Ann Inst Stat Math (2012) 64:715–736
DOI 10.1007/s10463-011-0330-0

Optimal inferences for proportional hazards model with parametric covariate transformations

Chunpeng Fan · Jason P. Fine · Jong-Hyeon Jeong

Received: 10 June 2010 / Revised: 16 November 2010 / Published online: 12 March 2011

Abstract The traditional Cox model assumes a log-linear relationship between covariates and the underlying hazard function. However, the linearity may be invalid in real data. We study a Cox model which employs unknown parametric covariate transformations. This model is applicable to observational studies or randomized trials when a treatment effect is investigated after controlling for a confounding variable that may have non-log-linear relationship with the underlying hazard function. While the proposed generalization is simple, the inferential issues are challenging due to the loss of identifiability under no effects of transformed covariates. Optimal tests are derived for certain alternatives. Rigorous parametric inference is established under regularity conditions and non-zero transformed covariate effects. The estimates perform well in simulation studies with realistic sample size, and the proposed tests are more powerful than the usual partial likelihood ratio test, which is no longer optimal. Data from a breast cancer trial are used to illustrate the model building strategy and the better fit of the proposed model, comparing to the traditional Cox model.

Keywords Cox proportional hazards model · Optimal test · Semi-parametric model

C. Fan (✉)
Department of Biostatistics and Programming, Sanofi-aventis, Bridgewater, NJ 08807, USA
e-mail: Chunpeng.Fan@sanofi-aventis.com

J. P. Fine
Department of Biostatistics, The University of North Carolina, Chapel Hill, NC 27599, USA
e-mail: jfine@bios.unc.edu

J.-H. Jeong
Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA
e-mail: jeong@nsabp.pitt.edu

1 Introduction

When dealing with failure time data, a famous model is the Cox proportional hazards model (Cox 1972):

$$h(t; z) = h_0(t) \exp\left\{\boldsymbol{\beta}^\tau z(t)\right\}, \quad t \geq 0 \tag{1}$$

where the hazard rate or intensity of failure $h(t) = \lim_{\Delta t \downarrow 0} Pr\left[T \leq t + \Delta t | T > t\right]$ for the survival time T of an individual, z is a $p \times 1$ covariate vector which may depend on the time t, $\boldsymbol{\beta}$ is a p-vector of unknown regression coefficients and $h_0(t)$, the underlying hazard, is an unknown and unspecified nonnegative function.

Two assumptions are implied in Cox model, the proportional hazard assumption and the linear relationship between log hazard and covariates. In some circumstances, the log-linear relationship assumption may be invalid and may then induce invalidated proportional hazard assumption. A natural idea to recover the linearity is to allow parametric transformations to the covariates.

Parametric covariate transformations have been widely used in many kinds of models. They were first considered in the classic normal linear model (Draper and Smith 1981; Neter et al. 1985) with constant variance; see, for example, Box and Tidwell (1962). A natural extension is to relax the homogenous normal error distribution; see Carroll and Ruppert (1988) for a discussion of non-normal error distributions (exponential, Laplace, etc) and variance function models. Another extension of the linear model is to incorporate unknown transformations of the response. Parametric (Draper and Cox 1969; Lindsey 1972) and non-parametric [including but not limited to Doksum (1987); Pettitt (1982, 1984); Cheng et al. (1995); Fine et al. (1998)] response transformations have been studied. In addition to their utility in assessing classical linear model assumptions, nonparametric response transformations have been thoroughly studied with censored survival data because the Cox model (1) is a special case with homogeneous extreme value errors.

Response and covariate transformations have been considered jointly (Cook and Wang 1983; Atkinson 1986, 1988). These approaches are fully parametric with both response and covariates permitted to follow different parametric transformations. In related work on transform-both-sides models, the response and the linear predictor in the linear model are transformed using a single transformation, which may be either parametric (Carroll and Ruppert 1988) or nonparametric (Wang and Ruppert 1995), but untransformed covariates.

Theoretical challenges arise with parametric covariate transformations in linear models, with and without response transformations. The covariate's transformation cannot be identified with zero covariate effect, which means that standard likelihood ratio tests are not applicable with unknown transformation. The issue has been glossed over in most of the above-cited papers on parametric covariate transformations. Inferences which are valid under this loss of identifiability can be traced to early work by Davies (1977, 1987) on likelihood inferences for parametric models and later work by Andrews and Ploberger (1994), which studied the admissibility of the likelihood ratio test and the construction of optimal tests with better power properties than the likelihood ratio test. All existing work focuses on likelihood based methods and their

extension to the semi-parametric Cox model where inferences involve partial likelihood has not been established.

Nonparametric covariate transformations [as in generalized additive models, see Hastie and Tibshirani (1986)] have also received much attention. The loss of identifiability does not occur with nonparametric covariate transformations where the covariate's transformation and effect are not distinguished, as in Hastie and Tibshirani (1990), where the Cox model is specified with nonparametric covariate effects. Of course, smoothing is generally needed, and the usual parametric rates of convergence do not hold, making inference tricky. Such difficulties have hindered the adoption of such methods for formal testing. In practice, such models are generally used in checking goodness-of-fit of parametric assumptions and in exploratory analysis of covariate effects.

In this paper, we generalize the Cox model (1) to allow parametric covariate transformations and establish its inferences under regularity conditions and non-zero effects of the transformed covariates, and we also develop optimal tests for the effects of the transformed covariates adapting the framework in Andrews and Ploberger (1994) for likelihood inferences for parametric models to partial likelihood for proportional hazards model.

The generalized Cox model with parametric covariate transformations can be defined as

$$h(t; z) = h_0(t) \exp\left\{\boldsymbol{\beta}^\tau g_{\boldsymbol{\lambda}}(z(t))\right\}, \quad t \geq 0 \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\tau \in \mathscr{R}^p$ is a p-dimensional parameter, $g_{\boldsymbol{\lambda}}(z)$ is the transformation function, and if we denote $z = (z_1, \ldots, z_p)^\tau \in \mathscr{Z} \subset \mathscr{R}^p$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)^\tau \in \boldsymbol{\Lambda} \subset \mathscr{R}^q$, $g_{\boldsymbol{\lambda}}(z) : \mathscr{Z} \times \boldsymbol{\Lambda} \mapsto \mathscr{R}^p$ is defined as $g_{\boldsymbol{\lambda}}(z) = \left\{g_1(z_1, \lambda_1), \ldots, g_q(z_q, \lambda_q), g_{q+1}(z_{q+1}), \ldots, g_p(z_p)\right\}^\tau$, where $\left\{g_i(z_i, \lambda_i), i = 1, \ldots, q\right\}$ are real-valued transformation functions with parameters λ_i; $\{g_j(z_j), j = q+1, \ldots, p\}$ are known functions. For non-negative covariates, a commonly used unspecified transformation is the Box–Cox transformation (Box and Cox 1964): $g_i(z_i, \lambda_i) = (z_i^{\lambda_i} - 1)/\lambda_i$ if $\lambda_i \neq 0$, and $\log(z_i)$ if $\lambda_i = 0$, for $i = 1, \ldots, q$. A commonly used example of $g_j(z_j)$ is just taking $g_j(z_j) = z_j$. General conditions for $g_{\boldsymbol{\lambda}}(z)$ will be discussed in Sect. 3.

When the model is identifiable, i.e., when regularity conditions are satisfied and the true value of any $\beta_i, i = 1, \ldots, q$, is not 0, estimates of the true value of $\boldsymbol{\theta} = (\boldsymbol{\beta}^\tau, \boldsymbol{\lambda}^\tau)^\tau$ can be obtained by maximizing the partial likelihood function

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \frac{\exp\{\boldsymbol{\beta}^\tau g_{\boldsymbol{\lambda}}(z_i(T_i))\}}{\sum_{j \in \mathscr{R}_i} \exp\{\boldsymbol{\beta}^\tau g_{\boldsymbol{\lambda}}(z_j(T_i))\}} \right\}^{\delta_i} \tag{3}$$

where $\mathscr{R}_i = \{j : T_j \geq T_i\}$ and $1 - \delta_i$ is an indicator for censoring. We can modify Breslow's estimator (Breslow 1972, 1974) to estimate the cumulative hazard $H_0 = \int_0^t h_0(s)ds : \hat{H}(t) = \sum_{T_i \leq t} \left[\delta_i / \sum_{j \in \mathscr{R}_i} \exp\left\{\hat{\boldsymbol{\beta}}^\tau g_{\hat{\boldsymbol{\lambda}}}(z_j(T_i))\right\}\right]$. Counting process setup similar to Andersen and Gill (1982) can be employed to investigate the asymptotic properties of the estimates of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ and the cumulative baseline hazard $H_0(t)$; see Sects. 2 and 3.

In model (2), if the true value of β_i is 0 for some $i \in \{1, \ldots, q\}$, corresponding transformation parameter λ_i is not identifiable. This is the so-called model identifiability problem investigated in Davies (1977, 1987). Ideas from Andrews (1993) and Andrews and Ploberger (1994) motivate us to test $H_{i0} : \beta_i = 0$ versus $H_{ia} : \beta_i \neq 0$, and transformation parameter λ_i exists, for $i = 1, \ldots, q$.

To adapt the methods and results in Andrews and Ploberger (1994), a real likelihood instead of a partial likelihood is needed. When covariate z is time independent, Doksum (1987) shows that partial likelihood can be interpreted as rank likelihood which is a real likelihood function with integral equal to 1. Section 4 adapts the optimal theorems in Andrews and Ploberger (1994) to the rank likelihood to achieve the proposed test.

Because the limiting distribution of our optimal tests is nonstandard, its critical values need to be decided. In Sect. 5, we develop a Gaussian multiplier method to generate these critical values by adapting ideas from Hansen (1996, 2000). Because the partial likelihood is not based on independent terms, we use the martingale structure to implement the generation procedure.

Section 6 gives simulation results to compare the rejection rates with theoretical values, to compare the power of the optimal test and naive tests, and to verify the parameter estimations.

We apply the proposed model to a data from a breast cancer clinical trial in Sect. 7. Model building procedure including tests for covariate effects, covariate transformation selection, and estimates of parameters is shown. When assessing the proposed model and the Cox model without parametric covariate transformations, AIC is used to evaluate the goodness-of-fit, and estimated explained variation is used to evaluate the prediction.

The rest of the paper is organized as the follows. Section 2 defines counting process setup for the Cox model with covariate transformations (2). Section 3 gives the asymptotic properties of the estimates of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, and baseline hazard $h_0(t)$ under some regularity conditions including model identifiability. In Sect. 4, optimal tests for the covariate effects are derived. Section 5 develops a Gaussian multiplier method to generate the limiting distribution of the optimal test statistic. Section 6 investigates simulation studies. The application to the breast cancer data is shown in Sect. 7. Discussions are located in Sect. 8.

2 Notation and assumptions

In this section we shall formulate the model (2) in the framework of multivariate counting process which is similar to the framework for the Cox model in Andersen and Gill (1982). For simplicity, we shall be working on the time interval [0,1], and the extension to process on $[0, \infty]$ would be similar to Andersen and Gill (1982) thus omitted. Background theories including multivariate counting processes, stochastic integrals, and local martingales would be used without further comment.

Suppose we have a sequence of models, indexed by $n = 1, 2, \ldots$. First, we generate the possibly censored observation of the lifetimes of n individuals to the observation (in the nth model) of an n-component multivariate counting process

$N^{(n)} = (N_1^{(n)}, \ldots, N_n^{(n)})$, where $N_i^{(n)}$ counts observed events in the life of the ith individual, $i = 1, \ldots, n$, over the time interval [0,1]. Therefore, the sample paths of $N_1^{(n)}, \ldots, N_n^{(n)}$ are step functions, zero at time zero, with jumps of size $+1$ only, no two component processes jumping at the same time. We assume $N_i^{(n)}(1)$ is almost surely finite.

The basic assumption is that for each n, $N^{(n)}$ has random intensity process $h^{(n)} = (h_i^{(n)}, \ldots, h_n^{(n)})$ such that $h_i^{(n)}(t) = Y_i^{(n)}(t)h_0(t)\exp\{\boldsymbol{\beta}_0^\tau g_{\boldsymbol{\lambda}_0}(\boldsymbol{Z}_i^{(n)}(t))\}$. Here, $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^\tau, \boldsymbol{\lambda}_0^\tau)^\tau$ is a fixed column vector of $(p+q)$ components, h_0 a fixed underlying hazard function, g the specified transformation function, and $Y_i^{(n)}$ a predictable process taking values in {0, 1} indicating (by value 1) when the ith individual is under observation (so in particular, $N_i^{(n)}$ jumps only when $Y_i^{(n)} = 1$). Finally, $\boldsymbol{Z}_i^{(n)} = (Z_{i1}^{(n)}, \ldots, Z_{ip}^{(n)})^\tau$ is a column vector of p covariate processes for the ith individual. We suppose that $\boldsymbol{Z}_i^{(n)}$ is predictable and locally bounded. Assumptions about the continuous function $g_{\boldsymbol{\lambda}}(\cdot)$ will be introduced in detail in the next section.

By stating that $N^{(n)}$ has intensity process $h^{(n)}$ we mean that the process $M_i^{(n)}$ defined by

$$M_i^{(n)}(t) = N_i^{(n)}(t) - \int_0^t h_i^{(n)}(u)\mathrm{d}u, \quad i = 1, \ldots, n, \ t \in [0, 1],$$

is a local martingale on the time interval [0,1]. As a consequence, they are in fact local square integrable martingales, with $\langle M_i^{(n)}, M_i^{(n)}\rangle(t) = \int_0^t h_i^{(n)}(u)\mathrm{d}u$ and $\langle M_i^{(n)}, M_j^{(n)}\rangle(t) = 0, i \neq j$, i.e., $M_i^{(n)}$ and $M_j^{(n)}$ are orthogonal when $i \neq j$.

For simplicity, in the following, we will drop the superscript (n). Only $\boldsymbol{\theta}_0$ and h_0 are the same in all models (i.e., for each n). Convergence in probability ($\rightarrow_p$) and convergence in distribution ($\rightarrow_d$) are always relative to the probability measures $P^{(n)}$ parameterized by $\boldsymbol{\theta}_0$ and h_0.

3 Asymptotic properties

For simplicity, we denote $\dot{g}_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\cdot) = \partial\{\boldsymbol{\beta}^\tau g_{\boldsymbol{\lambda}}(\cdot)\}/\partial\boldsymbol{\lambda}$, $\ddot{g}_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\cdot) = \partial^2\{\boldsymbol{\beta}^\tau g_{\boldsymbol{\lambda}}(\cdot)\}/\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^\tau$, and $\dot{g}_{\boldsymbol{\lambda}}(\cdot) = \partial g_{\boldsymbol{\lambda}}(\cdot)/\partial\boldsymbol{\lambda}^\tau$ with dimensions $q \times 1$, $q \times q$, and $p \times q$, respectively.

$$S^{(0)}(\boldsymbol{\theta}, t) = \frac{1}{n}\sum_{i=1}^n Y_i(t)e^{\boldsymbol{\beta}^\tau g_{\boldsymbol{\lambda}}(\boldsymbol{Z}_i(t))}, \quad S^{(1)}(\boldsymbol{\theta}, t) = \frac{\partial}{\partial\boldsymbol{\theta}}S^{(0)}(\boldsymbol{\theta}, t),$$

$$S^{(2)}(\boldsymbol{\theta}, t) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\tau}S^{(0)}(\boldsymbol{\theta}, t),$$

$$S_0^{(2)}(\boldsymbol{\theta}, t) = \frac{1}{n}\sum_{i=1}^n Y_i(t)e^{\boldsymbol{\beta}_0^\tau g_{\boldsymbol{\lambda}_0}(\boldsymbol{Z}_i(t))} \cdot \begin{bmatrix} 0 & \dot{g}_{\boldsymbol{\lambda}}(\boldsymbol{Z}_i(t)) \\ \dot{g}_{\boldsymbol{\lambda}}^\tau(\boldsymbol{Z}_i(t)) & \ddot{g}_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\boldsymbol{Z}_i(t)) \end{bmatrix},$$

$$E(\boldsymbol{\theta}, t) = \frac{S^{(1)}(\boldsymbol{\theta}, t)}{S^{(0)}(\boldsymbol{\theta}, t)},$$

$$V(\boldsymbol{\theta}, t) = \frac{S^{(2)}(\boldsymbol{\theta}, t) - S_0^{(2)}(\boldsymbol{\theta}, t)}{S^{(0)}(\boldsymbol{\theta}, t)} - E(\boldsymbol{\theta}, t)E^{\tau}(\boldsymbol{\theta}, t).$$

Note that $S^{(0)}$ is scalar, $S^{(1)}$ and E are $(p+q)$ vectors (i.e., with dimension $(p+q)\times 1$), and $S^{(2)}$, $S_0^{(2)}$, and V are $(p+q)\times(p+q)$ matrices.

As an extension of the regular Cox partial likelihood estimator, the estimate of the parameter $\boldsymbol{\theta}_0$ can be reasonably obtained by maximizing the partial likelihood function $L(\boldsymbol{\theta})$ in (3). Denoting

$$C(\boldsymbol{\theta}, t) = \sum_{i=1}^{n} \int_0^t \boldsymbol{\beta}^{\tau} g_{\boldsymbol{\lambda}}(\boldsymbol{Z}_i(s)) \mathrm{d}N_i(s) - \int_0^t \log \left\{ \sum_{i=1}^{n} Y_i(s) e^{\boldsymbol{\beta}^{\tau} g_{\boldsymbol{\lambda}}(\boldsymbol{Z}_i(s))} \right\} \mathrm{d}\bar{N}(s)$$

where $\bar{N} = \sum_{i=1}^{n} N_i$, then we have that $C(\boldsymbol{\theta}, 1) = \log L(\boldsymbol{\theta})$, and the estimator $\hat{\boldsymbol{\theta}}$ is defined as the solution to the likelihood equation $U(\boldsymbol{\theta}, 1) = (\partial/\partial\boldsymbol{\theta})C(\boldsymbol{\theta}, 1) = 0$, where

$$U(\boldsymbol{\theta}, t) = \frac{\partial}{\partial \boldsymbol{\theta}} C(\boldsymbol{\theta}, t) = \sum_{i=1}^{n} \int_0^t \begin{bmatrix} g_{\boldsymbol{\lambda}}(\boldsymbol{Z}_i(s)) \\ \dot{g}_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\boldsymbol{Z}_i(s)) \end{bmatrix} \mathrm{d}N_i(s) - \int_0^t \frac{S^{(1)}(\boldsymbol{\theta}, s)}{S^{(0)}(\boldsymbol{\theta}, s)} \mathrm{d}\bar{N}(s).$$

CONDITIONS.

A. (Finite interval). $\int_0^1 h_0(t)\mathrm{d}t < \infty$.
B. (Regular transformations). Transformation $g_{\boldsymbol{\lambda}}(z)$ satisfies the differentiability and boundedness condition: $g_{\boldsymbol{\lambda}}(z)$ is two-order differentiable in $\boldsymbol{\lambda}$ and z, and $g_{\boldsymbol{\lambda}}(z)$, and all its first and second order derivatives are uniformly bounded and continuous in $\boldsymbol{\lambda}$ across $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ and $z \in \mathscr{Z}$, where $\boldsymbol{\Lambda} \in \mathscr{R}^q$ and $\mathscr{Z} \in \mathscr{R}^p$ are bounded sets.
C. (Asymptotic stability). There exists a neighborhood $\boldsymbol{\Theta}$ of $\boldsymbol{\theta}$ and scalar function $s^{(0)}$, vector function $s^{(1)}$, and matrix function $s^{(2)}$, $s_0^{(2)}$ defined on $\boldsymbol{\Theta} \times [0, 1]$ such that for $j = 0, 1, 2$, $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, t\in[0,1]} ||S^{(j)}(\boldsymbol{\theta}, t) - s^{(j)}(\boldsymbol{\theta}, t)|| \to_p 0$, and $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, t\in[0,1]} ||S_0^{(2)}(\boldsymbol{\theta}, t) - s_0^{(2)}(\boldsymbol{\theta}, t)|| \to_p 0$.
D. (Lindeberg condition). There exists $\delta > 0$ such that

$$n^{-1/2} \sup_{i,t} \left\| \begin{bmatrix} g_{\boldsymbol{\lambda}_0}(\boldsymbol{Z}_i(t)) \\ \dot{g}_{\boldsymbol{\beta}_0,\boldsymbol{\lambda}_0}(\boldsymbol{Z}_i(t)) \end{bmatrix} \right\| Y_i(t) I \left\{ \boldsymbol{\beta}_0^{\tau} g_{\boldsymbol{\lambda}_0}(\boldsymbol{Z}_i(t)) > -\delta \left\| \begin{bmatrix} g_{\boldsymbol{\lambda}_0}(\boldsymbol{Z}_i(t)) \\ \dot{g}_{\boldsymbol{\beta}_0,\boldsymbol{\lambda}_0}(\boldsymbol{Z}_i(t)) \end{bmatrix} \right\| \right\} \to_p 0.$$

E. (Asymptotic regularity conditions). Let $\boldsymbol{\Theta}$, $s^{(0)}$, $s^{(1)}$, $s^{(2)}$, and $s_0^{(2)}$ be as in Condition C and define $e = s^{(1)}/s^{(0)}$ and $v = (s^{(2)} - s_0^{(2)})/s^{(0)} - ee^{\tau}$. For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $t \in [0, 1]$, $s^{(1)}(\boldsymbol{\theta}, t) = \partial s^{(0)}(\boldsymbol{\theta}, t)/\partial\boldsymbol{\theta}$, $s^{(2)}(\boldsymbol{\theta}, t) = \partial^2 s^{(0)}(\boldsymbol{\theta}, t)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\tau}$. $s^{(0)}(\cdot, t)$, $s^{(1)}(\cdot, t)$, $s^{(2)}(\cdot, t)$ and $s_0^{(2)}(\cdot, t)$ are continuous functions of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, uniformly in $t \in [0, 1]$, $s^{(0)}$, $s^{(1)}$, $s^{(2)}$, and $s_0^{(2)}$ are bounded on $\boldsymbol{\Theta} \times [0, 1]$; $s^{(0)}$ is bounded away from zero on $\boldsymbol{\Theta} \times [0, 1]$, and the matrix $\Sigma = \int_0^1 v(\boldsymbol{\theta}_0, t)s^{(0)}(\boldsymbol{\theta}_0, t)h_0(t)\mathrm{d}t$ is positive definite.

Note that the partial derivative conditions on $s^{(0)}$, $s^{(1)}$, $s^{(2)}$, and $s_0^{(2)}$ are satisfied by $S^{(0)}$, $S^{(1)}$, $S^{(2)}$, and $S_0^{(2)}$; and that Σ is automatically positive semi-definite. Furthermore, the interval [0,1] in the conditions may everywhere be replaced by the set $\{t : h_0(t) > 0\}$. Consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$ are in Theorem 1. Corollary 1 gives an estimate of the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Theorem 2 gives the weak convergence of the estimated cumulative hazard function $\hat{H}$. Corollary 2 gives the estimate of its limiting covariance function. The proofs for these results are similar to those for Theorems 3.1, 3.2, and 3.3 in Andersen and Gill (1982) and therefore omitted.

Theorem 1 $\hat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_0$ *and* $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d N(0, \Sigma^{-1})$.

Corollary 1 (Consistent estimate of Σ). $n^{-1}\mathscr{I}(\hat{\boldsymbol{\theta}}, 1) \to_p \Sigma$, *where the positive semi-definite matrix is minus the second derivative of* $C(\boldsymbol{\theta}, t)$ *w.r.t.* $\boldsymbol{\theta}$*:*

$$\mathscr{I}(\boldsymbol{\theta}, t) = -\int_0^t \sum_{i=1}^n \begin{bmatrix} 0 & \dot{g}_{\boldsymbol{\lambda}}(\boldsymbol{Z}_i(s)) \\ \dot{g}_{\boldsymbol{\lambda}}^{\tau}(\boldsymbol{Z}_i(s)) & \ddot{g}_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\boldsymbol{Z}_i(s)) \end{bmatrix} dN_i(s) + \int_0^t \left[\frac{S^{(2)}(\boldsymbol{\theta}, s)}{S^{(0)}(\boldsymbol{\theta}, s)} - \frac{S^{(1)}(\boldsymbol{\theta}, s)\{S^{(1)}(\boldsymbol{\theta}, s)\}^{\tau}}{\{S^{(0)}(\boldsymbol{\theta}, s)\}^2} \right] d\bar{N}(s).$$

Theorem 2 (Weak convergence of $n^{1/2}(\hat{H} - H_0)$). $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ *and the process equal in the point* t *to* $n^{1/2}\{\hat{H}(t) - H_0(t)\} + n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\tau} \int_0^t e(\boldsymbol{\theta}_0, u)h_0(u)du$ *are asymptotically independent, the latter being asymptotically distributed as a Gaussian martingale with variance function* $\int_0^t h_0(u)/s^{(0)}(\boldsymbol{\theta}_0, u)du$.

Corollary 2 (Consistent estimate of limiting covariance function of $n^{1/2}(\hat{H} - H_0)$).

$$\sup_{t\in[0,1]} \left\| K(\hat{\boldsymbol{\theta}}, t) + \int_0^t e(\boldsymbol{\theta}_0, u)h_0(u)du \right\| \to_p 0,$$

where $K(\boldsymbol{\theta}, t) = -\int_0^t S^{(1)}(\boldsymbol{\theta}, t)/(S^{(0)})^2(\boldsymbol{\theta}, t)d\bar{N}(s)$.

4 Optimal tests for covariate effects

When there exists $i \in \{1, \ldots, q\}$ such that the true value of β_i is 0, λ_i is not identifiable. Although we exclude this case when we derive the asymptotic properties by assuming matrix Σ in condition E is positive definite, we need the formal test to decide whether there is effect by covariate Z_i. Without loss of generality, we make our hypotheses be

$$H_{0d} : \boldsymbol{\beta}_d = \mathbf{0} \quad \text{vs} \quad H_{ad} : \boldsymbol{\beta}_d \neq \mathbf{0}, \text{ and there exists transformation parameter } \boldsymbol{\lambda}_d,$$

where $1 \leq d \leq q$, $\boldsymbol{\beta}_d = (\beta_1, \ldots, \beta_d)^{\tau}$, $\boldsymbol{\lambda}_d = (\lambda_1, \ldots, \lambda_d)^{\tau}$. When $d \neq 1$, we can get that there exist the effects of at least one of $Z_1, \ldots, Z_d$. When $d = 1$, we can tell if Z_1 has effect.

Because nuisance parameter $\boldsymbol{\lambda}_d$ appears only under the alternative, we may use the idea of profiling $\boldsymbol{\lambda}_d$ over range $\boldsymbol{\Lambda}_d \subset \mathscr{R}^d$ which was investigated by Andrews and Ploberger (1994) in parametric case. We assume the true values of $\beta_{d+1}, \ldots, \beta_q$ under the null hypothesis $\boldsymbol{\Theta}_{0d}$ to be non-zero to ensure $\lambda_{d+1}, \ldots, \lambda_q$ are estimatable, where the null hypothesis space $\boldsymbol{\Theta}_{0d} = \{\boldsymbol{\theta}_d \in \boldsymbol{\Theta}_d : \boldsymbol{\theta}_d = (0, \ldots, 0, \beta_{d+1}, \ldots, \beta_q, \beta_{q+1}, \ldots, \beta_p, \lambda_{d+1}, \ldots, \lambda_q)^\tau\} \subset \boldsymbol{\Theta}_d$, and the union of null and alternative hypotheses spaces $\boldsymbol{\Theta}_d = \{\boldsymbol{\theta}_d \in \boldsymbol{\Theta}_d : \boldsymbol{\theta}_d = (\beta_1, \ldots, \beta_p, \lambda_{d+1}, \ldots, \lambda_q)^\tau\} \subset \boldsymbol{\Theta} \cap \mathscr{R}^{p+q-d}$.

For fixed $\boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d \subset \mathscr{R}^d$, the model is an identifiable Cox model with covariate transformations which has been discussed in Sect. 2. To utilize similar methods investigated by Andrews and Ploberger (1994), we may need the partial likelihood for fixed $\boldsymbol{\lambda}_d$, $L_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d) = L(\boldsymbol{\theta})$, to be a true likelihood. When the covariates $Z_1, \ldots, Z_p$ are time-independent, Pettitt (1983) introduced the rank likelihood $l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d) = P_{\boldsymbol{\lambda}_d}(\boldsymbol{R} = \boldsymbol{r}, \boldsymbol{\delta} = \boldsymbol{\delta}|\boldsymbol{\theta}_d)$ where $\boldsymbol{r}$ is the rank vector for all event and censored times. Doksum (1987) tells that the relationship between $L_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d)$ and $l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d)$ is

$$L_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d) = l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d) = P_{\boldsymbol{\lambda}_d}(\boldsymbol{R} = \boldsymbol{r}, \boldsymbol{\delta} = \boldsymbol{\delta}|\boldsymbol{\theta}_d).$$

We can now construct the optimal tests. Let $\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d) = \log l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d)$, $D\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d) = \partial \mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d)/\partial \boldsymbol{\theta}_d$, and $D^2\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d) = \partial^2 \mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d)/\partial \boldsymbol{\theta}_d \partial \boldsymbol{\theta}_d^\tau$. Note that in general, $D\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d})$ and $D^2\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d})$ depend on $\boldsymbol{\lambda}_d$ although $l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d})$ and $\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d})$ do not.

Because the optimal test would be based on maximum rank likelihood estimators, i.e., the maximum partial likelihood estimator, with fixed $\boldsymbol{\lambda}_d$, we may need Conditions A–E hold for each fixed $\boldsymbol{\lambda}_d$, and some of these convergence properties need to be uniform in $\boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d$.

Assumptions 1. For each fixed $\boldsymbol{\lambda}_d$, Conditions A–E hold for model (2) with parameter $\boldsymbol{\theta}_d$.

2. $\boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d$, a closed set in $\mathscr{R}^d$, and $\boldsymbol{\lambda}_d \sim J(\boldsymbol{\lambda}_d)$.
3. $-n^{-1}D^2\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d) \to_p \Sigma(\boldsymbol{\theta}_d, \boldsymbol{\lambda}_d)$ uniformly over $\boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d$ and $\boldsymbol{\theta}_d \in \boldsymbol{\Theta}_{0d}$ under $\boldsymbol{\theta}_{0d}$; $\Sigma(\boldsymbol{\theta}_d, \boldsymbol{\lambda}_d)$ is uniformly continuous in $(\boldsymbol{\theta}_d, \boldsymbol{\lambda}_d)$ over $\boldsymbol{\Theta}_{0d} \times \boldsymbol{\Lambda}_d$; $\Sigma(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_d)$ is uniformly positive definite over $\boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d$, where $\Sigma(\boldsymbol{\theta}_d, \boldsymbol{\lambda}_d)$ is the $(p+q-d) \times (p+q-d)$ sub-matrix of $\Sigma(\boldsymbol{\theta}) = \int_0^1 v(\boldsymbol{\theta}, t) S^{(0)}(\boldsymbol{\theta}, t) h_0(t) \mathrm{d}t$ and does not involve rows or columns corresponding to $\boldsymbol{\lambda}_d$.
4. (*Local alternative*). Local alternatives to H_{0d} is of the form $l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d} + \boldsymbol{h}/\sqrt{n})$ for $\boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d$, and $\boldsymbol{h} \in \mathscr{R}^{p+q-d}$. $\boldsymbol{h}$ has pre-set distribution $Q_{\boldsymbol{\lambda}_d}(\boldsymbol{h}) = N(\boldsymbol{0}, c\Sigma_{\boldsymbol{\lambda}_d})$ where c is a scalar constant and $N(\boldsymbol{0}, \Sigma)$ denotes a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix Σ.

The local alternative assumption involves a weight function $Q_{\boldsymbol{\lambda}_d}(\cdot)$ on $\mathscr{R}^{p+q-d}$ that concentrates on the orthogonal complement of $\boldsymbol{V}$ with respect to the inner product $\langle \boldsymbol{h}, \boldsymbol{l} \rangle_{\boldsymbol{\lambda}_d} = \boldsymbol{h}^\tau \mathscr{I}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d}) \boldsymbol{l}$ for $\boldsymbol{h}, \boldsymbol{l} \in \mathscr{R}^{p+q-d}$, where $\boldsymbol{V}$ is the linear subspace of $\mathscr{R}^{p+q-d}$: $\boldsymbol{V} = \{\boldsymbol{\theta}_d \in \mathscr{R}^{p+q-d} : \boldsymbol{\theta}_d = (0, \ldots, 0, \beta_{d+1}, \ldots, \beta_p, \lambda_{d+1}, \ldots, \lambda_q)^\tau\}$. We call this orthogonal complement $\boldsymbol{V}^\perp_{\boldsymbol{\lambda}_d}$.

Since $\boldsymbol{V}$ is $(p+q-2d)$-dimensional subspace of $\mathscr{R}^{p+q-d}$, $\boldsymbol{V}^\perp_{\boldsymbol{\lambda}_d}$ is a d-dimensional subspace of $\mathscr{R}^{p+q-d}$. Let $\{a_{1\boldsymbol{\lambda}_d}, a_{2\boldsymbol{\lambda}_d}, \ldots, a_{d\boldsymbol{\lambda}_d}\}$ be some basis of $\boldsymbol{V}^\perp$ and define $A_{\boldsymbol{\lambda}_d} = [a_{1\boldsymbol{\lambda}_d}, \ldots, a_{p\boldsymbol{\lambda}_d}] \in \mathscr{R}^{(p+q-d)\times(p+q-2d)}$. An example would be $A_{\boldsymbol{\lambda}_d} =$

$\{I_d, -(\mathscr{I}_{3\boldsymbol{\lambda}_d}^{-1}\mathscr{I}_{2\boldsymbol{\lambda}_d}^{\tau})^{\tau}\}^{\tau}$, if we denote $\mathscr{I}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d}) = \begin{bmatrix} \mathscr{I}_{1\boldsymbol{\lambda}_d} & \mathscr{I}_{2\boldsymbol{\lambda}_d} \\ \mathscr{I}_{2\boldsymbol{\lambda}_d}^{\tau} & \mathscr{I}_{3\boldsymbol{\lambda}_d} \end{bmatrix}$ with $\mathscr{I}_{1\boldsymbol{\lambda}_d} \in \mathscr{R}^{d\times d}$, $\mathscr{I}_{2\boldsymbol{\lambda}_d} \in \mathscr{R}^{d\times(p+q-2d)}$, and $\mathscr{I}_{3\boldsymbol{\lambda}_d} \in \mathscr{R}^{(p+q-2d)\times(p+q-2d)}$. Consequently, $\boldsymbol{V}_{\boldsymbol{\lambda}_d}^{\perp} = \{\boldsymbol{h} \in \mathscr{R}^{p+q-d} : \boldsymbol{h} = (\boldsymbol{s}^{\tau}, -(\mathscr{I}_{3\boldsymbol{\lambda}_d}^{-1}\mathscr{I}_{2\boldsymbol{\lambda}_d}^{\tau}\boldsymbol{s})^{\tau})^{\tau}$ for some $\boldsymbol{s} \in \mathscr{R}^{d}\}$.

Next, we can define $\Sigma_{\boldsymbol{\lambda}_d} = A_{\boldsymbol{\lambda}}\{A_{\boldsymbol{\lambda}_d}^{\tau}\mathscr{I}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d})A_{\boldsymbol{\lambda}_d}\}^{-1}A_{\boldsymbol{\lambda}_d}^{\tau} = \begin{bmatrix} \Sigma_{\boldsymbol{\lambda}_d 11} & \Sigma_{\boldsymbol{\lambda}_d 12} \\ \Sigma_{\boldsymbol{\lambda}_d 12}^{\tau} & \Sigma_{\boldsymbol{\lambda}_d 22} \end{bmatrix}$, where $\Sigma_{\boldsymbol{\lambda}_d 11} = (\mathscr{I}_{1\boldsymbol{\lambda}_d} - \mathscr{I}_{2\boldsymbol{\lambda}_d}\mathscr{I}_{3\boldsymbol{\lambda}_d}^{-1}\mathscr{I}_{2\boldsymbol{\lambda}_d}^{\tau})^{-1}$, $\Sigma_{\boldsymbol{\lambda}_d 22} = \mathscr{I}_{3\boldsymbol{\lambda}_d}^{-1}\mathscr{I}_{2\boldsymbol{\lambda}_d}^{\tau}\Sigma_{\boldsymbol{\lambda}_d 11}\mathscr{I}_{2\boldsymbol{\lambda}_d}\mathscr{I}_{3\boldsymbol{\lambda}_d}^{-1}$, and $\Sigma_{\boldsymbol{\lambda}_d 12} = -\Sigma_{\boldsymbol{\lambda}_d 11}\mathscr{I}_{2\boldsymbol{\lambda}_d}\mathscr{I}_{3\boldsymbol{\lambda}_d}^{-1}$.

Note that the d-dimensional covariance matrix of $Q_{\boldsymbol{\lambda}_d}$ is singular when there is unknown parameter under the null, i.e., $p+q-d > d$.

Let $\hat{\boldsymbol{\theta}}_d(\boldsymbol{\lambda}_d)$ be the unrestricted maximum rank likelihood estimator (also maximum partial likelihood estimator) of $\boldsymbol{\theta}_d$ for fixed $\boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d$, i.e., $\hat{\boldsymbol{\theta}}_d(\boldsymbol{\lambda}_d)$ satisfies $\mathscr{L}_{\boldsymbol{\lambda}_d}(\hat{\boldsymbol{\theta}}_d(\boldsymbol{\lambda}_d)) = \max_{\boldsymbol{\theta}_d\in\boldsymbol{\Theta}_d}\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d)$, for any $\boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d$ with probability $\to 1$ under $\boldsymbol{\theta}_{0d}$. Let $\tilde{\boldsymbol{\theta}}_d$ be the restricted maximum rank likelihood estimator (also restricted maximum partial likelihood estimator) of $\boldsymbol{\theta}_d$, i.e., $\tilde{\boldsymbol{\theta}}_d$ does not depend on $\boldsymbol{\lambda}_d$ and satisfies $\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d) = \max_{\boldsymbol{\theta}_d\in\boldsymbol{\Theta}_{0d}}\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d)$ with probability $\to 1$ under $\boldsymbol{\theta}_{0d}$.

For known $\boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d$, the standard LM, Wald, and LR test statistics for testing H_{0d} against H_{ad} are given by

$$\begin{aligned} \mathrm{LM}(\boldsymbol{\lambda}_d) &= \left\{\frac{1}{\sqrt{n}}D\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d)\right\}^{\tau}\mathscr{I}_{\boldsymbol{\lambda}_d}^{-1}(\tilde{\boldsymbol{\theta}}_d)\left\{\frac{1}{\sqrt{n}}D\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d)\right\}, \\ W(\boldsymbol{\lambda}_d) &= \left\{H\sqrt{n}\hat{\boldsymbol{\theta}}_d(\boldsymbol{\lambda}_d)\right\}^{\tau}\left\{H\mathscr{I}_{\boldsymbol{\lambda}_d}^{-1}(\hat{\boldsymbol{\theta}}_d(\boldsymbol{\lambda}_d))H^{\tau}\right\}^{-1}\left\{H\sqrt{n}\hat{\boldsymbol{\theta}}_d(\boldsymbol{\lambda}_d)\right\}, \\ \mathrm{LR}(\boldsymbol{\lambda}_d) &= -2\left\{\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d) - \mathscr{L}_{\boldsymbol{\lambda}_d}(\hat{\boldsymbol{\theta}}_d(\boldsymbol{\lambda}_d))\right\}, \end{aligned} \tag{4}$$

where $H = [I_d : 0] \subset \mathscr{R}^{d\times(p+q-d)}$ and $\mathscr{I}_{\boldsymbol{\lambda}_d} = -D^2\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d)/n$.

The exponential LM test can be defined as

$$\mathrm{Exp}-\mathrm{LM} = (1+c)^{-d/2}\int \exp\left\{\frac{1}{2}\frac{c}{1+c}\mathrm{LM}(\boldsymbol{\lambda}_d)\right\}\mathrm{d}J(\boldsymbol{\lambda}_d),$$

where c is the scalar constant which may depend on the local alternative hypothesis. Exponential Wald (LR) test is just replacing $\mathrm{LM}(\boldsymbol{\lambda}_d)$ by $W(\boldsymbol{\lambda}_d)$ ($\mathrm{LR}(\boldsymbol{\lambda}_d)$).

Now we can summarize the optimal results in Theorems 3 and 4.

Theorem 3 (Asymptotic distribution). *Under the null hypothesis and Assumptions* 1–4, *(a)* $Exp-LM \to_d \chi(\boldsymbol{\theta}_{0d}, c)$, *(b)* $Exp-W \to_d \chi(\boldsymbol{\theta}_{0d}, c)$, *and (c)* $Exp-LR \to_d \chi(\boldsymbol{\theta}_{0d}, c)$, *where*

$$\begin{aligned} \chi(\boldsymbol{\theta}_{0d}, c) = (1+c)^{-d/2} \times \int \exp\Bigg[&\frac{1}{2}\frac{c}{1+c}\left\{H\mathscr{I}_{\boldsymbol{\lambda}_d}^{-1}(\boldsymbol{\theta}_{0d})G_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d})\right\}^{\tau} \\ &\times\left\{H\mathscr{I}_{\boldsymbol{\lambda}_d}^{-1}(\boldsymbol{\theta}_{0d})H^{\tau}\right\}^{-1}\left\{H\mathscr{I}_{\boldsymbol{\lambda}_d}^{-1}(\boldsymbol{\theta}_{0d})G_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d})\right\}\Bigg]dJ(\boldsymbol{\lambda}_d) \end{aligned}$$

and $\{G_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d}) : \boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d\}$ *is a mean zero* $\mathscr{R}^{p+q-d}$*-valued Gaussian process with the covariance* $EG_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d})G_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d})^\tau = \Sigma(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_d)$ *for any* $\boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d$ *and has bounded uniformly continuous sample paths (as functions of* $\boldsymbol{\lambda}_d$ *for fixed* $\boldsymbol{\theta}_{0d}$*) with probability 1.*

Theorem 4 (Optimality). *Under Assumptions* 1–4*, for any sequence of asymptotically level* α *tests based on ranks* $\boldsymbol{r} = (r_1, \ldots, r_n)^\tau$*,* $\{\psi_n : n \geq 1\}$*, a sequence of asymptotically level* α *exponential LM (Wald or LR) tests* $\{\xi_n : n \geq 1\}$ *satisfies*

$$\begin{aligned}&\overline{\lim_{n\to\infty}} \int \left[\int \psi_n l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d} + \boldsymbol{h}/\sqrt{n}) d\mu_n\right] dQ_{\boldsymbol{\lambda}_d}(\boldsymbol{h}) dJ(\boldsymbol{\lambda}_d)\\ &\quad \leq \lim_{n\to\infty} \int \left[\int \xi_n l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d} + \boldsymbol{h}/\sqrt{n}) d\mu_n\right] dQ_{\boldsymbol{\lambda}_d}(\boldsymbol{h}) dJ(\boldsymbol{\lambda}_d),\end{aligned}$$

where a test ψ_n *is a rank-based test of asymptotic significance level* α *if* $\int \psi_n l(\boldsymbol{\theta}_{0d}) d\mu_n \to \alpha$ *for all* $\boldsymbol{\theta}_{0d}$ *that satisfy the null hypothesis* H_{0d}*;* $\int \psi_n l(\boldsymbol{\theta}_{0d}) d\mu_n$ *denotes the probability of rejection of* H_{0d} *using* ψ_n*; the power of* ψ_n *against the local alternative* $l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d} + \boldsymbol{h}/\sqrt{n})$ *is denoted as* $\int \psi_n l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d} + \boldsymbol{h}/\sqrt{n}) d\mu_n$*.*

Remark 1 One may interpret the optimality results in Theorem 4 in two ways. First, it provides a greatest asymptotic weighted average power result for the exponential LM test against the alternatives $\left\{l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d} + \boldsymbol{h}/\sqrt{n}) : \boldsymbol{h} \in \mathscr{R}^{p+q-d}, \boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d\right\}$. Second, it shows that the exponential LM test has the greatest asymptotic power against the single sequence of local alternatives $\left\{\int l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_{0d} + \boldsymbol{h}/\sqrt{n}) dQ_{\boldsymbol{\lambda}_d}(\boldsymbol{h}) \mathrm{d}J(\boldsymbol{\lambda}_d)\right\}$ amongst all rank-based tests of asymptotic level α.

If we replace $c/(1+c)$ by r and replace $1+c$ by 1, when $r \to \infty$, Exp $-$ LR test becomes the partial likelihood ratio test. Since above three conditions cannot hold at the mean time, the partial likelihood ratio test is not optimal in our proposed model.

5 Implementing the optimal tests

Ideas from Hansen (1996, 2000) motivate us to utilize the Gaussian multiplier method to generate the limiting distribution $\chi(\boldsymbol{\theta}_{0d}, c)$. However, it is more difficult here since partial likelihood is not based on independent terms, and we adapt the Gaussian multiplier technique for statistics defined by a counting process martingale used in Lin et al. (1993, 1994).

As in Andrews et al. (1996), we illustrate these issues for $c = 0$, 1, and ∞ with the explicit form of the Exp $-$ LM test statistics being $\int \mathrm{LM}(\boldsymbol{\lambda}_d) \mathrm{d}J(\boldsymbol{\lambda}_d)$, $\int \exp\{\mathrm{LM}(\boldsymbol{\lambda}_d)/4\} \mathrm{d}J(\boldsymbol{\lambda}_d)$, and $\log \int \exp\{\mathrm{LM}(\boldsymbol{\lambda}_d)/2\} \mathrm{d}J(\boldsymbol{\lambda}_d)$, respectively. To get corresponding Exp $-$ W and Exp $-$ LR, we only need to change $\mathrm{LM}(\boldsymbol{\lambda}_d)$ in above three statistics to $W(\boldsymbol{\lambda}_d)$ and $\mathrm{LR}(\boldsymbol{\lambda}_d)$, respectively.

For each fixed $\boldsymbol{\lambda}_d$, (4) gives the usual LM test statistic with $\tilde{\boldsymbol{\theta}}_d$ which does not depend on $\boldsymbol{\lambda}_d$. If denoting $(p+q-d) \times (p+q-d)$ matrix R with upper-left sub-matrix I_d, a robust LM test statistic for fixed $\boldsymbol{\lambda}_d$ equivalent to (4) is

$$\mathrm{LM}_r(\boldsymbol{\lambda}_d) = \left\{R \cdot D\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d)/\sqrt{n}\right\}^\tau \mathscr{I}_{\boldsymbol{\lambda}_d}^{-1}(\tilde{\boldsymbol{\theta}}_d) \left\{R \cdot D\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d)/\sqrt{n}\right\} = \mathrm{LM}(\boldsymbol{\lambda}_d).$$

By counting process setup, if denoting $G_{\boldsymbol{\theta}_d,\boldsymbol{\lambda}_d}(\boldsymbol{Z}_i(s)) = \left[g^{\tau}_{\boldsymbol{\lambda}}(\boldsymbol{Z}_i(s)), \dot{g}^{\tau}_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\boldsymbol{Z}_i(s))\right]^{\tau}$, then

$$D\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d) = U(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, 1) = \sum_{i=1}^{n} \int_0^1 \left\{ G_{\tilde{\boldsymbol{\theta}}_d,\boldsymbol{\lambda}_d}(\boldsymbol{Z}_i(s)) - \frac{S^{(1)}(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, s)}{S^{(0)}(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, s)} \right\} \mathrm{d}M_i(s).$$

For $i = 1, \ldots, n$, since $E\{M_i(s)\} = 0, \mathrm{Var}\{M_i(s)\} = E\{N_i(s)\}$, if denoting $v_i \sim N(0, 1)$ independent with any quantity in $U(\boldsymbol{\theta}_d, \boldsymbol{\lambda}_d, 1)$, replacing $\{M_i(s)\}$ with $\{N_i(s)v_i\}$ can yield

$$D^*\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d) = U^*(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, 1) = \sum_{i=1}^{n} \int_0^1 \left\{ G_{\tilde{\boldsymbol{\theta}}_d,\boldsymbol{\lambda}_d}(\boldsymbol{Z}_i(s)) - \frac{S^{(1)}(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, s)}{S^{(0)}(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, s)} \right\} v_i \mathrm{d}N_i(s).$$

A generated conditional process is then

$$\mathrm{LM}^*_r(\boldsymbol{\lambda}_d) = \left\{ R \cdot D^*\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d)/\sqrt{n} \right\}^{\tau} \mathscr{I}^{-1}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d) \left\{ R \cdot D^*\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d)/\sqrt{n} \right\}.$$

The equivalence of $\mathrm{LM}_r(\boldsymbol{\lambda}_d)$ and $\mathrm{LM}^*_r(\boldsymbol{\lambda}_d)$ under H_{0d} can be shown in Theorem 5. Asymptotic equivalence of Exp − LM* and Exp − LM follows.

Theorem 5 *Under H_{0d}, $LM_r(\boldsymbol{\lambda}_d)$ and $LM^*_r(\boldsymbol{\lambda}_d)$ converge weakly to the same limiting distribution.*

For large sample size n, we can generate J replications of Exp − LM*, and they approximate the distribution of Exp − LM. The details are as follows. If we denote

$$\boldsymbol{D}_{\boldsymbol{\lambda}_d,i}(\boldsymbol{\theta}_d) = \frac{\partial}{\partial \boldsymbol{\theta}_d} \left[\boldsymbol{\beta}^{\tau} g_{\boldsymbol{\lambda}}(z_i) - \log \sum_{j \in \mathscr{R}_i} \exp\{\boldsymbol{\beta}^{\tau} g_{\boldsymbol{\lambda}}(z_j)\} \right]$$

$$\text{and} \quad \boldsymbol{D}^*_{\boldsymbol{\lambda}_d,i}(\boldsymbol{\theta}_d) = \begin{pmatrix} \{\boldsymbol{D}_{\boldsymbol{\lambda}_d,i}(\boldsymbol{\theta}_d)\}_d \\ \mathbf{0} \end{pmatrix},$$

where $\left\{\boldsymbol{D}_{\boldsymbol{\lambda}_d,i}(\boldsymbol{\theta}_d)\right\}_d$ denotes the dimensions of $\boldsymbol{D}_{\boldsymbol{\lambda}_d,i}(\boldsymbol{\theta}_d)$ corresponding to $\beta_1, \ldots, \beta_d$, we have $D\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d) = \sum_{i=1}^{n} \boldsymbol{D}_{\boldsymbol{\lambda}_d,i}(\boldsymbol{\theta}_d)$ and $D\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d) = \sum_{i=1}^{n} \boldsymbol{D}_{\boldsymbol{\lambda}_d,i}(\tilde{\boldsymbol{\theta}}_d) = \sum_{i=1}^{n} \boldsymbol{D}^*_{\boldsymbol{\lambda}_d,i}(\tilde{\boldsymbol{\theta}}_d)$.

Pre-specifying the large number J and noticing that $\boldsymbol{D}_{\boldsymbol{\lambda}_d,i}(\boldsymbol{\theta}_d)$ is $(p + q - d)$ dimensional, the simulation procedure for the limiting distribution follows the four steps below.

1. for $j = 1, \ldots, J$, generate $\{v_{ij}\}_{i=1}^{n}$ iid $N(0, 1)$ random variables;
2. set $D^*_j\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d) = \sum_{i=1}^{n} \boldsymbol{D}^*_{\boldsymbol{\lambda}_d,i}(\tilde{\boldsymbol{\theta}}_d)v_{ij}$ for each fixed $\boldsymbol{\lambda}_d$ in $J(\boldsymbol{\lambda}_d)$, note here only first d dimensions of $D_j\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d)$ are non-zero;
3. set $\mathrm{LM}^*_j(\boldsymbol{\lambda}_d) = \left\{ D^*_j\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d)/\sqrt{n} \right\}^{\tau} \mathscr{I}^{-1}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d) \left\{ D^*_j\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d)/\sqrt{n} \right\}$;
4. set $\mathrm{Exp} - \mathrm{LM}^*_j = (1 + c)^{-d/2} \int \exp\left[c\mathrm{LM}^*_j(\boldsymbol{\lambda}_d)/\{2(1 + c)\}\right] \mathrm{d}J(\boldsymbol{\lambda}_d)$.

This gives a random sample $\text{Exp}-\text{LM}_1^*, \ldots, \text{Exp}-\text{LM}_J^*$ of J observations from the distribution of $\text{Exp}-\text{LM}^*$ conditional on observed data, thus from the distribution of $\text{Exp}-\text{LM}$ asymptotically. When sample size n is large, it is a random sample from the limiting distribution $\chi(\boldsymbol{\theta}_{0d}, c)$. Asymptotic p value is the proportion of $\text{Exp}-\text{LM}_j^*$'s which are larger than the test statistic value $\text{Exp}-\text{LM}$, and critical values can also be tabulated.

6 Simulation studies

Simulations were carried out to check the performance of the estimates and the proposed tests. First, we check the estimates for $\boldsymbol{\theta} = (\boldsymbol{\beta}^\tau, \boldsymbol{\lambda}^\tau)^\tau$ under some specific cases where identifiability holds. For simplicity, suppose $p = 1$, then $\boldsymbol{\beta} = \beta \in \mathscr{R}$, $\boldsymbol{\lambda} = \lambda \in \mathscr{R}$. Also suppose the Box–Cox transformation $g_\lambda(z) = (z^\lambda - 1)/\lambda$ when $\lambda \neq 0$, and $\log(z)$ when $\lambda = 0$, where z is a real-valued covariate with distribution Uniform(0,2). Failure time T was assumed to be exponentially distributed; thus baseline hazard rate $h_0(t)$ was a constant, and we assumed it to be 0.15. The random censoring proportion was selected to be 25%. β and λ would be varied to conduct several scenarios. Now the assumed model is

$$h(t; z) = h_0(t) \exp\{\beta g_\lambda(z)\}, \quad \text{where } h_0(t) = 0.15 \text{ and } g_\lambda(z) = \begin{cases} (z^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log(z), & \lambda = 0 \end{cases}.$$

For $i = 1, \ldots, 500$, we generated covariate $Z_i \sim U(0.01, 2)$. Time T_i was generated from an exponential distribution with parameter $h(t; Z_i)$. Censoring indicator δ_i (0 if censoring) was generated from a Bernoulli distribution with mean 0.75. By this generation method, the distribution for event time X_i is exponential distribution with parameter $0.75h(t; Z_i)$, and the distribution for censoring time C_i is exponential distribution with parameter $0.25h(t; Z_i)$.

Based on the observed data, we can get $(\hat{\beta}, \hat{\lambda})$ by maximizing (3). Applying Theorem 1, estimated variances of $\hat{\beta}$ and $\hat{\lambda}$ can be obtained, together with 95% confidence intervals for β and λ. Repeat this procedure for k times. We can then get k copies of $(\hat{\beta}, \hat{\lambda})$, their estimated variances, and k confidence intervals for β and λ. Comparisons between the empirical variance and the mean estimated variance obtained from Theorem 1 can be conducted. The empirical coverage of the 95% confidence interval may also be computed. Table 1 below gives the results when $k = 1{,}000$. True values β_0 and λ_0 were chosen to be combinations of 1, 2, 3, and -1, -0.5, 0, 0.5, 1.

Table 1 shows that the point estimates of β and λ are close to their true values. For both β and λ, the empirical variances which were calculated from $k = 1{,}000$ $\hat{\beta}$'s are roughly equal to the estimated variances which are the mean of $k = 1{,}000$ variance estimators from Theorem 1. 95% confidence interval coverage ranges from 94 to 96% which also supports the accuracy.

Performances of the optimal tests can be assessed by checking the size of the tests and comparing the power to that of sup test and naive tests. Here, sup test is the supreme LM test (score test); see Liang et al. (1990) for such supreme score test in Cox model with one time structural change covariate. Using the asymptotic equivalence of

Table 1 Simulation studies for β and λ (sample size $n = 500$, 1,000 simulations)

β	λ	$\hat{\beta}$	Emp.$V(\hat{\beta})$	Est.$V(\hat{\beta})$	CI (%)	$\hat{\lambda}$	Emp.$V(\hat{\lambda})$	Est.$V(\hat{\lambda})$	CI (%)
1	−1	1.001	0.0145	0.0146	95.1	−1.011	0.0103	0.0107	95.6
1	−0.5	1.004	0.0142	0.0133	94.2	−0.506	0.0090	0.0082	94.4
1	0	1.006	0.0141	0.0137	94.7	0.002	0.0118	0.0115	95.7
1	0.5	1.002	0.0143	0.0139	94.4	0.514	0.0344	0.0347	95.6
1	1	0.993	0.0121	0.0120	95.3	1.024	0.0837	0.0871	95.7
2	−1	2.005	0.0317	0.0314	95.1	−1.006	0.0113	0.0111	95.6
2	−0.5	2.003	0.0269	0.0263	94.7	−0.506	0.0061	0.0059	94.5
2	0	2.007	0.0237	0.0232	94.1	−0.001	0.0057	0.0056	95.6
2	0.5	2.006	0.0216	0.0209	94.2	0.503	0.0107	0.0108	95.5
2	1	2.001	0.0182	0.0177	94.6	1.006	0.0229	0.0235	95.6
3	−1	3.007	0.0498	0.0495	94.5	−1.004	0.0094	0.0091	94.6
3	−0.5	3.006	0.0434	0.0428	94.6	−0.504	0.0052	0.0052	95.9
3	0	3.008	0.0375	0.0368	94.4	−0.002	0.0043	0.0043	95.7
3	0.5	3.008	0.0330	0.0322	94.1	0.501	0.0064	0.0065	95.6
3	1	3.006	0.0284	0.0276	94.2	1.003	0.0122	0.0127	96.0

$\mathrm{LM}^*(\cdot)$ and $\mathrm{LM}(\cdot)$, we can generate the p value for sup LM test adapting the method of generating the p value for Exp − LM test.

The sample size n was set to be 250. Set $\beta = 0$, using the generation method above, we generated observations $\{(T_i, \delta_i) : i = 1, \ldots, n\}$ under $H_0 : \beta = 0$. For each of the generated $k = 1{,}000$ data sets, we used the Gaussian multiplier method to get a p value with the number of resampling $J = 1{,}000$. The test should have these p values uniformly distributed in [0, 1].

The choice of $J(\lambda)$ may be misspecified. However, this does not affect the validity of the test although it may no longer be optimal. Even when $J(\lambda)$ is misspecified, combining information across λ using a vague prior may protect against misspecification of λ in naive tests which uses fixed transformations. Here, we choose $J(\lambda)$ to be uniform distribution in intervals $[-2, 2]$ and $[-1, 1]$, and the significance level for the tests to be 5%. For each case, the proportion of the p values less than 5% is the rejection rate. They are shown in the first row of Table 2. For 5% tests, the rates range from 4 to 5% which means that the Gaussian multiplier method generates appropriate limiting distribution, and the test is valid under the null.

The upper half of Table 2 also compares powers for the optimal tests, sup LM test, and naive tests. The naive tests were chosen with fixed $\lambda = -2, -1, 0, 1, 2$ and the true value under alternative. Comparing to the naive tests, our proposed tests are more robust. With a moderately assumed distribution of λ, the proposed test can achieve good power which improves on sup test or naive tests with greatly misspecified λ and comparable power to the test with true λ. When $\beta = 0.1$, $\lambda = -0.5$, with moderately wide distribution of λ, say $[-2, 2]$ or $[-1, 1]$, the proposed tests give power greater than 70%. Sup test gives power 63%, less than the optimal tests. For naive tests, the power is 80% when true λ is used, and if a mis-specified λ value is used, the power

Table 2 Comparisons of the type I errors and powers (percentages of rejection rates for Exp − LM 5% test, sample size $n = 250$, 1,000 simulations)

	$J(\lambda_d) = U[-2, 2]$			$J(\lambda_d) = U[-1, 1]$			Sup	Naive tests with fixed $\lambda =$					
(β, λ)	$c = 0$	1	∞	$c = 0$	1	∞	Test	-2	-1	0	1	2	True
Univariate analysis													
$(0, \cdot)$	5	4	4	5	5	5							
$(0.002, -2)$	77	77	77	73	74	74	75	89	87	65	23	13	89
$(0.05, -1)$	92	92	92	91	91	91	89	94	95	89	58	37	95
$(0.1, -0.5)$	72	71	71	75	74	74	63	64	75	74	52	35	80
$(0.2, 0)$	69	68	67	72	72	72	56	40	56	73	63	52	73
$(0.3, 0.5)$	69	68	67	72	72	72	56	28	41	72	73	66	76
$(0.4, 1)$	78	77	77	77	79	80	67	24	35	76	85	82	85
$(0.5, 2)$	88	89	89	84	87	89	85	21	30	80	96	97	97
Multivariate analysis: true $\beta_2 = 1$													
$(0, \cdot)$	4	4	4	5	4	4							
$(0.002, -2)$	78	78	79	74	76	76	76	88	87	69	27	14	88
$(0.05, -1)$	93	93	92	92	92	92	91	94	95	91	63	38	95
$(0.1, -0.5)$	74	73	73	76	76	76	66	66	76	75	49	33	79
$(0.2, 0)$	70	68	67	73	73	73	54	44	58	74	63	49	74
$(0.3, 0.5)$	70	69	68	73	73	73	56	31	44	72	73	64	76
$(0.4, 1)$	78	77	77	78	79	80	66	27	38	77	85	82	85
$(0.5, 2)$	88	88	88	84	87	88	84	23	32	81	95	97	97

may be as low as only 35% ($\lambda = 2$, square transformation). When no transformation is implemented ($\lambda = 1$), the power is only 52%. For other combination of β, λ, the advantage of optimal tests can also be observed.

Another simulation example was carried out for the case when nuisance parameter presents and we need to estimate it under the null. The model is selected to be

$$h(t; z_1, z_2) = h_0(t) \exp(\beta_1 z_1^{\lambda} + \beta_2 z_2), \quad \text{where} \quad h_0(t) = 0.15,$$

with $p = 2, q = d = 1$, and the test is $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ and λ exists.

The censoring proportion was again assumed to be 25%, and the technique to generate censoring was the same with previous scenario. After setting sample size $n = 250$, we generated the observed data $\{(T_i, \delta_i) : i = 1, \ldots, 250\}$ by considering $h(t; Z_{1i}, Z_{2i}) = h_0(t) \exp(\beta_{10} Z_{1i}^{\lambda_0} + \beta_{20} Z_{2i})$ and $Z_{1i}, Z_{2i} \sim U(0.01, 2)$ for true values of β_{10}, β_{20}, and λ_0.

We chose true values of β_2 to be 1. Under null, we generated $k = 1{,}000$ data sets. Similar to the previous scenario, the rejection rates can be shown in the first row of the lower half of Table 2. Similar to the upper half of Table 2, the lower half also gives the power of the 5% Exp − LM test for selected H_a. In terms of both rejection rate and power, similar improvements of the optimal tests can be observed.

7 Application to breast cancer data

These data come from a National Surgical Adjuvant Breast and Bowel Project (NSABP) study (B-20) designed to test, in women with estrogen receptor (ER)-positive breast cancer and histologically negative axillary lymph nodes, whether the addition of chemotherapy to tamoxifen would result in a greater benefit than that achieved with tamoxifen alone. In this study, a total of 2,363 patients were randomized over a period of 5 years to tamoxifen (TAM), MFT [tamoxifen(T) plus sequential methotrexate(M) and fluorouracil(F)], or CMFT [MFT plus a chemotherapy regimen containing the alkylating agent cyclophosphamide(C)], and have been followed for an additional 10 years. One of the primary endpoints of interest was disease-free survival (DFS). For this endpoint, a patient is considered to have an event when she recurs, has a second primary cancer, or dies (whichever occurs first). Fisher et al. (1997) reported that the administration of chemotherapy with tamoxifen resulted in a significantly better DFS than that achieved with tamoxifen alone. Only 2,183 eligible patients with positive follow-up and known tumor size and hormonal status information (735 patients in the tamoxifen group, 717 patients in the MFT group, and 731 patients in the CMFT group) are used for statistical analysis.

We focus on developing risk indices for DFS, using other non-treatment covariates. Prognostic factors on DFS include treatment, age, clinical tumor size (CTSIZE), pathological tumor size (PTSIZE), and progesterone receptor level (PgR). The proposed model is used to assess the effects of these risk factors both individually and jointly, with the results compared to those from a traditional Cox model without covariate transformations. We first examine the effects of individual factors separately, without adjusting for other covariates. The Box–Cox transformation is chosen as the nonlinear transformation. We focus on transformations for biological prognostic factors CTSIZE, PTSIZE and PgR.

For each covariate Z with transformation (Z can be CTSIZE, PTSIZE, or PgR), the model is

$$h(t; Z) = h_0(t) \exp\left\{\beta \times g_\lambda(Z)\right\} \quad \text{where } g_\lambda(z) = \begin{cases} (z^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log(z), & \lambda = 0 \end{cases}.$$

To get estimates of the parameters β and λ, we need to decide if the covariates have effects or not; thus, we test $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ and there exists λ.

When conducting naive tests (the upper portion of Table 3), as suggested by Atkinson (1987), we use $\lambda = -2, -1, -0.5, 0, 0.5, 1$, and 2. We choose the prior distribution of λ as uniform distribution in $[-2, 2]$, $[-1.5, 1.5]$, and $[-1, 1]$ and then Exp − LM test gives test statistic values as shown in the upper portion of Table 4.

When using naive tests with fixed transformation parameter λ, for PTSIZE, we can still reject the null, but for CTSIZE, log transformation or negative-valued transformation parameters give insignificant effect, with traditional cutoff point 0.05. For PgR, $\lambda = 1$ (no transformation) or <0 result in insignificant effect. However, in the proposed tests, all p values for testing effects of PTSIZE are less than 0.05, under $\alpha = 0.05$, we should reject H_0 of no effects of PTSIZE. Similar results can be obtained for CTSIZE and PgR with most p values <0.05.

Table 3 Breast cancer data: p values of naive tests

λ	-2	-1	-0.5	0	0.5	1	2
Univariate analysis $H_0 : \beta = 0$							
CTSIZE	0.970	0.873	0.553	0.117	0.011	0.003	0.004
PTSIZE	0.154	0.001	<0.001	<0.001	<0.001	<0.001	<0.001
PgR	0.485	0.361	0.142	0.009	0.004	0.084	0.984
Multivariate analysis $H_0 : \beta_4 = 0$							
CTSIZE	0.942	0.934	0.566	0.096	0.006	0.001	0.002
PTSIZE	0.111	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
PgR	0.649	0.510	0.232	0.014	0.006	0.082	0.847

Table 4 Breast cancer data: p values of optimal tests

	CTSIZE			PTSIZE			PgR		
$J(\boldsymbol{\lambda}_d)$	$c = 0$	1	∞	$c = 0$	1	∞	$c = 0$	1	∞
Univariate analysis $H_0 : \beta = 0$									
$U[-2, 2]$	0.03	0.01	0.01	0.00	0.00	0.00	0.06	0.04	0.03
$U[-1.5, 1.5]$	0.04	0.02	0.01	0.00	0.00	0.00	0.03	0.02	0.02
$U[-1, 1]$	0.06	0.03	0.02	0.00	0.00	0.00	0.01	0.01	0.01
Multivariate analysis $H_0 : \beta_4 = 0$									
$U[-2, 2]$	0.02	0.01	0.01	0.00	0.00	0.00	0.08	0.05	0.03
$U[-1.5, 1.5]$	0.03	0.01	0.01	0.00	0.00	0.00	0.04	0.03	0.02
$U[-1, 1]$	0.04	0.01	0.01	0.00	0.00	0.00	0.02	0.02	0.01

Next, we fit univariate models to estimate the covariate effects. To make the magnitude of the effects comparable across covariates, we divide the covariate by 10^t where 10^t is greater than the largest value of corresponding covariate, but 10^{t-1} is less than it. The upper portion of Table 5 gives the used covariates, their estimates and standard errors of their coefficients, and corresponding p values (Treatment and Age are omitted here). Note that the coefficients of the transformed covariates (CTSIZE, PTSIZE, or PgR) cannot be tested by simply using $\hat{\beta}/SE_{\hat{\beta}}$.

The p value for the transformation of PgR is less than 0.01, and for that of PTSIZE is 0.10, a suggestive but not significant p value. For CTSIZE, p value 0.73 means that the transformation to CTSIZE is unnecessary.

We now fit models for the effects of CTSIZE, PTSIZE, and PgR, respectively, with simultaneous adjustments for treatment and age. For each covariate Z_4 (Z_4 can be CTSIZE, PTSIZE, or PgR), the model is

$$h(t; \mathbf{Z}) = h_0(t) \exp\left\{\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 \times \text{Age} + \beta_4 \times g_\lambda(Z_4)\right\}$$

where Z_1 and Z_2 are dummy variables: $Z_1 = 1$ if treatment $=$MFT, 0 otherwise; $Z_2 = 1$ if treatment$=$CMFT, 0 otherwise, and $g_\lambda(z) = (z^\lambda - 1)/\lambda$ when $\lambda \neq 0$ and

Table 5 Breast cancer data: parameter estimates and standard errors

Covariate	$\hat{\beta}$	$SE_{\hat{\beta}}$	p value	$\hat{\lambda}$	$SE_{\hat{\lambda}}$	p value ($H_0 : \lambda = 1$)
Univariate analysis						
CTSIZE/10^2	1.23	0.97	–	1.19	0.55	0.73
PTSIZE/10^3	0.96	1.69	–	0.26	0.45	0.10
PgR/10^4	−0.39	0.35	–	0.38	0.17	<0.01
Multivariate analysis						
CTSIZE/10^2	1.34	0.95	–	1.19	0.50	0.70
PTSIZE/10^3	1.01	1.67	–	0.26	0.43	0.09
PgR/10^4	−0.42	0.40	–	0.41	0.18	<0.01

Table 6 AIC Values for the univariate models (smaller is better, real AIC values should be the values in this table plus 10,600)

	Proposed	Naive models with fixed transformation λ							Fisher et al.
		−2	−1	−0.5	0	0.5	1	2	
CTSIZE	47.79	56.21	56.19	55.85	53.62	49.74	47.92	49.19	48.33
PTSIZE	35.43	51.15	42.62	38.28	35.82	35.74	38.06	47.15	37.90
PgR	47.29	55.74	55.40	54.13	49.14	47.64	52.89	56.22	53.44

$\log(z)$when $\lambda = 0$. We still use the Gaussian multiplier method to test if the effect of PgR (CTSIZE, PTSIZE) is significant, as shown in the lower portion of Table 4. Comparing to p values from the naive tests in the lower portion of Table 3, the optimal tests suggest rejecting H_0 for model with CTSIZE, with PTSIZE or with PgR.

As shown in the lower portion of Table 5, the estimate for the transformation of PgR is 0.41, with standard error 0.18. This means that when there exist effects by PgR, the power transformation is significant (reject null hypothesis $\lambda = 1$). But for CTSIZE or PTSIZE, the test for null hypothesis $\lambda = 1$ has p value 0.70, 0.09, respectively; thus, it seems unnecessary transformation to CTSIZE and suggestive but not significant transformation to PTSIZE.

As a summary, we conclude that the effects of PgR is non-linear to DFS, either univariate or after adjustment for treatment and age. Thus, traditional Cox model may cause lack-of-fit.

AIC based on partial likelihood and estimated explained variation can be used to assess the proposed model and traditional Cox model. Table 6 gives AIC values for the proposed model and traditional models in univariate analysis, for both Cox model with continuous CTSIZE, PTSIZE, or PgR values and Cox model with dichotomized CTSIZE ($\leq$20, >20), PTSIZE ($\leq$20, >20), or PgR (<10, $\geq$10) values, as did in Fisher et al. (1997). Transformation implemented to PgR decreases AIC value to 10,647.29, while the model without covariate transformation has AIC 10,652.89 if considering PgR as continuous value and 10,653.44 if as dichotomized value.

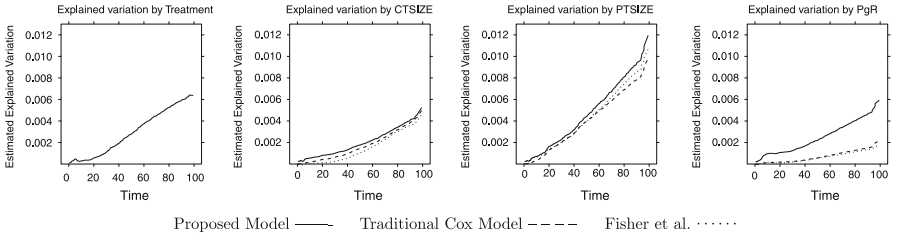

Fig. 1 Estimated explained variation. Compare the proposed model and traditional Cox models, univariate models (unit of time: month). Explained variations by treatment, CTSIZE, PTSIZE, and PgR are comparable. For CTSIZE or PTSIZE, the proposed model can only improve a little comparing to the traditional Cox model with continuous or dichotomized covariate values. For PgR, the proposed model leads to twofold improvement in terms of the explained variation. *Solid line* proposed model, *dashed line* Cox model with continuous covariate values, *dotted line* Cox model with dichotomized covariate values

Estimated explained variation (Graf and Schumacher 1995) can be used to assess the improvement in prediction. It is defined as a function of time t. Suppose the Kaplan–Meier estimate for the survival curve is $\hat{S}_0(t)$ and for a patient with covariate $\boldsymbol{Z}_i$, model-based estimate of the survival curve is $\hat{S}(t|\boldsymbol{Z}_i)$, then we can define

$$\text{estimated explained variation}(t) = 1 - \frac{(1/n)\sum_{i=1}^{n} \hat{S}(t|\boldsymbol{Z}_i)\{1 - \hat{S}(t|\boldsymbol{Z}_i)\}}{\hat{S}_0(t)\{1 - \hat{S}_0(t)\}},$$

and it is a criterion comparable to R^2 and larger value means better prediction.

Figure 1 shows comparisons between the proposed model and traditional models. The estimated explained variation by treatment, CTSIZE, PTSIZE, and PgR are plotted with the same scale; thus, we can conclude that the explained variation by them are comparable. For CTSIZE and PTSIZE, the proposed model can lead to a small improvement in terms of estimated explained variation, comparing to Cox model with continuous or dichotomized covariate values. Interestingly, for PGR, the proposed model leads to a substantial improvement in explained variation, with a roughly twofold improvement in prediction.

8 Discussion

The original motivation for this work is to recover the log-linear relationship between independent covariates and the underlying hazard function in the Cox model. Allowing covariate transformations works well toward this goal. This model can be used to develop the risk indices or to better control the confounding variables when a treatment effect is investigated. For the transformed covariates, inference and the corresponding interpretations can be naturally carried out on the transformed scale. The proposed model performs well in estimating parameters, and the proposed tests for the null of no effects of transformed covariates are more powerful than partial likelihood ratio tests with fixed transformation. Even though in current analysis, the optimal tests function when only time-independent covariates are involved in the model, it should be valid in general, but the proof of optimality is unclear without likelihood connections. More

broad usage of the proposed model needs us to extend the optimal tests to also allow time-dependent covariates.

Appendix: Proofs of Theorems 3–5

Proof of Theorem 3 and 4 Following Theorems 1 and 2 in Andrews and Ploberger (1994), we only need to verify their Assumptions 1–5 with $B = \sqrt{n}I$. Since the rank likelihood is just the partial likelihood, we will use the counting process set-up for partial likelihood to verify these assumptions. In this proof, even though the covariates are written as $\boldsymbol{Z}_1(t), \ldots, \boldsymbol{Z}_n(t)$, they are only allowed to be time independent for the equality of partial likelihood and rank likelihood. The reason using these notations is to keep consistent notations with previous sections.

Assumption 1 (a), (b), (d), (e), and (f) are obvious.

(c): By Conditions B, $\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d) = \log l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d) = \log L(\boldsymbol{\theta})$ is twice continuously differentiable in $\boldsymbol{\theta}$; thus, $l_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d) = L(\boldsymbol{\theta})$ whose derivatives are continuous function of $\mathscr{L}_{\boldsymbol{\lambda}_d}(\boldsymbol{\theta}_d)$ and $L(\boldsymbol{\theta})$ is also twice continuously differentiable in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta}_d \in \boldsymbol{\theta}_{0d}$ and $\boldsymbol{\lambda}_d \in \boldsymbol{\Lambda}_d$ under H_{0d}.

Assumption 2 Denote $\hat{\boldsymbol{\theta}}_d(\lambda_d)$ as the solution of $U(\boldsymbol{\theta}_d, \boldsymbol{\lambda}_d, 1) = 0$, where $U(\boldsymbol{\theta}_d, \boldsymbol{\lambda}_d, 1) =$ dimensions of $U(\boldsymbol{\theta}, 1)$ corresponding to $\boldsymbol{\theta}_d$. Theorem 1 gives the consistency of $\hat{\boldsymbol{\theta}}_d(\boldsymbol{\lambda}_d)$ to $\boldsymbol{\theta}_{0d}$ for fixed $\boldsymbol{\lambda}_d$ under H_{0d}. Since $U(\boldsymbol{\theta}_d, \boldsymbol{\lambda}_d, 1)$ is continuously differentiable in $\boldsymbol{\Theta}_{0d} \times \boldsymbol{\Lambda}_d$, by Implicit Function Theorem, $\hat{\boldsymbol{\theta}}_d(\boldsymbol{\lambda}_d)$ is continuously differentiable in $\boldsymbol{\lambda}_d$. Since $\boldsymbol{\Lambda}_d$ is a closed set, $\hat{\boldsymbol{\theta}}_d'(\boldsymbol{\lambda}_d)$ is bounded. Thus, $\hat{\boldsymbol{\theta}}_d(\boldsymbol{\lambda}_d) \to_p \boldsymbol{\theta}_{0d}$ uniformly under $\boldsymbol{\theta}_{0d}$.

Assumptions 3 and 4 are obvious.

Assumption 5 we need to check $n^{-1/2}U(\boldsymbol{\theta}_{0d}, \cdot, 1) \to_d G(\boldsymbol{\theta}_{0d}, \cdot)$ under $\boldsymbol{\theta}_{0d}$.

To establish the weak convergence, we need to prove the finite dimensional convergence and the tightness. Proving for the finite dimensional convergence is similar to the first part of that of Theorem 1 with proving for any $\boldsymbol{\lambda}_{d1}, \ldots, \boldsymbol{\lambda}_{dk} \in \boldsymbol{\Lambda}_d$, $n^{-1/2}[U(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_{d1}, 1), \ldots, U(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_{dk}, 1)]^\tau \to_d N(0, \Sigma^*)$, and let $H_{il}(t) = n^{-1/2}\{G^\tau_{\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_{d1}}(\boldsymbol{Z}_l(s)), \ldots, G^\tau_{\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_{dk}}(\boldsymbol{Z}_l(s))\}^\tau_i$ when we denote $U(\boldsymbol{\theta}_d, \boldsymbol{\lambda}_d, t) = \sum_{i=1}^n \int_0^t G_{\boldsymbol{\theta}_d, \boldsymbol{\lambda}_d}(\boldsymbol{Z}_i(s))\mathrm{d}M_i(s)$. Tightness can be verified by checking a tightness criterion in Billingsley (1968, p. 95). □

Proof of Theorem 5. We only need to show that under H_{0d}, $R \cdot D\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d)/\sqrt{n} = R \cdot n^{-1/2}U(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, 1)$ and $R \cdot D^*\mathscr{L}_{\boldsymbol{\lambda}_d}(\tilde{\boldsymbol{\theta}}_d)/\sqrt{n} = R \cdot n^{-1/2}U^*(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, 1)$ converge weakly to the same mean zero Gaussian process indexed by $\boldsymbol{\lambda}_d$.

Under H_{0d}, $\tilde{\boldsymbol{\theta}}_d$ is a strongly consistent estimator of $\boldsymbol{\theta}_{0d}$, thus $n^{-1/2}U(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, 1)$ and $n^{-1/2}U(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_d, 1)$ have the same limiting distribution. By checking Assumption 5 of Theorem 4, the limiting distribution for $n^{-1/2}U(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_d, 1)$ is a mean zero Gaussian process. If we denote $\boldsymbol{H}_{\boldsymbol{\lambda}_d, i}(s) = G_{\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_{dj}}(\boldsymbol{Z}_i(s)) - S^{(1)}(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_d, s)/S^{(0)}(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_d, s)$ which is a predictable process and locally bounded, then for $\boldsymbol{\lambda}_{dj} \neq \boldsymbol{\lambda}_{dk}$, by

Theorem II.3.1 in Andersen et al. (1993), an unbiased estimate of the covariance between $\int_0^1 \boldsymbol{H}_{\boldsymbol{\lambda}_{dj},i}(s)\mathrm{d}M_i(s)$ and $\int_0^1 \boldsymbol{H}_{\boldsymbol{\lambda}_{dk},i'}(s)\mathrm{d}M_{i'}(s)$ is

$$\left[\int_0^1 \boldsymbol{H}_{\boldsymbol{\lambda}_{dj},i}(s)\mathrm{d}M_i(s), \int_0^1 \boldsymbol{H}_{\boldsymbol{\lambda}_{dk},i'}(s)\mathrm{d}M_{i'}(s)\right] = \int_0^1 \boldsymbol{H}_{\boldsymbol{\lambda}_{dj},i}(s)\boldsymbol{H}^{\tau}_{\boldsymbol{\lambda}_{dk},i'}(s)d[M_i, M_{i'}](s),$$

where $d[M_i, M_{i'}](s)$ is optional covariation process of martingale M_i and $M_{i'}$, and $[M_i, M_i](s) = N_i(s)$ by (5.24) in Section 5.3 of Kalbfleisch and Prentice (2002), and $[M_i, M_{i'}](s) = 0$ when $i \neq i'$ because they are orthogonal. Then it is easy to see that an unbiased estimate of the covariance between $n^{-1/2}U(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_{dj}, 1) = n^{-1/2}\sum_{i=1}^n \int_0^1 \boldsymbol{H}_{\boldsymbol{\lambda}_{dj},i}(s)\mathrm{d}M_i(s)$ and $n^{-1/2}U(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_{dk}, 1) = n^{-1/2}\sum_{i=1}^n \int_0^1 \boldsymbol{H}_{\boldsymbol{\lambda}_{dk},i}(s)\mathrm{d}M_i(s)$ is

$$\begin{aligned}
\hat{\sigma}_{jk} &= n^{-1}\sum_{i=1}^n \int_0^1 \boldsymbol{H}_{\boldsymbol{\lambda}_{dj},i}(s)\boldsymbol{H}^{\tau}_{\boldsymbol{\lambda}_{dk},i}(s)d[M_i, M_i](s) \\
&= n^{-1}\sum_{i=1}^n \int_0^1 \boldsymbol{H}_{\boldsymbol{\lambda}_{dj},i}(s)\boldsymbol{H}^{\tau}_{\boldsymbol{\lambda}_{dk},i}(s)\mathrm{d}N_i(s) \\
&= n^{-1}\sum_{i=1}^n \left\{G_{\boldsymbol{\theta}_{0d},\boldsymbol{\lambda}_{dj}}(\boldsymbol{Z}_i(T_i)) - \frac{S^{(1)}(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_d, T_i)}{S^{(0)}(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_d, T_i)}\right\} \\
&\quad \times \left\{G_{\boldsymbol{\theta}_{0d},\boldsymbol{\lambda}_{dj}}(\boldsymbol{Z}_i(T_i)) - \frac{S^{(1)}(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_d, T_i)}{S^{(0)}(\boldsymbol{\theta}_{0d}, \boldsymbol{\lambda}_d, T_i)}\right\}^{\tau} \delta_i.
\end{aligned}$$

The conditional process $n^{-1/2}U^*(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, 1)$ consists of a sum of n independent random variables at each fixed $\boldsymbol{\lambda}_d$. If we regard $\{v_i\}$ as random and $\{T_i, \delta_i, \boldsymbol{Z}_i(\cdot)\}$ as fixed in $U^*(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, 1)$, it can be shown to converge weakly to a mean zero Gaussian process by applying the Lindeberg-Feller theorem and by verifying a tightness criterion (Billingsley 1968, p. 95). Furthermore, the covariance matrix when $\boldsymbol{\lambda}_{dj} \neq \boldsymbol{\lambda}_{dk}$ is

$$\begin{aligned}
\sigma^*_{jk} &= E[n^{-1/2}U^*(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_{dj}, 1) \cdot n^{-1/2}U^*(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_{dk}, 1)|\{\boldsymbol{T}_i, \delta_i, \boldsymbol{Z}_i(\cdot)\}] \\
&= n^{-1}\sum_{i=1}^n \left\{G_{\tilde{\boldsymbol{\theta}}_d,\boldsymbol{\lambda}_{dj}}(\boldsymbol{Z}_i(T_i)) - \frac{S^{(1)}(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_{dj}, T_i)}{S^{(0)}(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_{dj}, T_i)}\right\} \\
&\quad \times \left\{G_{\tilde{\boldsymbol{\theta}}_d,\boldsymbol{\lambda}_{dk}}(\boldsymbol{Z}_i(T_i)) - \frac{S^{(1)}(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_{dk}, T_i)}{S^{(0)}(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_{dk}, T_i)}\right\}^{\tau} \delta_i.
\end{aligned}$$

When $n \to \infty$, $\tilde{\boldsymbol{\theta}}_d$ is strongly consistent with $\boldsymbol{\theta}_{d0}$, thus σ^*_{jk} and $\hat{\sigma}_{jk}$ have the same unconditional limit. This shows that $n^{-1/2}U(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, 1)$ and $n^{-1/2}U^*(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, 1)$ have the same limiting distribution along all possible sample paths. Since $\mathrm{LM}_r(\boldsymbol{\lambda}_d)$ and $\mathrm{LM}^*_r(\boldsymbol{\lambda}_d)$ are generated from $n^{-1/2}U(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, 1)$ and $n^{-1/2}U^*(\tilde{\boldsymbol{\theta}}_d, \boldsymbol{\lambda}_d, 1)$ by the same function, respectively, they have the same limiting distribution. □

References

Andersen, P. K., Borgan, O., Gill, R. D., Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer.

Andersen, P. K., Gill, R. D. (1982). Cox's regression model for counting process: a large sample study. *Annals of Statistics, 10*, 1100–1120.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica, 61*, 821–856.

Andrews, D. W. K., Lee, I., Ploberger, W. (1996). Optimal changepoint tests for normal linear regressions. *Journal of Econometrics, 70*, 9–38.

Andrews, D. W. K., Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica, 62*, 1384–1414.

Atkinson, A. C. (1986). Diagnostic tests for transformations. *Technometrics, 28*, 29–38.

Atkinson, A. C. (1987). *Plots, transformations, and regressions: an introduction to graphical methods of diagnostic regression analysis*. New York: Oxford University Press.

Atkinson, A. C. (1988). Transformation unmasked. *Technometrics, 30*, 311–318.

Billingsley, P. (1968). *Convergence of probability measures*. New York: Wiley.

Box, G. E. P., Cox, D. R. (1964). An analysis of transformation (with discussion). *Journal of the Royal Statistical Society: Series B, 26*, 211–252.

Box, G. E. P., Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics, 4*, 531–550.

Breslow, N. (1972). Contribution to the discussion of the paper by d.r.cox. *Journal of the Royal Statistical Society: Series B, 34*, 187–220.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics, 30*, 89–99.

Carroll, R. J., Ruppert, D. (1988). *Transformation and weighting in regression*. New York: Chapman and Hall.

Cheng, S. C., Wei, L. J., Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika, 82*, 835–845.

Cook, R. D., Wang, P. C. (1983). Transformations and influential cases in regression. *Technometrics, 25*, 337–343.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B, 34*, 187–220.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika, 64*, 247–254.

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

Doksum, K. A. (1987). An extension of partial likelihood methods for proportional hazard model to general transformation models. *Annals of Statistics, 15*, 325–345.

Draper, N. R., Cox, D. R. (1969). On distributions and their transformations to normality. *Journal of the Royal Statistical Society: Series B, 31*, 472–476.

Draper, N. R., Smith, H. (1981). *Applied regression analysis*. John Wiley & Sons, New York, 2Edn.

Fine, J. P., Ying, Z., Wei, L. J. (1998). On the linear transformation model for censored data. *Biometrika, 85*, 980–986.

Fisher, B., Dignam, J., Wolmark, N., DeCillis, A., Emir, B., Wickerham, D. L., Bryant, J., Dimitrov, N. V., Abramson, N., Atkins, J. N., Shibata, H., Deschenes, L., Margolese, R. G. (1997). Tamoxifen and chemotherapy for lymph node-negative, estrogen receptor-positive breast cancer. *Journal of the National Cancer Institute, 89*, 1673–1682.

Graf, E., Schumacher, M. (1995). An investigation on measures of explained variation in survival analysis. *The Statistician: Journal of the Institute of Statisticians, 44*, 497–507.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica, 64*, 413–430.

Hansen, B. E. (2000). Testing for structural change in conditional models. *Journal of Econometrics, 97*, 93–115.

Hastie, T., Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*, 297–310.

Hastie, T., Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics, 46*, 1005–1016.

Kalbfleisch, J. D., Prentice, R. L. (2002). *The statistical analysis of failure time data*. New York: Wiley.

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

Lin, D. Y., Fleming, T. R., Wei, L. J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika, 81*, 73–81.

Lin, D. Y., Wei, L. J., Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika, 80*, 557–572.

Lindsey, J. K. (1972). Fitting response surfaces with power transformations. *Applied Statistics, 21*, 234–247.

Neter, J., Wasserman, W., Kutner, M. H. (1985). *Applied linear statistical models: regression, analysis of variance, and experimental designs* (2nd Ed.). Homewood: Richard D. Irwin Inc.

Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *Journal of the Royal Statistical Society, Series B: Statistical Methodology, 44*, 234–243.

Pettitt, A. N. (1983). Approximate methods using ranks for regression with censored data. *Biometrika, 70*, 121–132.

Pettitt, A. N. (1984). Proportional odds models for survival data and estimates using ranks. *Applied Statistics, 33*, 169–175.

Wang, N., Ruppert, D. (1995). Nonparametric estimation of the transformation in the transform-both-sides regression model. *Journal of the American Statistical Association, 90*, 522–534.