*, **47**, 1-17.

Buckley, J. (1984). Additive and multiplicative models for relative survival rates. *Biometrics*, **40**, 51-62.

Cheng, G., Zhou, L. and Huang, J. Z. (2014). Efficient semiparametric estimation in generalized partially linear additive models for longitudinal/clustered data. *Bernoulli*, **20**, 141-163.

Engebretsen, S. and Glad, I. K. (2020). Partially linear monotone methods with automatic variable selection and monotonicity direction discovery. *Statistics in Medicine*, **39**, 3549-3568.

Eriksson, F., Gerds, T. A. and Lesaffre, E. (2014). Unobserved confounder effects in models for clustered dental failure time data. *Statistical Modelling*, **6**, 549-566.

Fu, L., Yang, Z., Zhou, Y. and Wang, Y. G. (2021). An efficient gehan-type estimation for the accelerated failure time model with clustered and censored data. *Lifetime Data Analysis*, 1-31.

Geerdens, C., Acar, E. and Janssen, P. (2018). Conditional copula models for right-censored clustered event time data. *Biostatistics*, **19**, 247-262.

Geraci, M. (2019). Modelling and estimation of nonlinear quantile regression with clustered data. *Computational Statistics & Data Analysis*, **136**, 30-46.

Guo, X. and Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, **58**, 16-24.

Johnson, L. M. and Strawderman, R. L. (2009). Induced smoothing for the semiparametric accelerated failure time model: asymptotics and extensions to clustered data. *Biometrika*, **3**, 577-590.

Kulich, M. and Lin, D.Y. (2000). Additive hazards regression for case-cohort studies. *Biometrika*, **1**, 73-87.

Li, H. and Yin, G. (2009). Generalized method of moments estimation for linear regression with clustered failure time data. *Biometrika*, **2**, 293-306.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61-71.

Liu, J., Zhang, R., Zhao, W. and Lv, Y. (2016). Variable selection in partially linear hazard regression for multivariate failure time data. *Journal of Nonparametric Statistics*, **2**, 375-394.

Lu, M. and McMahan, C. S. (2018). A partially linear proportional hazards model for current status data. *Biometrics*, **4**, 1240-1249.

Mandal, S., Wang, S. and Sinha, S. (2019). Analysis of linear transformation models with covariate measurement error and interval censoring. *Statistics in Medicine*, **38**, 4642-4655.

Moertel, C., Fleming, T., Macdonald, J., Haller, D., Laurie, J., et al. (1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *The New England Journal of Medicine*, **6**, 352-358.

Pan, D., Liu, Y. Y. and Wu, Y. S. (2015). Additive hazards regression with random effects for clustered failure times. *Acta Mathematica Sinica*, **31**, 511-525.

Schumaker, L. (1981). *Spline Functions: Basic Theory*, New York: Wiley.

Song, X., Wang, L., Ma, S. and Huang, H. (2019). Variable selection for partially linear proportional hazards model with covariate measurement error. *Journal of Nonpara-*

metric Statistics, **1**, 196-220.

Wang, X., Song, Y. and Zhang, S. (2020). An efficient estimation for the parameter in additive partially linear models with missing covariates. *Journal of The Korean Statistical Society*, **49**, 779-801.

Yin, G. (2007). Model checking for additive hazards model with multivariate survival data. *Journal of Multivariate Analysis*, **98**, 1018-1032.

Yin, G. and Cai, J. (2004). Additive hazards model with multivariate failure time data. *Biometrika*, **91**, 801-818.

Zeng, D. and Cai, J. (2010). Additive transformation models for clustered failure time data. *Lifetime Data Analysis*, **3**, 333-352.

Zhang, R. and Kwun, C. G. C. (2014). A marginalizable frailty model for correlated right-censored data. ArXiv Preprint ArXiv:1403.6744.

Zou, Y., Fan, G. and Zhang, R. (2020). Quantile regression and variable selection for partially linear single-index models with missing censoring indicators. *Journal of Statistical Planning and Inference*, **204**, 80-95.

Received: August 17, 2022
Revised: October 12, 2022
Accept: October 17, 2022