Proportional Hazards Regression with Unknown Link Function

By WEI WANG

Harvard Medical School and Brigham and Women's Hospital, Boston, USA

wwang@partners.org

JANE-LING WANG

University of California, Davis, USA

wang@wald.ucdavis.edu

AND QIHUA WANG

Chinese Academy of Science, Beijing, China

qhwang@amss.ac.cn

The University of Hong Kong, Hong Kong

qhwang@hku.hk

SUMMARY

Proportional hazards regression model assumes that the covariates affect the hazard function through a link function and an index which is a linear function of the covariates. Traditional approaches, such as the Cox proportional hazards model, focus on estimating the unknown index by assuming a known link function between the log-hazard function and covariates. A linear link function is often employed for convenience without any validation. This paper provides an approach to estimate the link function, which can then be used to guide the choice of a proper parametric link function. This is accomplished through a two-step algorithm to estimate the link function and the effects of the covariates iteratively without involving the baseline hazard estimate. The link function is estimated by a smoothing method based on a local version of partial likelihood, and the index function is then estimated using a full version of partial likelihood. Asymptotic properties of the nonparametric link function estimate are derived, which facilitates model checking of the adequacy of the Cox Proportional hazards model. The approach is illustrated through a survival data and simulations.

Keywords: Partial likelihood, local partial likelihood; Nonparametric smoothing; Dimension reduction.

1. Introduction

Proportional hazards regression model has played a pivotal role in survival analysis since Cox proposed it in 1972. Let T represent survival time and Z its associate covariate vector. Under the proportional hazards model, the hazard function for T, given a particular value z for the covariate Z, is defined as

$$\lambda\{t \mid z\} = \lambda_0(t) \exp\big\{\psi({\beta_0}^T z)\big\}, \tag{1}$$

where $\lambda_0(t)$ is an unknown baseline hazard function corresponding to $z = (0, \cdots, 0)$, and $\psi(\cdot)$ is called the link function with $\psi(0) = 0$. With fully specified link function ψ, the partial likelihood method was introduced in [4, 5] to estimate the regression parameters, β_0, with the option to accommodate censored data. The most common choice for ψ is the identity function, which corresponds to the time-honored Cox model. In reality the link function is unknown and needs to be estimated. This is especially useful to validate a preferred choice, as an erroneous link function could dramatically distort risk assessment or interpretation of odds ratios. When the link function is known, such as in the Cox model, model (1) is a special case of the transformation model first proposed in [7] and subsequently studied in [6], [3] etc. Our goal in this paper is to consider model (1) with an unknown link function. This problem was first studied in an unpublished Ph.D. thesis [19]. However, the procedure there was less efficient and we propose an improved estimate, studying its asymptotic properties.

Previous work focuses on the special case when the covariate is one-dimensional, or equivalently when β is known in (1). Under this special one-dimensional case, a local partial likelihood technique in [18] and a variation of the local scoring algorithm of [12] can be used to estimate the unknown link function in (1). Gentleman and Crowley [10] proposed a local version of the full likelihood instead of partial likelihood by alternating between estimating the baseline hazard function and estimating the covariate effects. The local likelihood methods in these papers were based on data whose covariate values fall in a neighborhood of the targeted location. Fan, Gijbels and King [8] used instead a local polynomial method to approximate the local partial likelihood, and derived rigorous asymptotic results for their approach. Simulation studies there showed that the local partial likelihood method is comparable to the local full likelihood method in [10]. Two spline approaches have also been considered, with smoothing splines resulting from a penalized partial likelihood in O'Sullivan [16] and regression splines from [17].

While the aforementioned approaches can be easily extended to q-dimensional covariates by estimating a multivariate unknown link function $\psi(z_1, \cdots, z_q)$, such nonparametric approaches are subject

to the curse of dimensionality and may not be suitable for $q \geq 3$. Moreover, the resulting model would be different from model (1), which has the attractive dimension reduction feature that the covariate information is succinctly summarized in a single index and is a nonparametric extension of the Cox proportional hazards model. Model (1) could also be used as an exploratory tool to guide the choice of a suitable parametric link function.

A two-step iterative algorithm to estimate the link function and the covariate effects is proposed in Section 2. In the first step, an initial estimate of the regression parameter β is plugged in model (1) so that the link function can be estimated by a smoothing method based on a local version of partial likelihood ([4, 5]). The second step involves updating the regression parameters using the full partial likelihood with the estimated link function in step 1 inserted. These two steps will be iterated until the algorithm converges. Asymptotic results for the link estimators are stated in Section 2. In particular, Theorem 2 provides the building blocks to check the link function and inference for the individual risk, $\psi(\beta^T z)$. It also reveals that the nonparametric estimate of the link function is as efficient as the one for model (1) but with a known regression parameter β. Thus, there is no efficiency loss to estimate the link function even if β is unknown in our setting. This is also reflected in the simulation studies in Section 3. The approach in Section 2 is further illustrated in Section 4 through a data set from the Worcester Heart Attach Study. All the proofs of the main results are relegated to an appendix.

We remark here that [15] also studied model 1 with a different approach. They assumed that the link function is in a finite dimensional subspace spanned by polynomial spline bases functions and the dimension of this subspace is known. This leads to a flexible parametric model where the spline coefficients corresponding to the link functions and β can then be estimated directly through traditional partial likelihood approaches. While this has the benefit of simplicity as everything is in the parametric framework, it tends to underestimate the standard errors of the estimates. Two sources of bias arise, one derives from the fact that in reality the number of spline bases depends on the data and is a function of the sample size, so the standard errors are underestimated by the simple parametric inference. In addition, the link estimation might be biased ,as in theory an infinite number of spine bases might be required to span the unknown link function. These biases could significantly affect the asymptotic results. In contrast, our approach provides correct asymptotic theory and efficient estimation of the link function.

2. Estimation procedure and main results

Since there are three different unknown parameters, $\lambda_0(\cdot)$, $\psi(\cdot)$ and β in model (1), we need to

impose some conditions to ensure identifiability. To identify λ_0, it suffices to set $\psi(v) = 0$ at some point v, a common choice is $v = 0$. Since only the direction of β is identifiable if ψ is unknown, we assume that $\|\beta\|$=1 (here $\|\cdot\|$ represents the Euclidean norm) and that the sign of the first component of β is positive. As for the sampling plan, we assume an independent censoring scheme, in which the survival time T and censoring time C are conditionally independent, given the covariate vector Z. Let $X = min(T, C)$ be the observed event-time and $\Delta = I\{T \leq C\}$ be the censoring indicator. The data $\{X_i, Z_i, \delta_i\}$ is an i.i.d. sample of $\{X, Z, \Delta\}$. We use the notation $t_i < \cdots < t_N$ to denote the N distinctive ordered failure times, and (j) to denote the label of the item failing at time t_j. The risk set at time t_j is denoted by $\mathcal{R}_j = \{i : X_i \geq t_j\}$.

For a fixed parametric value β, one can estimate the link function $\psi(\cdot)$ by any smoothing method, such as those cited in Section 1 when β is assumed known. We adopt the local partial likelihood approach in [8] and assume, for a given point v, that the p-th order derivative of $\psi(v)$ at point v exists. A Taylor expansion for $\beta^T Z$ in a neighborhood of v then yields,

$$\begin{aligned} \psi(\beta^T Z) &\approx \psi(v) + \psi^{'}(v)(\beta^T Z - v) + \cdots + \frac{\psi^{(p)}(v)}{p!}(\beta^T Z - v)^P \\ &= \psi(v) + (\beta^T \mathbf{Z})^T \gamma(v), \end{aligned} \tag{2}$$

where $\gamma(v) = \{\psi'(v), \cdots, \psi^{(p)}(v)/p!\}^T$ is the $p-$dimensional vector associated with the derivatives of ψ and $\beta^T \mathbf{Z} = \{\beta^T Z - v, \cdots, (\beta^T Z - v)^p\}^T$.

Let K be a kernel function, h be a bandwidth, and define $K_h(u) = h^{-1}K(u/h)$. Applying kernel weights to the logarithm of the global partial likelihood

$$\sum_{j=1}^{N} \psi(\beta^T Z_{(j)}) - \log\left\{ \sum_{i \in \mathcal{R}_j} \exp\Big\{\psi(\beta^T Z_{(j)})\Big\} \right\}$$

and replacing $\psi(\beta^T z)$ by the local approximation in (2), we arrive at (similarly to [8]) the local version of the log partial likelihood:

$$\sum_{j=1}^{N} K_h\big(\beta^T Z_{(j)} - v\big) \left[\big(\beta^T \mathbf{Z}_{(j)}\big)^T \gamma(v) - \log\left\{ \sum_{i \in \mathcal{R}_j} \exp\Big\{ \big(\beta^T \mathbf{Z}_i\big)^T \gamma(v) \Big\} K_h\{\beta^T Z_i - v\} \right\} \right], \tag{3}$$

where $\beta^T \mathbf{Z}_i$ and $\beta^T \mathbf{Z}_{(j)}$ are defined as $\beta^T \mathbf{Z}$ with $\mathbf{Z}$ replaced by $\mathbf{Z}_i$ and $\mathbf{Z}_{(j)}$ respectively. It can be shown that the local log partial likelihood in (3) is strictly concave with respect to $\gamma(\cdot)$, so for a fixed β, it has a unique maximizer with respect to γ. Let $\hat{\gamma}(v)$ be the local partial likelihood estimate with $\hat{\gamma}_k(v)$ denoting its k-th component, then $\psi^{(k)}(v)$ can be estimated by $\hat{\psi}^{(k)}(v) = k!\ \hat{\gamma}_k(v)$, for $k = 1, \cdots, p$. In principle, one could maximize (3) with respect to both β and γ, and this corresponds to maximizing the real local log likelihood. But we choose to maximize (3) only with respect to γ for

a fixed estimated value of β, and this corresponds to maximizing a pseudo local log likelihood as the true β in (3) is replaced by an estimate. There are two reasons for our choice. First (3) is concave in γ, but not necessarily in β. Second, maximizing with respect to both parameters is probably not worth the additional computational cost, as the local likelihood procedures mainly serve as a smoother and the choice of the smoother is usually not crucial.

To estimate the link function, we use

$$\hat{\psi}(v) = \int_0^v \hat{\psi}'(w)dw,$$

where $\hat{\psi}'(v)$ is the first component of $\hat{\gamma}(v)$ at the last iteration step. There are several ways to approximate this integral, such as the trapezoidal rule or Gaussian quadrature. For computational simplicity, we apply the trapezoidal rule in the simulation studies, as suggested in [18], and this appears to be satisfactory.

2.1. Algorithm and computational issues

The procedure described in the previous subsection requires a certain choice of β in equation (2). This can be done either independently or iteratively as once an estimate of ψ is obtained, one can then estimate β through the global partial likelihood. An iterative algorithm, as shown below, can be established by alternatingly updating the estimates for β and ψ. Such an iteration procedure may improve the link estimate as a better estimate of β will lead to a better estimate of ψ.

Step 1. (a) Assign a nonzero initial value to β, and call it $\hat{\beta}$.

(b) For a given v, plug $\hat{\beta}$ into the pseudo log local partial likelihood and maximize

$$\sum_{j=1}^{N} K_h\{\hat{\beta}^T Z_{(j)} - v\} \cdot \left[[\hat{\beta}^T \mathbf{Z}_{(j)}]^T \gamma(v) - \log\left\{ \sum_{i \in \mathcal{R}_j} \exp\{[\hat{\beta}^T \mathbf{Z}_i]^T \gamma(v)\} K_h\{\hat{\beta}^T Z_i - v\} \right\} \right]$$

with respect to $\gamma(v)$ to get the estimate $\hat{\gamma}(v)$.

(c) Obtain the values of $\hat{\gamma}(v)$, for $v = \hat{\beta}^T Z_i, i = 1, \cdots, n$.

(d) Apply the trapezoidal rule to obtain $\{\hat{\psi}(\hat{\beta}^T Z_i) : i = 1, \cdots, n\}$.

Step 2. Plug $\hat{\psi}(\cdot)$ into the log (global) partial likelihood

$$l_G(\beta, \hat{\psi}) = \sum_{j=1}^{N} \left[\hat{\psi}(\beta^T Z_{(j)}) - \log\left\{ \sum_{i \in \mathcal{R}_j} \exp\{\hat{\psi}(\beta^T Z_i)\} \right\} \right],$$

and maximize it with respect to β to update the estimate $\hat{\beta}$. We use the angle between two estimated $\hat{\beta}$ at two consecutive iterations as the convergence criterion.

Remark 1. The Newton-Raphson method is used to find the estimators in Step 1 and 2. The initial value of β can be set in different ways but cannot be zero as a nonzero value is needed in step 1 to estimate the link function. However, this restriction does not exclude the final β-estimate to be zero or close to zero. A simple choice is to fit the usual Cox model and use this estimator in the first step. To accelerate the computation, one can also use alternative estimates as described below.

Remark 2. It is possible to accelerate the computation by using a $\sqrt{n}$-consistent initial estimator, as Theorems 1 and 2 below imply that no iteration is required for the link estimate and that it will converge efficiently at the usual nonparametric rate. Namely, the link function can be estimated with the same efficiency as when β is known. In practice, we find that one iteration helps to improve the numerical performance but further iteration is usually not necessary. There are two choices for a $\sqrt{n}$-consistent initial estimator, one is the estimator in [2] that extends the sliced inverse regression (SIR) approach to censored data. Specifically, this approach requires a certain design condition as listed in (2.3) there but has the advantage that it leads to a $\sqrt{n}$-consistent estimator for β without the need to estimate the link function. Another initial estimate which does not rely on the design conditions (2.3) in [2] is provided in a Ph. D. thesis [19]. Specifically, this involves replacing the ψ function in step 2 above by its local version (2), which leads to the cancelation of the term ψ and results in a version of log (global) partial likelihood that involves only the derivative $\gamma(v)$ of ψ but not ψ itself. Thus, Step 2 above is replaced by

Step 2*. Maximize the following approximate log (global) partial likelihood with respect to β:

$$\sum_{j=1}^{N}\left[[\beta^T Z_{(j)}]^T\hat{\gamma}\{\hat{\beta}^T Z_{(j)}\} - \log\left\{\sum_{i\in\mathcal{R}_j}\exp\Big([\beta^T Z_i]^T\hat{\gamma}\{\hat{\beta}^T Z_i\}\Big)\right\}\right].$$

This approximation may result in some efficiency loss, but has computational advantages over the estimate in Step 2, since we do not need to estimate $\psi(v)$ and thus can skip Step 1(d). The resulting estimate for β was shown in [19] to be $\sqrt{n}$-consistent, consequently an ideal choice as the initial estimate for β.

Remark 3. In step 1, the local log partial likelihood in (3) is replaced by a pseudo log partial likelihood with β replaced by $\hat{\beta}$. As this $\hat{\beta}$ approaches β, the link estimate resulting from maximizing the pseudo log partial likelihood can be expected to approach the true link function at the usual nonparametric rate. This is because the parametric estimate $\hat{\beta}$ converge to its target at the root-n rate, which is faster than the nonparametric link estimate. A rigorous proof is provided in Theorem 1 and Theorem 2.

Remark 4. For large sample sizes, it is unnecessary to estimate the link function at each data

point. An alternative way is to estimate the link function at equal-distance grid points, then using interpolation or smoothing methods to obtain the estimated value at each data point. Our simulation results show that this short-cut is computationally economical while retaining similar accuracy.

2.2. Main results

Let $f(\cdot)$ be the probability density of $\beta^T Z$, for a given v, let $P(t \mid v) = P(X \geq t \mid \beta^T Z = v)$, $Y(t) = I\{X \geq t\}$, $H = diag\{h, \cdots, h^p\}^T$ and $\mathbf{u} = \{u, \cdots, u^p\}^T$.

Theorem 1. *Under conditions (C1)-(C5) in the Appendix, for any $\sqrt{n}$ consistent estimator $\hat{\beta}$ of the true parameter β_0, let $\hat{\gamma}(\cdot)$ be the corresponding estimator for the derivatives $\gamma_0(\cdot)$ of the true link ψ and $\hat{\psi}(v) = \int_0^v \hat{\psi}'(w)dw$, where $\hat{\psi}'(\cdot)$ is the first component of $\hat{\gamma}(\cdot)$. If $h \to 0$, $nh/\log n \to \infty$, $nh^4 \to \infty$ then*

$$\sup_v |\hat{\gamma}(v) - \gamma_0(v)| \to_p 0,$$

and

$$\sup_z |\hat{\psi}(\hat{\beta}^T z) - \psi({\beta_0}^T z)| \to_p 0$$

Theorem 2. *Under the conditions in Theorem 1 and for bounded nh^{2p+3},*

$$(a) \qquad \sqrt{nh}\left\{H(\hat{\gamma}(v) - \gamma_0(v)) - \frac{\psi^{(p+1)}(v)}{(p+1)!}A^{-1}bh^{p+1}\right\} \to_D N\left\{0, \frac{\sigma^2(v)}{f(v)}A^{-1}DA^{-1}\right\}.$$

Furthermore, we have

$$(b) \qquad \sqrt{nh}\left\{H(\hat{\gamma}(\hat{\beta}^T z) - \gamma_0({\beta_0}^T z)) - \frac{\psi^{(p+1)}({\beta_0}^T z)}{(p+1)!}A^{-1}bh^{p+1}\right\} \to_D N\left\{0, \frac{\sigma^2({\beta_0}^T z)}{f({\beta_0}^T z)}A^{-1}DA^{-1}\right\},$$

where $A = \int \mathbf{u}\mathbf{u}^T K(u)du - \nu_1\nu_1^T$, $b = \int u^{p+1}(\mathbf{u} - \nu_1)K(u)du$, $D = \int K^2(u)(\mathbf{u} - \nu_1)^{\otimes 2}du$, $\nu_1 = \int \mathbf{u}K(u)du$, and $\sigma^2(v) = E\{\delta|\beta^T Z = v\}^{-1}$.

Theorem 1 establishes the uniform consistency of the local partial likelihood estimator of γ_0 and Theorem 2 provides the joint asymptotic normality of the derivative estimators. The limiting distribution of $\hat{\gamma}$ is identical to the one in [8], where β is assumed to be known. Thus, there is no efficiency loss as long as β can be estimated at the usual $\sqrt{n}$-rate.

2.3. Model Checking and Selection

While an estimated link function is of interest to correctly reflect the risk associated with a covariate, a parametric link function is often preferable to a nonparametric one to lend a parsimonious model with more interpretable results. Thus, a main incentive to estimate the link function could be for exploratory model selection to facilitate the choice of a proper parametric link function in the proportional hazards model (1). If so, the β estimate in Step 2 only aids in the link estimation and need not be the end product. Once a suitable link function has been selected, Theorem 2 can be used for model checking. For instance, to check the identity link function under the Cox model, one can test H_0: $\psi'(v) = 1$. Since the first component of $\gamma(v)$ is $\psi'(v)$, a local polynomial of order $p = 2$ is usually employed to estimate such a derivative, and the resulting asymptotic distribution of the corresponding estimate is given below.

Corollary 1. *Under the condition of Theorem 2, and with $p = 2$ there, we have*

$$\sqrt{nh^3}\left(\hat{\psi}'(v) - \psi'(v) - \frac{1}{6}\psi^{(3)}(v)h^2\frac{\int u^4K(u)du}{\int u^2K(u)du}\right) \to_D N\left(0, \frac{\sigma^2(v)\int u^2K^2(u)du}{f(v)\int u^2K(u)du}\right).$$

Corollary 1 facilitates the construction of testing procedures and asymptotic simultaneous confidence bands for the link function, but rigorous asymptotic theory requires much further work and is not available yet. In principle, one could check the appropriateness of the link function at all data points v that falls in the range of $\beta^T Z$. Since the true value of β is unknown, it is natural to replace it with an estimate. However, one must bear in mind the precision of this estimate as well as the low precision of $\psi'(v)$ for v in the boundary region of $\hat{\beta}^T Z$. Here boundary region is defined as within one bandwidth of the data range, where a smoothing procedure is employed. Since the bandwidth h is usually of a higher order than $n^{-\frac{1}{2}}$, the anticipated rate of convergence for $\hat{\beta}$, we recommend to restrict inference on $\psi'(v)$ for v that is in the interior and at least one bandwidth h away from either boundary of the range of $\hat{\beta}$.

Short of such a rigorous inference procedure for model checking, pointwise confidence intervals have often been used as a substitute for exploratory purposes. In the example in Section 4, we illustrate how to check the appropriateness of the Cox model, i.e. identity link function, using pointwise confidence intervals developed from Corollary 1. Readers should bear in mind that this is only an exploratory data analysis tool rather than a formal inference procedure.

3. Simulation studies

To see how the algorithm in Section 2 works for the proposed model, we conducted simulation studies where a quadratic link function $\psi(\beta^T Z) = (\beta^T Z)^2$ with $\beta = (1, 3)^T$ and a constant baseline,

$\lambda_0 = 0.005$, were employed. The design for the two-dimensional covariate, $Z = (Z_1, Z_2)^T$ is: $Z_1 \sim U(-1, 1)$ and Z_2 is a truncated $N(0, 1)$ with values in $[-1, 1]$. Parameters of β were chosen in such a way that the simulation generates a reasonable signal to noise ratio (cf. Figures 1). If we take ε to have the standard exponential distribution, exp(1), the resulting hazard function will be $\lambda_0 \exp\{\psi(\beta^T Z)\}$, and survival times from this model can be generated as $T = \exp\{\psi(-\beta^T Z)\}\varepsilon/\lambda_0$. Different uniform distributions were utilized to generate three independent censoring times so that the censoring rates were 0%, and roughly 25% and 50%. The Epanechnikov kernel was adopted in the link estimation. Two sample sizes, 200 and 50, were selected to see whether the methods are flexible for moderate to small samples. For $n = 200$ we used 25 equal-distance grid points to estimate the link function to save computational time as elucidated in Remark 4 of Section 2. Piece-wise spline interpolation was then used to get the link estimate during each iteration of the algorithm.

Due to the complication from the identifiability problem, the link function can only be identified up to a constant. Thus, $\hat{\psi}(v)$ and $\hat{\psi}(v) + c$, for any constant c, are considered to be equivalent procedures, and any measures of performance would declare these two procedures identical. This points to selecting a measure which measures the variation instead of the real difference. We adopt a measure proposed in [10] which is the standard deviation of the differences between the fitted values $\hat{\psi}(\hat{\beta}^T Z)$ and the true values $\psi(\beta^T Z)$ at all data points. More specifically, this measure, denoted by d, is the standard deviation of the difference $\{\hat{\psi}(\hat{\beta}^T Z) - \psi(\beta^T Z) : i = 1, \cdots, n\}$. We report in Table 1 the average values for this measure and its standard deviation based on 100 simulation runs.

Since at each estimating step, $\hat{\beta}$ was updated and the range of $\hat{\beta}^T Z$ might be different, we used a bandwidth h^*, which took a certain portion of the range of $\hat{\beta}^T Z$. For instance, an $h^* = 0.3$ means that the actual bandwidth is 0.3 times the range of the values of $\hat{\beta}^T Z$. Various bandwidths were explored, but we report only the results for bandwidth h^* varying from 0.1 to 0.4 (with 0.1 increment) times the data range of $\beta^T Z$ at each iteration stage. Results for other bandwidths were inferior and are not reported here.

Four procedures were compared and the results for $n = 200$ are shown in Table 1. Method 1 assumes that the link is identity (which is incorrect here) and the regression coefficient estimate $\hat{\beta}$ is therefore the Cox estimate based on the partial likelihood estimate. The aim is to see the effect of erroneously assuming the conventional Cox proportional hazards model. Method 2 assumes that β is known and estimates the unknown link function as in [8]. Method 3 is the new procedure where both the link function and regression coefficient β are estimated. Method 4 assumes that the true quadratic

Table 1: Comparison of different methods for estimating $\psi(\beta^T Z) = (\beta^T Z)^2$, where $\beta = (1,3)^T$ and $Z = (Z_1, Z_2)^T$ with $Z_1 \sim U(-1,1)$ and $Z_2 \sim N(0,1)$ (truncated at [-1,1]), n=200. The numbers before and after "/" are the means and standard deviations of d based on 100 simulations.

		h^*				Optimal
Censoring	†	0.1	0.2	0.3	0.4	MSE
No	1	3.152/0.181	3.152/0.181	3.152/0.181	3.152/0.181	9.970
	4	0.191/0.114	0.191/0.114	0.191/0.114	0.191/0.114	0.049
	2	1.234/2.554	0.439/0.142	0.843/0.150	1.417/0.151	0.213
	3	1.210/3.522	0.490/0.144	0.907/0.168	1.476/0.168	0.261
25%	1	3.151/0.181	3.151/0.181	3.151/0.181	3.151/0.181	9.961
	4	0.201/0.112	0.201/0.112	0.201/0.112	0.201/0.112	0.053
	2	1.256/2.548	0.425/0.157	0.684/0.172	1.197/0.178	0.206
	3	0.981/1.991	0.468/0.156	0.746/0.185	1.256/0.194	0.244
50%	1	3.149/0.181	3.149/0.181	3.149/0.181	3.149/0.181	9.947
	4	0.210/0.117	0.210/0.117	0.210/0.117	0.210/0.117	0.056
	2	1.361/2.529	0.525/0.215	0.535/0.161	0.729/0.210	0.322
	3	1.236/2.138	0.536/0.201	0.571/0.169	0.808/0.226	0.327

† Method 1 is under identity link and unknown β. Method 2 is under unknown link and true β.

Method 3 is under unknown link and unknown β. Method 4 is under quadratic link and unknown β.

link function is known and the regression coefficient estimate, $\hat{\beta}$, is the partial likelihood estimate. The comparisons for the distance d are reported in Table 1. The results of the best procedures together with the corresponding optimal bandwidths are highlighted with boxes. It is not surprising that the best results came from the procedures with true quadratic link function and unknown β. Our estimators are close to those from method 2 ([8]) with known β, while the estimators based on the identity link model have much larger d.

To demonstrate the effect of an erroneous link function on regression estimates, we report in Table 2 the results of the various estimates for β. Since there is no regression parameter estimate for method 2, only three procedures are compared in Table 2. There are several ways to compare the regression estimates, one way is to set the first component of β to the true value, then to compare the difference between $\hat{\beta}$ and β based on the second component. Table 2 shows the results of the difference between the true β_2 and the estimate $\hat{\beta}_2$ for various procedures. The best procedures for the profile estimator in method 3 are shown in boxes under the optimal bandwidths. Another way to compare the different estimators is to calculate the angles between these estimators and the true parameter. To save space only the MSEs based on the optimal bandwidths are listed in the last column of Table 2 for the angle measure (degrees). Based on both optimal MSE measures reported in the last two columns of Table 2, the differences between the new profile estimators and the true parameters are way smaller than those from the identity link model, and reasonably close to those under the true link.

We can see that using the wrong link will lead to huge bias and MSE under all censoring patterns. The average angles between the β estimates assuming identity link and the true parameters are around 90°, which suggests that the β estimates with identity link are perpendicular to the true parameter space, indicating a total inability to estimate the regression parameter. This is in addition to the problem that the link function itself has been misspecified. Both underscore the importance to check the shape of the link function at the beginning of data analysis.

Four typical simulated curves are shown in Figure 1. The procedure (method 4) with known quadratic link function and unknown β performed the best. The procedure (method 2) with known β and our procedure captured the shape of the true curve well, but the procedure (method 1) based on the Cox model failed to capture the shape of the link function. Results for $n = 50$ summarized in Table 3 are consistent with the findings for $n = 200$ in Table 1.

4. Data analysis

In this section, we illustrate the proposed model and estimation algorithm in Section 2 through

Table 2: Differences between the estimated $\hat{\beta}_2$ and the true β_2 where $\hat{\beta}_1$ is set equal to the true β_1, $\psi(\beta^T Z) = (\beta^T Z)^2$, where $\beta = (1,3)^T$ and $Z = (Z_1, Z_2)^T$ with $Z_1 \sim U(-1,1)$ and $Z_2 \sim N(0,1)$ (truncated at [-1,1]), n=200. The numbers before and after "/" are the biases and standard deviations of the $\hat{\beta}_2$ based on 100 simulations.

		h*				Optimal	Optimal
Censoring	†	0.1	0.2	0.3	0.4	MSE‡	MSE§
No	1	5.344/50.847	5.344/50.847	5.344/50.847	5.344/50.847	2613.949	11742.108
	4	-0.015/0.142	-0.015/0.142	-0.015/0.142	-0.015/0.142	0.021	0.703
	3	0.023/0.276	-0.051/0.154	-0.096/0.158	-0.215/0.190	0.026	0.968
25%	1	-1.229/24.742	-1.229/24.742	-1.229/24.742	-1.229/24.742	613.684	11320.063
	4	-0.013/0.154	-0.013/0.154	-0.013/0.154	-0.013/0.154	0.024	0.801
	3	0.038/0.307	-0.055/0.170	-0.089/0.171	-0.162/0.195	0.032	1.144
50%	1	-1.472/6.486	-1.472/6.486	-1.472/6.486	-1.472/6.486	44.237	11266.339
	4	-0.006/0.170	-0.006/0.170	-0.006/0.170	-0.006/0.170	0.029	0.968
	3	0.055/0.371	-0.050/0.188	-0.075/0.181	-0.110/0.190	0.038	1.362

† Method 1 is under identity link and unknown β. Method 4 is under true link and unknown β.

Method 3 is under unknown link and unknown β.

‡ is the optimal MSE for $\hat{\beta}_2$ when $\hat{\beta}_1$ is set to be 1. § is the optimal MSE for the angles between $\hat{\beta}$ and the true β.

Table 3: Comparison of different methods for estimating $\psi(\beta^T Z) = (\beta^T Z)^2$, where $\beta = (1,3)^T$ and $Z = (Z_1, Z_2)^T$ with $Z_1 \sim U(-1,1)$ and $Z_2 \sim N(0,1)$ (truncated at [-1,1]), n=50. The reported quantities are the smallest MSEs under optimal bandwidths.

	No Censoring		25% Censoring		50% Censoring	
Method	d	Optimal h^*	d	Optimal h^*	d	Optimal h^*
Identity link and unknown β	9.199	-	9.168	-	9.102	-
Quadratic link and unknown β	0.184	-	0.203	-	0.295	-
Unknown link and true β	0.808	0.3	0.833	0.3	1.359	0.4
Unknown link and unknown β	0.945	0.4	1.286	0.4	1.423	0.4

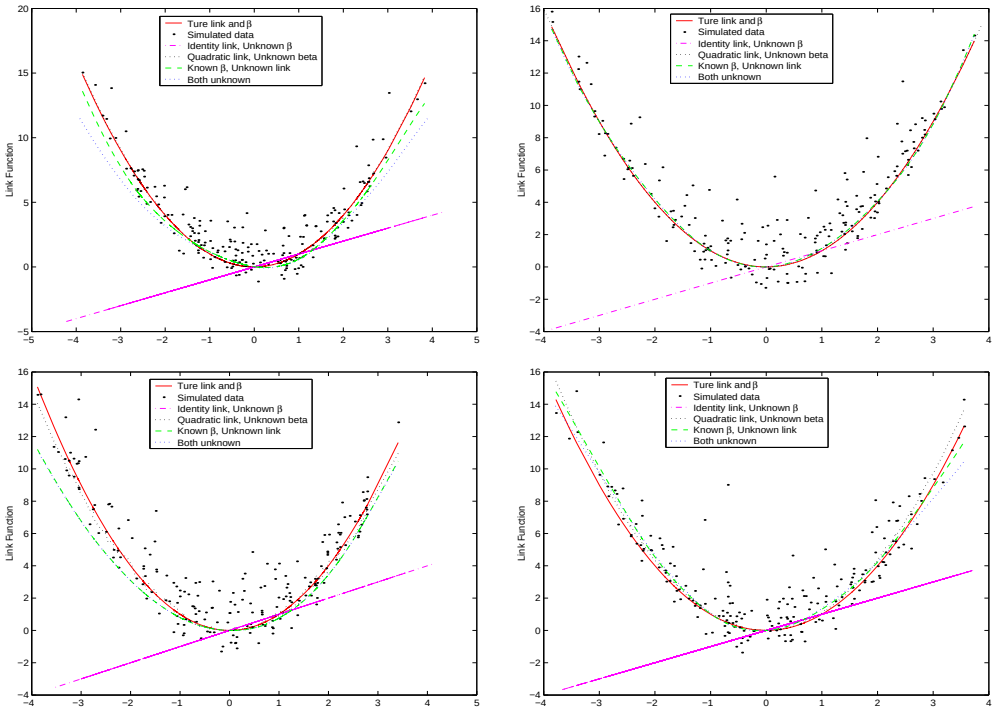

Figure 1: Four typical simulated sets of data and estimated curves from the model $\psi(\beta^T Z) = (\beta^T Z)^2$ with $\beta = (1,3)^T, Z = (Z_1, Z_2)^T$ where $Z_1 \sim U(-1,1)$ and $Z_2 \sim N(0,1)$ (truncated at [-1,1]), 25% censoring and $h^* = 0.2$. The simulated data is indicated as black dots. The true curve $\psi(\cdot)$ is indicated as a red solid line, and the estimated curves $\hat{\psi}(\cdot)$ as a magenta dash line with identity link and unknown β, as a black dash-dotted line with quadratic link and unknown β, as a green dotted line with unknown link and true β, and as a blue dash-dotted line with both unknown link and unknown β.

the Worcester Heart Attack Study (WHAS) data. One of the goals of this study is to identify factors associated with the survival rates following hospital admission for acute myocardial infarction. The main data set has more than 11000 admissions, but we used only a random sample of 500 patients as listed in [14]. This data set is chosen because the proportionality assumption has been carefully examined and is reasonably satisfied. Our goal here is to check the adequacy of the identity link function in the Cox proportional hazards regression model.

There were more than 10 covariates in the data set. After detailed model selection procedure, Hosmer et al. [14] included 6 variables age (AGE), initial heart rate (HR), initial diastolic blood pressure (DIASBP), body mass index (BMI), gender (GENDER), congestive heart complications (CHF), and the interaction between age and gender (AGEGENDER) in their model. After examining the linearity assumption using fractional polynomials, they decided to apply a two-term fractional polynomial model to the variable BMI. We thus begin with the univariate covariate BMI.

We tried different bandwidths and found similar patterns of the estimated link functions. In Figure 2 we report two of the results, which exhibit reasonable level of smoothness. The estimated link function in Figure 2 suggests clear nonlinearity. We then constructed a 95% point-wise approximated confidence interval of $\gamma(v)$ (Figure 3) to see whether it would cover the constant function 1. The results suggest that the estimated link functions have some curvature and further investigation is needed.

Next we applied the proposed procedure to the multivariate model with all 7 covariates. We tried different bandwidths ranging from 1/10, 1/8, 1/7, 1/6, 1/5, 1/4 ,1/3, to 1/2 of the single index range, and plot the results of three bandwidths in Figure 4. The estimated link functions for $h^* = 1/4$ appear oversmoothed but all three estimates exhibit two bumps. The 95% point-wise approximated confidence intervals for $\gamma(v)$ as shown in Figure 5 also reveal curvature away from the constant ($= 1$) horizontal line. Although it is arguable that the Cox model could be rejected at a low level, the suitability of an identity link function seems questionable.

5. Conclusion and Future Research

The proposed estimating procedures for the extended proportional hazards regression model with unknown link function and multi-dimensional covariates seem to be reliable for moderate to large sample sizes. Once the link function and the parameters of the index have been established, one can proceed to estimate the unknown baseline hazard function in (1) using a Breslow-type estimate ([1]).

The cost of a misspecified link function has been demonstrated through the simulation studies in Section 3. As a consequence, the risk of an individual may be misinterpreted. It is thus important

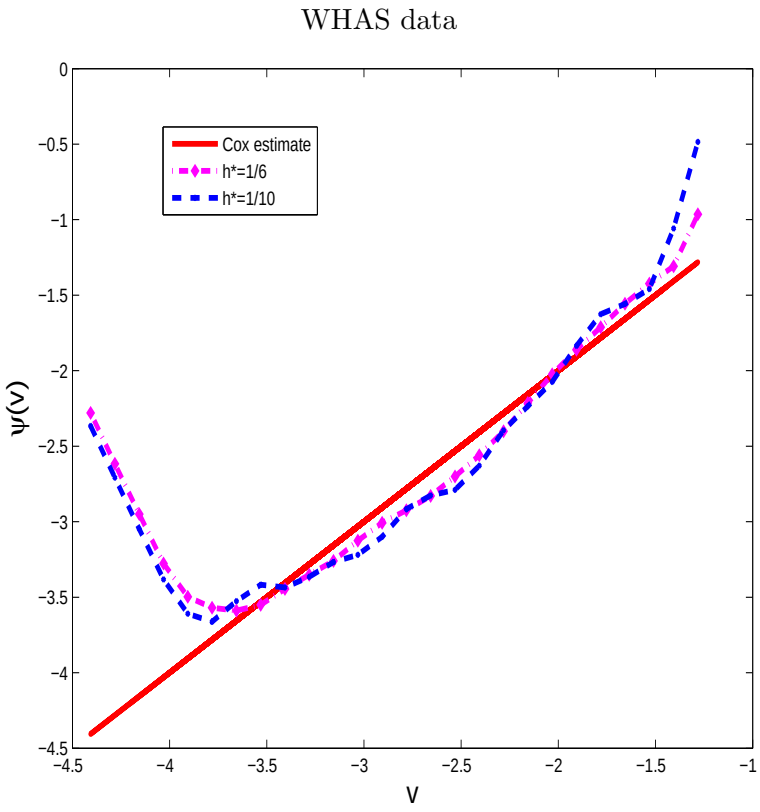

Figure 2: The estimated link function, under two bandwidths, for the WHAS data with BMI as a covariate .

to at least estimate the link function in the initial model fitting stage as a model checking tool or guidance to a suitable class of parametric link functions. A rigorous test of parametric link function will be a worthwhile future project, as is the asymptotic theory for simultaneous inference of the link function and regression parameters.

The choice of automatic smoothing parameters, the bandwidth h in this case, is a challenging problem for proportional hazards model when a likelihood based smoother, such as the local partial likelihood estimate, is employed in the link estimate. This is because the components of the partial likelihood are dependent, hence the usual automatic bandwidth selection methods for linear models are not applicable here. The usual least square cross-validation procedure in nonparametric regressions also cannot be easily adapted to hazard based models such as the proportional hazards model. An alternative criterion, less computational intensive than cross validation methods, was proposed in [18], based on a variation of the Akaike's information criterion for span selection, when the nearest neighborhood method was used for smoothing. However, the interpretation of AIC is not clear here since partial likelihood involves dependent components. The authors also acknowledged that the

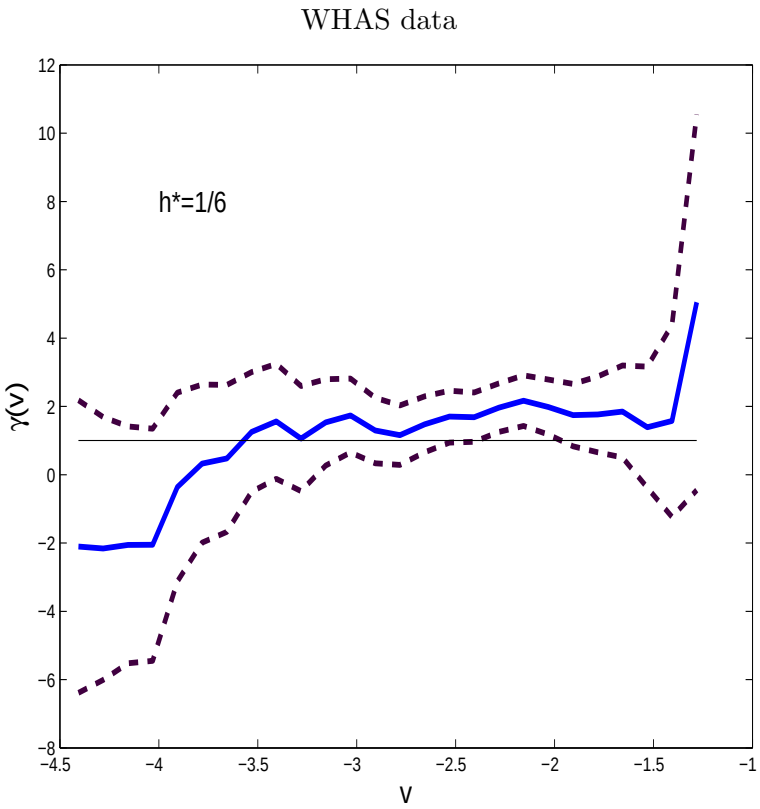

Figure 3: The estimated confidence interval of γ at bandwidth $h^* = 1/6$ for the WHAS data based on BMI.

asymptotic correctness of the AIC criterion has not been established. Thus, automatic bandwidth choice remains an open question when the link function is being estimated. Meanwhile, we recommend to try several bandwidths and choose one that yields a moderately smooth link function as we did for the WHAS data. This subjective choice based on the visual degree of smoothness is commonly adopted as an ad hoc tool.

While this paper deals with time-independent covariates, it would be desirable to extend model (1) to time-dependent covariates as well. One complication is that the entire history of the covariate process would be required or some kind of imputation needs to be performed to get even an initial estimate of β. Preliminary results were reported in [20] by imputing the covariate process through a functional principal components approach, and then proceeding with the estimation of the survival components at the second stage. Such a two-stage procedure is prone to bias as is well known in the joint modelling literature. Further investigations to correct the bias would be desirable, and joint modelling the longitudinal and survival process offers some hope if one can resolve the additional complication of an unknown link function. This is yet another worthwhile project to pursue in the

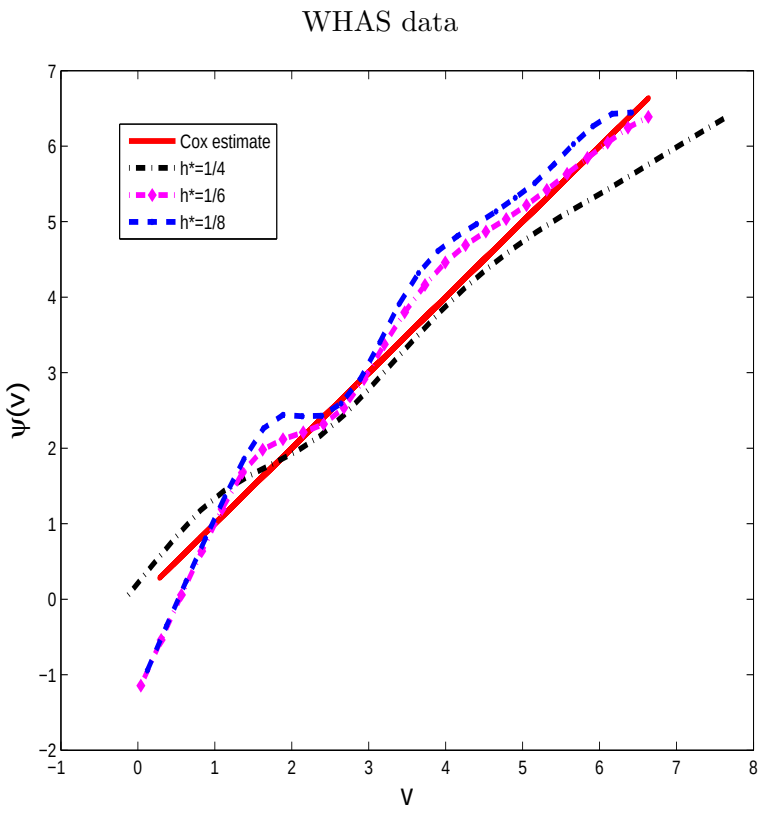

Figure 4: The estimated link function for the WHAS data under three different bandwidths.

future.

Acknowledgement

The work of Jane-Ling Wang and Wei Wang was completed with the partial support of National Science Foundation grants DMS04-06430, National Institutes of Health grant R01-DK-45939, and National Institute of Allergy and Infectious Diseases grant AI24643. Qihua Wang's research was supported by the National Science Fund for Distinguished Young Scholars in China (10725106), the National Natural Science Foundation of China (10671198), the National Science Fund for Creative Research Groups in China and the Research Grants Council of the Hong Kong (HKU 7050/06P). The authors are grateful for the invaluable suggestions of two reviewers and the editor.

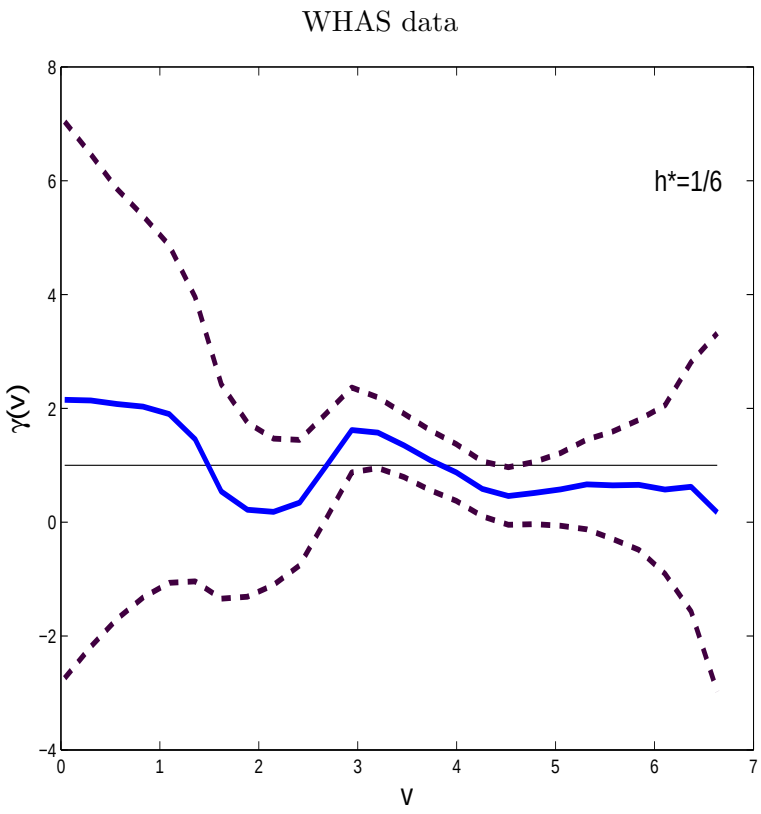

Figure 5: The estimated confidence interval of γ for the WHAS data at bandwidth $h^* = 1/6$.

Appendix

Let $f(\cdot)$ be the probability density of $\beta^T Z$, for a given v, let $P(t \mid v) = P(X \geq t \mid \beta^T Z = v)$, $\Lambda(t, v) = \int_0^t P(u \mid v)\lambda_0(u)du$, $Y(t) = I\{X \geq t\}$, $Y_i(t) = I\{X_i \geq t\}$, $H = diag\{h, \cdots, h^p\}^T$ and $\mathbf{u} = \{u, \cdots, u^p\}$.

We begin with some regularity conditions needed for the results.

$(C1)$ $K \geq 0$ is a bounded density with compact support, and it has bounded first and second derivative.

$(C2)$ $\psi(\cdot)$ has a continuous $(p+1)$th derivative around v.

$(C3)$ The density $f(\cdot)$ of $\beta^T Z$ is continuous at point v and $\inf_w f(w) > 0$.

$(C4)$ The conditional probability $P(t \mid \cdot)$ is equicontinuous at v.

$(C5)$ $\int_0^\tau \lambda_0(u)du < \infty$.

Denote

$s_0(\beta, \psi, u) = E\big[Y(u)\exp\{\psi(\beta^T Z)\}\big]$,

$s_1(\beta, \psi, u) = E\big[Y(u)\exp\{\psi(\beta^T Z)\}\psi'(\beta^T Z)Z\big]$,

$s_2(\beta, \psi, u) = E\big[Y(u)\exp\{\psi(\beta^T Z)\}\{\psi''(\beta^T Z) + [\psi'(\beta^T Z)]^2\}ZZ^T\big]$,

$s_2^*(\beta,\psi,u) = E\big[Y(u)\exp\{\psi(\beta^T Z)\}[\psi'(\beta^T Z)]^2 ZZ^T\big]$.

(C6) The functions $s_r, r = 0, 1$,and 2, and s_2^* are bounded and s_0 is bounded away from 0 on $\mathcal{B}\times[0,\tau]$; the family of functions $s_r(\cdot,\psi,u)$ and $s_2^*(\cdot,\psi,u)$ is an equicontinuous family at β_0.

The following lemma is used repeatedly in the later proofs. The proof will be omitted and can be found in [19].

Lemma 1. *Let* $c_n(\beta_0,t) = n^{-1}\sum_{i=1}^n Y_i(t)g({\beta_0}^T Z_i)K_h({\beta_0}^T Z_i - v)$ *and* $c(t) = f(v)g(v)P(t \mid v)\int K(u)du$, *Under conditions (C1) and (C4), if* $g(\cdot)$ *is continuous at the point* v, *then*

$$\sup_{0\le t\le\tau}|c_n(\beta_0,t) - c(t)| \to_P 0,$$

provided that $h\to 0, nh/\log n\to\infty, 0<\tau\le+\infty$.

If furthermore, $\hat\beta$ *is a* $\sqrt{n}$*-consistent estimate of* β_0 *and* $nh^4\to\infty$,*then*

$$\sup_{0\le t\le\tau}|c_n(\hat\beta,t) - c(t)| \to_P 0.$$

Proof of Theorem 1

Proof. For notation simplicity, we will use γ to represent $\gamma(v)$ and γ_0 for $\gamma_0(v)$.

The log local partial likelihood function at point v is given as

$$l\{\beta,\gamma,v\} = \frac{1}{n}\sum_{j=1}^N K_h(\beta^T Z_{(j)} - v)\bigg[\big[(\beta^T\mathbf{Z}_{(j)})^T\gamma\big] - \log\Big\{\sum_{i\in\mathcal{R}_j}\exp\{(\beta^T\mathbf{Z}_i)^T\gamma\}K_h(\beta^T Z_i - v)\Big\}\bigg].$$

Using counting process notation $N(t) = I\{X\le t,\delta=1\}$ and $N_i(t) = I\{X_i\le t,\delta_i=1\}$, under the independent censoring,

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)\exp\{\psi({\beta_0}^T Z_i)\}\lambda_0(u)du,$$

is a martingale with respect to the filtration $\mathcal{F}_t = \sigma\{N(u), I_{\{X\le u,\delta=0\}} : 0\le u\le t\}$.

The empirical counterpart of $l\{\beta,\gamma,v\}$ up to time t is

$$\begin{aligned} l_n(\beta,\gamma,t,v) = \int_0^t \frac{1}{n}\sum_{i=1}^n K_h(\beta^T Z_i - v)\bigg[&\big[(\beta^T\mathbf{Z}_i)^T\gamma\big] \\ &- \log\Big\{\sum_i^n Y_i(u)\exp\{(\beta^T\mathbf{Z}_i)^T\gamma\}K_h(\beta^T Z_i - v)\Big\}\bigg]dN_i(u). \end{aligned}$$

Denote $S_{h,0}(\beta,\gamma,u,v) = \frac{1}{n}\sum_{i=1}^n K_h(\beta^T Z_i - v)Y_i(u)\exp\{(\beta^T\mathbf{Z}_i)^T\gamma\}$.

Let $\hat\beta$ be a $\sqrt{n}-$consistent estimate of the true parameter β_0, and $\hat\gamma$ be the corresponding estimate of the true γ_0, we can write

$$
\begin{aligned}
& l_n(\hat{\beta},\gamma,\tau,v) - l_n(\beta_0,\gamma_0,\tau,v) \\
=& \int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h({\beta_0}^T Z_i - v)\left[[(\beta_0^T\mathbf{Z}_i)^T\gamma - (\beta_0^T\mathbf{Z}_i)^T\gamma_0] - \log\frac{S_{h,0}(\beta_0,\gamma,u,v)}{S_{h,0}(\beta_0,\gamma_0,u,v)}\right]dM_i(u) \\
+& \int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v)\left[[(\hat{\beta}^T\mathbf{Z}_i)^T\gamma - (\beta_0^T\mathbf{Z}_i)^T\gamma] - \log\frac{S_{h,0}(\hat{\beta},\gamma,u,v)}{S_{h,0}(\beta_0,\gamma,u,v)}\right]dM_i(u) \\
+& \int_0^\tau \frac{1}{n}\sum_{i=1}^n \Big(K_h(\hat{\beta}^T Z_i - v) - K_h({\beta_0}^T Z_i - v)\Big)\left[(\beta_0^T\mathbf{Z}_i)^T\gamma - \log S_{h,0}(\beta_0,\gamma,u,v)\right]dM_i(u) \\
+& \int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h({\beta_0}^T Z_i - v)\left[[(\beta_0^T\mathbf{Z}_i)^T\gamma - (\beta_0^T\mathbf{Z}_i)^T\gamma_0] - \log\frac{S_{h,0}(\beta_0,\gamma,u,v)}{S_{h,0}(\beta_0,\gamma_0,u,v)}\right] \\
& \qquad \times Y_i(u)\exp\{\psi({\beta_0}^T Z_i)\}\lambda_0(u)du \\
+& \int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v)\left[[(\hat{\beta}^T\mathbf{Z}_i)^T\gamma - (\beta_0^T\mathbf{Z}_i)^T\gamma] - \log\frac{S_{h,0}(\hat{\beta},\gamma,u,v)}{S_{h,0}(\beta_0,\gamma,u,v)}\right] \\
& \qquad \times Y_i(u)\exp\{\psi({\beta_0}^T Z_i)\}\lambda_0(u)du \\
+& \int_0^\tau \frac{1}{n}\sum_{i=1}^n \Big(K_h(\hat{\beta}^T Z_i - v) - K_h({\beta_0}^T Z_i - v)\Big)\left[(\beta_0^T\mathbf{Z}_i)^T\gamma - \log S_{h,0}(\beta_0,\gamma,u,v)\right] \\
\equiv& X_n(\beta_0,\gamma,\tau,v) + I + II + A_n(\beta_0,\gamma,\tau,v) + III + IV.
\end{aligned}
$$

Under the regularity conditions and from Lemma 1, it can be shown that

(1) $X_n(\beta_0,\gamma,\tau,v)$ is a locally square integrable martingale with the predictable variation process

$$
\begin{aligned}
& \langle X_n(\beta_0,\gamma,\tau,v), X_n(\beta_0,\gamma,\tau,v)\rangle \\
=& \int_0^\tau \frac{1}{n^2}\sum_{i=1}^n K_h^2({\beta_0}^T Z_i - v)\left[(\beta_0^T\mathbf{Z}_i)^T(\gamma-\gamma_0) - \log\frac{S_{h,0}(\beta_0,\gamma,u,v)}{S_{h,0}(\beta_0,\gamma_0,u,v)}\right]^2 Y_i(u)\exp\{\psi({\beta_0}^T Z_i)\}\lambda_0(u)du \\
=& O_p(\frac{1}{nh}).
\end{aligned}
$$

(2) $A_n(\beta_0,\gamma,\tau,v) \to_p$

$$
\begin{aligned}
& f(v)\exp\{\psi(v)\}\Lambda(\tau,v)\times\left[(\int \mathbf{u}K(u)du)^T H(\gamma-\gamma_0) - \log\{\int \exp\{\mathbf{u}^T H(\gamma-\gamma_0)\}K(u)du\}\right] + o_p(1) \\
& \equiv A(\beta_0,\gamma,\tau,v) + o_p(1),
\end{aligned}
$$

(3) $I = O_p(\frac{1}{nh})$, $II = O_p(\frac{1}{nh^2})$, $III = O_p(\frac{1}{\sqrt{nh^2}})$, and $IV = O_p(\frac{1}{\sqrt{nh^4}})$.

This means $X_n(\beta_0,\gamma,\tau,v)$, I, II, III and IV converge to zero at a faster rate than $A_n(\beta_0,\gamma,\tau,v)$. By Lemma 8.2.1(2) in [9], $l_n(\hat{\beta},\gamma,\tau,v) - l_n(\beta_0,\gamma_0,\tau,v)$ has the same limiting distribution as $A_n(\beta_0,\gamma,\tau,v)$. Thus, we have

$$
l_n(\hat{\beta},\gamma,\tau,v) - l_n(\beta_0,\gamma_0,\tau,v) \to_p A(\beta_0,\gamma,\tau,v). \tag{4}
$$

It is obvious that $A(\beta_0, \gamma, \tau, v)$ is strictly concave, with a maximum at $\gamma = \gamma_0$. Hence, the right-hand side of (4) is maximized at $\gamma = \gamma_0$. The left-hand side of (4) is maximized at $\gamma = \hat{\gamma}$, since $\hat{\gamma}$ maximizes $l_n(\hat{\beta}, \gamma, \tau, v)$. Therefore, $\sup_v \left|\hat{\gamma}(v) - \gamma_0(v)\right| \to_p 0$.

By Dominated Convergence Theorem, we have

$$\sup_v \left|\hat{\psi}(v) - \psi(v)\right| \to_p 0. \tag{5}$$

This implies

$$\sup_z \left|\hat{\psi}(\hat{\beta}^T z) - \psi({\beta_0}^T z)\right| \leq \sup_z \left|\hat{\psi}(\hat{\beta}^T z) - \psi(\hat{\beta}^T z)\right| + \sup_z \left|\psi(\hat{\beta}^T z) - \psi({\beta_0}^T z)\right| \to_p 0,$$

where the second term converges to zero by continuity of ψ and $\sqrt{n}$-consistency of $\hat{\beta}$. Theorem 1 is thus proved. □

Proof of Theorem 2

Proof. Let $\eta = H\gamma$, we can write the log local partial likelihood function in terms of β and η

$$\begin{aligned} l_n(\beta, \eta, \tau, v) = \int_0^\tau \frac{1}{n} \sum_{i=1}^n K_h(\beta^T Z_i - v) \Big[& (\beta^T \mathbf{Z}_i)^T H^{-1} \eta \\ & - \log\{\sum_i^n Y_i(u) \exp\{(\beta^T \mathbf{Z}_i)^T H^{-1} \eta\} K_h(\beta^T Z_i - v)\} \Big] dN_i(u). \end{aligned}$$

Accordingly let $S_{h,0}(\beta, \eta, u, v) = \frac{1}{n} \sum_{i=1}^n K_h(\beta^T Z_i - v) Y_i(u) \exp\{(\beta^T \mathbf{Z}_i)^T H^{-1} \eta\}$, and $S_{h,1}(\beta, \eta, u, v) = \frac{1}{n} \sum_{i=1}^n K_h(\beta^T Z_i - v) Y_i(u) \exp\{(\beta^T \mathbf{Z}_i)^T H^{-1} \eta\} (\beta^T \mathbf{Z}_i)^T H^{-1}$, and for a $\sqrt{n}$-consistent estimate $\hat{\beta}$ of β_0, Lemma 1 implies

$$\sup_{0 \leq u < \tau} \left| \frac{S_{h,1}(\hat{\beta}, \eta, u, v)}{S_{h,0}(\hat{\beta}, \eta, u, v)} - \nu_1 \right| \to_p 0,$$

where $\nu_1 = \int \mathbf{u} K(u) du$.

The derivative of $l_n(\beta, \eta, \tau, v)$ with respect to η evaluated at $\hat{\beta}$ and $\eta_0 = H\gamma_0$ is

$$\begin{aligned} & l_n'(\hat{\beta}, \eta_0, \tau, v) \\ = & \int_0^\tau \frac{1}{n} \sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v) \left[(\hat{\beta}^T \mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\hat{\beta}, \eta_0, u, v)}{S_{h,0}(\hat{\beta}, \eta_0, u, v)} \right] dN_i(u) \\ = & \int_0^\tau \frac{1}{n} \sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v) \left[(\hat{\beta}^T \mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\hat{\beta}, \eta_0, u, v)}{S_{h,0}(\hat{\beta}, \eta_0, u, v)} \right] dM_i(u) \\ + & \int_0^\tau \frac{1}{n} \sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v) \left[(\hat{\beta}^T \mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\hat{\beta}, \eta_0, u, v)}{S_{h,0}(\hat{\beta}, \eta_0, u, v)} \right] Y_i(u) \exp\{\psi({\beta_0}^T Z_i)\} \lambda_0(u) du \\ \equiv & U_n(\hat{\beta}, \eta_0, \tau, v) + B_n(\hat{\beta}, \eta_0, \tau, v). \end{aligned}$$

The first term

$$
\begin{aligned}
&U_n(\hat{\beta},\eta_0,\tau,v)\\
=&\int_0^t \frac{1}{n}\sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v)\left[(\hat{\beta}^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\hat{\beta},\eta_0,u,v)}{S_{h,0}(\hat{\beta},\eta_0,u,v)}\right]dM_i(u)\\
=&\int_0^t \frac{1}{n}\sum_{i=1}^n K_h({\beta_0}^T Z_i - v)\left[(\beta_0^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\beta_0,\eta_0,u,v)}{S_{h,0}(\beta_0,\eta_0,u,v)}\right]dM_i(u)\\
+&\int_0^t \frac{1}{n}\sum_{i=1}^n \Big(K_h(\hat{\beta}^T Z_i - v) - K_h({\beta_0}^T Z_i - v)\Big)\left[(\beta_0^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\beta_0,\eta_0,u,v)}{S_{h,0}(\beta_0,\eta_0,u,v)}\right]dM_i(u)\\
+&\int_0^t \frac{1}{n}\sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v)\left[(\hat{\beta}^T\mathbf{Z}_i - \beta_0^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\hat{\beta},\eta_0,u,v)}{S_{h,0}(\hat{\beta},\eta_0,u,v)} - \frac{S_{h,1}(\beta_0,\eta_0,u,v)}{S_{h,0}(\beta_0,\eta_0,u,v)}\right]dM_i(u)\\
\equiv&U_n(\beta_0,\eta_0,\tau,v) + V + VI.
\end{aligned}
$$

It is clear that $\sqrt{nh}U_n(\beta_0,\eta_0,t,v)$ is a martingale with predictable variation

$$
\begin{aligned}
&\langle\sqrt{nh}U_n(\beta_0,\eta_0,t,v), \sqrt{nh}U_n(\beta_0,\eta_0,t,v)\rangle\\
=&\frac{nh}{n^2}\sum_{i=1}^n\int_0^\tau K_h^2({\beta_0}^T Z_i - v)\left[(\beta_0^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\beta_0,\eta_0,u,v)}{S_{h,0}(\beta_0,\eta_0,u,v)}\right]^{\otimes 2} Y_i(u)\exp\{\psi({\beta_0}^T Z_i)\}\lambda_0(u)du.\\
=&f(v)\exp\{\psi(v)\}\Lambda(t,v)\int K^2(u)(\mathbf{u}-\nu_1)^{\otimes 2}du + o_p(1) \equiv \Sigma_U(t,v) + o_p(1),
\end{aligned}
$$

where the last step follows from Lemma 1.

The Lindberg conditions are satisfied (see [19] for details), we have thus proven that

$$
\sqrt{nh}U_n(\beta_0,\eta_0,\tau,v) \to_D N\big(0,\Sigma_U(\tau,v)\big). \tag{6}
$$

As for the term V and VI, similarly to the proof of Theorem 1, we have

$$
V = \int_0^\tau \frac{1}{n}\sum_{i=1}^n \Big(K_h(\hat{\beta}^T Z_i - v) - K_h({\beta_0}^T Z_i - v)\Big)\left[(\beta_0^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\beta_0,\eta_0,u,v)}{S_{h,0}(\beta_0,\eta_0,u,v)}\right]dM_i(u) = O_p(\frac{1}{nh^2}), \tag{7}
$$

and

$$
\begin{aligned}
VI =& \int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v)\left[(\hat{\beta}^T\mathbf{Z}_i - \beta_0^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\hat{\beta},\eta_0,u,v)}{S_{h,0}(\hat{\beta},\eta_0,u,v)} - \frac{S_{h,1}(\beta_0,\eta_0,u,v)}{S_{h,0}(\beta_0,\eta_0,u,v)}\right]dM_i(u)\\
=&O_p(\frac{1}{nh}).
\end{aligned} \tag{8}
$$

Applying Lemma 1 again and by Taylor expansion we get

$$
\begin{aligned}
B_n(\hat{\beta},\eta_0,\tau,v) =& f(v)\exp\{\psi(v)\}\frac{\psi^{(p+1)}(v)}{(p+1)!}\Lambda(\tau,v)\int K(u)(\mathbf{u}-\nu_1)u^{p+1}duh^{p+1} + o_p(h^{p+1}) + O_p(\frac{1}{\sqrt{n}})\\
=&b(\tau,v) + o_p(h^{p+1}) + O_p(\frac{1}{\sqrt{n}}).
\end{aligned} \tag{9}
$$

We have thus shown that, under (6), (7), (8) and (9),

$$\sqrt{nh}l_n'(\hat{\beta}, \eta_0, \tau, v) \to_D N\Big(b(\tau, v), \Sigma_U(\tau, v)\Big). \tag{10}$$

Next we focus on the property of the second derivative $l_n''(\hat{\beta}, \eta, t, v)$. Let $\hat{\eta} = H\hat{\gamma}$, by Taylor expansion and Lemma 1 we have

$$0 = l_n'(\hat{\beta}, \hat{\eta}, \tau, v) = l_n'(\hat{\beta}, \eta_0, \tau, v) + l_n''(\hat{\beta}, \eta^{**}, \tau, v)(\hat{\eta} - \eta_0), \tag{11}$$

where η^{**} lies in between $\hat{\eta}$ and η_0. Theorem 1 implies $\hat{\eta} \to_p \eta_0$, hence $\eta^{**} \to_p \eta_0$.

Using condition (*C1*) and boundedness of $\hat{\beta}^T Z$, we arrive at

$$l_n''(\hat{\beta}, \eta^{**}, \tau, v) = l_n''(\hat{\beta}, \eta_0, \tau, v) + o_p(1) = \Sigma_l(\tau, v) + o_p(1). \tag{12}$$

By (10), (11), (12) and Slutsky's theorem,

$$\begin{aligned}\sqrt{nh}(\hat{\eta} - \eta_0) &= \sqrt{nh}\Big[-l_n''(\hat{\beta}, \eta^{**}, \tau, v)^{-1} l_n'(\hat{\beta}, \eta_0, \tau, v)\Big] + o_p(1) \\ &\to_D N\Big(b(\tau, v), \Sigma_l(\tau, v)^{-1}\Sigma_U(\tau, v)\Sigma_l(\tau, v)^{-1}\Big).\end{aligned}$$

Simple calculations lead to the result in Theorem 2. □

References

[1] N. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30:89–99, 1974.

[2] C. H. Chen, K. C. Li, and J. L. Wang. Dimension reduction for censored regression data. *The Annals of Statistics*, 27:1–23, 1999.

[3] S. C. Cheng, L. J. Wei, and Z. Ying. Analysis of transformation models with censored data . *Biometrika*, 82:835–845, 1995.

[4] D. R. Cox. Regression models and lift-table (with discussion). *Journal of the Royal Statistcal Society, Ser.*, B. 4:187–220, 1972.

[5] D. R. Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975.

[6] D. M. Dabrowska and K. A. Doksum. Partial likelihood in transformation models with censored-data. *Scandinavian Journal of Statistics*, 15:1–23, 1988.

[7] K. A. Doksum. An extension of partial likelihood methods for proportional hazard models to general transformation models. *The Annals of Statistics*, 15:325–345, 1987.

[8] J. Fan, I. Gijbels, and M. King. Local likelihood and local partial likelihood in hazard regression. *The Annals of Statistics*, 25:1661–1690, 1997.

[9] T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis.* John Wiley & Sons, Inc., New York, 1991.

[10] R. Gentleman and J. Crowley. Local full likelihood estimation for the proportional hazards model. *Biometrics*, 47:1283–1296, 1991.

[11] R. J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 420:942–951, 1992.

[12] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 3:297–318, 1986.

[13] T. Hastie and R. Tibshirani. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 46:1005–1016, 1990.

[14] D. W. Hosmer, S. Lemeshow, and S. May. *Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition.* John Wiley & Sons, Inc., New York, 2008.

[15] J. Huang and L. Liu. Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics*, 62:793–802, 2006.

[16] F. O'Sullivan. Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing*, 9:531–542, 1988.

[17] L. A. Sleeper and D. P. Harrington. Regression splines in the cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, 85:941–949, 1990.

[18] R. Tibshirani and T. Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82:559–567, 1987.

[19] W Wang. *Proportional Hazards Model With Unknown Link Function and Applications To Time-to-event Longitudinal Data.* Ph.D. Thesis, 2001.

[20] W Wang. Proportional hazards regression with unknown link function and time-dependent covariates. *Statistica Sinica*, 14:885–905, 2004.