Published in final edited form as:
Stat Sin. 2012 January 1; 22(1): 47–68. doi:10.5705/ss.2010.161.

ANALYSIS ON CENSORED QUANTILE RESIDUAL LIFE MODEL VIA SPLINE SMOOTHING

Yanyuan Ma and
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, U.S.A.

Ying Wei
Department of Biostatistics, Columbia University, New York, NY, U.S.A.

Yanyuan Ma: ma@stat.tamu.edu; Ying Wei: ying.wei@columbia.edu

Abstract

We propose a general class of quantile residual life models, where a specific quantile of the residual life time, conditional on an individual has survived up to time *t*, is a function of certain covariates with their coefficients varying over time. The varying coefficients are assumed to be smooth unspecified functions of *t*. We propose to estimate the coefficient functions using spline approximation. Incorporating the spline representation directly into a set of unbiased estimating equations, we obtain a one-step estimation procedure, and we show that this leads to a uniformly consistent estimator. To obtain further computational simplification, we propose a two-step estimation approach in which we estimate the coefficients on a series of time points first, and follow this with spline smoothing. We compare the two methods in terms of their asymptotic efficiency and computational complexity. We further develop inference tools to test the significance of the covariate effect on residual life. The finite sample performance of the estimation and testing procedures are further illustrated through numerical experiments. We also apply the methods to a data set from a neurological study.

Key words and phrases

Censored data; nonparametric regression; quantile regression; residual life; spline

1. Introduction

Residual life is defined as the remaining time to event given the fact that the survival time T of a patient is at least t, i.e., $T - t|T \geq t$. In many clinical studies, especially when the associated diseases are chronic or/and incurable, knowing residual life is the major concern to patients. Modeling and estimating the mean of residual life has generated a large literature, for example, Oakes and Dasu (1990, 2003), Chen and Cheng (2005, 2006), Chen, Jewell and Cheng (2005), Müller and Zhang (2005), and Chen (2007). Compared with mean residual life models, quantile residual life models provide more complete and informative interpretation, especially when the distribution of the residual life is non-symmetric or skewed. Researches in this area are fairly recent, and include Jeong, Jung, and Costantino (2008), Jung, Jeong, and Bandos (2009), and Ma and Yin (2010). The quantile residual life models considered in the current literature focus on modeling and estimation at a single fixed t. Our interest here is in investigating the covariate effects along a range of times t. We take the covariate effects to be time variant, smooth functions of t in a varying coefficient quantile residual life model.

Our research is initially motivated by a clinical study on MELAS (mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes), which is a rare genetically-inherited neuroglial disease. Once the disease starts, MELAS patients suffer from progressive encephalopathy and stroke-like episodes that lead to disability and early death. There is as yet no effective treatment for this devastating condition, hence at each patient's hospital visit, both the patient and the clinician are mainly interested in how much longer the patient can survive. When a patient is known to be a carrier of such genotype as yet without the disease, the time to disease onset becomes of central interest. Quantile analysis is more informative comparing to the classical mean approach. For example, the patients might be more interested in knowing how long their remaining time is with a 90% probability, rather than in knowing the average residual time. Our proposed method answers such questions, taking into consideration the patient's characteristics.

We first represent the coefficient functions in the quantile residual life models by normalized B-splines, and estimate the spline coefficients using the residual life model jointly at different time points. This is what we refer to as one-step estimation. A second approach is a modification, in which we estimate the time varying coefficient function values at a set of different time points first, and then use a spline representation to approximate the coefficient functions based on estimated function values. This is what we refer to as two-step estimation. A similar two-step estimation is also used in a longitudinal data setting in Fan and Zhang (2000). We show a close link between the two estimation procedures, and point out computational advantage of the two-step procedure. We also study the large sample properties of the estimation procedures. To the best of our knowledge, this is the first time the residual life model has been considered simultaneously over a range of times.

The remainder of the paper is organized as follows. In Section 2, we present the quantile residual life model in its general form and show that the model is well-defined. We introduce two estimation procedures in Section 3. The one-step estimation procedure is discussed in Section 3.1, where we establish its root-*n* consistency and asymptotic normality. We further develop a simplified two-step estimation procedure in Section 3.2, and point out how the two estimation procedures are related in Section 3.3. Testing procedures are subsequently developed in Section 4, and we perform numerical analysis through simulation studies and a data analysis on the MELAS study in Section 5. We finish the paper with some discussion in Section 6, and collect the technical details and proofs in a web Appendix.

2. Censored Quantile Residual Life Model

Let (X_i, T_i, C_i), $i = 1, \ldots, n$, be identical and independently distributed (i.i.d.), where X_i is a covariate vector, T_i is the event (death) time, and C_i is a competing censoring time. Assume the censoring time C_i to be independent of the event time T_i and the covariate X_i. Let $Y_i = \min(T_i, C_i)$ and $D_i = I(T_i \le C_i)$, the binary index of censoring. As a typical situation in survival data analysis, we take the actual observations to be (X_i, Y_i, D_i) for $i = 1, \ldots, n$. For notational convenience, we assume the observations are sorted in increasing order, $0 < Y_1 \cdots \le Y_n$. The quantile residual life model we consider has the general form

$$Q_\tau(T_i - t | X_i, T_i \ge t) = m\{X_i, \boldsymbol{\beta}(t)\}, t \ge 0, \quad (2.1)$$

where $Q_\tau(T|A)$ denotes the τth conditional quantile function of a random variable T conditional on an event A, τ is a quantile level ranging between 0 and 1, and t is the time at which the residual life is considered. Here, $m(\cdot)$ is a parametric function of covariate X, while the parameter $\boldsymbol{\beta}(t) = \{\beta_1(t), \beta_2(t), \ldots, \beta_p(t)\}^T$ consists of p unknown smooth functions of t. Model (2.1) basically assumes that, given the covariate X_i, and the fact that $T_i > t$, the τth conditional quantile of the residual life $T_i - t$ can be characterized by a parametric function m with its coefficient $\beta(t)$ varying with time t. Our main interest is in estimating $\beta(t)$, as well as testing the effect of certain components in the covariate vector X. A special case of the model is the familiar linear varying-coefficient model,

$$Q_\tau(T_i - t|X_i, T_i \geq t) = X_i^{\mathrm{T}}\boldsymbol{\beta}(t), t \geq 0.$$

Here we let the first component of X_i be 1, hence the model includes a time-dependent intercept term. By taking into consideration that $\beta(t)$ is a smooth function of t, we can obtain a unified presentation of the residual life over a period of time, which is of interest in many applications. Moreover, compared to estimating the residual life at given times separately, we can achieve a more efficient estimator by estimating $\beta(t)$ globally.

Before proceeding to the estimation of $\beta(t)$, we first establish that there indeed exists a survival model that satisfies the quantile restriction in (2.1), simultaneously for all $t \geq 0$. Note that if the model is only required to hold at an arbitrary fixed t, identifiability is not an issue. If $S(t|X) = \Pr(T \geq t|X)$ is the survival function of T given the covariate X, then (2.1) can be written as

$$S[t+m\{X, \boldsymbol{\beta}(t)\}|X] = (1-\tau)S(t|X)$$

for any $t \geq 0$. This functional equation can be recognized as a special case of a Schröder's equation, and a solution for S exists as long as for all $t \geq 0$, $m\{X, \beta(t)\}$ is positive and continuous with respect to t, and $t + m\{X, \beta(t)\}$ is strictly increasing as a function of t (Gupta and Langford (1984)). These are moderate conditions and are easily satisfied for a large class of m functions. Hence the model in (2.1) is well defined and self-coherent. In the next section, we proceed to describe the estimation algorithm for $\beta(t)$.

Here and throughout the text, the o_p or O_p notation is component-wise in the case of vectors; $\|\cdot\|$ refers to the L_2 or l_2 norm according to the content.

3. Estimation

3.1. One-step estimation of β(*t*)

In this section, we propose one-step estimation equations for $\beta(t)$ based on normalized B-spline approximation. Specifically, we take $\mathbf{b}(t) = [\pi_1(t), \ldots, \pi_{k_n}(t)]^{\mathrm{T}}$ as k_n B-spline basis functions given a set of internal knots and the order of spline, and then approximate $\beta(t)$ by $\beta(t) \approx \alpha\mathbf{b}(t)$, where α is a $p \times k_n$ matrix of unspecified parameters. Although many other nonparametric methods exist in the literature, we use B-spline approximation due to its convenience in implementation. In this notation, (2.1) can be approximated by

$$Q_\tau(T_i - t|X_i, T_i \geq t) = m\{X_i, \boldsymbol{\alpha}\mathbf{b}(t)\}, t \geq 0.$$

For a fixed basis $\mathbf{b}(t)$, this can be treated as a parametric model. At any fixed $t = t_0$ and for a general m function, a slight modification of the estimator in Jung, Jeong, and Bandos (2009) yields the estimating equation

$$\sum_{i=1}^{n}\frac{\partial m\{X_i,\boldsymbol{\alpha}\mathbf{b}(t_0)\}}{\partial\alpha}\left(\frac{I[Y_i\geq t_0{+}m\{X_i,\boldsymbol{\alpha}\mathbf{b}(t_0)\}]}{G[t_0{+}m\{X_i,\boldsymbol{\alpha}\mathbf{b}(t_0)\}]}-(1-\tau)\frac{I(Y_i\geq t_0)}{G(t_0)}\right)=0. \quad (3.1)$$

Here α is a length pk_n vector formed by concatenating all the rows of $\boldsymbol{\alpha}$, G is the censoring process survival function, $G(t) = \Pr(C \geq t)$. In practice, G is typically estimated by a Kaplan-Meier estimator.

A careful inspection of $\partial m\{X_i, \alpha\mathbf{b}(t_0)\}/\partial\alpha$ reveals that it equals $\partial m\{X_i, \alpha\mathbf{b}(t_0)\}/\partial\{\alpha\mathbf{b}(t_0)\} \otimes \mathbf{b}(t_0)$, where $\otimes$ denotes a Kronecker product. Hence (3.1) includes only p independent estimating equations, hence does not suffice to estimate all the pk_n elements in $\boldsymbol{\alpha}$. However, since (2.1) holds for all $t > 0$, one can estimate $\boldsymbol{\alpha}$ by assembling a collection of equations of type (3.1) at ($t_j : j = 1, \ldots, J$), a set of distinctive values of the observed Y_i's. Specifically, we propose to obtain $\boldsymbol{\alpha}$ through minimizing

$$s(\boldsymbol{\alpha}){=}\sum_{j=1}^{J}\left\{\sum_{i=1}^{n}\frac{\partial m\{X_i,\boldsymbol{\alpha}\mathbf{b}(t_j)\}}{\partial\{\boldsymbol{\alpha}\mathbf{b}(t_j)\}}\left(\frac{I[Y_i\geq t_j{+}m\{X_i,\boldsymbol{\alpha}\mathbf{b}(t_j)\}]}{\hat{G}[t_j{+}m\{X_i,\boldsymbol{\alpha}\mathbf{b}(t_j)\}]}-(1-\tau)\frac{I(Y_i\geq t_j)}{\hat{G}(t_j)}\right)\right\}^{\otimes 2}, \quad (3.2)$$

where $\upsilon^{\otimes 2}$ denotes $\upsilon^{\mathrm{T}}\upsilon$ for any vector υ. Using $\hat{\boldsymbol{\alpha}}$ to denote the estimate obtained from minimizing (3.2), the estimator of $\beta(t)$ is $\hat{\beta}(t) = \hat{\boldsymbol{\alpha}}\mathbf{b}(t)$. Obviously, several tuning parameters need to be decided in this procedure. First, to select the number of basis functions k_n, we could use some standard selection criterion such as BIC. Specifically, we estimate $\hat{\alpha}(k_n)$ for a fixed candidate k_n, and form $s\{\alpha(k_n)\}$. We then select the optimal k_n through minimizing $s\{\alpha(k_n)\} + \log(n)k_n$. In the following, we discuss the selection of the other tuning parameters J and the t_j's that are more specific to the residual life model.

The choice of t_j's. One can choose an arbitrary set of times $t_1,\ldots, t_J$ in (3.2) as long as at least k_n of the J corresponding equations of the form (3.1) are linearly independent, for then the estimator given in (3.2) is uniquely defined. Note that this requires that $J \geq k_n$, so the number of distinctive event/censor times is larger than the number of B-spline basis functions. Our subsequent theoretical development further requires that there exist $\varepsilon > 0$ so that $J = o(n^{1/2-\varepsilon})$. A natural choice is to let $t_1 = 0$ and $t_{j+1} = Y_j$, the jth event or censoring time, for $j = 1, \ldots, J - 1$. Since the distribution of the Y_i is usually continuous over $[0, T]$, this choice generally satisfies the requirement. Computationally, when t_j increases, fewer observations contribute to the corresponding estimating equation. In addition, the estimated $\hat{G}$ also becomes less reliable. Hence, we recommend in practice to stop the summation over j in (3.2) at a value between k_n and one third of the total number of distinct Y_i values. The same rule is also applied to the two-step estimation approach introduced later in Section 3.2.

Asymptotic properties—In this section, we give the convergence rate and asymptotic distribution of $\hat{\beta}(t)$ and $\hat{\alpha}$ obtained from minimizing (3.2). Let $\mathbf{t} = (t_1, \ldots, t_J)$ and

$$s_i\{t,\boldsymbol{\beta}(t),G\}{=}\frac{\partial m\{X_i,\boldsymbol{\beta}(t)\}}{\partial\boldsymbol{\beta}(t)}\left(\frac{I[Y_i\geq t{+}m\{X_i,\boldsymbol{\beta}(t)\}]}{G[t{+}m\{X_i,\boldsymbol{\beta}(t)\}]}-(1-\tau)\frac{I(Y_i\geq t)}{G(t)}\right),\ f_i\{\boldsymbol{\alpha}\mathbf{b}(\mathbf{t}),G\}$$

$$=\left(\frac{\partial E[s_i^{\mathrm{T}}\{t_1,\boldsymbol{\alpha}\mathbf{b}(t_1),G\}]}{\partial\alpha},\ldots,\frac{\partial E[s_i^{\mathrm{T}}\{t_J,\boldsymbol{\alpha}\mathbf{b}(t_J),G\}]}{\partial\alpha}\right)\begin{bmatrix}s_i\{t_1,\boldsymbol{\alpha}\mathbf{b}(t_1),G\}\\ \vdots\\ s_i\{t_J,\boldsymbol{\alpha}\mathbf{b}(t_J),G\}\end{bmatrix}.$$

It is easy to see that the estimator in (3.2) satisfies an estimating equation of the form

$$n^{-1/2}\sum_{i=1}^{n} f_i\{\hat{\alpha}\mathbf{b}(\mathbf{t}), \hat{G}\} = o_p(Jn^{\varepsilon}),$$

for any $\varepsilon > 0$. In what follows, we outline conditions under which we derive the asymptotic properties of $\hat{\alpha}$.

A1 : The true time-varying coefficient vector $\beta_0(t)$ consists of p smooth functions defined on a closed interval $[0, T]$, and each of them has a bounded rth derivative with $r \geq 2$.

Under Condition **A1**, there exists a B-spline approximation $\alpha_0\mathbf{b}(t)$ and a constant C_1 such that $\sup_{t\in[0,T]} |\boldsymbol{\beta}_0(t) - \boldsymbol{\alpha}_0\mathbf{b}(t)| < C_1 k_n^{-r}$ (Schumaker (1981)).

A2 : The quantile function $m(x, \beta)$ has a bounded second derivative with respect to β.

Let

$$S_n\{\boldsymbol{\beta}(t)\} = \sum_{i=1}^{n} s_i\{t, \boldsymbol{\beta}(t), G\} = \sum_{i=1}^{n} \frac{\partial m\{X_i, \boldsymbol{\beta}(t)\}}{\partial \boldsymbol{\beta}(t)} \left(\frac{I[Y_i \geq t + m\{X_i, \boldsymbol{\beta}(t)\}]}{G[t + m\{X_i, \boldsymbol{\beta}(t)\}]} - (1-\tau)\frac{I(Y_i \geq t)}{G(t)} \right),$$

be the functional estimation equations for $\beta(t)$ (without B-spline approximations).

A3 : The functional estimating equation $ES_n\{\beta(t)\} = 0$ has a unique solution $\beta_0(t)$. In addition, there exist a compact set $\Omega \in R^{p+1}$ such that the p curves contained in $\beta_0(t)$ form an interior point of Ω. Note that this implies each curve in $\beta_0(t)$ is uniformly bounded.

A4 : The censoring survival function $G(t)$ and the event survival function $S(t)$ are differentiable; $g(t) = G'(t)$, and $s(t) = S'(t)$ are bounded away from zero and infinity and are bounded for all $t \in [0, T]$.

A5 : $\max_i \sup_t E\|s_i\{t, \beta(t), G\}\|^2 = O(1)$.

A6 : The first derivative of each component of $f_i\{\beta(t), G\}$ with respect to G is uniformly bounded. That is, there exists a constant $C > 0$ such that $|\partial f_i\{\beta(t), G\}/\partial G| < C$ for all $\beta(t)$, G and $i = 1, \ldots, n$.

With those conditions, we summarize the asymptotic properties of $\hat{\alpha}$ in two theorems. The proofs are deferred to a web Appendix.

Theorem 1. *Under* **A1–A6**, *if the number of B-spline basis function, k_n, satisfies $n^{1/4r} << k_n << n^{1/4}$ for $r \geq 2$, then*

$$\|\hat{\alpha} - \boldsymbol{\alpha}_0\|^2 = O_p\left(\frac{k_n}{n}\right). \quad (3.3)$$

It follows from the Theorem 1 that $\hat{\beta}(t)$ is uniformly consistent for $t \in [0, T]$, i.e. $\sup_{t\in[0,T]} \|\hat{\beta}(t) - \beta(t)\|^2 = O_p(k_n/n)$. We now define the following notations relate to the asymptotic distribution of $\hat{\alpha}$. Let

$$M=\left(\frac{\partial E[s_i^{\mathrm{T}}\{t_1,\boldsymbol{\alpha}_0\mathbf{b}(t_1),G\}]}{\partial\alpha},\ldots,\frac{\partial E[s_i^{\mathrm{T}}\{t_J,\boldsymbol{\alpha}_0\mathbf{b}(t_J),G\}]}{\partial\alpha}\right),\ \mathscr{D}=\mathrm{cov}\left[\{\nu_i^{\mathrm{T}}(t_1,\boldsymbol{\beta}_1,G),\ldots,\nu_i^{\mathrm{T}}(t_J,\boldsymbol{\beta}_J,G)\}^{\mathrm{T}}\right],$$

where

$$\begin{aligned}&\nu_i(t_j,\boldsymbol{\beta}_j,G)\\&\quad=s_i(t_j,\boldsymbol{\beta}_j,G)\\&\quad-\mathbf{q}_2(\boldsymbol{\beta}_j,t_j)\int_{-\infty}^{t_j}h^{-1}(s)\{dI(Y_i\le s,D_i\\&\quad=0)-I(Y_i\ge s)d\Lambda_G(s)\}+\int_{-\infty}^{\infty}G^{-1}(s)\int_{-\infty}^{s}h^{-1}(v)\{dI(Y_i\le v,D_i\\&\qquad=0)-I(Y_i\ge v)d\Lambda_G(v)\}d\mathbf{q}_1(\boldsymbol{\beta}_j,s).\end{aligned}$$

Here $h(s) = E(Y_1 \geq s)$, Λ_G is the cumulative hazard function of the censoring process, and

$$\mathbf{q}_1(\boldsymbol{\beta}_j,s)=E\left[\frac{\partial m(X_i,\boldsymbol{\beta}_j)}{\partial\boldsymbol{\beta}_j}I\{t_j+m(X_i,\boldsymbol{\beta}_j)\le\min(s,Y_i)\}\right],\ \mathbf{q}_2(\boldsymbol{\beta}_j,t_j)=(1-\tau)G(t_j)^{-1}E\left\{I(Y_i\ge t_j)\frac{\partial m(X_i,\boldsymbol{\beta}_j)}{\partial\boldsymbol{\beta}_j}\right\}.$$

The difference of s_i and ν_i is a consequence of the Kaplan-Meier estimation of $\check{G}(t)$. With this notation, the following theorem summarizes the asymptotic distribution of α.

Theorem 2. *Under* **A1–A6**, *for any* $\eta \in R^{pk_n}$ *and* $\|\eta\| = 1$, $n^{1/2}\eta^{\mathrm{T}}(\hat{\alpha} - \alpha_0)/\sigma \to N(0, 1)$ *in distribution when* $n \to \infty$, *where* $\sigma^2 = \eta^{\mathrm{T}}\mathscr{V}$, *and* $\mathscr{V} = \mathscr{B}\mathscr{D}^{\mathrm{T}}$, $\mathscr{B} = (MM^{\mathrm{T}})^{-1}M$.

From Theorem 2, we can see that estimating the censoring process survival function $\hat{G}(t)$ does not bring additional bias while it has an impact on the estimation variance. With $\hat{\beta}(t) = \hat{\alpha}\mathbf{b}(t) = \hat{\alpha}^{\mathrm{T}}\{I_p \otimes \mathbf{b}(t)\}$, it follows from Theorem 2 that, for any given time t, $\hat{\beta}(t)$ is asymptotically normal with mean $\beta(t)$ and variance-covariance matrix $\{I_p \otimes \mathbf{b}(t)\}^{\mathrm{T}}\mathscr{V} I_p \otimes \mathbf{b}(t)\}/n$.

The one-step estimation procedure introduced here requires intensive computation, since the dimension of the unknown parameter α in (3.2) is pk_n. In Section 3.2, we propose a two-step approach to reduce the computational burden. A discussion comparing the two approaches is provided in Section 3.3.

3.2. An alternative two-step estimation approach

It is worth noting that the nonparametric function estimation for $\beta(t)$ differs from the conventional one in an important aspect. In a classical regression, $\beta(t)$ contributes to a relation as a function of all t's in a valid range; in the residual life model, $\beta(t)$ contributes to a relation only as a function value at a fixed t. At different t_j, $\beta(t_j)$ is subject to a different requirement. Other than $\beta(t)$ being sufficiently smooth, $\beta(t_j)$ at different t_j values are not inherently related via these requirements. Thus, intuitively, one can estimate $\beta(t)$ at a large set of t values to obtain $\{t_j, \check{\beta}(t_j)\}$, $j = 1, \ldots, J$, and then use these as pseudo observations to perform a nonparametric fitting using, say, splines.

To be precise, select t_j, $j = 1, \ldots, J$, as in Section 3.1. At each t_j, obtain $\check{\beta}(t_j)$ from

$$\sum_{i=1}^{n} s_i\{t_j,\boldsymbol{\beta}(t_j),\hat{G}\}=\sum_{i=1}^{n}\frac{\partial m\{X_i,\boldsymbol{\beta}(t_j)\}}{\partial\boldsymbol{\beta}(t_j)}\left(\frac{I[Y_i\geq t_j+m\{X_i,\boldsymbol{\beta}(t_j)\}]}{\hat{G}[t_j+m\{X_i,\boldsymbol{\beta}(t_j)\}]}-(1-\tau)\frac{I(Y_i\geq t_j)}{\hat{G}(t_j)}\right)=0, \tag{3.4}$$

then obtain an estimator of α from minimizing $\sum_{j=1}^{J}\{\boldsymbol{\alpha}\mathbf{b}(t_j)-\check{\beta}(t_j)\}^{\otimes 2}$. The minimizer $\tilde{\alpha}$ has the explicit form

$$\tilde{\alpha}=\left\{\sum_{j=1}^{J}\check{\beta}(t_j)\mathbf{b}^{\mathrm{T}}(t_j)\right\}\left\{\sum_{j=1}^{J}\mathbf{b}(t_j)\mathbf{b}^{\mathrm{T}}(t_j)\right\}^{-1}.$$

As before, we construct the two-step estimator of β(*t*) using $\tilde{\beta}(t)=\tilde{\alpha}\mathbf{b}(t)$.

Note that the estimator $\tilde{\alpha}$ can be written as

$$\tilde{\alpha}=\sum_{j=1}^{J}I_p\otimes\left[\left\{\sum_{j=1}^{J}\mathbf{b}(t_j)\mathbf{b}^{\mathrm{T}}(t_j)\right\}^{-1}\mathbf{b}(t_j)\right]\check{\beta}(t_j)=\mathscr{C}\left(\check{\beta}(t_1)^{\mathrm{T}},\ldots,\check{\beta}(t_J)^{\mathrm{T}}\right)^{\mathrm{T}}, \tag{3.5}$$

where

$$\mathscr{C}\equiv\left(I_p\otimes\left[\left\{\sum_{j=1}^{J}\mathbf{b}(t_j)\mathbf{b}^{\mathbf{T}}(t_j)\right\}^{-1}\mathbf{b}(t_1)\right],\ldots,I_p\otimes\left[\left\{\sum_{j=1}^{J}\mathbf{b}(t_j)\mathbf{b}^{\mathrm{T}}(t_j)\right\}^{-1}\mathbf{b}(t_J)\right]\right),$$

and I_p stands for *p*-dimensional identity matrix. Since $\check{\beta}(t_j)\to\beta(t_j)$ for any $1\leq j\leq J$, as long as the B-spline basis is adequate, $\tilde{\alpha}\mathbf{b}(t)$ is a consistent estimate of the true coefficient function β(*t*). Let $\mathscr{M}=\mathrm{diag}(\partial E\,[s_i\{t_1,\beta(t_1),G\}]/\partial\beta(t_1)^{\mathrm{T}},\ldots,\partial E\,[s_i\{t_J,\beta(t_J),G\}]/\partial\beta(t_J)^{\mathrm{T}})$,,

Theorem 3. *Under the regularity conditions of Theorem 1, for any* $\eta\in R^{k_n}$ *and* $\|\eta\|=1$, *we have* $n^{1/2}\eta^{\mathrm{T}}(\tilde{\alpha}-\alpha_0)/\sigma\to N(0,1)$ *in distribution when* $n\to\infty$, *where* $\sigma^2=\eta^{\mathrm{T}}\mathscr{W}$, *where* $\mathscr{W}=\mathscr{C}^{-1}\mathscr{D}\mathscr{M}^{-1})^{\mathrm{T}}\mathscr{C}^{\mathrm{T}}$.

Theorem 3 can be proved similarly as Theorem 2 by grouping *J* estimating equations in (3.4) together and using the relation (3.5). We omit the proof.

As before, it follows from Theorem 3 that, for any given time *t*, $\tilde{\beta}(t)$ is asymptotically normal with mean β(*t*) and variance-covariance matrix $\{I_p\otimes\mathbf{b}(t)\}^{\mathrm{T}}\mathscr{W}I_p\otimes\mathbf{b}(t)\}$.

3.3. Relation between one-step and two-step approaches

The two estimation approaches are essentially two ways of linking point-wise curve estimation and the spline curve representation. Specifically, the one-step estimation imposes the spline representation before forming an estimate, with $\hat{\alpha}$ obtained through a one-step optimization. In contrast, the two-step approach forms an estimate at various time points first, then links the results to the spline representation. In terms of computational cost, the one-step estimation is more expensive, since it means solving a pk_n-dimensional estimation equation, while the two-step approach solves *J* separate *p*-dimensional estimating equations, followed by a simple matrix-vector multiplication. Both estimators are consistent and enjoy asymptotically normality. We now investigate in detail the estimation efficiency of the two approaches.

Recall that $\mathscr{V}$ and $\mathscr{W}$ are limiting variance-covariance matrices for the one-step estimator $\hat{\alpha}$ and two-step estimator $\tilde{\alpha}$, respectively. The two matrices share the same pivotal component $\mathscr{D}$ To understand the it differences, we first establish the association between M and $\mathcal{M}$.

Let $(M^{\mathrm{T}})_{jl}$ be the (j, l)th size $p \times k_n$ block of M^{T}, $\mathcal{M}_{jj}$ the (j, j)th size $p \times p$ block of $\mathcal{M}$, $\mathscr{C}_{jl}$ the (j, l)th size $k_n \times p$ block of $\mathscr{C}$ e_l the length p vector with ith entry 1 and all others 0. Then

$$(M^{\mathrm{T}})_{jl}=\frac{\partial E\{s_i(t_j,\boldsymbol{\alpha}\mathbf{b}(t_j),G)\}}{\partial\alpha_l^{\mathrm{T}}}=\mathscr{M}_{jj}e_l\mathbf{b}^{\mathrm{T}}(t_j),$$

$$\mathscr{C}_{lj'}=e_l^{\mathrm{T}}\otimes\left[\left\{\sum_{j=1}^{J}\mathbf{b}(t_j)\mathbf{b}^{\mathrm{T}}(t_j)\right\}^{-1}\mathbf{b}(t_{j'})\right],$$

$$(M^{\mathrm{T}}\mathscr{C})_{jj'}=\mathscr{M}_{jj}\sum_{l=1}^{p}e_l\mathbf{b}^{\mathrm{T}}(t_j)\left(e_l^{\mathrm{T}}\otimes\left[\left\{\sum_{j=1}^{J}\mathbf{b}(t_j)\mathbf{b}^{\mathrm{T}}(t_j)\right\}^{-1}\mathbf{b}(t_{j'})\right]\right)=\mathbf{b}^{\mathrm{T}}(t_j)\left\{\sum_{j=1}^{J}\mathbf{b}(t_j)\mathbf{b}^{\mathrm{T}}(t_j)\right\}^{-1}\mathbf{b}(t_{j'})\mathscr{M}_{jj}.$$

Assembling the blocks of $M^{\mathrm{T}}\mathscr{C}$ and defining a hat matrix

$$H=\{\mathbf{b}(t_1),\ldots,\mathbf{b}(t_J)\}^{\mathrm{T}}\left\{\sum_{j=1}^{J}\mathbf{b}(t_j)\mathbf{b}^{\mathrm{T}}(t_j)\right\}^{-1}\{\mathbf{b}(t_1),\ldots,\mathbf{b}(t_J)\},$$

we obtain the relationship between M and $\mathcal{M}$ that $M^{\mathrm{T}}\mathscr{C}= \mathcal{M}(H\otimes I_p)$. Therefore,

$$\begin{aligned}\mathscr{A}(\mathscr{W}-\mathscr{V})\mathscr{A}^{\mathrm{T}}=&M\{M^{\mathrm{T}}\mathscr{C}\mathscr{M}^{-1}\mathscr{D}(\mathscr{M}^{-1})^{\mathrm{T}}\mathscr{C}^{\mathrm{T}}M\\&-\mathscr{D}\}M^{\mathrm{T}}\\=&M\{\mathscr{M}(H\otimes I_p)\mathscr{M}^{-1}\mathscr{D}(\mathscr{M}^{-1})^{\mathrm{T}}(H\otimes I_p)^{\mathrm{T}}\mathscr{M}^{\mathrm{T}}-\mathscr{D}\}M^{\mathrm{T}}=M\mathscr{M}\{(H\otimes I_p)\mathscr{M}^{-1}\mathscr{D}(\mathscr{M}^{-1})^{\mathrm{T}}(H\otimes I_p)\\&-\mathscr{M}^{-1}\mathscr{D}(\mathscr{M}^{-1})^{\mathrm{T}}\}\mathscr{M}^{\mathrm{T}}M^{\mathrm{T}}.\end{aligned}$$

Consequently, $\mathscr{W}$ can be written as

$$\mathscr{W}=\{\mathscr{A}^{-1}M\mathscr{M}(H\otimes I_p)\mathscr{M}^{-1}\}D\{\mathscr{A}^{-1}M\mathscr{M}(H\otimes I_p)\mathscr{M}^{-1}\}^{\mathrm{T}},$$

while the component $H \otimes I_p$ is simply replaced by the identity matrix in the expression for $\mathscr{V}$ We conclude the following,

1. If $J = k_n$, then $\mathscr{W} = \mathscr{V}$ and the two estimators are equivalent. For, in this case, $H = I_{k_n}$.
2. If $J > k_n$, then $\mathscr{W} - \mathscr{V}$ can have zero, positive, and negative eigenvalues, and hence there is no definitive winner between the two estimators in terms of efficiency.

In practice, a relative small k_n is often sufficient to approximate the smooth components in $\beta(t)$. However, to fully utilize the model structure in (2.1), as long as computational stability is retained, one would choose a large J. Thus, $J = k_n$ almost never happens in reality. The two-step estimator has appealing computational advantages over the one-step estimator and, at the same time, is not inferior in terms of estimation efficiency. We hence recommend using the two-step procedure in practice. In fact, both the one-step and the two-step procedures can be improved through better weighting, as we now discuss.

Equivalence between optimized $\hat{\alpha}$ and $\tilde{\alpha}$—We could view the one-step estimator $\hat{\alpha}$ and the two-step estimator $\tilde{\alpha}$ as special cases of two families of estimations. First, instead of forming the sum of squares of the estimating equation terms, we could form the sum of weighted squares using the Generalized Method of Moments (GMM). We define a family of GMM estimator by

$$\hat{\alpha}\mathbf{w}_1 = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \begin{bmatrix} s_i\{t_1, \boldsymbol{\alpha}\mathbf{b}(t_1), \hat{G}\} \\ \cdots \\ s_i\{t_J, \boldsymbol{\alpha}\mathbf{b}(t_J), \hat{G}\} \end{bmatrix} \right)^{\mathrm{T}} \mathbf{W}_1 \left(\sum_{i=1}^{n} \begin{bmatrix} s_i\{t_1, \boldsymbol{\alpha}\mathbf{b}(t_1), \hat{G}\} \\ \cdots \\ s_i\{t_J, \boldsymbol{\alpha}\mathbf{b}(t_J), \hat{G}\} \end{bmatrix} \right), \quad (3.6)$$

where $\mathbf{W}_1$ is a pJ × pJ-dimensional weight matrix. It is easy to see that, when $\mathbf{W}_1 = I_{pJ}$ is the identity matrix, the GMM estimating equations at (3.6) reduce those at (3.2), and consequently $\hat{\alpha}$ is a special case of a GMM estimator. Second, we define a family of Weighted Least Squares (WLS) estimators by

$$\tilde{\alpha}\mathbf{w}_2 = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \begin{Bmatrix} \boldsymbol{\alpha}\mathbf{b}(t_1) - \check{\beta}(t_1) \\ \cdots \\ \boldsymbol{\alpha}\mathbf{b}(t_J) - \check{\beta}(t_J) \end{Bmatrix}^{\mathrm{T}} \mathbf{W}_2 \begin{Bmatrix} \boldsymbol{\alpha}\mathbf{b}(t_1) - \check{\beta}(t_1) \\ \cdots \\ \boldsymbol{\alpha}\mathbf{b}(t_J) - \check{\beta}(t_J) \end{Bmatrix},$$

where $\mathbf{W}_2$ is a pJ × pJ-dimensional weight matrix. Similarly, the two-step estimator $\tilde{\alpha}$ is a special case of an WLS estimator when $\mathbf{W}_2 = I_{pJ}$ is the identity matrix.

It is well-known that the most efficient WLS estimator is reached when the weight matrix $\mathbf{W}_2$ is the inverse of the variance-covariance matrix of $\check{\beta}$. i.e. $\mathbf{W}_2 = \{\mathcal{W}^{-1}\mathcal{D}\mathcal{W}^{-1})^{\mathrm{T}}\}^{-1}$. The same results in Theorem 3 holds for the optimal WLS estimator with limiting matrix $\mathcal{W}$ replaced by $\tilde{\mathcal{W}} = (F\mathcal{W}^{\mathrm{T}}\mathcal{D}^{-1}\mathcal{W}F^{\mathrm{T}})^{-1}$, where $F = \{I_p \otimes \mathbf{b}(t_1), \ldots, I_p \otimes \mathbf{b}(t_J)\}$. On the other hand, the most efficient GMM estimator is the one with $\mathbf{W}_1 = \mathcal{D}^{-1}$, the inverse of the variance-covariance matrix of the estimating equations that invokes (3.1) at $t_1, \ldots, t_J$. The resulting optimal GMM estimator has a limiting variance-covariance matrix $\tilde{\mathcal{V}} = ((M \mathcal{D}^{-1}M^{\mathrm{T}})^{-1}$ to the first order. It is not difficult to verify that $M = F\mathcal{W}^{\mathrm{T}}$ to the first order, hence the estimation variance of the optimal one-step GMM estimator is the same to the first order as the optimal two-step WLS estimator, that is, they are effectively equivalent.

Although ideally a weighted approach should be used, it is not recommended in practice, because the optimal weights involve density estimation in the quantile regression framework. This is known to be unreliable. One could use the bootstrap to generate the variance estimation, but it is computationally undesirable. For these reasons we focus our discussion on the unweighed estimations $\hat{\beta}(t)$ and $\tilde{\beta}(t)$.

We summarize the differences between the two estimations as follows. First, their constructions are different. The one-step estimation incorporates the spline representations first to reduce the problem into a parameter estimation problem, it then constructs different estimating equations at various times. Because the number of parameters is likely smaller than the number of the resulting estimating equations, it falls in the category of GMM estimation. The two-step approach estimates the function values at various individual time points, then links the results to the spline representation. Because the number of spline coefficients is likely smaller than the function values obtained, this falls to the linear regression category and calls for a LS criterion. Computational complexities are also different. The two-step approach reduces a pk_n-dimensional estimation equation to J p-dimensional estimations. Hence it is computationally less challenging and more efficient. Finally, in terms of estimation variance, the two approaches are equally efficient when the respective optimal weights are used. The proposed $\hat{\alpha}$ and $\tilde{\alpha}$ use equal weights, which leads to different efficiency losses. The efficiency loss of $\hat{\alpha}$ is the type of loss seen in the regression context with heteroscedastic error when the error variance is treated as a constant. Hence the amount of loss in practice depends on the error variance structure. The efficiency loss of $\tilde{\alpha}$ is the type of loss seen under the GMM framework when the performance difference of available estimation equations and their correlations are ignored. Hence the amount of loss in practice depends on how different and how correlated these estimating equations are. Neither of $\hat{\alpha}$ and $\tilde{\alpha}$ is uniformly better than the other. When we reduce the dimension of the estimation equations to the parsimonious case of $J = k_n$, the two estimators are again equivalent. Accordingly, we recommend the two-step approach in practice.

4. Inference Tools

Once we obtain the estimate of $\beta(t)$, the next step is to test the covariate effect on the τth quantile of residual life. To this end, we write $\boldsymbol{\beta}(t)=\{\boldsymbol{\beta}_1^{\mathrm{T}}(t),\boldsymbol{\beta}_2^{\mathrm{T}}(t)\}^{\mathrm{T}}$, where $\beta_1(t)$ and $\beta_2(t)$ are p_1- and p_2-dimensional sub-vectors of $\beta(t)$, ($p_1 \le p$, $p_2 \le p$). We assume that, through proper parameterization, interest is in testing whether $\beta_2(t)$ is a zero function. Thus, the null and alternative hypotheses are respectively

$$H_0{:}\boldsymbol{\beta}_2(t)=0\forall_t \ \mathrm{and} H_1{:}\boldsymbol{\beta}_2(t) \neq 0 \mathrm{for\ \ at\ \ least\ \ one} t \in [0,T]. \quad (4.1)$$

Under the same spline representation for $\beta(t)$, testing (4.1) is equivalent to testing

$$H_0{:}\boldsymbol{\alpha}_2=0, \mathrm{vs} H_1{:}\boldsymbol{\alpha}_2 \neq 0.$$

Here, $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_1^{\mathrm{T}},\boldsymbol{\alpha}_2^{\mathrm{T}})^{\mathrm{T}}$, where α_1 and α_2 are sub-matrixes of α with dimensions $p_1 \times k_n$ and $p_2 \times k_n$, respectively, the spline coefficients associated with $\beta_1(t)$ and $\beta_2(t)$. Under Theorems 2 and 3, we could construct Wald-type statistics

$$T_{n,1}=n\hat{\alpha}_2^{\mathrm{T}}\mathscr{V}_{22}^{-1}\hat{\alpha}_2 \mathrm{and} T_{n,2}=n\tilde{\alpha}_2^{\mathrm{T}}\mathscr{W}_{22}^{-1}\tilde{\alpha}_2,$$

where $\mathscr{V}_{22}$ and $\mathscr{W}_{22}$ are, respectively, $p_2k_n \times p_2k_n$ lower-right sub-matrixes of $\mathscr{V}$ and $\mathscr{W}$ associated with α_2. Under the null hypothesis, $T_{n,1}$ and $T_{n,2}$ are asymptotically chi-square distributed with degrees of freedom p_2k_n.

Both $\mathscr{V}_{22}$ and $\mathscr{W}_{22}$ involve unknown components that need to be estimated empirically. Two estimation approaches exist. In large sample situations, we can use asymptotic results. To estimate M in $\mathscr{V}$ we use the sample average to replace the expectation, i.e., $E[s_i^{\mathrm{T}}\{t,\boldsymbol{\alpha}_0\mathbf{b}(t),G\}] \approx n^{-1}\sum_{i=1}^{n} s_i^{\mathrm{T}}\{t,\hat{\alpha}\mathbf{b}(t),\hat{G}\}$, and use a numerical difference for the derivative. Following the same line, we can estimate $\mathscr{M}$ in $\mathscr{W}$ In the web Appendix we show that $\mathscr{D}$can be approximated by

$$\hat{\mathscr{D}}=n^{-1}\sum_{i=1}^{n}\{\hat{\nu}_i(t_1,\hat{\beta}(t_1),\hat{G})^{\mathrm{T}},\ldots,\hat{\nu}_i(t_J,\hat{\beta}(t_J),\hat{G})^{\mathrm{T}}\}^{\mathrm{T}}\{\hat{\nu}_i(t_1,\hat{\beta}(t_1),\hat{G})^{\mathrm{T}},\ldots,\hat{\nu}_i(t_J,\hat{\beta}(t_J),\hat{G})^{\mathrm{T}}\},$$

where

$$\hat{\nu}_i^{\mathrm{T}}(t_j,\hat{\beta}(t_j),\hat{G})=s_i\{t_j,\hat{\beta}(t_j),\hat{G}\}+n^{-1}\sum_{l=1}^{n}\frac{\partial m\left\{X_l,\hat{\beta}(t_j)\right\}}{\partial\hat{\beta}(t_j)}\hat{G}^{-1}\left[t_j+m\left\{X_l,\hat{\beta}(t_j)\right\}\right]I\left[t_j+m\left\{X_l,\hat{\beta}(t_j)\right\}\leq Y_l\right]\times$$
$$\left\{\hat{h}^{-1}(Y_i)(1-D_i)I\left[Y_i\leq t_j+m\left\{X_l,\hat{\beta}(t_j)\right\}\right]-n^{-1}\sum_{k=1}^{n}\hat{h}^{-2}(Y_k)(1-D_k)I\left(Y_k\leq \min\left[Y_i,t_j+m\left\{X_l,\hat{\beta}(t_j)\right\}\right]\right)\right\}-\hat{\mathbf{q}}_2(\hat{\beta}(\qquad (4.2$$

$\hat{h}(s)=n^{-1}\sum_{i=1}^{n}I(Y_i\geq s)$, and

$$\hat{\mathbf{q}}_2\{\hat{\mathbf{b}}(t_j),t_j\}=(1-\tau)\hat{G}(t_j)^{-1}n^{-1}\sum_{i=1}^{n}\left[I(Y_i\geq t_j)\frac{\partial m\left\{X_i,\hat{\beta}(t_j)\right\}}{\partial\hat{\beta}(t_j)}\right].$$

We then assemble empirical estimates of $\mathscr{V}_{22}$ and $\mathscr{W}_{22}$, and substitute for the true ones in $T_{n,1}$ and $T_{n,2}$.

A more precise and stable approach is to use the bootstrap method to estimate $\mathscr{V}_{22}$ and $\mathscr{W}_{22}$, especially when sample size is not large. The cost of bootstrap is in computation intensity. This is the standard bootstrap used in quantile regression, so we omit implementation details. We also do not propose the score test for the quantile residual life model because, in the model we consider, the score test loses its typical advantage over Wald test. Since the residual life model holds for all t, we also need $E\{f_{T|X}(t|X)X_2X_1^{\mathrm{T}}\}=0$ for all t to enjoy the advantage of the score test. Such condition cannot be satisfied without violating the original model assumptions. For this reason, the score test is not recommended in our context.

5. Numerical Results

5.1. Simulation

We conducted simulation studies to investigate the finite sample performance of the proposed method. The first quantile residual life we study has the form $Q_\tau(T_i - t|X_i, T_i \geq t) = \beta_1(t) + \beta_2(t)X_i$, where X_i is the ith subject's covariate, $\tau = 0.5$, and $\beta_1(t)$ and $\beta_2(t)$ are time varying intercept and slope coefficients that are linear functions of t: $\beta_1(t) = \beta_{1c} + \beta_{1l}t$ and $\beta_2(t) = \beta_{2c} + \beta_{2l}t$. We write $\beta_c = \beta_{1c} + \beta_{2c}$ and $\beta_l = \beta_{1l} + \beta_{2l}$. To generate data sets, we adopt the model with survival function

$$S(t|X_i)=\left\{1+\frac{t(X_i^{\mathrm{T}}\beta_l)}{X_i^{\mathrm{T}}\beta_c)}\right\}^{\log(1-\tau)/\log(1+X_i^{\mathrm{T}}\beta_l)}=(1-\tau)^{\log\{1+tX_i^{\mathrm{T}}\beta_l/(X_i^{\mathrm{T}}\beta_c)\}/\log(1+X_i^{\mathrm{T}}\beta_l)}.$$

Data generated here satisfy the quantile residual life model as long as $\beta_1(t)$ and $\beta_2(t)$ do not simultaneously degenerate to a constant function. In fact, when the slopes in both $\beta_1(t)$ and $\beta_2(t)$ are zero, the above survival function does not exist, and we need to generate data from

$$S(t|X_i)=e^{t\log(1-\tau)/X_i^{\top}\beta_c}=(1-\tau)^{t/X_i^{\top}\beta_c}$$

in order to satisfy the corresponding quantile residual life model. We study all four situations, in which $\beta_1(t)$, $\beta_2(t)$ can be either a linear function or a constant function.

We generated the covariates X_i's from a uniform distribution in [0, 2], and we generated the censoring distribution from a mixture of infinity and an exponential distribution, so that the censoring rate was approximate 15%–20%. The sample sizes were n = 100, 200, 300, and 1,000, and we chose the first one-third of the y_i values to form the t_j values in calculating the $\hat{\beta}_j$'s. One could of course put more values into the collection of t_j's, but, as pointed in Section 3.1, the effective samples participating in the estimation of $\hat{\beta}_j$ are the ones with $Y_i \geq t_j$. Thus, a larger value of t_j yields less efficient estimation of $\hat{\beta}_j$. When the effective sample size is too small, the asymptotic results may not be relevant, and various numerical issues also occur. For computational stability, we chose the t_j values to ensure that there were at least one third of the observations contributing to the estimation of $\hat{\beta}_j$. The quadratic spline basis functions were selected, we put four knots at the boundary and equally spaced positions between the boundaries. A total of 1,000 simulations was conducted in each model.

To illustrate the performance of the method on a non-polynomial functional form of $\beta(t)$, we conducted a second simulation study with the quantile residual function $m\{X, \beta(t)\} = e^{-2X}\beta_1(t) + e^{-X}\beta_2(t)$. Here $\beta_1(t) = \{\log(1-\tau)\}^2$ and $\beta_2(t) = -2\sqrt{a+t}\log(1-\tau)$ at a = 0.01. It can be verified that the random process with survival function

$$S(t|X_i)=\exp\{-e^{X_i}(\sqrt{t+a}-\sqrt{a})\}$$

yields a τth quantile residual of the desired form. We generated the covariate X_i's from a uniform distribution in [−1.5, 0.5] and a censoring time from a mixture of infinity and exponential distribution as before to retain a similar censoring rate.

We present the mean squared error (MSE) of the estimation for different models and sample sizes in Table 1. Here for each function $\beta_j(t)$, the MSE was calculated using $\sum_i(\hat{\beta}_j(t_i) - \beta_j(t_i))^2$, where t_i's are equally spaced on the range of t considered. For comparison, we also present the MSE of the estimation with smoothing the pointwise estimates of $\beta(t)$. Clearly, for all the models and the sample sizes, our method yields an estimate with smaller MSE than the pointwise procedure. The improvement is especially dramatic when sample size is small or moderate. When sample sizes are 1,000, presented for comparison purposes, the improvement becomes less impressive, although still quite important. To provide a visual inspection of the estimation for both the linear and nonlinear model, we also plotted the mean estimated curves together with the true curves and 90% pointwise confidence bands in Figure 1. As can be seen, the estimated average curves are rather close to the truth, indicating the validity of our proposal. We point out here that the confidence bands in Figure 1 contain a constant curve, hence at the 10% level, one may not conclude that the varying coefficient model is really necessary. This is caused by the small sample size. With n =1,000, the 90% confidence bands corresponding to the nonlinear true function no longer contain any constant curve.

We also implemented the Wald test procedure, where the interest is in testing whether the slope function in the linear models or the coefficient function of e^{-X} in the nonlinear model $\beta_2(t)$ is the zero function. To test the level precision, we let $\beta_1(t) = 1$ and $1 + t$ for the linear models, and generated the data from $S(t|X_i) = \exp\{te^{2X_i}/\log(1-\tau)\}$ for the nonlinear model. We considered the levels 0.01, 0.05 and 0.1. The results for various sample sizes are given in Table 2. They indicate that the test levels are close to the nominal values when the sample size is n = 300 for the linear models, while they generally perform well even for smaller sample sizes for the nonlinear model. We point out that, because of the nature of the model, the residual life at t practically relies only on the observations that are both uncensored and still surviving at t; the sample size n = 200, for instance, only yields an effective sample size of about 100 in our simulation set up. Thus it is not a surprise to see this kind of level performance. To demonstrate the local power of the test, in the linear models we kept the same $\beta_1(t)$ with $\beta_2(t)=c/\sqrt{n}$ and $c(1+t)/\sqrt{n}$ for c = 5, 10. For the nonlinear model, we set $\beta_1(t) = \{c\log(1 - \tau)\}^2/n$, $\beta_2(t)=-2c\log(1-\tau)\sqrt{(t+a)/n}$ for c = 40 and a = 0.01. and generated data from $S(t|X_i)=\exp\{-e^{X_i}\sqrt{n}/c(\sqrt{t+a}-\sqrt{a})\}$. The local power results for various sample sizes are given in Table 3. One notices that the power does not necessarily increase as the sample size increases. This is because we are performing a local test where the local alternative is at a root n distance from the null, while the typical convergence rate of $\beta(t)$ is slower than root n. We view the results in Table 3 as a worst case scenario of the power result.

5.2. Application: MELAS study

For illustrative purpose, we applied Model (2.1) to part of the data from the aforementioned MELAS study, consisting of 135 MELAS mutation carriers followed up over the past 10 years (Kaufmann et al. (2009)). We chose the disease onset as the time when the patient fails to perform the daily activities of healthy people. The Karnofsky score is common measurement for functional impairment, ranging from 0 to 100. A healthy subject should be scored at 100. We take T_i to be first year that the ith patient's Karnofsky score is at 90 or lower. About 30% patients are censored since they are still neurologically fully functional at the end of the study. The researchers found that male patients tend to have earlier disease onset. It is of clinical interest to confirm whether MELAS affects male and female patients differently in terms of residual life time. Using gender of a patient as a predictor, we applied a varying coefficient linear median residual life model.

The estimation of the constant coefficient function $\beta_1(t)$ and the time varying gender effect $\beta_2(t)$, along with their upper and lower 5% quantile bootstrap confidence bands are given in Figure 2. Specifically, $\beta_1(t)$ depicts the median residual life time to disease onset of male MELAS patients at various ages. For example, at birth, the median time to disease onset of a male MELAS patient is about 34 years, while at year 16, the median residual time is about 20 years. Here $\beta_2(t)$ describes the difference in median residual time between male and female patients, with confidence bands largely situated above the zero level. Indeed, a formal testing procedure using the method developed in Section 4, yields a p-value $2.35e^{-7}$ that strongly suggests a gender difference. In Particular, the female residual survival is superior to that of the male, with the median residual time of female patients about 15 years longer than that of male patients. Such a difference slightly increases after birth, and decreases again after age 8.

We also estimated the male and female residual life time at the 0.25 and 0.1 quantile levels. As with the median, the female has later disease onset time at each age. For example, given that a patient survived to age 4, a female's residual life is 17 years longer than that of a male, with probability 0.9. This is the age at which the female 10% residual life advantage is the largest. This advantage slowly declines when the patient continues to survive. At age 16, 90% of the surviving females have at least 6 years advantage over males, and 75% of them have at least 11 years advantage. The plots of the 25% and 10% residual life and their corresponding confidence intervals are also in Figure 2. Similar to the median, a test on no gender differences at these two quantile levels yield p-values of $3.94e^{-4}$ and 0.0255 respectively, hence it is quite clear that a female has superior residual survival time at these two quantile levels as well.

One may notice from Figure 2 that, at the 80% confidence level, some confidence bands contain a constant curve. In other words, one might adopt a constant coefficient assumption to model the residual life in those cases. This is a rather typical trade-off between the model complexity and flexibility – whether or not to use a constant coefficient model here depending on how comfortable one feels to make this simplification at a 80% confidence level. One also needs to note that a constant coefficient model may be sufficient at one τ, but may not be so for other τ values. Our methods can help in the decision of whether or not to adopt a constant coefficient assumption.

6. Discussion

We have proposed a time-varying coefficient residual life quantile model. This model allows one to simultaneously model the residual life at different times, yet still ensure the self coherence of the model. Compared to modeling the survival time directly, it allows more flexibility and enables one to describe the residual life directly. We proposed a practically feasible estimation procedure using the spline representation to approximate the time-varying coefficient function, and demonstrated its validity through asymptotic properties. We emphasize here that slightly more complex, yet still feasible, estimation procedures based on quadratic inference functions (Qu, Lindsay, and Li (2000)) can be used to improve the efficiency of the unweighted estimation procedure. We further proposed inference procedures to test the covariate effect. We applied both the estimation and testing procedures in simulations studies as well as to MELAS data.

We have not included the special case where some coefficient might be fixed instead of varying with time. It is easy to see that this can be handled by restricting the spline approximation to the time varying coefficient functions that include the fixed unknown parameter in the set of α's. The developed inference tools can be used to determine whether a certain coefficient indeed varies with time. Specifically, we can reparameterize to a coefficient function $\beta_j(t)=c_0+\beta_j^*(t)$, where $\beta_j^*(0)=0$, and proceed to test $\beta_j^*(t)=0$. As in Jung, Jeong, and Bandos (2009) we have assumed the censoring process to be independent of the covariates *X*, a reasonable assumption for MELAS data. If, however, this assumption is violated, we only need replace the Kaplan-Meier estimator of the censoring survival function with a suitable local Keplan-Meier estimator for $G(t|x)$. For example, we can use

$$\hat{G}(t|x)=\prod_{i=1}^{n}\left[1-\frac{K\{(x_i-x)/h\}}{\sum_{j=1}^{n}I(Y_j\geq Y_i)K\{(x_j-x)/h\}}\right]^{I(Y_i\leq t,\delta_i=0)}$$

to replace $\hat{G}(t)$, where *K* is a kernel function and *h* is a bandwidth, and keep all the remaining procedures unchanged. In the simulation, we used quadratic spline basis functions with a fixed number of knots. If this is not sufficient and more sophisticated spline smoothing techniques, for example the P-spline or the regression spline, are needed, one can use smoothing parameter selection techniques on the pseudo observations $(t_j, \hat{\beta}_j)$, $j = 1, \ldots, J$. Because the $\hat{\beta}_j$'s are estimated at a root-*n* rate, while the spline smoothing rate is slower than that, the consistency of the estimated coefficient functions is preserved without any special treatment of the $\hat{\beta}_j$'s. Finally, instead of splines, other basis functions, such as wavelets or a Fourier basis, can be implemented. A kernel based approach can also be explored, but research in these areas is clearly beyond the scope of this paper.

Acknowledgments

Ma's research was supported by a grant from the National Science Foundation (DMS-0906341) and the National Institute of Neurological Disorders and Stroke (R01-NS073671). Wei's research was supported by the National Science Foundation (DMS-0906568) and a career award from NIEHS Center for Environmental Health in Northern Manhattan (ES009089).

References

Chen YQ. Additive expectancy regression. J. Amer. Statist. Assoc. 2007; 102:153–166.

Chen YQ, Cheng S. Semiparametric regression analysis of mean residual life with censored survival data. Biometrika. 2005; 92:19–29.

Chen YQ, Cheng S. Linear life expectancy regression with censored data. Biometrika. 2006; 93:303–313.

Chen YQ, Jewell NP, Cheng SC. Semiparametric estimation of proportional mean residual life model in presence of censoring. Biometrics. 2005; 61:170–178. [PubMed: 15737090]

Csorgo, S.; Horvath, L. The Rate of Strong Uniform Consistency for the Product-Limit Estimator. Berlin: Springer; 1983.

Fan J, Zhang JT. Two-step estimation of functional linear models with applications to longitudinal data. J. Roy. Statist. Soc. Ser. B. 2000; 62:303–322.

Fleming, TR.; Harrington, DP. Counting Processes and Survival Analysis. New York: Wiley; 1991.

Gupta RC, Langford ES. On the determination of a distribution by its median residual life function: a functional equation. J. Appl. Probab. 1984; 21:120–128.

He X, Shao XM. On parameters of increasing dimensions. J. Multivar. Anal. 2000; 73:120–135.

Jeong JH, Jung SH, Costantino JP. Nonparametric inference on median residual life function. Biometrics. 2008; 64:157–163. [PubMed: 17501936]

Jung SH, Jeong JH, Bandos H. Regression on quantile residual life. Biometrics. 2009; 65:1203–1212. [PubMed: 19432781]

Kaufmann P, Engelstad K, Wei Y, Kulikova R, Oskoui M, Battista V, Koenigsberger D, Pascual JM, Sano M, Hinton V, Hirano M, Millar WS, Shungu DC, Mao X, DiMauro S, De Vivo DC. Protean Phenotypic Features of the A3243G Mitochondrial DNA Mutation. Achieves Neuro. 2009; 66:85–91.

Müller HG, Zhang Y. Time-varying functional regression for predicting remaining lifetime distributions from longitudinal trajectories. Biometrics. 2005; 61:1064–1075. [PubMed: 16401280]

Ma Y, Yin G. Semiparametric median residual life model and inference. Canad. J. Statist. 2010; 38:665–679.

Oakes D, Dasu T. A note on residual life. Biometrika. 1990; 77:409–410.

Oakes, D.; Dasu, T. Inference for the proportional mean residual life model. In: Kolassa, JE.; Oakes, D., editors. Crossing Boundaries: Statistical Essays in Honor of Jack Hall. Vol. 43. Institute of Mathematical Statistics: Hayward, CA; 2003. p. 105-116.Institute of Mathematical Statistics Lecture Notes Monograph Series

Qu A, Lindsay BG, Li B. Improving generalized estimating equations using quadratic inference functions. Biometrika. 2000; 87:823–836.

Schumaker, L. Spline Functions: Basic Theory. New York: Wiley; 1981.

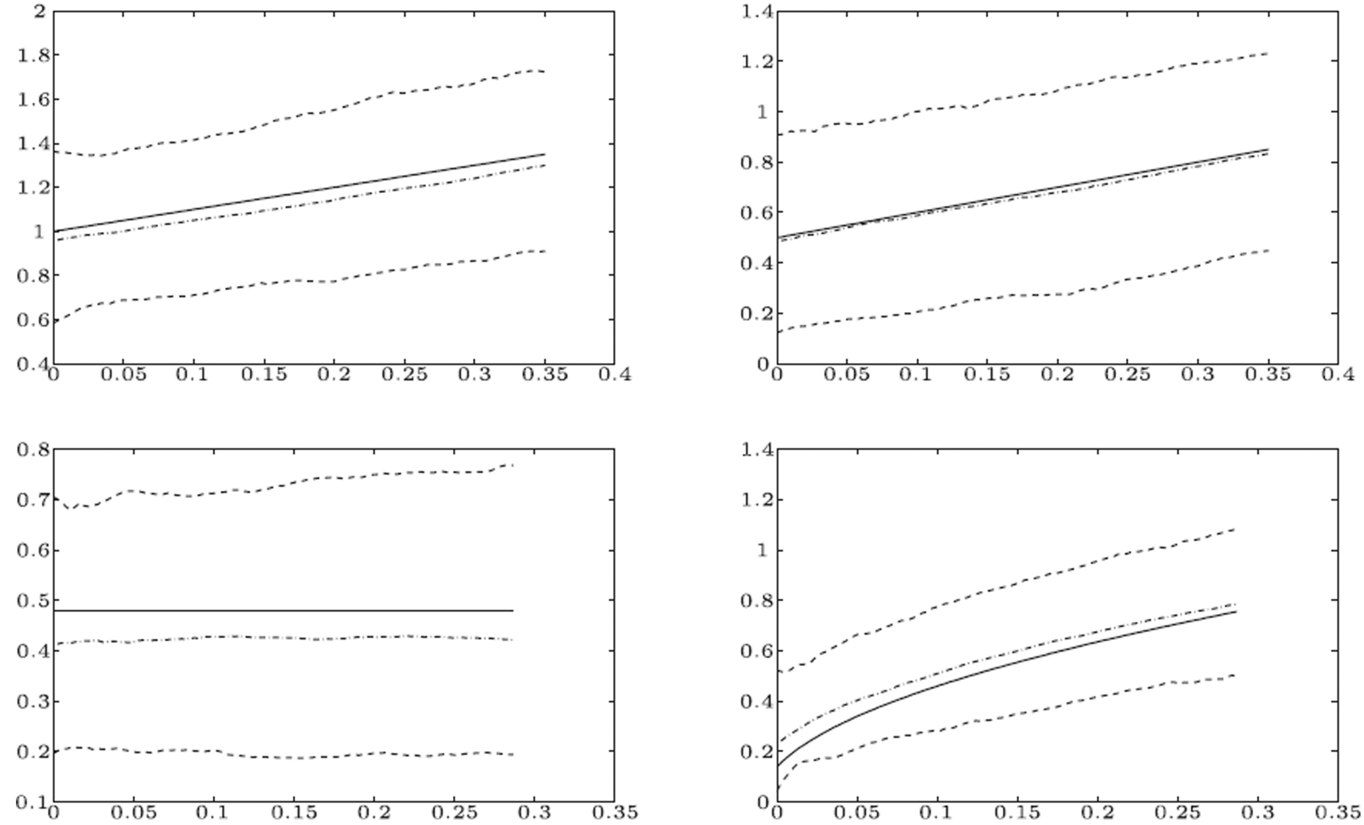

Figure 1.

Curve fitting for $\beta_1(t)$ and $\beta_2(t)$ in different models. Top row: $m(x, \beta) = \beta_1(t) + \beta_2(t)X$, $\beta_1(t) = 1 + t$, $\beta_2(t) = 0.5 + t$. Bottom row: $m(x, \beta) = e^{-2X} \beta_1(t) + e^{-X} \beta_2(t)$, $\beta_1(t) = (log(2))^2$, $\beta_2(t)=2\log 2\sqrt{0.01+t}$. True curve ('−'), median estimated curve ('−.'), and 90% pointwise confidence band ('−'). Sample size $n = 300$.

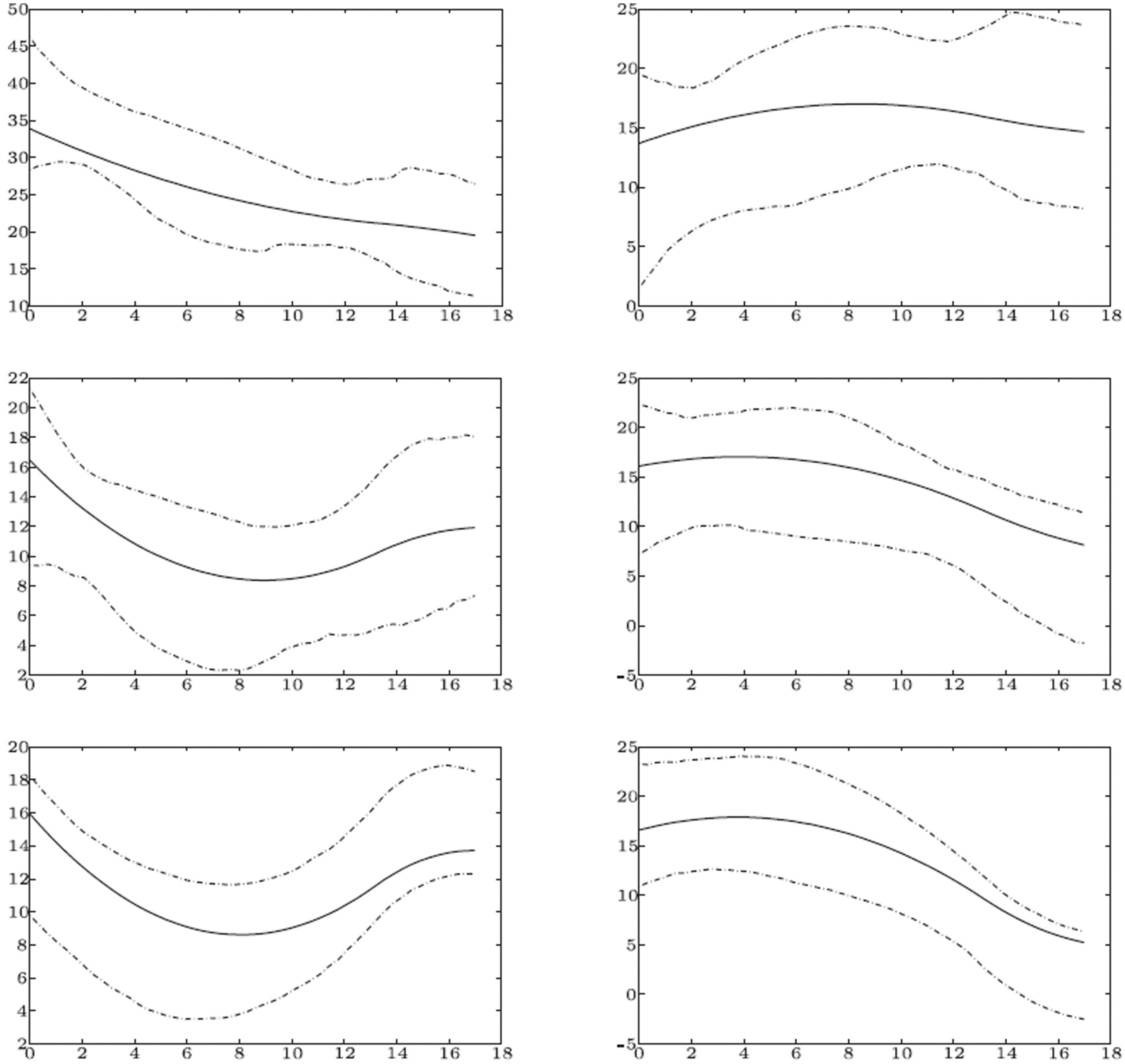

Figure 2.

Estimated time varying intercept function (left) and slope function (right), and their confidence bands (upper and lower 10% quantile) in MELAS study for median (upper), lower quartile (middle) and 10% quantile (lower) residual life time.

Table 1

MSE of un-smoothed and smoothed curve fitting.

	True functions		MSE un-smoothed		MSE smoothed	
n	**$\beta_1(t)$**	**$\beta_2(t)$**	**$\beta_1(t)$**	**$\beta_2(t)$**	**$\beta_1(t)$**	**$\beta_2(t)$**
100	1	0.5	0.0693	0.0553	0.0381	0.0302
	1	$0.5 + t$	0.1211	0.1154	0.0634	0.0498
	$1 + t$	0.5	0.1341	0.1014	0.0612	0.0448
	$1 + t$	$0.5 + t$	0.1700	0.1483	0.0769	0.0600
	$(\log 2)^2$	$2\log 2\sqrt{0.01+t}$	0.0561	0.1288	0.0347	0.0605
200	1	0.5	0.0456	0.0432	0.0296	0.0310
	1	$0.5 + t$	0.0800	0.0924	0.0425	0.0535
	$1 + t$	0.5	0.0903	0.0688	0.0447	0.0374
	$1 + t$	$0.5 + t$	0.1350	0.1029	0.0623	0.0453
	$(\log 2)^2$	$2\log 2\sqrt{0.01+t}$	0.0433	0.0946	0.0307	0.0410
300	1	0.5	0.0340	0.0352	0.0254	0.0290
	1	$0.5 + t$	0.0581	0.0828	0.0354	0.0574
	$1 + t$	0.5	0.0778	0.0633	0.0453	0.0423
	$1 + t$	$0.5 + t$	0.1078	0.0968	0.0526	0.0547
	$(\log 2)^2$	$2\log 2\sqrt{0.01+t}$	0.0376	0.0676	0.0287	0.0276
1,000	1	0.5	0.0150	0.0152	0.0118	0.0130
	1	$0.5 + t$	0.0238	0.0407	0.0185	0.0345
	$1 + t$	0.5	0.0287	0.0287	0.0200	0.0229
	$1 + t$	$0.5 + t$	0.0419	0.0561	0.0295	0.0440
	$(\log 2)^2$	$2\log 2\sqrt{0.01+t}$	0.0167	0.0226	0.0152	0.0139

Table 2

Level precision of the Wald tests for $H_0 : \beta_2(t) = 0$, $H_1 : \beta_2(t) \neq 0$.

	n	**0.01**	**0.05**	**0.1**
			linear model	
$\beta_1(t) = 1$	100	0.1170	0.2350	0.3050
	200	0.0650	0.1610	0.2270
	300	0.0090	0.0500	0.1050
$\beta_1(t) = 1 + t$	100	0.0590	0.1230	0.1790
	200	0.0460	0.1120	0.1580
	300	0.0150	0.0550	0.0940
			nonlinear model	
$\beta_1(t) = \{\log(1 - \tau)\}^2$	100	0.0240	0.0490	0.0800
	200	0.0260	0.0590	0.0960
	300	0.0160	0.0520	0.0900

Table 3

Power of the Wald tests for $H_0 : \beta_2(t) = 0$, $H_1 : \beta_2(t) \neq 0$.

	n	**0.01**	**0.05**	**0.1**
			$\beta_2(t)=5/\sqrt{n}$	
$\beta_1(t) = 1$	100	0.8640	0.9390	0.9650
	200	0.5470	0.7240	0.7980
	300	0.2840	0.5110	0.6270
			$\beta_2(t)=5(1+t)/\sqrt{n}$	
$\beta_1(t) = 1$	100	0.9660	0.9920	0.9970
	200	0.7470	0.8480	0.8920
	300	0.4900	0.7050	0.7850
			$\beta_2(t)=5/\sqrt{n}$	
$\beta_1(t) = 1 + t$	100	0.3510	0.5680	0.6760
	200	0.2530	0.4450	0.5750
	300	0.1140	0.2610	0.3830
			$\beta_2(t)=5(1+t)/\sqrt{n}$	
$\beta_1(t) = 1 + t$	100	0.4000	0.6160	0.7170
	200	0.2970	0.4850	0.6010
	300	0.1290	0.2790	0.3820
			$\beta_2(t)=10/\sqrt{n}$	
$\beta_1(t) = 1 + t$	100	0.7700	0.8590	0.8940
	200	0.7240	0.8530	0.8850
	300	0.5900	0.7410	0.8090
			$\beta_2(t)=10(1+t)/\sqrt{n}$	
$\beta_1(t) = 1 + t$	100	0.8330	0.9070	0.9340
	200	0.7520	0.8560	0.9050
	300	0.6000	0.7480	0.8100
			$\beta_2(t)=-80\log(1-\tau)\sqrt{(t+0.01)/n}$	
$\beta_1(t) = \{40\log(1-\tau)\}^2/n$	100	0.7470	0.8140	0.8340
	200	0.6600	0.8230	0.8820
	300	0.4080	0.5830	0.6930