Cox regression model with time-varying coefficients in nested case–control studies

MENGLING LIU*

Division of Biostatistics, Department of Environmental Medicine, New York University School of Medicine, New York, NY 10016, USA
mengling.liu@nyu.edu

WENBIN LU

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA

ROY E. SHORE

Radiation Effects Research Foundation, Hiroshima 732-0815, Japan

ANNE ZELENIUCH-JACQUOTTE

Division of Epidemiology, Department of Environmental Medicine, New York University School of Medicine, New York, NY 10016, USA

Summary

The nested case–control (NCC) design is a cost-effective sampling method to study the relationship between a disease and its risk factors in epidemiologic studies. NCC data are commonly analyzed using Thomas’ partial likelihood approach under Cox’s proportional hazards model with constant covariate effects. Here, we are interested in studying the potential time-varying effects of covariates in NCC studies and propose an estimation approach based on a kernel-weighted Thomas’ partial likelihood. We establish asymptotic properties of the proposed estimator, propose a numerical approach to construct simultaneous confidence bands for time-varying coefficients, and develop a hypothesis testing procedure to detect time-varying coefficients. The proposed inference procedure is evaluated in simulations and applied to an NCC study of breast cancer in the New York University Women’s Health Study.

Keywords: Kernel estimation; Martingale; Nested case–control study; Proportional hazards model; Risk-set sampling; Time-varying coefficient.

1. Introduction

Epidemiologic cohort studies of rare diseases are usually expensive to conduct because a large number of individuals need to be followed up for a long time in order to obtain an adequate number of cases. Moreover, the cost of assembling exposure variables of interest and confounders for the entire cohort can be financially prohibitive. Therefore, the nested case–control (NCC) design (Thomas, 1977) has been

*To whom correspondence should be addressed.
widely used as a cost-effective alternative to the full-cohort design. For example, an NCC study was conducted in the New York University Women’s Health Study (NYUWHS) to investigate the association between breast cancer risk and genetic variations in the nucleotide excision repair (NER) pathway (Shore and others, 2008). The NYUWHS is a prospective cohort study that enrolled 14,274 healthy women aged 35–65 between 1985 and 1991 at a breast cancer screening center and has followed these women since the enrollment for cancers and other health outcomes. The NER mechanism is important for cells to prevent unwanted mutations by removing DNA damage, and thus genes in the NER pathway are hypothesized to play a role in the development of cancers and other diseases involving DNA damage and genetic mutations. It would have been very costly to ascertain genetic information for the entire cohort. Thus, the NCC design was implemented, and for each of the 612 identified invasive breast cancer cases, one control was selected from the case’s risk set. Genotypic information on 2 genes, \( XPC \) and \( ERCC2 \) in the NER pathway, was obtained for the cases and their matched controls. Covariate information on demographics, smoking status, and pregnancy history was collected from baseline and follow-up questionnaires. Effects of NER genes, environmental exposure of smoking that causes DNA damage, and their interactions on breast cancer risk were of primary interest.

Cox’s proportional hazards model (Cox, 1972) has been commonly used to analyze NCC data with the successful implementation of the partial likelihood technique (Thomas, 1977; Oakes, 1981). Under the assumption of proportional hazards, that is, the ratio of hazard functions with different covariate values remains constant over time, the expression of Thomas’ partial likelihood function is equivalent to the conditional logistic likelihood for matched case–control studies. Theoretical properties of Thomas’ maximum partial likelihood estimator have been formally established using counting process and martingale theory (Goldstein and Langholz, 1992; Borgan and others, 1995).

Due to the nature of long-term observation and complex relationships to be explored in large epidemiologic studies, the proportional hazards assumption may be violated. In fact, researchers have extended Cox’s model to improve the modeling flexibility and one popular extension is to allow the coefficients to vary with time. Specifically, the Cox model with time-varying coefficients (Zucker and Karr, 1990; Hastie and Tibshirani, 1993) assumes that the individual hazard function has a multiplicative form

\[
\lambda[t|Z(t)] = \lambda_0(t) \exp\{a_0^T(t)Z(t)\},
\]

where \( \lambda_0(t) \) denotes an unspecified baseline hazard function, \( Z(t) \) denotes \( p \) covariate processes, and \( a_0(t) \) are \( p \) processes characterizing the covariates’ temporal effects. The estimation and inference for model (1.1) in cohort studies have been studied by many researchers using various techniques: for example, the penalized partial likelihood approach with smoothing splines (Tibshirani and Hastie, 1987; Zucker and Karr, 1990); the sieve maximum partial likelihood approach with histogram sieves (Murphy and Sen, 1991); the integrated Newton–Raphson equation for the cumulative coefficient functions \( \int_0^t a_0(u)du \) (Martinussen and others, 2002); the kernel-weighted partial likelihood approach (Cai and Sun, 2003; Tian and others, 2005). However, the use of model (1.1) in NCC studies remains limited.

In this article, we are interested in studying model (1.1) with NCC data. We show that the kernel-weighted local polynomial fitting technique (Fan and Gijbels, 1996) can be well coupled with Thomas’ partial likelihood to study the time-varying coefficients in NCC studies. The rest of the article is organized as follows. In Section 2, we propose a kernel-weighted partial likelihood estimation approach and establish the asymptotic properties for the proposed estimator. Pertinent to making inference for time-varying coefficients, we develop numerical approaches to constructing simultaneous confidence bands and to testing hypotheses of existing time-invariant coefficients. Furthermore, we consider inference for an extension to incorporate both time-varying and time-invariant coefficients. In Section 3, we present numerical studies including simulations to evaluate the finite-sample performance of our proposed approaches and the analysis of the NYUWHS breast cancer data. We conclude with some remarks in Section 4. All proofs are relegated to the supplementary material available at Biostatistics online.
2. Methods

Throughout the paper, let \( \{T^*, C, Z(\cdot)\} \) denote a random triplet of failure time, right-censoring time, and \( p \)-dimensional covariate processes. Consider a cohort of size \( n \) and refer the full-cohort data to \( n \) independent realizations of \( \{T_i, \delta_i, Z_i(\cdot), i = 1, \ldots, n\} \), where \( T_i = \min(T_i^*, C_i) \) is the observed failure time and \( \delta_i \) indicates the status of the observed event, taking value of 1 for observing a real failure event and 0 otherwise. At a specific time \( t \), let \( R(t) = \{i: T_i \geq t\} \) denote the risk set. When the NCC design is used to sample from the cohort, one identifies cases as subjects with \( \delta_i = 1 \) and, for each case, randomly samples \( (m-1) \) controls without replacement from the risk set at the case’s failure time excluding the case itself. For a given case \( i \), let \( R_i^* \) denote the indices of the \( (m-1) \) selected controls and define \( R_i = R_i^* \cup \{i\} \). The complete covariate information is ascertained for all the cases and selected controls.

2.1 Kernel-weighted partial likelihood estimation approach

Consider model (1.1) and assume time-varying coefficients \( a_0(\cdot) \) to be smooth functions with continuous first and second derivatives. Locally around a time point \( t \), we approximate \( a_0(\cdot) \) by a linearization using the first-order Taylor expansion,

\[
a_0(u) \approx a_0(t) + a_0'(t)(u - t).
\]

Let \( \beta_0(t) = (a_{01}(t), \ldots, a_{0p}(t), a_{01}'(t), \ldots, a_{0p}'(t))^T \) and \( \tilde{Z}_i(u, u - t) = (1, u - t)^T \otimes Z_i(u) \) with \( \otimes \) denoting the Kronecker product. Moreover, define a counting process representation for the observed samples and 0 otherwise. At a specific time \( t \), for a given case \( i \), let \( \tilde{R}_i \) denote the indices of the \((m-1)\) selected controls and define \( R_i = \tilde{R}_i \cup \{i\} \). The complete covariate information is ascertained for all the cases and selected controls.

\[
ln(\beta, t) = n^{-1} \sum_{i=1}^{n} \int_{0}^{\tau_0} K_h(u - t) \left[ \beta^T \tilde{Z}_i(u, u - t) - \log \left( \sum_{j \in R_i} e^{\beta^T \tilde{Z}_j(u, u - t)} \right) \right] dN_i(u), \tag{2.1}
\]

where \( K(\cdot) \) is a kernel function, \( h \) denotes a bandwidth parameter, \( K_h(\cdot) = h^{-1}K(\cdot/h) \) is the scaled kernel function, and \( \tau_0 \) is the upper bound of follow-up period and satisfies \( \tau_0 = \inf \{ t : \Pr(T > t) = 0 \} \). We assume that \( K(\cdot) \) is a symmetric density function with bounded support on \([-1, 1]\) and \( h = h_n = O(n^{-v}) \) for \( 0 < v < 1 \). This kernel function down-weights the contributions from subjects whose event times are far from \( t \) and the bandwidth parameter controls the size of local neighborhood. Thus, the local partial likelihood function (2.1) depends only on the case–control sets with case event times in the close vicinity of \( t \).

The score function of (2.1) can be easily derived:

\[
U_n(\beta, t) = n^{-1} \sum_{i=1}^{n} \int_{0}^{\tau_0} K_h(u - t) \left[ \tilde{Z}_i(u, u - t) - E_{R_i}\tilde{Z}_i(\beta, u, u - t) \right] dN_i(u), \tag{2.2}
\]

where for a set \( R \), \( s_{R_i}^{(k)}(\beta, u, u - t) = \sum_{j \in R} e^{\beta^T \tilde{Z}_j(u, u - t)} \tilde{Z}_j^{\otimes k}(u, u - t) \) and \( E_{R_i}\tilde{Z}_i(\beta, u, u - t) = s_{R_i}^{(1)}(\beta, u, u - t) \). For a vector \( a \), let \( a^{\otimes k} = 1, a \) and \( aa^T \) for \( k = 0, 1, \) and 2, respectively. Furthermore,
the Hessian matrix of (2.1) equals

\[ l''_n(\beta, t) = -n^{-1} \sum_{i=1}^{n} \int_0^T \frac{K_h(u-t)}{S_{R_i}(u-t)} \left[ \sum_{j \in R_i} e^{\beta^T Z_j(u-u-t)} (\tilde{Z}_j(u, u-t) \right. \\
\left. - E_{R_i}(\beta, u-u-t) \right) \right] dN_i(u), \] (2.3)

It is evident that \( l''_n(\beta, t) \) is semi-negative and thus the concavity of (2.1) assures a unique maximum. The Newton–Raphson method or other gradient-based search algorithms can be used to find \( \hat{\beta} \) that maximizes (2.1). We denote the kernel-weighted maximum partial likelihood estimate of \( a_0(t) \) by \( \hat{a}(t) \), which are the first \( p \) components of \( \hat{\beta} \).

2.2 Asymptotic properties of \( \hat{a}(t) \)

Because of the risk-set sampling mechanism of NCC design, even when the sample size increases, the size of each sampled risk set \( R_i \) for case \( i \) is always \( m \), rather than increasing to infinity as the size of risk sets in Cox’s partial likelihood function for the full-cohort data. Thus, we adopt similar theoretical arguments for Thomas’ maximum partial likelihood estimator used in Goldstein and Langholz (1992) and consider its kernel-weighted local version by using the kernel polynomial fitting technique (Fan and Gijbels, 1996).

Let \( r = \{1, \ldots, m\}, Y_r(t) = \prod_{i \in r} Y_i(t), \) where \( Y_i(t) = I(T_i \geq t) \) and \( p_0(t) = \Pr(Y_i(t) = 1) \). Denote

\[ \Sigma_0(t) = E \left[ m^{-1} \sum_{j \in r} e^{\hat{a}_0(t) Z_j(t)} \left( Z_j(t) - \frac{\sum_{j \in r} e^{\hat{a}_0(t)} Z_j(t) Z_j(t)}{\sum_{j \in r} e^{\hat{a}_0(t)} Z_j(t)} \right)^2 \bigg| Y_r(t) = 1 \right]. \]

Indeed, \( p_0(t) \lambda_0(t) \Sigma_0(t) \) is the local contribution to the asymptotic information matrix of Thomas’ partial likelihood function (Goldstein and Langholz, 1992). Here, we state the main asymptotic results. Regularity conditions A.1–A.5, remarks on the technical device used in Goldstein and Langholz (1992) to make the NCC sampling process predictable, and proofs are given in the supplementary material available at Biostatistics online. Let \( \mu_j = \int s^j K(s) ds \) and \( v_j = \int s^j K^2(s) ds \) for \( j = 0, 1, 2 \).

PROPOSITION 2.1 As \( n \to \infty \) (i) under Conditions A.1–A.4, \( \sup_{t \in (0, \tau_0)} |\hat{a}(t) - a(t)| \to_p 0 \) as \( n \to \infty \); (ii) under Conditions A.1–A.5, for \( t \in (0, \tau_0) \),

\[ (nh)^{1/2} \left\{ \hat{a}(t) - a_0(t) - \frac{h^2}{2} \mu_2 a''_0(t) \right\} \to_D N \left( 0, \frac{v_0}{p_0(t) \lambda_0(t)} \Sigma_0^{-1}(t) \right). \]

2.3 Point-wise confidence interval and simultaneous confidence band

From Proposition 1, it is evident that the optimal bandwidth that minimizes the mean squared error or mean integrated squared error is \( h = O(n^{-1/5}) \). This theoretical optimal bandwidth, however, will lead to an asymptotically biased estimator, that is, the bias term is \( O(1) \). In this paper, we prefer to use a slightly faster rate for bandwidth \( h = n^{-\nu} \), with \( 1/5 < \nu < 1 \), to obtain an unbiased estimator. The main reason is to avoid complications due to the estimation bias in constructing point-wise confidence intervals and simultaneous confidence bands. We refer the reader to Härdle and Marron (1991) for more discussion of handling bias in constructing simultaneous confidence bands for general nonparametric curve estimation.
The variance of \( \hat{a}(t) \) can be consistently estimated by the upper left \( p \times p \) matrix of

\[
\tilde{\Gamma}(t) = (nh)^{-1}v_0[-l''_0(\hat{\beta}; t)]^{-1},
\]

(2.4)

where \( l''_0(\beta, t) \) is as specified in (2.3). Thus, the 100\((1 - \alpha)\)% point-wise confidence interval for the \( j \)th element of \( a_0(t) \) at time \( t \) can be constructed as \([\hat{a}_j(t) \pm z_{1-\alpha/2}\tilde{\Gamma}_{jj}(t)]\), where \( \tilde{\Gamma}_{jj}(t) \) denotes the \( j \)th diagonal element of \( \tilde{\Gamma}(t) \) and \( z_{1-\alpha/2} \) is the 100\((1 - \alpha/2)\)th percentile of the standard normal distribution. However, to make inference for the underlying coefficient function over a specific time interval, it is desirable and more informative to consider the simultaneous confidence band than the point-wise confidence interval. In general, it is difficult to derive an analytic form for simultaneous confidence bands and they are usually estimated using numerical approaches that mimic the original data structure while assessing variability (Härdle and Marron, 1991; Tian and others, 2005). Here, we construct simultaneous confidence bands by approximating the distribution of

\[
S_j = \sup_{[t_1, t_2]} w_n(t) |\hat{a}_j(t) - a_0(t)|, \quad j = 1, \ldots, p,
\]

(2.5)

with the resampling method of Lin and others (1994), where the weight function \( w_n(t) \) can be a data-related positive function that uniformly converges to a deterministic function. In our numerical studies, we choose \( w_n(t) = \tilde{\Gamma}_{jj}^{1/2}(t) \) to take the variability into account. In the proof of Proposition 1, we show that, if \( h = n^{-\alpha} \) and \( 1/5 < \alpha < 1 \), the kernel-weighted score function (2.2) at the true coefficient values is asymptotically equivalent to

\[
U_n^*(\beta_0, t) = n^{-1} \sum_{i=1}^{n} \int_0^{t_0} K_h(u - t)[\tilde{Z}_i(u, u - t) - E_{R_i,z}(\beta_0, u, u - t)]dM_i(u),
\]

where \( dM_i(u) = dN_i(u) - Y_i(u) \exp(a_0^T(u)Z_i(u))\lambda_0(u) \) and consequently, under standard measurability assumptions \( M_i(u) \) is a local martingale. Substituting \( \beta_0 \) by \( \hat{\beta} \) and \( M_i(t) \) by \( N_i(t)G_i \), where \( G_i \) are independent standard normal random variables, we obtain a randomly perturbed version of \( U_n^*(\beta_0, t) \), denoted by \( \tilde{U}_n^*(\hat{\beta}, t) \). At each specific \( t \), the conditional limiting distribution of \( \tilde{U}_n^*(\hat{\beta}, t) \) given the observed data is the same as the unconditional limiting distribution of \( U_n^*(\beta_0, t) \) (Lin and others, 1994). Therefore, the distribution of \( S_j \) (2.5) can be numerically approximated by its randomly perturbed counterpart \( \tilde{S}_j = \sup_{[t_1, t_2]} w_n(t)|\{l''_0(\hat{\beta}, t)^{-1}\tilde{U}_n(\hat{\beta}, t)\}| \). Let \( c_{1-\alpha} \) denote the sample 100\((1 - \alpha)\)th percentile of a large number of realizations of \( \tilde{S}_j \)’s, and thus the simultaneous confidence band for \( a_0(t) \) over \([t_1, t_2]\) can be constructed as \([\hat{a}_j(t) \pm c_{1-\alpha}w_n^{-1}(t)]\).

### 2.4 Inference for mixed Cox model with time-varying and time-invariant coefficients

Model (2.1) is flexible by allowing all coefficients to be time-varying. But when there are indications of possible time-invariant coefficients for certain covariates, one may want to consider a mixed model with time-varying and time-invariant coefficients. Without loss of generality, we consider a mixed model with the first \( q \) components of \( a_0(t) \) being constants, that is, \( a_0^T(t) = \{\gamma_0^T, a_{0-q}^T(t)\} \), where \( a_{q} \) denotes the first \( q \) elements of a vector \( a \) and \( a_{[-q]} \) denotes the remaining \( p - q \) elements. Based on the proposed kernel-weighted partial likelihood estimates \( \hat{a}(t) \), we estimate the constant coefficients \( \gamma \) by

\[
\hat{\gamma} = \int_0^{t_0} w_\gamma(t)\hat{a}_{[q]}(t)dt,
\]

(2.6)

where \( w_\gamma(t) \) is a weight function converging to a deterministic function and also satisfies \( \int_0^{t_0} w_\gamma(t)dt = I_{q \times q} \) as an identity matrix. As suggested in Tian and others (2005), a natural choice for this weight
function is the standardized inverse covariance matrix of $\hat{a}(t)$, that is, $w_\gamma(t) = \{\int_0^\tau J(t)dt\}^{-1}J(t)$, where $J(t)$ is the standardized upper left $q \times q$ submatrix of $\hat{\Gamma}(t)$ defined in (2.4). To make inference for $r_0$, by arguments similar to Tian and others (2005), we can show that $n^{1/2}(\hat{\gamma} - r_0)$ converges weakly to a mean-zero normal distribution for $h = n^{-v}$ with $1/4 < v < 1/2$, and the limiting covariance matrix can be consistently estimated by $\hat{\Gamma}_\gamma = \{\int_0^\tau J(t)dt\}^{-1}$.

In practice, after we obtain the estimates $\hat{a}(t)$ and the simultaneous confidence bands, we can examine each plot for whether the confidence band encloses a horizontal line to check if a constant assumption for this coefficient is possible. The drawbacks for this procedure include the slow convergence rate of the local estimates and its low sensitivity, and it also does not take into account the uncertainty of the constant coefficient estimation. As shown in the paragraph above, we can consider the cumulative function of the time-varying coefficient estimates to achieve a better convergence rate (Martinussen and others, 2002). To check whether the $j$th component of $a(t)$ is independent of time, we consider the process $T_j(t) = n^{1/2}\int_0^\tau \{\hat{a}_j(u) - \hat{\gamma}_j\}du$, where $\hat{a}(t)$ and $\hat{\gamma}$ are the local and integrated estimates, respectively. Based on a similar resampling method described in Section 2.3, we can obtain randomly perturbed realizations of $\tilde{T}_j(t)$ to evaluate this hypothesis graphically and numerically.

More specifically, define $\tilde{\beta}(t)$ as $\beta(t)$ with the $j$th and the $(p + j)$th elements replaced by $(\hat{\gamma}_j, 0)$. When the hypothesis $H_{0j}: a_j(t) = \gamma$ is true, $T_j(t)$ converges weakly to a mean-zero Gaussian process (Tian and others, 2005). Its asymptotic distribution can be approximated by the conditional distribution of the randomly perturbed process

$$
\tilde{T}_j(t) = n^{-1/2} \sum_{i=1}^n \int_0^{r_0} \left( \int_0^t [\{-l''_u(\tilde{\beta}(s), s)\}]^{-1} K_h(u-s)\{\tilde{Z}_i(u, u - s) - E_{R_i,z}(\tilde{\beta}(s), u, u - s)\}ds \right.
- t[-l''_u(\tilde{\beta}(u), u)]^{-1}\{\tilde{Z}_i(u, 0) - E_{R_i,z}(\tilde{\beta}(u), u, 0)\})_j \right) dN_i(u)G_i.
$$

A large number of resampled realizations of $\tilde{T}_j(t)$ can be plotted with $T_j(t)$ to provide a visual examination for the constant hypothesis. Furthermore, a statistical test can be constructed based on $T_j^* = \max_{[0, r_0]}|T_j(t)|$, where the critical value can be obtained using the sample quantile of the resampled counterparts $\tilde{T}_j^* = \max_{[0, r_0]}|\tilde{T}_j(t)|$.

### 2.5 Bandwidth selection

In practice, we suggest to use the cross-validation method (Hastie and others, 2008) to select the bandwidth parameter $h$. Specifically, we first randomly split the case–control sets into $K$ subsets. Following Tian and others (2005), we may use minus the logarithm of the partial likelihood function as a measure for the prediction error. For each $h$ and the $k$th data part,

$$
PE_k(h) = -\sum_{i \in kth part} \int_0^{r_0} \left[\{\hat{a}^{(k)}(t)\}^T Z_i(t) - \log\left(\sum_{j \in R_i} \exp[\{\hat{a}^{(k)}(t)\}^T Z_j(t)]\right)\right]dN_i(t),
$$

where $\hat{a}^{(k)}(t)$ is estimated using $K - 1$ data sets excluding the $k$th part with bandwidth $h$. Then the total prediction error with $h$ is $PE(h) = \sum_{k=1}^K PE_k(h)$, and we can choose the bandwidth parameter $h$ that minimizes $PE(h)$.
3. Numerical studies

3.1 Simulations

Simulation studies were carried out to evaluate the performance of the proposed estimation and inference procedures under finite-sample sizes. The failure times were generated from a model with the hazard function of \( \lambda(t|Z) = \lambda_0(t)e^{a_1(t)Z_1 + a_2(t)Z_2} \), where the baseline hazard function \( \lambda_0(t) = b_0 + b_1t + b_2t \exp\{b_3(t-2)^2\} \) with \((b_0, b_1, b_2, b_3)\) to be specified later and covariates \(Z_1\) and \(Z_2\) were generated from the binomial distribution with success probability of 0.5. We considered 2 types of time-varying function for \(a_1(t)\): polynomial, where \(a_1(t) = 1 - 0.25(t-2)^2\) and \((b_0, b_1, b_2, d) = (0.0, 0.005, 0.002, 0.25)\); log-sinusoid, where \(a_1(t) = \log[1 - 0.8 \sin(t/12)]\) and \((b_0, b_1, b_2, d) = (0.02, 0.01, 0, 0)\). We assumed \(a_2(t) = \gamma\) to be a constant coefficient of \(-0.5\). In addition, the censoring time was generated as \(C = \min(5, C^*)\), where \(C^*\) was from a uniform distribution on \((3, 6)\). As NCC studies are commonly implemented when the disease incidence rate is low, our simulated incidence rates were all about 10–13%.

We simulated the NCC data from full cohorts sized 1000 and 2000 and selected 2 controls for each case. We used the Epanechnikov kernel and ran 1000 simulations for each setting. The proposed simultaneous confidence band and testing procedure were carried out with 5000 resampling runs.

Table 1 presents the results regarding the simultaneous confidence band for the time-varying coefficient \(a_1(t)\). The bandwidth parameter \(h\) was set to change from 0.6 to 1.4 by an increment of 0.2 and the simultaneous band was constructed over \([1−h, 1+h]\), where \(h\) was reached, for example, when \(h = 1.2\) for \(n = 1000\) and \(h = 1\) for \(n = 2000\), the simultaneous confidence band yielded satisfactory coverage probability. The overall performance improved with increased sample size. For example, when \(n = 2000\) and \(h = 1.0\), the resampling-based simultaneous coverage probabilities matched the nominal level reasonably well and the empirical quantiles of observed \(S\) based on 1000 simulated data sets were very close to the resampling-based threshold. Furthermore, when \(a_1(t)\) was the quadratic polynomial function, the curvature of this polynomial function, that is, \(\alpha_1''(t) = -0.5\) for all \(t\); while \(a_1(t)\) being the log-sinusoid function, its curvature ranged from \(-0.5\) to 2.8. Therefore, by comparing the results between these 2 different types of time-varying coefficients, we found that the performance of simultaneous confidence band for the polynomial coefficient was less sensitive to the

<table>
<thead>
<tr>
<th>(a_1(t))</th>
<th>(h)</th>
<th>(N = 1000)</th>
<th>(N = 2000)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S.CP(\dagger)</td>
<td>E.S(\ddagger)</td>
<td>R.S(\S)</td>
</tr>
<tr>
<td>Polynomial</td>
<td>0.6</td>
<td>98.4</td>
<td>2.524</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>97.9</td>
<td>2.526</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>97.2</td>
<td>2.511</td>
</tr>
<tr>
<td></td>
<td>1.2</td>
<td>96.0</td>
<td>2.529</td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td>95.3</td>
<td>2.609</td>
</tr>
<tr>
<td>Sinusoid</td>
<td>0.6</td>
<td>97.8</td>
<td>2.494</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>97.1</td>
<td>2.514</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>96.5</td>
<td>2.568</td>
</tr>
<tr>
<td></td>
<td>1.2</td>
<td>95.2</td>
<td>2.643</td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td>92.7</td>
<td>2.739</td>
</tr>
</tbody>
</table>

\(\dagger\)S.CP: Simultaneous 95% coverage probability; \(\ddagger\)E.S: Empirical 95% quantile of \(\hat{S}\) in 1000 simulations; \(\S\)R.S: average of resampling-based thresholds; \(\$\)SD(R.S): standard deviation of resampling-based thresholds.
selection of $h$ than with the log-sinusoid coefficient function. This observation is not surprising because it is usually more difficult to characterize a more variable function and the bias of local linear fitting depends on the magnitude of $a''(t)$.

Figures 1 and 2 show the estimated time-varying coefficient curves with sample size of 2000, the average 95% point-wise confidence intervals over 1000 simulations, and 95% confidence envelope constructed using the point-wise 2.5% and 97.5% quantiles of the estimated curves over 1000 simulations. We found that when $h$ was small, the estimated curves were very close to the true curves but the point-wise confidence intervals were wide. As $h$ increased, the estimated curves showed biases at “the valley” but the point-wise confidence intervals were narrower.

Fig. 1. Estimated coefficient curves with different bandwidth parameters for the polynomial curve. 1: the true underlying coefficient; 2: the average of 1000 estimated curves; 3: the median of 1000 estimated curves; 4: the average of 95% point-wise confidence intervals; 5: confidence envelope of 2.5% and 97.5% quantiles of 1000 estimated curves.
We summarize the estimation results for the constant coefficient $a_2(t) \equiv \gamma$ in Table 2. We report the bias, the sample standard deviation (SD) of the estimates over 1000 simulations, the average standard error (SE) using the asymptotic approximation, and the 95% coverage probability (CP) of Wald-type confidence intervals. The biases were all small and the SDs and SEs decreased as the sample size increased. Overall, the SEs and SDs matched well and the 95% CPs were close to the nominal level. The performance of the integrated estimator $\hat{\gamma}$ for the constant coefficient was stable with respect to $h$. Such an observation confirms the theoretical result that the integrated estimator has rate of $n^{-1/2}$ and is independent of $h$ (given the rate of $h$ in certain range).
Table 2. Simulation results: estimation of constant coefficient $\gamma$

<table>
<thead>
<tr>
<th>$a(t)$</th>
<th>$h$</th>
<th>$N = 1000$</th>
<th></th>
<th>$N = 2000$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>SD$^\dagger$</td>
<td>SE$^\ddagger$</td>
<td>CP$^§$</td>
<td>Bias</td>
</tr>
<tr>
<td>Polynomial</td>
<td>0.6</td>
<td>0.002</td>
<td>0.247</td>
<td>0.252</td>
<td>0.964</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>−0.002</td>
<td>0.244</td>
<td>0.248</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>−0.003</td>
<td>0.242</td>
<td>0.248</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>1.2</td>
<td>−0.002</td>
<td>0.241</td>
<td>0.251</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td>−0.002</td>
<td>0.240</td>
<td>0.255</td>
<td>0.971</td>
</tr>
<tr>
<td>Sinusoid</td>
<td>0.6</td>
<td>0.005</td>
<td>0.262</td>
<td>0.276</td>
<td>0.969</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>−0.002</td>
<td>0.257</td>
<td>0.270</td>
<td>0.969</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>−0.004</td>
<td>0.253</td>
<td>0.270</td>
<td>0.969</td>
</tr>
<tr>
<td></td>
<td>1.2</td>
<td>−0.005</td>
<td>0.250</td>
<td>0.272</td>
<td>0.973</td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td>−0.005</td>
<td>0.249</td>
<td>0.276</td>
<td>0.976</td>
</tr>
</tbody>
</table>

$^\dagger$SD: Sample standard deviation of the proposed estimates in 1000 simulations; $^\ddagger$SE: average of standard error estimates; $^§$CP: coverage probability of the 95% Wald-type confidence interval.

Table 3. Simulation results: identifying time-varying coefficients with $N = 2000$

| $a(t)$ | $h$ | Constant coefficient | | Time-varying coefficient | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 5% rate | E.T$^\dagger$ | R.T$^\ddagger$ | SD(R.T)$^§$ | 5% rate | E.T$^\dagger$ | R.T$^\ddagger$ | SD(R.T)$^§$ |
| Polynomial | 0.6 | 0.065 | 1.220 | 1.111 | 0.160 | 0.412 | 2.053 | 1.123 | 0.160 |
| | 0.8 | 0.058 | 1.099 | 1.038 | 0.140 | 0.401 | 1.797 | 1.050 | 0.148 |
| | 1.0 | 0.050 | 1.084 | 1.008 | 0.137 | 0.363 | 1.649 | 1.020 | 0.147 |
| | 1.2 | 0.042 | 1.043 | 0.996 | 0.136 | 0.314 | 1.605 | 1.008 | 0.146 |
| | 1.4 | 0.039 | 1.024 | 0.993 | 0.134 | 0.277 | 1.562 | 1.004 | 0.143 |
| Sinusoid | 0.6 | 0.097 | 1.614 | 1.337 | 0.182 | 0.897 | 4.558 | 1.358 | 0.182 |
| | 0.8 | 0.068 | 1.369 | 1.237 | 0.157 | 0.882 | 3.820 | 1.263 | 0.155 |
| | 1.0 | 0.061 | 1.259 | 1.171 | 0.146 | 0.850 | 3.403 | 1.208 | 0.149 |
| | 1.2 | 0.055 | 1.171 | 1.124 | 0.141 | 0.798 | 2.997 | 1.171 | 0.149 |
| | 1.4 | 0.048 | 1.086 | 1.093 | 0.139 | 0.732 | 2.653 | 1.146 | 0.151 |

$^\dagger$E.T: Empirical 95% percentile of the test statistics in 1000 simulations; $^\ddagger$R.T: average of resampling-based thresholds; $^§$SD(R.T): standard deviation of resampling-based thresholds.

Finally, we assessed the performance of our proposed testing procedure for testing time-varying coefficients. Table 3 reports 5% error rates for testing $a_1(t)$ and $a_2(t)$, respectively, with sample size of 2000. The empirical threshold was estimated by the 95th quantile of the sample test statistics from 1000 runs of simulations, and the average resampling-based threshold was defined as the mean of resampling-based 1000 thresholds. For the constant coefficient, the proposed testing procedure showed good error rate control; for the time-varying coefficient, it yielded reasonable power. Similar to the observation in Table 1, the power was higher for detecting the log-sinusoid coefficient.

3.2 Breast cancer study in the NYUWHS

The details of this NCC study have been reported in Shore and others (2008). As an illustration, we estimated the gene–environmental interaction effects of gene XPC with the smoking exposure which can induce DNA damage. Based on the results of Shore and others (2008), we assumed a recessive model.
for gene $XPC$ and fitted model (1.1) with 4 covariates: Ethnicity (1 for Caucasian; 0 otherwise), XPC-Smoking-10 (1 for $XPC$-PAT allele $+/+$ nonsmokers; 0 otherwise), XPC-Smoking-01 (1 for $XPC$-PAT allele $−/−$ or $−+/+$ smokers; 0 otherwise), and XPC-Smoking-11 (1 for $XPC$-PAT allele $+/+$ smokers; 0 otherwise). Because there were very few incident cases before age 45, we focused our analysis on the NCC data with age of diagnosis between 45 and 75.

We fitted the kernel-weighted partial likelihood approach using the Epanechnikov kernel function with bandwidth parameter $h = 10$. The estimated coefficient curves, the point-wise confidence intervals, and the simultaneous confidence bands are presented in the top panel of Figure 3. We found that the Caucasian group had lower risk in early age than the non-Caucasian group but then the risk increased and peaked around 60–65. Both the 90% point-wise confidence interval and the simultaneous confidence

Fig. 3. Analysis results of the breast cancer NCC study. The solid curves are the estimated coefficients; the dashed lines are 90% point-wise confidence intervals; the dotted lines are 90% simultaneous confidence bands.
band excluded zero. There was no significantly increased risk in the XPC-PAT +/- nonsmokers group (XPC-Smoking-10) comparing to the reference group of XPC wild-type nonsmokers as the point-wise and simultaneous confidence intervals all included zero. In XPC-Smoking-01 group, the risk seemed to be elevated at early ages then diminished but the overall effect was not significant according to the simultaneous confidence band. Lastly, the risk of breast cancer in the group of XPC-PAT +/- smokers was uniformly elevated across all ages, and the effect was borderline significant, which agreed with the findings in Shore and others (2008).

We next applied the proposed resampling testing procedure to examine whether each covariate effect can be sufficiently described by a constant. We plot $T_j(t)$ and 10 resampling realizations in the lower panel of Figure 3, and the p-values based on 5000 resampling runs are also presented in the plots. The constant coefficient assumption for Ethnicity variable was rejected at 0.05 level ($p$-value = 0.0497), but this assumption seemed to be reasonable for all other covariates. Therefore, we further estimated the constant coefficients using the proposed integrated estimator (2.6). Comparing to XPC wild-type nonsmokers, there was no significantly increased risk in XPC-Smoking-10 group (OR = 1.01, 95% CI: 0.57–1.76, $p$ = 0.99) or in XPC-Smoking-01 group (OR = 0.94, 95% CI: 0.68–1.30, $p$ = 0.70), but a significant increase in XPC-Smoking-11 group (OR = 1.93, 95% CI: 1.15–3.25, $p$ = 0.01). Again, these observations confirmed the results of Shore and others (2008).

4. DISCUSSION

As a way of enhancing modeling flexibility and providing an alternative and diagnostic tool for Cox’s proportional hazards model, we developed inference procedures for the Cox regression model with time-varying coefficients in NCC studies. The NCC design is a member of a general class of cohort risk-set sampling designs that include counter-matching design (Langholz and Borgan, 1995), quota-matching design, and many others. Borgan and others (1995) studied Cox’s proportional hazards model in this general framework using the marked point process to characterize the failure time process and the sampling scheme simultaneously. As a referee pointed out, the proposed inference procedure for NCC studies may be generalized to accommodate other types of cohort risk-set sampling designs by a formulation based on marked point process theory (Brémaud, 1981).

In cohort studies, the data can often present not only right censoring but also left truncation. For example, left-truncated data occur when only subjects who are disease-free enter the study because the disease can only happen afterward. To incorporate the left truncation, controls need to be drawn from an adjusted risk set defined as $\tilde{R}(t) = \{i: X_i < t \leq T_i\}$, where $X_i$ denotes the left-truncation time for subject $i$. The extension to accommodate left-truncated and right-censored NCC data is certainly of interest and may be handled in a unified manner using the marked point process and martingale theory (Brémaud, 1981). We shall further investigate this extension elsewhere.

Our analysis of the breast cancer data in the NYUWHS has confirmed the finding of interaction between gene XPC in the NER pathway and smoking, a DNA-damaging agent (Shore and others, 2008). In addition, the finding that the Caucasian group has lower risk at earlier age and then elevated risk at older age compared to the non-Caucasian group is consistent with the well-described black-to-white ethnic crossover in breast cancer incidence rate (Gray and others, 1980; Joslyn and others, 2005; Anderson and others, 2008) and demonstrates the issue of race disparity in disease risk. In practice, it is also common to add covariate-by-time interaction terms to Cox’s proportional hazards model to examine possible time-varying coefficients. We have also considered this approach by adding an interaction term of ethnicity with time and fitted the model with Thomas’ partial likelihood estimation approach. However, the interaction term did not reach statistical significance ($p$-value = 0.127). One possible reason is that as shown in Figure 3, the shape of the ethnic effect is like a $U$-shape rather than linear or monotone, and the
simple covariate–time interaction term may not be able to catch such a trend. In summary, the Cox model with time-varying coefficients can elucidate the effect of risk factors on the disease and provide intuitive tool to visualize such effects.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The authors would like thank the co-editor, an associate editor, and a referee for their valuable suggestions. Conflict of Interest: None declared.

FUNDING

National Cancer Institute (CA098661, CA091892, CA16087, CA140632); Department of Defense (DAMD17-01-1-0578); Susan G. Komen Breast Cancer Research Foundation (BCTR 2000 685).

REFERENCES


[Received December 18, 2009; revised May 4, 2010; accepted for publication May 5, 2010]