Statistica Sinica **34** (2024), 111-132
doi:https://doi.org/10.5705/ss.202022.0017

THRESHOLD ESTIMATION IN PROPORTIONAL MEAN RESIDUAL LIFE MODEL

Bing Wang and Xinyuan Song*

Anhui University and Chinese University of Hong Kong

Abstract: The mean residual life model is vital for its ability to investigate the association between covariates and patient life expectancy. In certain circumstances, a patient's lifespan may change when a covariate exceeds a particular threshold value, which is critical to predicting the patient's life expectancy and preventing diseases. This study considers a threshold regression analysis of a proportional mean residual life model with a continuous thresholding variable. We construct martingale-based smoothed estimating equations to obtain parameter estimators, and establish the large-sample properties of the proposed estimators. Furthermore, we propose a supremum test to examine the existence of the threshold. Finally, we assess the finite-sample performance of the proposed method using simulation studies, and then apply the methodology to data from colorectal and breast cancer studies.

Key words and phrases: Proportional mean residual model, smoothed estimation equation, subgroup identification, threshold test.

1. Introduction

The mean residual life (MRL) function measures the remaining life expectancy of a subject who has survived up to a specific time point. As a valuable alternative to the hazard-based approach, the MRL model directly examines how potential covariates affect the MRL function, and is widely applied in biomedical sciences, industrial reliability research, and actuarial studies. For a nonnegative survival time $\tilde{T}$ with finite expectation, the MRL function at time $t \geq 0$ is defined as

$$m(t) = E(\tilde{T} - t|\tilde{T} > t) = S^{-1}(t) \int_t^\infty S(u)du,$$

where $S(t)$ represents the survival function of $\tilde{T}$. The MRL function reveals how long a subject can survive, given his/her current life status. Using a simple calculation, $S(t)$ can be derived from $m(t)$ by the inversion formula as

$$S(t) = \frac{m(0)}{m(t)} \exp\left\{-\int m^{-1}(u)du\right\},$$

*Corresponding author.

and the corresponding hazard function is

$$\lambda(t) = \left\{ \frac{m'(t) + 1}{m(t)} \right\},$$

where $m'(t)$ denotes the first derivative of $m(t)$. Notably, $\lambda(t)$ is always nonnegative. Therefore, $m'(t) + 1 \geq 0$, and $m(t) + t$ is nondecreasing, which is an important property of the MRL function. For other properties of the MRL function, refer to Balkema and de Haan (1974), Hollander and Proschan (1975), Kotz and Shanbhag (1980), and Arnold and Zahedi (1988).

Assessing the effects of covariates on the MRL function is of considerable interest in clinical studies. As a result, numerous works have investigated regression analyses of the MRL function. Oakes and Dasu (1990) and Maguluri and Zhang (1994) proposed proportional MRL (PMRL) models with dichotomous and continuous covariates, respectively, in the absence of censoring. Chen and Cheng (2005) and Chen et al. (2005) developed semiparametric estimation procedures for PMRL models with censoring. Chen and Cheng (2006) and Chen (2007) considered additive MRL models and discussed various estimation procedures, with and without censoring. Sun and Zhang (2009) studied a class of transformed MRL models, and Sun, Song and Zhang (2011) extended the transformation models to incorporate time-dependent covariates. However, the aforementioned studies assume linear covariate effects, thus disregarding the situation in which a covariate effect on the MRL function may change substantially when the covariate exceeds a particular threshold.

The present study fills this gap by considering a threshold PMRL model in the presence of censoring. This kind of threshold regression can be used as a parsimonious strategy for nonparametric function estimation (Guallar and Pastor (1998); Hansen (2000); Fong et al. (2017)), and can be used to identify critical subgroups of a population who may require highly personalized treatment recommendations (Goldberg and Kosorok (2012); Zhao et al. (2014b)). Threshold regression models have been widely applied to substantive studies in economics. Deidda and Fattouh (2002) used a threshold model to specify the nonlinear relationship between financial and economic development. Baum, Checherita-Westphal and Rother (2013) proposed a dynamic threshold panel model to analyze the nonlinear impact of public debt on GDP growth, against the background of the euro area sovereign debt crisis. Interested readers can refer to Hansen (2000), Gonzalo and Wolf (2005), Andrews, Kitagawa and McCloskey (2021), and the references therein. Threshold covariate effects have also received considerable attention in clinical studies, including the fasting plasma glucose effect in the Australian Diabetes Obesity and Lifestyle Study (Tapp et al. (2006)), midthigh muscle cross-sectional area effect in the COPD Study (Marquis et al. (2002)), and leukocyte telomere length effect in the Strong Heart Family Study (Zhao et al. (2014a)). One type of threshold model assumes

a threshold at an unknown time in order to detect the lag effects of covariates (Liang, Self and Liu (1990); Luo (1996); Pons (2002)). Some of these models examines a continuous change in the regression coefficient when a covariate crosses a threshold (Gandy and Jensen (2005); Gandy, Jensen and Lütkebohmert (2005); Jensen and Lütkebohmert (2008)). Another class of models investigates discontinuous changes in covariate effects (Pons (2003); Kosorok and Song (2007); Deng et al. (2017); Wang, Li and Wang (2021)). The present study aims to examine threshold covariate effects in the context of the PMRL model, assuming that the covariate effects change discontinuously, and that the threshold lies in the range of a continuous covariate.

Estimating threshold regression models is complicated, because the models are not smooth in the threshold parameter. Two common approaches for estimating the threshold are the grid-search method (Hansen (2000); Pons (2003); Kosorok and Song (2007)) and the smoothing method (Seo and Linton (2007); He, Lin and Tu (2018)). The grid-search method selects a grid of candidate thresholds on the thresholding covariate. Given a candidate threshold, a threshold model reduces to a regular regression model. The threshold estimate can then be obtained by maximizing the likelihood of the reduced regression model. However, the threshold estimator obtained by the grid-search method has a nonstandard limiting distribution, making statistical inference highly complicated. In addition, the likelihood-based grid-search procedure has difficulty estimating the semiparametric PMRL model. In contrast, the smooth method approximates the step function using a smooth function with a bandwidth. Thus, we propose using martingale-based smoothed estimating equations to estimate the threshold. We prove that the resulting threshold and regression parameter estimators are asymptotically independent and normally distributed. We also show that the convergence rate of the smoothed estimator of the threshold is $h/\sqrt{n}$, where $h \to 0$ is the bandwidth in the smoothing of the indicator function. Furthermore, we propose a supremum test that relies on Wald test statistics to examine the existence of the threshold effect.

The remainder of the paper is organized as follows. Section 2 outlines the threshold estimation of the PMRL model. Section 3 establishes the asymptotic theory for the estimators of the threshold, regression parameters, and the baseline MRL function. Section 4 describes a test procedure for testing the existence of the threshold. In Section 5, we use simulation studies to assess the finite-sample performance of the proposed method. In Section 6, we apply the proposed method to colorectal cancer data from the United States National Cancer Institute Surveillance Epidemiology and End Results (SEER) database, and to breast cancer data from The Cancer Genome Atlas Program (TCGA). Section 7 concludes the paper. All technical proofs are relegated to the Supplementary Material.

2. Method

2.1. PMRL model with structure breaks

Let $\tilde{T}$ be the failure time, X be a continuous covariate with an effect on the response that may have a threshold, and $\boldsymbol{Z}$ denote other p-dimensional covariates. The PMRL model for $\tilde{T}$ given $(\boldsymbol{Z}, X)$ takes the form

$$m(t|\boldsymbol{Z}, X) = m_0(t) \exp\{r^*_{\boldsymbol{\theta}}(\boldsymbol{Z}, X)\}, \tag{2.1}$$

and

$$r^*_{\boldsymbol{\theta}}(\boldsymbol{Z}, X) = \boldsymbol{\beta}^T \boldsymbol{Z} + (\alpha + \boldsymbol{\eta}^T \boldsymbol{Z}) I(X > \zeta) = \boldsymbol{\xi}^T \tilde{\boldsymbol{Z}}^*,$$

where $m_0(t)$ is an unknown baseline MRL function, $\tilde{\boldsymbol{Z}}^* = (\boldsymbol{Z}^T, I(X > \zeta), I(X > \zeta)\boldsymbol{Z}^T)^T$, ζ is an unknown threshold, $\boldsymbol{\xi} = (\boldsymbol{\beta}^T, \alpha, \boldsymbol{\eta}^T)^T$ is a $(2p+1)$-dimensional vector of unknown parameters, $I(U)$ is the indicator of the set U, and $\boldsymbol{\theta} = (\zeta, \boldsymbol{\xi}^T)^T \in \boldsymbol{\Theta} \subset \mathbb{R}^{2p+2}$. We assume that the parameter space $\boldsymbol{\Theta}$ is compact, and that the true parameter $\boldsymbol{\theta}_* = (\zeta_*, \boldsymbol{\beta}_*^T, \alpha_*, \boldsymbol{\eta}_*^T)^T$ is an interior point of $\boldsymbol{\Theta}$. Denote the true value of $m_0(t)$ by $m_*(t)$. Model (2.1) indicates that the effect of $\boldsymbol{Z}$ is $\boldsymbol{\beta}$ when $X \le \zeta$, but changes to $\boldsymbol{\beta} + \boldsymbol{\eta}$ when $X > \zeta$. Additionally, given $\boldsymbol{Z}$, there is a difference of $\exp(\alpha + \boldsymbol{\eta}^T \boldsymbol{Z})$ for the MRL function between $X \le \zeta$ and $X > \zeta$.

2.2. Estimation

Let C be the potential censoring time, and let $T = \min(\tilde{T}, C)$. Conditional on $\boldsymbol{Z}$ and X, $\tilde{T}$ and C are assumed to be independent. To guarantee that the PMRL function is estimable, we assume that the support of C is longer than that of the survival time $\tilde{T}$ and $0 < \tau = \inf\{t : P(\tilde{T} > t) = 0\} < \infty$, which avoids a lengthy technical discussion on the tail behavior of the limiting distributions. Let $\{T_i, \delta_i, \boldsymbol{Z}_i, X_i; i = 1, \ldots, n\}$ be the observed data set, where $\delta_i = I(\tilde{T}_i \le C_i)$. In addition, let $N_i(t) = I(T_i \le t)\delta_i$, $Y_i(t) = I(T_i \ge t)$, and $\Lambda_i(t; \boldsymbol{\theta}_*, m_*)$ be the cumulative hazard function of T_i. The survival function of T given $\boldsymbol{Z}$ and X is

$$S(t|\boldsymbol{Z}, X) = \frac{m(0|\boldsymbol{Z}, X)}{m(t|\boldsymbol{Z}, X)} \exp\left\{ - \int_0^t \frac{1}{m(u|\boldsymbol{Z}, X) du} \right\},$$

and the density function is

$$f(t|\boldsymbol{Z}, X) = S(t\boldsymbol{Z}, X) \left[\frac{m'(t|\boldsymbol{Z}, X) + 1}{m(t|\boldsymbol{Z}, X)} \right],$$

where $m'(t|\boldsymbol{Z}, X)$ is the first derivative of $m(t|\boldsymbol{Z}, X)$ with respect to t. From above, we have $m_0(t) d\Lambda_i(t; \boldsymbol{\theta}_*, m_*) = \exp\{-r^*_{\boldsymbol{\theta}}(\boldsymbol{Z}, X)\} dt + dm_0(t)$. Given that $m_0(t)$ is unknown, it is difficult to estimate the parameter $\boldsymbol{\theta}$ using the likelihood method. Thus, we consider a martingale-based smoothed estimating equation procedure for the threshold estimation.

Let

$$M_i(t;\boldsymbol{\theta}, m_0) = N_i(t) - \int_0^t Y_i(s)d\Lambda_i(s;\boldsymbol{\theta}, m_0), \quad \text{for } i = 1, \ldots, n.$$

Then, $\{M_i(t;\boldsymbol{\theta}_*, m_*)\}$ are zero-mean stochastic processes, which we can use to construct the estimating equations. The threshold ζ is involved in the indicator function, making it difficult to estimate ζ. Hence, we construct smoothed estimating equations

$$\frac{1}{n}\sum_{i=1}^{n}[m_0(t)dN_i(t) - Y_i(t)\{\exp(-r_{\boldsymbol{\theta}}(\boldsymbol{Z}_i, X_i))dt + dm_0(t)\}] = 0,\ 0 \le t \le \tau, \quad (2.2)$$

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \tilde{\boldsymbol{Z}}_i(m_0(t)dN_i(t) - Y_i(t)[\exp\{-r_{\boldsymbol{\theta}}(\boldsymbol{Z}_i, X_i)\}dt + dm_0(t)]) = \mathbf{0}, \quad (2.3)$$

$$\frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau W_i(m_0(t)dN_i(t) - Y_i(t)[\exp\{-r_{\boldsymbol{\theta}}(\boldsymbol{Z}_i, X_i)\}dt + dm_0(t)]) = 0, \quad (2.4)$$

where

$$r_{\boldsymbol{\theta}}(\boldsymbol{Z}_i, X_i) = \boldsymbol{\beta}^T\boldsymbol{Z}_i + (\alpha + \boldsymbol{\eta}^T\boldsymbol{Z}_i)\Phi\left(\frac{X_i - \zeta}{h}\right),$$

in which $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$, $\tilde{\boldsymbol{Z}}_i = (\boldsymbol{Z}_i^T, \Phi\{(X-\zeta)/h\}, \boldsymbol{Z}_i\Phi\{(X-\zeta)/h\})^T$, $W_i = (\alpha+\boldsymbol{\eta}^T\boldsymbol{Z}_i)\phi\{(X-\zeta)/h\}$, and $\phi(\cdot)$ is the density function of $N(0,1)$.

From (2.2), $m_0(t)$ can be estimated as

$$\hat{m}(t;\boldsymbol{\theta}) = \hat{S}(t)\int_t^\tau \hat{S}(u)Q(u;\boldsymbol{\theta})du,$$

where $\hat{S}(t) = \exp\{-\int_0^t \sum_{i=1}^n dN_i(u)/\sum_{i=1}^n Y_i(u)\}$, which is the Nelson–Aalen estimator of the survival function, and $Q(t;\boldsymbol{\theta}) = \sum_{i=1}^n Y_i(t)\exp\{-r_{\boldsymbol{\theta}}(\boldsymbol{Z}_i, X_i)\}/\sum_{i=1}^n Y_i(t)$.

To obtain $\hat{\boldsymbol{\theta}}$, we replace $m_0(t)$ with $\hat{m}(t;\boldsymbol{\theta})$ in Equations (2.3) and (2.4). The resulting equations are

$$U_n^{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{\tilde{\boldsymbol{Z}}_i - \bar{\bar{\boldsymbol{Z}}}(t)\}[\hat{m}(t;\boldsymbol{\theta})dN_i(t) - Y_i(t)\exp\{-r_{\boldsymbol{\theta}}(\boldsymbol{Z}_i, X_i)\}dt] = \mathbf{0},$$

$$U_n^{\zeta}(\zeta) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau \{W_i - \bar{W}(t)\}[\hat{m}(t;\boldsymbol{\theta})dN_i(t) - Y_i(t)\exp\{-r_{\boldsymbol{\theta}}(\boldsymbol{Z}_i, X_i)\}dt] = 0,$$

where $\bar{\bar{Z}}(t) = \sum_{i=1}^n Y_i(t)\tilde{Z}_i/\sum_{i=1}^n Y_i(t)$, and $\bar{W}(t) = \sum_{i=1}^n Y_i(t)W_i/\sum_{i=1}^n Y_i(t)$.

3. Asymptotic Property

This section establishes the consistency and weak convergence of the estimators of the threshold, regression parameters, and the baseline MRL function. First, we define some notation.

We define

$$U_n^m(m(t);\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(m(t)dN_i(t) - Y_i(t)[\exp\{-r_{\boldsymbol{\theta}}(\boldsymbol{Z}_i, X_i)\}dt + dm(t)]) = 0,$$
$$0 \le t \le \tau,$$

and

$$U_n(\boldsymbol{\theta}) = \begin{pmatrix} U_n^{\boldsymbol{\xi}}(\boldsymbol{\xi}) \\ U_n^{\zeta}(\zeta) \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} U_{ni}.$$

Note that

$$\begin{aligned}\frac{\partial U_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} =& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\begin{pmatrix}\tilde{\boldsymbol{Z}}_i - \bar{\tilde{\boldsymbol{Z}}}(t) \\ \{W_i - \bar{W}(t)\}/h\end{pmatrix} \\ &\times\left[\frac{\partial \hat{m}(t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}dN_i(t) - Y_i(t)\begin{pmatrix}\tilde{\boldsymbol{Z}}_i \\ W_i/h\end{pmatrix}^T \times \exp\{-r_{\boldsymbol{\theta}}(\boldsymbol{Z}_i, X_i)\}dt\right] \\ &- \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\begin{bmatrix}0_{p\times p} & 0_{p\times 1} & 0_{p\times p} & 0_{p\times 1} \\ 0 & 0 & 0 & Q_1 \\ 0_{p\times p} & 0_{p\times 1} & 0_{p\times p} & Q_2 \\ 0 & Q_1 & Q_2 & Q_3\end{bmatrix} \\ &\times(\hat{m}(t;\boldsymbol{\theta})dN_i(t) - Y_i[\exp\{-r_{\boldsymbol{\theta}}(\boldsymbol{Z}_i, X_i)\}dt + d\hat{m}(t;\boldsymbol{\theta})]),\end{aligned}$$

where

$$\begin{aligned}Q_1 &= \frac{1}{h}\phi\left(\frac{X_i - \zeta}{h}\right), \\ Q_2 &= \frac{1}{h}\boldsymbol{Z}_i^T\phi\left(\frac{X_i - \zeta}{h}\right), \\ Q_3 &= \frac{1}{h^2}(\alpha + \boldsymbol{\eta}^T\boldsymbol{Z}_i)\phi'\left(\frac{X_i - \zeta}{h}\right),\end{aligned}$$

and $\phi'(\cdot)$ is the derivative function of $\phi(\cdot)$. Denote $\hat{A}(\boldsymbol{\theta}) = D(\partial U_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T)D$, where D is a $(2p+2)$-dimensional diagonal matrix, the first $2p+1$ elements of which are one and the last element is $\sqrt{h}$. For a vector $\boldsymbol{a}$, $|\boldsymbol{a}|$ and $\|\boldsymbol{a}\|$ represent its L_1 and L_2 norms, respectively.

To establish the asymptotic properties of the estimators, we require the following technical conditions.

C1 The true baseline MRL function $m_*(t)$ is continuously differentiable on $[0, \tau]$.

C2 There exists a constant $d_z > 0$, such that $P(\|\boldsymbol{Z}\| > d_z) = 0$.

C3 The limiting matrix of $\hat{A}(\boldsymbol{\theta}_*)$, denoted by $A(\boldsymbol{\theta}_*)$, is nonsingular.

C4 For all X in a neighborhood of ζ, and almost every $\boldsymbol{Z}$, the density function of X conditional on Z, $f_{X|\boldsymbol{Z}}(x|\boldsymbol{z})$, and its derivative function $f'_{X|\boldsymbol{Z}}(x|\boldsymbol{z})$ have positive density everywhere with respect to the Lebesgue measure and are bounded.

C5 As $n \to \infty$, $h \to 0$ and $nh^3 \to 0$.

Condition C1 indicates that $m_*(t)$ is bounded on $[0, \tau]$. Conditions C2 and C3 are necessary for parameter identifiability. Conditions C4 and C5 ensure the weak convergence of the estimator $\hat{\zeta}$. These conditions are common in the threshold detection and survival analysis literature.

Theorem 1. *Under Conditions* C1–C5, $\hat{\boldsymbol{\theta}}$ *uniquely exists and converges consistently to* $\boldsymbol{\theta}_*$ *as* $n \to \infty$*; for every* $t \in [0, \tau]$, $\hat{m}(t; \boldsymbol{\theta})$ *uniquely exists, and* $\hat{m}(t; \boldsymbol{\theta}) \to m_*(t)$ *almost surely uniformly in* $[0, \tau]$ *as* $n \to \infty$.

In the proof of Theorem 1, we first show that $U_n^m(m, \boldsymbol{\theta})$ converges uniformly to U^m in probability, where U^m is defined in the proof of Theorem 1. Next, we verify the identification of m in U^m. The uniform convergence of U_n^m to U^m and the implicit function theorem yield that, for any $\boldsymbol{\theta}$ in the neighborhood of $\boldsymbol{\theta}_*$, $\hat{m}(.; \boldsymbol{\theta})$ converges uniformly to the solution of $U^m(m, \boldsymbol{\theta}) = 0$, $m(.; \boldsymbol{\theta})$, with probability one. Then, by the convergence of U_n and because A is strictly positive definite, $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_*$ in probability.

Theorem 2. *Under Conditions* C1–C5, $\sqrt{n}D^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)$ *is asymptotically normal with mean zero and a covariance matrix that can be consistently estimated by* $\hat{\Sigma}(\hat{\boldsymbol{\theta}}) = \hat{A}^{-1}(\hat{\boldsymbol{\theta}})\hat{B}(\hat{\boldsymbol{\theta}})\hat{A}^{-1}(\hat{\boldsymbol{\theta}})$, *where*

$$\hat{B}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{ \begin{pmatrix} \tilde{\boldsymbol{Z}}_i - \bar{\tilde{\boldsymbol{Z}}}(t) \\ \{W_i - \bar{W}(t)\}/\sqrt{h} \end{pmatrix} - \hat{\boldsymbol{e}} \right\}^{\otimes 2} \hat{m}(t; \hat{\boldsymbol{\theta}})[\exp\{-r_{\hat{\boldsymbol{\theta}}}(\boldsymbol{Z}_i, X_i)\}dt + d\hat{m}(t; \hat{\boldsymbol{\theta}})],$$

and

$$\hat{\boldsymbol{e}} = \begin{pmatrix} \hat{\boldsymbol{e}}_{\tilde{\boldsymbol{Z}}^*} \\ \hat{e}_W \end{pmatrix} = \frac{\hat{S}(t)\int_0^t \hat{S}^{-1}(u)\sum_{j=1}^n \begin{pmatrix} \tilde{\boldsymbol{Z}}_j - \bar{\tilde{\boldsymbol{Z}}}(u) \\ \{W_j - \bar{W}(u)\}/\sqrt{h} \end{pmatrix} dN_j(u)}{\sum_{j=1}^n Y_j(t)}.$$

Theorem 2 shows that the asymptotic distribution of $\hat{\zeta}$ is a normal distribution, thus avoiding a complex statistical inference when using the grid-search method. The convergence rates of $\hat{\zeta}$ and $\hat{\boldsymbol{\xi}}$ are $h/\sqrt{n}$ and $1/\sqrt{n}$,

respectively. Moreover, we obtain the asymptotic covariance of $\hat{\boldsymbol{\theta}} = (\hat{\zeta}, \hat{\boldsymbol{\xi}}^T)^T$. In contrast, in the grid-search method, we obtain only the covariance of $\hat{\boldsymbol{\xi}}$ with fixed ζ, which can cause bias in the covariance estimation, as discussed in Hansen (2000).

Theorem 3. *Under Conditions* C1–C5, $\sqrt{n}\{\hat{m}(t;\hat{\boldsymbol{\theta}}) - m_*(t)\}$ *converges weakly on* $[0,\tau]$ *to a zero-mean Gaussian process, the covariance function of which at* (t,s) *can be consistently estimated by* $\hat{\Gamma}(t,s) = n^{-1}\sum_{i=1}^n \hat{O}_i(t)\hat{O}_i(s)$, *and*

$$\hat{O}_i(t) = \frac{\partial \hat{m}(t;\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \hat{A}(\hat{\boldsymbol{\theta}}) D U_{n_i}(\hat{\boldsymbol{\theta}}) D + \hat{S}(t)^{-1} \int_t^\tau \frac{\hat{S}(u)\hat{m}(u;\hat{\boldsymbol{\theta}}) dM_i(u;\hat{\boldsymbol{\theta}},\hat{m})}{\sum_{j=1}^n Y_j(u)}.$$

4. Threshold Test

Testing the existence of the threshold is essential in practice. In the proposed model, the null hypothesis is $H_0 : \alpha = 0, \boldsymbol{\eta} = \mathbf{0}$. Notably, in the estimating equation (2.4), the threshold is unidentifiable if both α and $\boldsymbol{\eta}$ are zero. We adopt a type of supremum test to tackle this problem. The test statistic relies on Wald statistics, and is defined as follows:

$$\mathrm{SUP}_K = \sup_{\zeta \in \{\zeta_1,\ldots,\zeta_K\}} \{\hat{\alpha}(\zeta), \hat{\boldsymbol{\eta}}^T(\zeta)\}^T \hat{\Sigma}_{\alpha\boldsymbol{\eta}}(\zeta) \{\hat{\alpha}(\zeta), \hat{\boldsymbol{\eta}}^T(\zeta)\},$$

where $\hat{\alpha}(\zeta)$ and $\hat{\boldsymbol{\eta}}^T(\zeta)$ are obtained from the estimating equations (2.2) and (2.3), respectively, with fixed ζ, $\hat{\Sigma}_{\alpha\boldsymbol{\eta}}(\zeta)$ is the element of $\hat{\Sigma}(\boldsymbol{\theta})$ corresponding to α and $\boldsymbol{\eta}$, $\{\zeta_1, \zeta_2, \ldots, \zeta_K\}$ are prespecified values in the range of X, and K is the number of grids. Theoretically, $\{\zeta_1, \zeta_2, \ldots, \zeta_K\}$ can take all distinct observed values of X, while excluding those below the 0.1th or above the 0.9th quantile to avoid edge effects. However, having too many grids increases the computational burden, and may reduce the power of the test, as shown in the simulation study (Table 9). Therefore, we suggest taking equispaced levels between the 0.1th and 0.9th quantiles of X, such as $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$th quantiles. Alternatively, one can take equidistant grids in the range of X, as shown in the simulation study, wherein we consider $K = 1$, 3, or 13 to assess the impact of small, moderate, and relatively large K, respectively, on the test performance.

This supremum-type test statistic does not follow a standard chi-squared distribution, as shown by Davies (1987). Therefore, we adopt a permutation procedure to obtain the critical value of SUP_K under the null hypothesis. Specifically, we shuffle the covariate X enough times to obtain the permutation distribution of SUP_K. Then, we can generate the critical value at a certain significance level. The test procedure is as follows:

Step 1 : Compute the statistic SUP_K for the original data.

Step 2: Generate $X_i^*, i = 1, \ldots, n$ by randomly sampling from $\{X_i, i = 1, \ldots, n\}$ without replacement, and construct a new sample $\{T_i, \Delta_i, \boldsymbol{Z}_i, X_i^*, i = 1 \ldots, n\}$.

Step 3: Generate a total of L (e.g., $L = 500$) simulated trials using **Step 2**. Compute the test statistics SUP_K^ℓ, for $\ell = 1, \ldots, L$.

Step 4: Reject the null hypothesis if SUP_K is larger than the 95% percentile of $\{\mathrm{SUP}_K^\ell, \ell = 1, \ldots, L\}$.

Step 2 is similar to the permutation in the two-sample test. The idea behind it is intuitive; given a test statistic, we compute its distribution under H_0 by permuting the two samples $\{i : X_i \le \zeta\}$ and $\{i : X_i > \zeta\}$. Replacing X_i with X_i^* extracted from $\{X_j; j = 1, \ldots, n\}$ without replacement means the new set $\{i : X_i^* \le \zeta\}$ contains observations that are originally in $\{i : X_i > \zeta\}$. Then, the samples of $\{i : X_i \le \zeta\}$ and $\{i : X_i > \zeta\}$ are mixed, thereby matching the distribution under H_0.

5. Simulation Study

We conduct simulation studies to evaluate the finite-sample performance of the proposed estimation and test procedures. The first simulation evaluates the estimation performance using the bias (Bias), sample standard deviation (SSD), standard error estimate (SEE), and coverage probability (CP) of the 95% confidence interval. We consider covariates $\boldsymbol{Z} = (Z_1, Z_2)$, where Z_1 is a Bernoulli random variable with a success probability of 0.5 and Z_2 is a uniform random variable on $[0, 1]$, and a thresholding variable $X \sim \mathrm{Uniform}(-1, 1)$ with a true threshold at 0 or 0.5. The survival time $\tilde{T}$ is generated according to Model (2.1). The true population values of the parameters are assigned as follows: $\boldsymbol{\beta} = (\beta_1, \beta_2)$ is set to $(0.2, 0.2)$, $(\alpha, \boldsymbol{\eta}) = (\alpha, \eta_1, \eta_2)$ is set to Case 1: $(-0.3, 0.2, 0.2)$ and Case 2: $(-0.5, 0.5, 0.5)$ to assess the effect of the jump size on the parameter estimation, and the baseline MRL function is set to $m_0(t) = 1$ or $m_0(t) = 1/(1+t)$. The censoring time follows $\mathrm{Exp}(c)$, and the censoring rate (CR) is controlled at approximately 15% or 30% by adjusting c. In addition, we consider the sample size $n = 400$ or 800 and the bandwidth $h = sd(X)n^{-1/2}\log(n)$, which meets Condition C5. All results are based on 1,000 replications.

Tables 1–4 summarize the simulation results. We have the following observations. The proposed method provides approximately unbiased estimates and similar SSD and SEE. Increasing the jump size $(\alpha, \boldsymbol{\eta})$ decreases the SSD and SEE of the threshold estimator $\hat{\zeta}$, but has little effect on the other parameter estimators. In contrast, increasing the threshold ζ from 0 to 0.5 reduces the SSD and SSE of $\hat{\boldsymbol{\beta}}$, but increases the SSD and SEE of $(\alpha, \boldsymbol{\eta})$, because the jump

Table 1. Simulation results for the threshold and regression parameters in the simulation study ($\zeta = 0$, $m_0(t) = 1$).

			$n = 400$				$n = 800$			
CR	Case	Para	Bias	SEE	SSD	CP	Bias	SEE	SSD	CP
15%	1	ζ	0.001	0.148	0.139	0.919	−0.010	0.136	0.114	0.917
		β_1	0.002	0.162	0.151	0.958	0.006	0.116	0.112	0.959
		β_2	−0.012	0.282	0.273	0.959	0.002	0.202	0.194	0.953
		α	−0.001	0.085	0.089	0.934	−0.002	0.060	0.060	0.953
		η_1	−0.007	0.232	0.211	0.966	−0.018	0.166	0.153	0.964
		η_2	0.005	0.393	0.384	0.955	0.001	0.281	0.258	0.964
15%	2	ζ	0.007	0.085	0.083	0.938	0.001	0.045	0.044	0.951
		β_1	0.002	0.167	0.156	0.960	0.009	0.117	0.115	0.958
		β_2	−0.002	0.291	0.285	0.952	0.002	0.204	0.202	0.951
		α	0.001	0.089	0.091	0.942	0.002	0.062	0.063	0.949
		η_1	−0.015	0.238	0.216	0.967	−0.031	0.167	0.157	0.961
		η_2	−0.006	0.399	0.396	0.941	−0.007	0.281	0.266	0.959
30%	1	ζ	0.004	0.183	0.159	0.911	0.006	0.127	0.114	0.920
		β_1	−0.001	0.178	0.167	0.957	0.005	0.126	0.120	0.965
		β_2	−0.017	0.309	0.297	0.957	−0.002	0.219	0.213	0.947
		α	−0.001	0.092	0.097	0.934	−0.003	0.064	0.066	0.935
		η_1	−0.005	0.255	0.232	0.958	−0.019	0.178	0.163	0.973
		η_2	0.006	0.431	0.407	0.956	0.001	0.304	0.284	0.960
30%	2	ζ	−0.009	0.091	0.075	0.930	−0.002	0.049	0.046	0.958
		β_1	0.001	0.176	0.172	0.950	0.006	0.125	0.123	0.951
		β_2	−0.015	0.307	0.299	0.953	0.002	0.217	0.217	0.948
		α	0.001	0.092	0.096	0.940	0.001	0.065	0.066	0.948
		η_1	−0.013	0.248	0.238	0.957	−0.018	0.176	0.166	0.968
		η_2	−0.002	0.419	0.412	0.952	−0.008	0.299	0.285	0.961

Note: CR, Para, SSD, SEE, and CP denote the censoring rate, parameter, sample standard deviation, standard error estimate, and coverage probability of the 95% confidence interval, respectively.

size estimator is related only to the sample with $X > \zeta$. However, increasing n from 400 to 800 or decreasing CR from 30% to 15% decreases the SSD and SEE of the estimators. Moreover, the estimators have smaller SSD and SEE when $m_0(t) = (1+t)^{-1}$ than when $m_0(t) = 1$. Finally, under different settings, the CP remains stable and close to the nominal level of 95%.

Moreover, we investigate the effect of a varying bandwidth h on the estimation. We fix $m_0(t) = 1$, $\zeta = 0$, $(\boldsymbol{\beta}, \alpha, \boldsymbol{\eta}) = (0.5, 0.5, 0.5, 0.5, 0.5)$, and $n = 400$, and set h as $\{0.01, 0.05, 0.1, 0.15, 0.164, 0.2, 0.3\}$, where 0.164 is obtained from the proposed value of $h = sd(X)n^{-1/2}\log(n)$. Tables 5 and 6 present the Bias, SSD, and root mean squared error (RMSE) for the parameter estimators under CR = 15% and CR = 30%, respectively. The estimates of $\hat{\boldsymbol{\beta}}$, α, and $\hat{\boldsymbol{\eta}}$ are not sensitive to h, but an extremely small h can cause slight instability when estimating ζ. Furthermore, an h near 0.164 is preferred for $\hat{\zeta}$. Therefore, our choice of $h = sd(X)n^{-1/2}\log(n)$ is suitable.

Table 2. Simulation results for the threshold and regression parameters in the simulation study ($\zeta = 0.5$, $m_0(t) = 1$).

			$n = 400$				$n = 800$			
CR	Case	Para	Bias	SEE	SSD	CP	Bias	SEE	SSD	CP
15%	1	ζ	−0.015	0.181	0.128	0.906	−0.014	0.135	0.094	0.932
		β_1	−0.001	0.130	0.124	0.964	0.002	0.092	0.091	0.951
		β_2	−0.008	0.225	0.221	0.954	−0.001	0.159	0.155	0.959
		α	−0.002	0.125	0.123	0.937	−0.002	0.085	0.087	0.949
		η_1	0.001	0.273	0.246	0.956	−0.018	0.189	0.180	0.957
		η_2	−0.005	0.436	0.446	0.944	0.015	0.308	0.289	0.953
15%	2	ζ	−0.012	0.099	0.079	0.923	0.005	0.044	0.066	0.954
		β_1	−0.001	0.134	0.129	0.966	0.004	0.096	0.094	0.948
		β_2	−0.005	0.233	0.236	0.950	−0.001	0.167	0.163	0.955
		α	0.005	0.124	0.129	0.936	0.005	0.088	0.092	0.941
		η_1	−0.015	0.273	0.252	0.953	−0.033	0.195	0.186	0.951
		η_2	−0.013	0.440	0.459	0.937	0.004	0.317	0.299	0.955
30%	1	ζ	−0.013	0.202	0.162	0.903	−0.020	0.129	0.112	0.925
		β_1	−0.003	0.143	0.137	0.953	0.006	0.101	0.098	0.957
		β_2	−0.014	0.248	0.241	0.957	−0.002	0.175	0.171	0.958
		α	−0.001	0.131	0.133	0.936	−0.004	0.091	0.093	0.941
		η_1	0.004	0.294	0.267	0.956	−0.014	0.204	0.192	0.960
		η_2	−0.001	0.474	0.472	0.943	0.001	0.337	0.320	0.957
30%	2	ζ	−0.014	0.089	0.089	0.934	−0.006	0.048	0.047	0.954
		β_1	−0.002	0.143	0.140	0.960	0.001	0.101	0.100	0.952
		β_2	−0.009	0.249	0.242	0.953	0.001	0.176	0.174	0.959
		α	0.007	0.129	0.135	0.931	0.004	0.092	0.095	0.946
		η_1	−0.009	0.289	0.268	0.940	−0.031	0.204	0.193	0.960
		η_2	−0.017	0.468	0.475	0.940	0.003	0.334	0.321	0.954

Note: CR, Para, SSD, SEE, and CP denote the censoring rate, parameter, sample standard deviation, standard error estimate, and coverage probability of the 95% confidence interval, respectively.

The second simulation assesses the performance of the proposed test statistic SUP_K. We choose $K \in \{1, 3, 13\}$ to examine the effect of the number of grids, and set the true threshold to 0 or 0.5 and the grids for SUP_1, SUP_3, and SUP_{13} to $\{0\}$, $\{-0.3, 0, 0.3\}$, and $\{-0.6, -0.5, -0.4, \ldots, 0.5, 0.6\}$, respectively, to evaluate the effect of the distance between the threshold and the grids on the performance of the test procedure. Thus, SUP_1 is the optimal test if the true threshold is the same as the preassigned threshold zero. We compare SUP_1, SUP_3, and SUP_{13} in terms of their type-I error and power, with a significance level of 5% when $n = 400$ and CR = 30%.

Table 7 summarizes the results obtained based on 500 replications. The left panel shows that the type-I errors are close to 0.05 when the true model has no threshold, and the power increases with the jump size. In addition, the power may be affected by the threshold location, because it decreases when the

threshold location is close to the boundary. Furthermore, the number of grids and the distance between the threshold and the grids synergistically affect the performance of the test. When $\zeta = 0.5$, SUP_{13} has the highest power, because it has the shortest distance between the threshold and the grids. When $\zeta = 0$, the distances between the threshold and the grids are the same for all the tests. In this case, the optimal test is SUP_1, with the highest power, and SUP_{13} has the lowest power. Finally, all the tests have higher power when $m_0(t) = (1+t)^{-1}$ than when $m_0(t) = 1$.

The third simulation checks the performance of the proposed estimation and test procedures in the case of heavy censoring. We mimic the setting of the second real data set by considering the covariates $\boldsymbol{Z} = (Z_1, Z_2)$, where Z_1 and Z_2 are as in Simulation 1, and the thresholding variable $X \sim \text{Uniform}(-1, 1)$ with the true threshold at zero. The survival time $\tilde{T}$ is generated from the following model:

$$m(t|\boldsymbol{Z}, X) = m_0(t) \exp\{\beta Z_2 + (\alpha + \eta Z_1) I(X > \zeta)\}.$$

We set $\beta = 0.3$, $\alpha = 0.3$, $\eta = -0.3$, and $m_0(t) = 1$. The censoring time follows Exp(3), and CR is approximately 80%. In addition, we consider the sample size $n = 900$, and implement the estimation procedure similarly as before. Table 8 shows the results summarized based on 1,000 replications. The estimation performance is not as good as that obtained in the case of light censoring, but is still acceptable.

We also examine the performance of the test statistic with heavy censoring, and compare SUP_{13} and SUP_7. We recommend using SUP_7 with grids at $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$th quantiles of X for substantive studies; see Section 6. Table 9 reports the results obtained based on 500 replications. Here, the proposed test statistic still performs acceptably in the case of heavy censoring. Moreover, SUP_7 has a significantly lower computational burden and a slightly higher power than SUP_{13}.

6. Real-data Analysis

6.1. Colorectal cancer data

We first apply the proposed procedure to colorectal cancer data collected from SEER. Colorectal cancer is a disease in which malignant cells form in the tissues of the colon or rectum, and is the third leading cause of cancer in both men and women in the United States. Established risk factors of colorectal cancer do not include the sex variable. However, the report on colorectal cancer shows differences in deaths between men and women for each race. Therefore, we investigate sex as a potential risk factor for colorectal cancer.

We extract the 2010–2014 San Francisco colorectal cancer data from SEER. There are 5,410 patients, and about 77.4% of the observations are subject to right

Table 3. Simulation results for the threshold and regression parameters in the simulation study ($\zeta = 0$, $m_0(t) = (1+t)^{-1}$).

			$n = 400$				$n = 800$			
CR	Case	Para	Bias	SEE	SSD	CP	Bias	SEE	SSD	CP
15%	1	ζ	−0.004	0.096	0.083	0.946	−0.001	0.050	0.051	0.953
		β_1	0.002	0.087	0.082	0.962	0.004	0.061	0.060	0.960
		β_2	−0.004	0.151	0.146	0.957	0.002	0.107	0.104	0.950
		α	−0.004	0.050	0.049	0.950	0.003	0.036	0.037	0.945
		η_1	−0.002	0.132	0.122	0.967	−0.013	0.094	0.088	0.960
		η_2	0.007	0.223	0.219	0.958	−0.001	0.157	0.149	0.962
15%	2	ζ	0.002	0.043	0.041	0.945	−0.003	0.027	0.027	0.958
		β_1	0.006	0.088	0.084	0.959	0.007	0.062	0.061	0.960
		β_2	−0.002	0.123	0.122	0.949	0.005	0.108	0.106	0.948
		α	0.006	0.055	0.054	0.956	0.008	0.039	0.041	0.930
		η_1	−0.016	0.138	0.130	0.967	−0.023	0.098	0.093	0.957
		η_2	−0.009	0.229	0.234	0.945	−0.011	0.162	0.157	0.954
30%	1	ζ	−0.005	0.118	0.084	0.941	−0.003	0.052	0.050	0.954
		β_1	0.002	0.092	0.087	0.965	0.004	0.065	0.064	0.954
		β_2	−0.005	0.161	0.157	0.955	0.001	0.114	0.111	0.946
		α	0.001	0.054	0.052	0.960	0.001	0.038	0.039	0.948
		η_1	−0.006	0.140	0.131	0.966	−0.013	0.099	0.093	0.966
		η_2	−0.001	0.238	0.234	0.951	0.001	0.169	0.159	0.962
30%	2	ζ	−0.001	0.058	0.042	0.956	0.002	0.029	0.028	0.954
		β_1	0.005	0.093	0.088	0.962	0.007	0.066	0.065	0.948
		β_2	−0.002	0.161	0.157	0.951	0.004	0.114	0.111	0.950
		α	0.005	0.057	0.055	0.960	0.006	0.040	0.042	0.940
		η_1	−0.015	0.145	0.136	0.965	−0.022	0.103	0.097	0.958
		η_2	−0.009	0.241	0.244	0.948	−0.009	0.171	0.164	0.952

Note: CR, Para, SSD, SEE, and CP denote the censoring rate, parameter, sample standard deviation, standard error estimate, and coverage probability of the 95% confidence interval, respectively.

censoring. Figure 1(a) displays the Kaplan–Meier (KM) curves for males and females, and their 95% confident bands. Based on the log-rank test, the difference between the two gender groups is not statistically significant. Therefore, there is no evidence that sex is a vital risk factor for the survival rate. However, there might be a subgroup that exhibits a gender difference.

Published medical reports of colorectal cancer show that the death rate varies among age groups. Hence, we examine whether a specific age subgroup exists in which sex is a significant factor. For example, suppose we set the cut-point as the median age of 72. Then, as shown in Figure 1(b), sex becomes a significant risk factor for individuals older than 72. This motivates us to apply the proposed method to detect an objective threshold from the age distribution. The covariates we consider include sex, tumor size, and their interaction with the dichotomized age at diagnosis with an unknown threshold to be identified. We code females as one and males as zero, and standardize the tumor size.

Table 4. Simulation results for the threshold and regression parameters in the simulation study ($\zeta = 0.5$, $m_0(t) = (1+t)^{-1}$).

			$n = 400$				$n = 800$			
CR	Case	Para	Bias	SEE	SSD	CP	Bias	SEE	SSD	CP
15%	1	ζ	−0.001	0.080	0.079	0.950	−0.004	0.050	0.053	0.961
		β_1	0.001	0.070	0.067	0.959	0.002	0.050	0.049	0.952
		β_2	−0.002	0.121	0.118	0.950	0.000	0.086	0.084	0.953
		α	−0.004	0.074	0.072	0.950	0.005	0.051	0.053	0.946
		η_1	−0.002	0.159	0.146	0.957	−0.023	0.112	0.106	0.957
		η_2	0.007	0.254	0.264	0.943	0.007	0.189	0.176	0.951
15%	2	ζ	−0.003	0.040	0.040	0.945	0.001	0.027	0.027	0.956
		β_1	0.003	0.071	0.068	0.963	0.004	0.050	0.049	0.953
		β_2	−0.002	0.123	0.122	0.949	0.002	0.087	0.085	0.961
		α	0.012	0.079	0.082	0.939	0.015	0.055	0.061	0.926
		η_1	−0.017	0.169	0.158	0.957	−0.028	0.119	0.116	0.946
		η_2	−0.021	0.265	0.290	0.929	−0.007	0.188	0.189	0.947
30%	1	ζ	0.011	0.094	0.082	0.931	−0.007	0.062	0.062	0.954
		β_1	0.000	0.074	0.070	0.959	0.001	0.053	0.052	0.954
		β_2	−0.006	0.129	0.125	0.953	0.000	0.091	0.089	0.957
		α	0.004	0.078	0.076	0.953	0.003	0.054	0.055	0.946
		η_1	−0.002	0.169	0.157	0.959	−0.014	0.119	0.113	0.958
		η_2	−0.003	0.273	0.280	0.943	0.007	0.194	0.190	0.944
30%	2	ζ	−0.009	0.091	0.075	0.930	−0.001	0.029	0.028	0.957
		β_1	0.002	0.074	0.071	0.960	0.004	0.053	0.052	0.951
		β_2	−0.003	0.129	0.126	0.951	0.003	0.092	0.090	0.954
		α	0.013	0.082	0.084	0.947	0.012	0.057	0.062	0.924
		η_1	−0.014	0.176	0.166	0.966	−0.028	0.124	0.120	0.951
		η_2	−0.018	0.280	0.301	0.939	−0.005	0.200	0.200	0.947

Note: CR, Para, SSD, SEE, and CP denote the censoring rate, parameter, sample standard deviation, standard error estimate, and coverage probability of the 95% confidence interval, respectively.

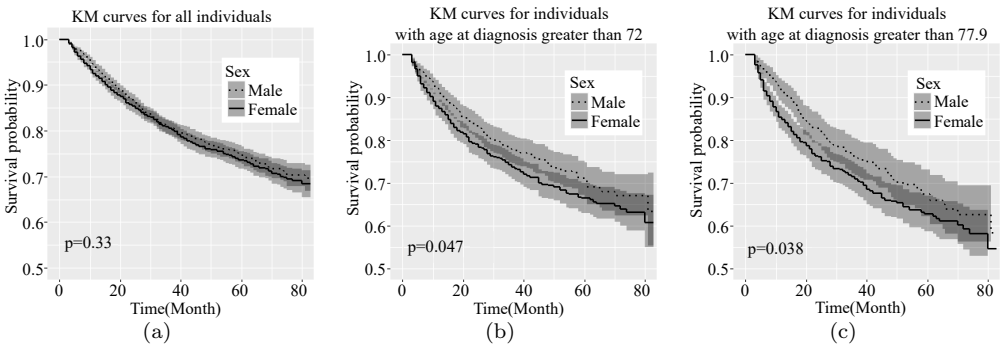

Figure 1. The KM curves for colorectral cancer and their 95% confidence bands. The p-value is calculated using the log-rank test.

Table 5. Simulation results with varying h and CR = 15%.

	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\alpha}$		
	Bias	SSD	RMSE	Bias	SSD	RMSE	Bias	SSD	RMSE
$h = 0.01$	0.006	0.147	0.147	−0.016	0.289	0.289	0.000	0.085	0.085
$h = 0.05$	0.006	0.147	0.147	0.017	0.290	0.290	0.000	0.084	0.084
$h = 0.1$	0.006	0.147	0.147	0.017	0.290	0.290	−0.002	0.086	0.086
$h = 0.15$	0.006	0.147	0.147	0.018	0.290	0.290	−0.004	0.084	0.084
$h = 0.164$	0.006	0.147	0.147	0.018	0.290	0.290	−0.004	0.084	0.084
$h = 0.2$	0.006	0.147	0.147	0.018	0.290	0.290	−0.004	0.084	0.084
$h = 0.3$	0.006	0.147	0.147	0.018	0.290	0.290	−0.005	0.081	0.081
	$\hat{\eta}_1$			$\hat{\eta}_2$			$\hat{\zeta}$		
	Bias	SSD	RMSE	Bias	SSD	RMSE	Bias	SSD	RMSE
$h = 0.01$	−0.037	0.219	0.222	−0.037	0.401	0.403	0.006	0.179	0.178
$h = 0.05$	−0.033	0.219	0.222	−0.038	0.403	0.405	0.005	0.086	0.086
$h = 0.1$	−0.033	0.219	0.222	−0.038	0.402	0.404	−0.006	0.067	0.067
$h = 0.15$	−0.033	0.219	0.222	−0.038	0.402	0.404	−0.009	0.066	0.066
$h = 0.164$	−0.033	0.219	0.222	−0.038	0.402	0.404	−0.009	0.068	0.068
$h = 0.2$	−0.033	0.219	0.222	−0.038	0.402	0.404	−0.009	0.073	0.073
$h = 0.3$	−0.033	0.219	0.222	−0.038	0.402	0.404	−0.010	0.083	0.084

Note: SSD and RMSE denote the sample standard deviation and root mean squared error, respectively.

Table 6. Simulation results with varying h and CR = 30%.

	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\alpha}$		
	Bias	SSD	RMSE	Bias	SSD	RMSE	Bias	SSD	RMSE
$h = 0.01$	−0.005	0.157	0.157	0.000	0.343	0.343	0.004	0.100	0.100
$h = 0.05$	−0.006	0.157	0.157	0.001	0.342	0.342	0.001	0.098	0.098
$h = 0.1$	−0.004	0.159	0.159	−0.002	0.344	0.344	−0.006	0.097	0.097
$h = 0.15$	−0.005	0.159	0.159	0.002	0.343	0.343	−0.007	0.096	0.096
$h = 0.164$	−0.004	0.159	0.159	0.002	0.343	0.343	−0.007	0.095	0.095
$h = 0.2$	−0.005	0.159	0.159	0.002	0.343	0.343	−0.008	0.094	0.094
$h = 0.3$	−0.004	0.159	0.159	0.003	0.344	0.344	−0.010	0.088	0.089
	$\hat{\eta}_1$			$\hat{\eta}_2$			$\hat{\zeta}$		
	Bias	SSD	RMSE	Bias	SSD	RMSE	Bias	SSD	RMSE
$h = 0.01$	−0.031	0.235	0.237	−0.018	0.435	0.435	0.016	0.154	0.154
$h = 0.05$	−0.030	0.234	0.265	−0.021	0.435	0.435	0.015	0.122	0.123
$h = 0.1$	−0.031	0.235	0.237	−0.021	0.436	0.437	−0.017	0.075	0.077
$h = 0.15$	−0.031	0.235	0.237	−0.021	0.435	0.436	−0.015	0.068	0.070
$h = 0.164$	−0.031	0.235	0.237	−0.021	0.435	0.436	−0.015	0.072	0.074
$h = 0.2$	−0.031	0.235	0.237	−0.021	0.435	0.436	−0.016	0.076	0.078
$h = 0.3$	−0.032	0.236	0.238	−0.023	0.433	0.434	−0.022	0.103	0.105

Note: SSD and RMSE denote the sample standard deviation and root mean squared error, respectively.

Table 7. Size and power with or without change point.

		$m_0(t)=1$			$m_0(t)=(1+t)^{-1}$		
ζ	(α,η_1,η_2)	SUP_1	SUP_3	SUP_{13}	SUP_1	SUP_3	SUP_{13}
0	$(0,0,0)$	5.2%	5.8%	4.4%	5.4%	5.0%	4.8%
	$(-0.1,0.1,0.1)$	14.6%	14.4%	13.4%	43.0%	48.2%	40.6%
	$(-0.2,0.2,0.2)$	59.0%	55.4%	49.2%	96.6%	97.4%	98.0%
	$(-0.3,0.3,0.3)$	91.2%	90.2%	86.8%	100%	100%	100%
	$(-0.4,0.4,0.4)$	99.4%	99%	98.2%	100%	100%	100%
	$(-0.5,0.5,0.5)$	100%	100%	100%	100%	100%	100%
0.5	$(0,0,0)$	4.4%	5.8%	4.2%	4.6%	5.0%	4.8%
	$(-0.1,0.1,0.1)$	8.2%	10.6%	12.4%	17.8%	26.4%	34.4%
	$(-0.2,0.2,0.2)$	18.8%	30.4%	38.6%	60.8%	80.2%	93.4%
	$(-0.3,0.3,0.3)$	42.2%	62.4%	71.8%	92%	98.8%	100%
	$(-0.4,0.4,0.4)$	65.0%	82.2%	93.4%	98.8%	100%	100%
	$(-0.5,0.5,0.5)$	80.8%	94.6%	94.8%	100%	100%	100%

Table 8. Simulation results for the case of heavy censoring

Para	Bias	SEE	SSD	CP
ζ	-0.007	0.296	0.218	90.6%
β	-0.085	0.185	0.195	93.7%
α	-0.007	0.178	0.195	92.2%
η	0.014	0.196	0.226	93.7%

Note: Para, SEE, SSD, and CP stand for parameter, standard error estimate, sample standard deviation, and the coverage probability of the 95% confidence interval, respectively.

Table 9. Size and power with and without change points.

(α,η)	SUP_7	SUP_{13}
$(0,0)$	4.4%	4.2%
$(-0.1,0.1)$	13.4%	12.4%
$(-0.2,0.2)$	41.2%	40.0%
$(-0.3,0.3)$	70.6%	70.0%
$(-0.4,0.4)$	77.6%	75.4%
$(-0.5,0.5)$	85.8%	84.6%

We use the proposed test procedure to determine the existence of a threshold. The threshold search set is $\{52,57,61,64,69,73,78\}$, corresponding to the $\{0.2,0.3,0.4,0.5,0.6,0.7,0.8\}$th quantiles of age. We obtain a p-value much lower than 0.05, indicating significant evidence to reject the null hypothesis. Therefore, we conclude that a threshold does exist.

Next, we apply the estimation method described in Section 2 to perform the parameter estimation. The estimated cut-point location is $\hat{\zeta}=77.9$, which

Table 10. Analysis results for the colorectal cancer data ($\hat{\zeta} = 77.9$).

Parameter	Estimate	SSE	p-value
Sex	−0.038	0.026	0.430
Tumor size	−0.064	0.008	<0.001
Age (> 77.9)	−0.102	0.033	0.002
Sex×Age (> 77.9)	−0.113	0.036	0.002
Tumorsize×Age (> 77.9)	0.003	0.160	0.852
Threshold	77.904	1.411	<0.001

is close to the result of Wang, Li and Wang (2021), who analyzed this data set using a proportional hazards mixture cure model with a single threshold. Table 10 presents the parameter estimates. Sex is not significant when patients are below 77.9. However, for those older than 77.9, females have a significantly shorter MRL than males. This finding is in line with the results shown in Figure 1(c), which shows that the survival probability is higher for males than for females in the cohort of age at diagnosis greater than 77.9. Moreover, tumor size has a significant adverse effect on the MRL function, but its effect becomes negligible when age crosses the threshold.

6.2. Breast cancer data

Breast cancer is cancer that develops from breast tissue. Based on the 2016–2018 data from the National Cancer Institute, around 12.9% of women are diagnosed with breast cancer at some point during their lifetime. Therefore, clinicians are interested in improving prognostic prediction. The established risk factors include obesity, old age, and lack of physical exercise. Moreover, as suggested by (Borcherding et al. (2018)), protein-level data have particular advantages in assessing putative prognostic or therapeutic targets in tumors. We apply the proposed procedure to the breast cancer data extracted from the TCGA. We consider age at diagnosis and the proteins BLC2A1 and CDK1 obtained from the TCPA as covariates (Li et al. (2013)). After deleting samples with missing data, the sample size is 874, and the censoring rate is 86%.

We first preprocess the data before analysis. Note that because the data are encrypted, we can obtain only level 3 or 4 protein data from the TCPA. However, the order of the numerical values in the data remains unchanged, even though the encryption conceals the data. By dichotomizing CKD1 into a binary variable, we can disregard the unknown data transformations. Thus, we consider a model with only CDK1 as a thresholding variable and obtain the estimated threshold −0.07, which determines whether CDK1 is highly expressed. Then, we convert CDK1 to a binary variable, using the value one for high expression (> -0.07) and zero for low expression (≤ -0.07). As shown by Piao et al. (2019), the expression of CDK1 is important for the prognosis of breast cancer. However, based on the

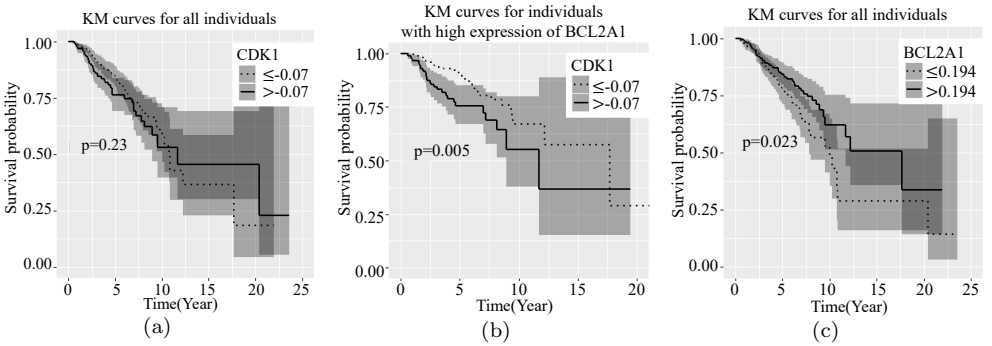

Figure 2. The KM curves for breast cancer and their 95% confidence bands. The p-value is calculated using the log-rank test.

information provided in the TCPA database, CDK1 in the univariant Cox model exerts a nonsignificant effect (p-value $= 0.64$). Even though we convert CDK1 to a binary variable, its effect on the survival probability is still not apparent, as shown in Figure 2(a). Therefore, we set the expression of BCL2A1 as the thresholding covariate to identify a subgroup in which CDK1 may have a significant effect on the MRL function. The covariates we consider include standardized age at diagnosis, the converted CDK1, and the interaction between the converted CDK1 and the dichotomized BCL2A1, with an unknown threshold to be identified.

Next, we apply the proposed test procedure to determine the existence of a threshold. The threshold search set is $\{0.024, 0.080, 0.142, 0.210, 0.286, 0.382, 0.511\}$, corresponding to the $\{0.2, 0.3, 0.4, 0.5, 0.6.0.7, 0.8\}$th quantiles of the expression level of BCL2A1. We obtain a p-value much lower than 0.05, providing strong evidence of the existence of a threshold.

Finally, we use the suggested estimation procedure to obtain the parameter estimates. The estimated cut-point location is 0.194. Table 11 presents the parameter estimates. CDK1 is nonsignificant when the expression of BCL2A1 is below 0.194, but becomes significantly negative when it exceeds 0.194. Thus, for patients with BCL2A1 greater than 0.194, a high CDK1 expression is a poor prognosis for breast cancer. This finding is consistent with the result shown in Figure 2(b), which indicates that the survival probability is higher when CDK1 ≤ -0.07 than when CDK1 > -0.07 in the cohort of high BCL2A1 expression. Furthermore, the binary BCL2A1 has a significant positive effect on the MRL function, aligning with the finding shown in Figure 2(c).

The above subgroup analyses demonstrate the utility of the proposed method and reveal new insights into potential risk factors for cancer and other diseases. The computer code is written in R and is available at `https://github.com/caterpillar-star/TEPMRL`.

Table 11. Analysis results for the colorectal cancer data ($\hat{\zeta} = 0.194$).

Parameter	Estimate	SSE	p-value
Age	−0.202	0.036	<0.001
CDK1 (> -0.07)	0.003	0.113	0.976
BCL2A1 (> 0.194)	0.291	0.073	<0.001
CDK1 (> -0.07)×BCL2A1 (> 0.194)	−0.311	0.131	0.018
Threshold	0.194	0.068	<0.001

7. Conclusion

Threshold models are common in many fields, and identifying a meaningful threshold usually leads to discovering essential subgroups in the population. This study considers a semiparametric PMRL model, and develops a smoothed estimating equation approach to estimate the threshold, regression coefficients, and baseline MRL function. We also develop a test procedure to examine the existence of the threshold. The proposed method is guaranteed theoretically using large-sample theory and is supported empirically by means of simulation studies and two real-life applications. Notably, if there is an interaction between $\boldsymbol{Z}$ and continuous X, the problem can be regarded as sample splitting based on a continuous variable (Hansen (2000)). Deciding on an appropriate cut-off at which to split the sample is often of great interest and practical value. It enables informative comparisons between two subgroups and facilitates group-specific recommendations. Alternatively, one may consider using a varying coefficient model to capture fine and smooth details of local effect changes. However, estimating infinite "parameters" for an unknown function typically requires a large sample size and more sophisticated methods. Thus, threshold models are a parsimonious strategy for nonparametric function estimation or can be used as a preliminary step for investigating complicated data structures.

Although we consider only the PMRL model, our approach can be extended to additive and transformation MRL models without much difficulty. Moreover, we focus only on a single threshold in the present study. Algorithms such as the binary segmentation method in the Gaussian framework and ℓ_1 penalization methods have been developed to reduce a multiple threshold problems to several single threshold problems. Therefore, we can adapt these methods under our framework. Nevertheless, when dividing a finite sample into many subgroups, each subgroup may contain a limited number of observations. Consequently, we may introduce high-level heterogeneity into the estimated results. Such multigroup results may also be over-trained, and thus difficult to generalize to external samples. Therefore, a single-threshold or two-subgroup analysis is still valuable in many scientific applications. In addition, it would be interesting to consider multiple comparisons for detecting multiple thresholds. Finally, existing studies (e.g., Lee and Lam (2020)) combine the detection and estimation

of the threshold. Two necessary conditions must be satisfied to achieve this purpose. First, the test must be based on the likelihood function. Second, $\{\zeta_1, \zeta_2, \ldots, \zeta_K\}$ must include all distinct observed values of X. Then, if the threshold effect is detected, we can estimate the threshold as ζ_ℓ that maximizes the test statistic. However, such a combination is difficult in the proposed model framework, because the likelihood function-based method cannot be applied to the current semiparametric PMRL model. The feasibility of such an extension requires further investigation.

Supplementary Material

The supplementary materials contain proofs of the theoretical results.

Acknowledgments

Xinyuan Song's research was supported by GRF grants 14302519 and 14302220 from the Research Grant Council of the Hong Kong Special Administrative Region. Bing Wang's research was supported by NSFC grant 12301327.

References

Andrews, I., Kitagawa, T. and McCloskey, A. (2021). Inference after estimation of breaks. *Journal of Econometrics* **224**, 39–59.

Arnold, B. C. and Zahedi, H. (1988). On multivariate mean remaining life functions. *Journal of Multivariate Analysis* **25**, 1–9.

Balkema, A. A. and de Haan, L. (1974). Residual Life Time at Great Age. *The Annals of Probability* **2**, 792–804.

Baum, A., Checherita-Westphal, C. and Rother, P. (2013). Debt and growth: New evidence for the euro area. *Journal of International Money and Finance* **32**, 809–821.

Borcherding, N., Bormann, N. L., Voigt, A. P. and Zhang, W. (2018). Trgated: A web tool for survival analysis using protein data in the Cancer Genome Atlas. *F1000Res* **7**, 1235.

Chen, Y. Q. (2007). Additive expectancy regression. *Journal of the American Statistical Association* **102**, 153–166.

Chen, Y. Q. and Cheng, S. (2005). Semiparametric regression analysis of mean residual life with censored survival data. *Biometrika* **92**, 19–29.

Chen, Y. Q. and Cheng, S. (2006). Linear life expectancy regression with censored data. *Biometrika* **93**, 303–313.

Chen, Y. Q., Jewell, N. P., Lei, X. and Cheng, S. C. (2005). Semiparametric estimation of proportional mean residual life model in presence of censoring. *Biometrics* **61**, 170–178.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternatives. *Biometrika* **74**, 33–43.

Deidda, L. and Fattouh, B. (2002). Non-linearity between finance and growth. *Economics Letters* **74**, 339–345.

Deng, Y., Zeng, D., Zhao, J. and Cai, J. (2017). Proportional hazards model with a change point for clustered event data. *Biometrics* **73**, 835–845.

Fong, Y., Di, C., Huang, Y. and Gilbert, P. B. (2017). Model-robust inference for continuous threshold regression models. *Biometrics* **73**, 452–462.

Gandy, A. and Jensen, U. (2005). On goodness-of-fit tests for Aalen's additive risk model. *Scandinavian Journal of Statistics* **32**, 425–445.

Gandy, A., Jensen, U. and Lütkebohmert, C. (2005). A cox model with a change-point applied to an actuarial problem. *Brazilian Journal of Probability and Statistics* **19**, 93–109.

Goldberg, Y. and Kosorok, M. R. (2012). Q-learning with censored data. *Annals of statistics* **40**, 529–560.

Gonzalo, J. and Wolf, M. (2005). Subsampling inference in threshold autoregressive models. *Journal of Econometrics* **127**, 201–224.

Guallar, E. and Pastor, R. (1998). Use of two-segmented logistic regression to estimate change-points in epidemiologic studies. *American Journal of Epidemiology* **148**, 631–642.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica* **68**, 575–603.

He, Y., Lin, H. and Tu, D. (2018). A single-index threshold cox proportional hazard model for identifying a treatment-sensitive subset based on multiple biomarkers. *Statistics in Medicine* **37**, 3267–3279.

Hollander, M. and Proschan, F. (1975). Tests for the mean residual life. *Biometrika* **62**, 585–593.

Jensen, U. and Lütkebohmert, C. (2008). A cox-type regression model with change-points in the covariates. *Lifetime Data Analysis* **14**, 267–285.

Kosorok, M. R. and Song, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *The Annals of Statistics* **35**, 957–989.

Kotz, S. and Shanbhag, D. N. (1980). Some new approaches to probability distributions. *Advances in Applied Probability* **12**, 903–921.

Lee, C. Y. and Lam, K. (2020). Survival analysis with change-points in covariate effects. *Statistical Methods in Medical Research* **29**, 3235–3248.

Li, J., Lu, Y., Akbani, R., Ju, Z., Roebuck, P. L., Liu, W. et al. (2013). TCPA: A resource for cancer functional proteomics data. *Nature Methods* **10**, 1046–1047.

Liang, K. Y., Self, S. G. and Liu, X. (1990). The cox proportional hazards model with change point: An epidemiologic application. *Biometrics* **46**, 783–793.

Luo, X. (1996). The asymptotic distribution of MLE of treatment lag threshold. *Journal of Statistical Planning and Inference* **53**, 33–61.

Maguluri, G. and Zhang, C.-H. (1994). Estimation in the mean residual life regression model. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 477–489.

Marquis, K., Debigaré, R., Lacasse, Y., LeBlanc, P., Jobin, J., Carrier, G. et al. (2002). Midthigh muscle cross-sectional area is a better predictor of mortality than body mass index in patients with chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine* **166**, 809–813.

Oakes, D. and Dasu, T. (1990). A note on residual life. *Biometrika* **77**, 409–410.

Piao, J., Zhu, L., Sun, J., Li, N., Dong, B., Yang, Y. et al. (2019). High expression of CDK1 and BUB1 predicts poor prognosis of pancreatic ductal adenocarcinoma. *Gene* **701**, 15–22.

Pons, O. (2002). Estimation in a cox regression model with a change-point at an unknown time. *Statistics* **36**, 101–124.

Pons, O. (2003). Estimation in a cox regression model with a change-point according to a threshold in a covariate. *The Annals of Statistics* **31**, 442–463.

Seo, M. H. and Linton, O. (2007). A smoothed least squares estimator for threshold regression models. *Journal of Econometrics* **141**, 704–735.

Sun, L., Song, X. and Zhang, Z. (2011). Mean residual life models with time-dependent coefficients under right censoring. *Biometrika* **99**, 185–197.

Sun, L. and Zhang, Z. (2009). A class of transformed mean residual life models with censored survival data. *Journal of the American Statistical Association* **104**, 803–815.

Tapp, R., Zimmet, P., Harper, C., de Courten, M., McCarty, D., Balkau, B. et al. (2006). Diagnostic thresholds for diabetes: The association of retinopathy and albuminuria with glycaemia. *Diabetes Research and Clinical Practice* **73**, 315–321.

Wang, B., Li, J. and Wang, X. (2021). Change point detection in cox proportional hazards mixture cure model. *Statistical Methods in Medical Research* **30**, 440–457.

Zhao, J., Zhu, Y., Lin, J., Matsuguchi, T., Blackburn, E., Zhang, Y. et al. (2014a). Short leukocyte telomere length predicts risk of diabetes in American Indians: The strong heart family study. *Diabetes* **63**, 354–362.

Zhao, L., Feng, D., Bellile, E. L. and Taylor, J. M. G. (2014b). Bayesian random threshold estimation in a cox proportional hazards cure model. *Statistics in Medicine* **33**, 650–661.

Bing Wang

School of Big Data and Statistics, Anhui University, Hefei, Anhui 230601, China.

E-mail: 23051@ahu.edu.cn

Xinyuan Song

Department of Statistics, Chinese University of Hong Kong, Shatin, NT, Hong Kong, China.
Shenzhen Research Institute, Chinese University of Hong Kong, Shenzhen, Guangdong, China.

E-mail: xysong@cuhk.edu.hk

(Received January 2022; accepted April 2022)