 Vol. 23: 157–177, 1996

Model Checking Techniques for Parametric Regression with Censored Data

D. Y. LIN and C. F. SPIEKERMAN

University of Washington

ABSTRACT. We develop a broad class of graphical and numerical methods for checking individual components of parametric survival models as well as omnibus tests for assessing the overall goodness of fit. The general goodness-of-fit process for testing the distributional assumption is the difference between the parametric maximum likelihood estimator and the Aalen–Breslow type estimator for the baseline survival function. We approximate the null distribution of this process with a zero-mean Gaussian process whose distribution can be easily generated through simulation, which enables us to compute the *P*-value for the supremum test. In order to examine the deterministic model components and to construct global omnibus tests, we study cumulative sums of martingale-based residuals over the failure time or/and covariates. Under the assumed model, the distributions of these partial-sum processes can again be approximated through simulating certain zero-mean Gaussian processes. One may then plot each observed process along with a few realizations from the corresponding Gaussian process to assess visually how unusual the observed residual pattern is. Supremum tests may also be performed. Extensive Monte Carlo studies demonstrate that the proposed test have proper sizes and are highly sensitive to model mis-specification. Applications to two well-known real data sets yield some new insights.

Key words: failure time data, goodness of fit, martingale residual, model mis-specification, omnibus test, regression diagnostic, residual plot, survival analysis

1. Introduction

The Cox (1972) proportional hazards model has been commonly used to study the effects of (possibly time-varying) covariates on the failure rate when the failure time is subject to censoring. This model is semi-parametric in that the dependence of the failure time on covariates is represented by a limited number of unknown regression parameters while the baseline distribution is unspecified. In contrast, a (fully) parametric regression model assumes that the baseline distribution belongs to a given parametric family. The Cox model is analysed by the partial likelihood principle (Cox, 1975) whereas the analysis of parametric models is based on the ordinary likelihood theory. Both the Cox model and parametric regression models were carefully treated in standard survival analysis texts such as Kalbfleisch & Prentice (1980), Lawless (1982), Cox & Oakes (1984) and Andersen *et al.* (1993).

Although overshadowed by the fashionable Cox model, parametric regression models are valuable tools in survival analysis for several reasons. First, precise specification of the baseline distribution leads to more efficient estimation of regression parameters and related quantities. The partial likelihood will have low efficiency relative to the full likelihood if regression parameters are far from zero, if the censoring is strongly dependent on covariates or if there are strong time trends in covariates (Kablfleisch & Prentice, 1980, p. 141; Cox & Oakes, 1984, p. 123). Secondly, parametric analysis tends to be simpler than semi-parametric analysis, especially for non-proportional hazards model (e.g. the accelerated failure time model) and for non-standard setups (e.g. interval censoring). Thirdly, parametric formulation of the baseline distribution provides the basis for more concise summarization of the data and enhances our understanding of the underlying failure mechanism. In some applications, it is of primary interest to ascertain the form of the failure time distribution.

In order for parametric survival models to play a prominent role in applications, it is imperative to develop goodness-of-fit techniques for examining the distributional assumption. There is now a very rich literature on testing the parametric distribution of an homogeneous population in the presence of right censoring, the most notable recent contributions being Pearson-type chi-squared tests due to Habib & Thomas (1986), Akritas (1988), Hjort (1990), Hollander & Peña (1992) and Li & Doss (1993). In contrast, very few procedures are available for the regression setting. Gray & Pierce (1985) embedded the parametric model in a larger parametric family and used score tests. Naturally, such tests are only sensitive to specific alternatives. To our knowledge, the only test against a general alternative is the chi-squared test for the baseline distribution of the proportional hazards model proposed by Hjort (1990), which is based on the comparison of a Breslow-type estimator and the parametric estimator for the cumulative baseline hazard function. There are two widely recognized limitations with the chi-squared test. First, the test requires arbitrary partition of the time axis. The rather unpleasant scenario may arise where different partitions lead to conflicting results. Secondly, there is a clear loss of information (relative to a supremum test) since the chi-squared test compares the "observed" and "expected" numbers over broad intervals rather than at individual time points. The main reason for resorting to the chi-squared test (or score test) was that natural goodness-of-fit processes such as that of Hjort's do not have independent increments asymptotically (except in the one-sample case with a one-dimensional parameter), which makes the limiting distribution of the Kolmogorov–Smirnov-type test analytically intractable.

In this article, we offer a new solution to the general problem of checking the distributional form of the parametric regression model under right censorship. Our idea is to approximate the null distribution of the difference between the parametric estimator and the Aalen–Breslow type estimator for the baseline distribution function by some zero-mean Gaussian process whose distribution can be easily generated through simulation. This approach enables us to calculate the P-value for the supremum test via simulation, avoiding the formidable task of a formulaic evaluation.

We shall also develop a class of numerical and graphical techniques for examining the deterministic aspects of the parametric regression model as well as omnibus tests for assessing the overall fit of the model. These procedures are derived from cumulative sums of martingale-based residuals over the failure time or/and covariates. Under the assumed model, the distributions of these (possibly multi-parameter) partial-sum processes can again be approximated through simulating certain zero-mean Gaussian processes. One may then plot each observed (one-dimensional) process along with a few realizations from the corresponding Gaussian process to assess visually how unusual the observed residual pattern is. Numerical test may also be performed.

The rest of this article is organized as follows. The next section contains some essential background material for parametric and semi-parametric survival models. In section 3, we deal with the problem of testing the distributional assumption. Model checking techniques based on martingale residuals are described in section 4. Section 5 reports the results from our simulation studies to evaluate the finite-sample performance of the proposed tests. Finally, two real examples are provided in section 6.

2. Data and models

2.1. Basic notation and assumptions

Let T and C denote the failure time and censoring time, and let $Z(\cdot) = \{Z_1(\cdot), \ldots, Z_p(\cdot)\}'$ denote a p-vector of possibly time-varying covariates. Assume that T and C are independent

conditional on Z and that Z has bounded total variation, i.e. $\int_0^\infty |dZ_j(t)| + |Z_j(0)| \leqslant \mathscr{K}$ for some $\mathscr{K} > 0$ and all j. Suppose that the data consist of n independent replicates of $\{X, \Delta, Z(\cdot)\}$, where $X = \min(T,C)$, $\Delta = I(T \leqslant C)$, and $I(\cdot)$ is the indicator function. In the counting process notation, the data consist of n independent replicates of $\{Y(\cdot), N(\cdot), Z(\cdot)\}$, where $Y(t) = I(X \geqslant t)$ and $N(t) = \Delta I(X \leqslant t)$. The counting process $N(\cdot)$ can be uniquely decomposed such that $N(t) = M(t) + \int_0^t Y(v)\lambda(v \mid Z)\, dv$ for all t, where $\lambda(\cdot \mid Z)$ is the hazard function for T given $Z(\cdot)$ and $M(t)$ is a (local square-integrable) martingale process with respect to the σ-filtration consisting of all the information up to the failure time point t.

2.2. *Parametric models*

Let $\lambda(t \mid Z, \theta_0)$ be the parametric hazard function for T given the p-vector of covariates $Z(\cdot)$, where θ_0 is a q-vector ($q \geqslant p$) of unknown parameters. We assume that cond. VI.1.1 of Andersen *et al.* (1993, pp. 420–421) holds. Discussion of this condition in the regression context can be found in Andersen *et al.* (1993, pp. 585–586). We write $\theta_0 = (\beta_0', \gamma_0')'$, where β_0 is a p-vector of regression parameters characterizing the dependence of T on Z, and γ_0 is a $(q-p)$-vector of unknown parameters for the baseline failure time distribution. The two most important classes of parametric regression models are the proportional hazards (PH) model

$$\lambda(t \mid Z, \theta_0) = \lambda_0(t; \gamma_0) \exp(\beta_0' Z(t), \tag{2.1}$$

where $\lambda_0(\cdot; \gamma_0)$ is a parametric baseline hazard function, and the accelerated failure time (AFT) model

$$\log T = \beta_0' Z + \varepsilon, \tag{2.2}$$

or

$$\lambda(t \mid Z, \theta_0) = \lambda_0(t \exp(-\beta_0' Z); \gamma_0) \exp(-\beta_0' Z),$$

where ε is an error term with a specified distribution, and $\lambda_0(\cdot; \gamma_0)$ is the hazard function for e^ε.

Given the data $\{X_i, \Delta_i, Z_i(\cdot)\}$ $(i = 1, \ldots, n)$, the likelihood score function and information matrix are, respectively,

$$U(\theta) = \sum_{i=1}^n \left\{ \Delta_i \partial \log \lambda(X_i \mid Z_i, \theta)/\partial\theta - \int_0^{X_i} \partial\lambda(t \mid Z_i, \theta)/\partial\theta\, dt \right\},$$

and

$$\mathscr{I}(\theta) = -n^{-1} \sum_{i=1}^n \left\{ \Delta_i \partial^2 \log \lambda(X_i \mid Z_i, \theta)/\partial\theta^2 - \int_0^{X_i} \partial^2\lambda(t \mid Z_i, \theta)/\partial\theta^2\, dt \right\}.$$

The maximum likelihood estimator $\hat{\theta} = (\hat{\beta}', \hat{\gamma}')'$ is the solution to $\{U(\theta) = 0\}$. Note that

$$U(\theta_0) = \sum_{i=1}^n \int_0^\infty \partial \log \lambda(t \mid Z_i, \theta_0)/\partial\theta\, dM_i(t), \tag{2.3}$$

and

$$n^{1/2}(\hat{\theta} - \theta_0) = \Omega^{-1} n^{-1/2} U(\theta_0) + o_p(1), \tag{2.4}$$

where Ω is the limit of $\mathscr{I}(\theta_0)$. It follows from (2.3) and (2.4) that $n^{1/2}(\hat{\theta} - \theta_0)$ is asymptotically zero-mean normal with covariance matrix Ω^{-1}.

2.3. Semi-parametric models

2.3.1. The PH model

Under the Cox proportional hazards model,

$$\lambda(t \mid Z, \beta_0) = \lambda_0(t) \exp(\beta_0' Z(t)), \tag{2.5}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function. Let $S^{(r)}(\beta, t) = n^{-1} \Sigma_{i=1}^n Y_i(t) \exp(\beta' Z_i(t)) Z_i(t)^{\otimes r}$ and $s^{(r)}(\beta, t) = E\{S^{(r)}(\beta, t)\}$, where $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa'$. Then the partial likelihood score function for β_0 is

$$\tilde{U}(\beta) = \sum_{i=1}^n \Delta_i \{Z_i(X_i) - \bar{Z}(\beta, X_i)\},$$

where $\bar{Z}(\beta, t) = S^{(1)}(\beta, t)/S^{(0)}(\beta, t)$. The information matrix is

$$\tilde{\mathscr{I}}(\beta) = n^{-1} \sum_{i=1}^n \Delta_i \left\{ \frac{S^{(2)}(\beta, X_i)}{S^{(0)}(\beta, X_i)} - \bar{Z}(\beta, X_i)^{\otimes 2} \right\}.$$

The maximum partial likelihood estimator $\tilde{\beta}$ is the solution to $\{\tilde{U}(\beta) = 0\}$. The random vector $n^{1/2}(\tilde{\beta} - \beta_0)$ is asymptotically zero-mean normal with covariance matrix $\tilde{\Omega}^{-1}$, where $\tilde{\Omega}$ is the limit of $\tilde{\mathscr{I}}(\beta_0)$. The cumulative baseline hazard function $\Lambda_0(t) = \int_0^\infty \lambda_0(v)\, dv$ is commonly estimated by

$$\tilde{\Lambda}_0(t; \tilde{\beta}) = \sum_{i=1}^n \frac{I(X_i \leqslant t)\, \Delta_i}{nS^{(0)}(\tilde{\beta}, X_i)} \tag{2.6}$$

(Breslow, 1972), which is consistent and asymptotically normal. As shown by Andersen *et al.* (1993, VII.2.2),

$$\tilde{U}(\beta_0) = \sum_{i=1}^n \int_0^\infty \{Z_i(t) - \bar{Z}(\beta_0, t)\}\, dM_i(t) \tag{2.7}$$

and

$$n^{1/2}(\tilde{\beta} - \beta_0) = \tilde{\Omega}^{-1} n^{-1/2} \tilde{U}(\beta_0) + o_p(1). \tag{2.8}$$

2.3.2. The AFT model

One may estimate β_0 in model (2.2) without parameterizing the error distribution. Let $e_i(\beta) = \log X_i - \beta' Z_i$ $(i = 1, \ldots, n)$. The log rank statistic based on $\{e_i(\beta), \Delta_i, Z_i\}$ $(i = 1, \ldots, n)$ is

$$U_R(\beta) = \sum_{i=1}^n \Delta_i \{Z_i - \bar{Z}_R(\beta, e_i(\beta))\},$$

where $\bar{Z}_R(\beta, u) = \Sigma_{j=1}^n I(e_j(\beta) \geqslant u) Z_j / \Sigma_{j=1}^n I(e_j(\beta) \geqslant u)$. The random vector $n^{-1/2} U_R(\beta_0)$ is asymptotically zero-mean normal with covariance matrix $B = \int_{-\infty}^\infty D(u) \lambda_\varepsilon(u)\, du$, where λ_ε is the hazard function for ε and $D(u)$ is the limit of $n^{-1} \Sigma_{i=1}^n I(e_i(\beta_0) \geqslant u) \{Z_i - \bar{Z}_R(\beta_0, u)\}^{\otimes 2}$. We define the rank estimator $\hat{\beta}_R$ as a minimizer of $\|U_R(\beta)\|$ (Lin & Geyer, 1992). It can be shown that

$$n^{1/2}(\hat{\beta}_R - \beta_0) = A^{-1} n^{-1/2} \sum_{i=1}^n \int_{-\infty}^\infty \{Z_i - \bar{Z}_R(\beta_0, u)\}\, dM_i(u; \beta_0) + o_p(1), \tag{2.9}$$

where $A = \int_{-\infty}^\infty D(u)\, d\lambda_\varepsilon(u)$, $M_i(u; \beta) = N_i(u; \beta) - \int_{-\infty}^u I(e_i(\beta) \geqslant v) \lambda_\varepsilon(v)\, dv$ and $N_i(u; \beta) = \Delta_i I(e_i(\beta) \leq u)$ (Tsiatis, 1990; Wei *et al.*, 1990; Ying, 1993). Note that $M_i(u; \beta_0)$ $(i = 1, \ldots, n)$

are martingale processes with respect to the σ-filtration generated by the failure and censoring information of the residuals $e_i(\beta_0)$ $(i = 1, \ldots, n)$ up to the error term time u. From (2.9) it follows that $n^{1/2}(\hat{\beta}_R - \beta_0)$ is asymptotically zero-mean normal with covariance matrix $A^{-1}BA^{-1}$.

3. Checking distributional assumptions

In this section, we show how to verify the parametric form of the baseline distribution. The general strategy is to compare the parametric maximum likelihood estimator of the survival function with a counterpart which is constructed without the distributional assumption. We shall concentrate on parametric PH and AFT models, though the basic ideas developed here apply to other survival models as well. In addition, we assume that the deterministic components are correctly specified; therefore, the null hypothesis means that the whole model is correct.

3.1. PH models

Define $W(t) = n^{1/2}\{\Lambda_0(t; \hat{\gamma}) - \tilde{\Lambda}_0(t; \tilde{\beta})\}$, where $\Lambda_0(t; \hat{\gamma}) = \int_0^t \lambda_0(v; \hat{\gamma})\, dv$ and $\tilde{\Lambda}_0(t; \tilde{\beta})$ is the Breslow estimator given in (2.6). Also, let $\tau < \inf\{t: EY(t) = 0\}$. In order to derive the asymptotic distribution of $W(\cdot)$ under model (2.1), we make the following decomposition

$$W(t) = n^{1/2}\{\Lambda_0(t; \hat{\gamma}) - \Lambda_0(t; \gamma_0)\} - n^{1/2}\{\tilde{\Lambda}_0(t; \tilde{\beta}) - \tilde{\Lambda}_0(t; \beta_0)\}$$
$$- n^{1/2}\{\tilde{\Lambda}_0(t; \beta_0) - \Lambda_0(t; \gamma_0)\} = W_1(t) - W_2(t) - W_3(t),$$

say.

It follows from Taylor series expansions and simple probability arguments that, uniformly in $t \leqslant \tau$, $W_1(t) = g'(t)n^{1/2}(\hat{\theta} - \theta_0) + o_p(1)$ and $W_2(t) = h'(t)n^{1/2}(\tilde{\beta} - \beta_0) + o_p(1)$, where

$$g(t) = \begin{bmatrix} 0 \\ \partial \Lambda_0(t; \gamma_0)/\partial \gamma \end{bmatrix},$$

$h(t) = -\int_0^t \bar{z}(\beta_0, v)\lambda_0(v)\, dv$ and $\bar{z}(\beta, t) = s^{(1)}(\beta, t)/s^{(0)}(\beta, t)$. In view of (2.3), (2.4), (2.7) and (2.8),

$$W_1(t) = g'(t)\Omega^{-1}n^{-1/2}\sum_{i=1}^{n}\int_0^{\infty} \partial \log \lambda(v \mid Z_i, \theta_0)/\partial\theta\, dM_i(v) + o_p(1), \tag{3.1}$$

$$W_2(t) = h'(t)\tilde{\Omega}^{-1}n^{-1/2}\sum_{i=1}^{n}\int_0^{\infty} \{Z_i(v) - \bar{Z}(\beta_0, v)\}\, dM_i(v) + o_p(1). \tag{3.2}$$

On the other hand,

$$W_3(t) = n^{-1/2}\left\{\int_0^t \frac{\sum_{i=1}^{n} dN_i(v)}{S^{(0)}(\beta_0, v)} - \int_0^t \frac{\sum_{i=1}^{n} Y_i(v)\exp(\beta_0' Z_i(v))\lambda_0(v; \gamma_0)\, dv}{S^{(0)}(\beta_0, v)}\right\}$$
$$= n^{-1/2}\sum_{i=1}^{n}\int_0^t \frac{dM_i(v)}{S^{(0)}(\beta_0, v)}. \tag{3.3}$$

Combining (3.1)–(3.3), we obtain

$$W(t) = g'(t)\Omega^{-1}n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \partial \log \lambda(v \mid Z_i, \theta_0)/\partial\theta \, dM_i(v)$$
$$- h'(t)\tilde{\Omega}^{-1}n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \{Z_i(v) - \bar{Z}(\beta_0, v)\}\, dM_i(v)$$
$$- n^{-1/2}\sum_{i=1}^{n}\int_0^t \frac{dM_i(v)}{S^{(0)}(\beta_0, v)} + o_p(1). \tag{3.4}$$

By Rebolledo's central limit theorem (Andersen *et al.*, 1993, th. II.5.1), the process $W(t)$ $(0 \leqslant t \leqslant \tau)$ converges weakly to a zero-mean Gaussian process with covariance function

$$\xi(t, t^\dagger) = g'(t)\Omega^{-1}g(t^\dagger) + h'(t)\tilde{\Omega}^{-1}h(t^\dagger) + \int_0^{\min(t, t^\dagger)} \lambda_0(v)\, dv/s^{(0)}(\beta_0, v)$$
$$- g'(t)\Omega^{-1}\{\Gamma\tilde{\Omega}^{-1}h(t^\dagger) + \eta(t^\dagger)\} - g'(t^\dagger)\Omega^{-1}\{\Gamma\tilde{\Omega}^{-1}h(t) + \eta(t)\}, \tag{3.5}$$

where

$$\Gamma = \int_0^\infty E[Y(v)\exp(\beta_0'Z(v))\partial \log \lambda(v \mid Z, \theta_0)/\partial\theta\{Z(v) - \bar{z}(\beta_0, v)\}']\lambda_0(v)\, dv,$$

and

$$\eta(t) = \int_0^t E\{Y(v)\exp(\beta_0'Z(v))\partial \log \lambda(v \mid Z, \theta_0)/\partial\theta\}\lambda_0(v)\, dv/s^{(0)}(\beta_0, v).$$

The covariance function ξ can be consistently estimated by replacing the unknown quantities in (3.5) by their respective sample estimators. Expression (3.5) indicates that $W(\cdot)$ does not have an independent increment structure asymptotically. Thus, one cannot transform (only) the time and space axes of $W(\cdot)$ to achieve a Brownian bridge limit.

We shall approximate the null distribution of $W(\cdot)$ by a zero-mean Gaussian process whose distribution can be easily generated through simulation. By replacing $\{M_i(\cdot)\}$ $(i = 1, \ldots, n)$ in (3.4) with $\{N_i(\cdot)G_i\}$ $(i = 1, \ldots, n)$, where $\{G_i\}$ $(i = 1, \ldots, n)$ are independent standard normal variables which are independent of $\{Y_i(\cdot), N_i(\cdot), Z_i(\cdot)\}$ $(i = 1, \ldots, n)$, and also replacing other unknown quantities in (3.4) with their respective sample estimators, we obtain

$$\hat{W}(t) = \hat{g}'(t)\mathscr{I}^{-1}(\hat{\theta})n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \partial \log(\lambda(v \mid Z_i, \hat{\theta})/\partial\theta G_i\, dN_i(v)$$
$$- \hat{h}'(t)\mathscr{I}^{-1}(\hat{\beta})n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \{Z_i(v) - \bar{Z}(\hat{\beta}, v)\}G_i\, dN_i(v)$$
$$- n^{-1/2}\sum_{i=1}^{n}\int_0^t \frac{G_i\, dN_i(v)}{S^{(0)}(\hat{\beta}, v)}, \tag{3.6}$$

where

$$\hat{g}(t) = \begin{bmatrix} 0 \\ \partial\Lambda_0(t; \hat{\gamma})/\partial\gamma \end{bmatrix} \quad \text{and} \quad \hat{h}(t) = -n^{-1}\sum_{i=1}^{n}\int_0^t \bar{Z}(\hat{\beta}, v)\, dN_i(v)/S^{(0)}(\hat{\beta}, v).$$

When approximating the distribution of W, we regard $\{G_i\}$ $(i = 1, \ldots, n)$ as random and $\{Y_i(\cdot), N_i(\cdot), Z_i(\cdot)\}$ $(i = 1, \ldots, n)$ as fixed in (3.6). It is shown in the Appendix that $\hat{W}(\cdot)$ and $W(\cdot)$ have the same limiting distribution. To approximate the distribution of $W(\cdot)$, we simply obtain a large number of realizations from $\hat{W}(\cdot)$ by repeatedly generating normal random

samples $\{G_i;\ i = 1, \ldots, n\}$ while fixing the data $\{Y_i(\cdot), N_i(\cdot), Z_i(\cdot)\}$ $(i = 1, \ldots, n)$ at their observed values.

It is useful to consider the transformed process $B(t) = n^{1/2}\pi(t)[\phi\{\Lambda_0(t; \hat{\gamma})\} - \phi\{\tilde{\Lambda}_0(t; \tilde{\beta})\}]$, where ϕ is a known function whose derivatives $\dot{\phi}$ is continuous and non-zero in the time interval $[t_1, t_2]$ $(0 \leqslant t_1 \leqslant t_2 \leqslant \tau)$, and the weight process $\pi(\cdot)$ converges in probability to a non-negative bounded function uniformly on $[t_1, t_2]$. By the functional delta-method (Andersen *et al.*, 1993, II.8), $B(t) = \pi(t)\dot{\phi}\{\tilde{\Lambda}_0(t; \tilde{\beta})\}W(t) + o_p(1)$. Thus the distribution of $B(t)$ can be approximated by that of $\hat{B}(t) = \pi(t)\dot{\phi}\{\tilde{\Lambda}_0(t; \tilde{\beta})\}\hat{W}(t)$.

The supremum test statistic is $Q = \sup_{t_1 \leqslant t \leqslant t_2} |B(t)|$. Let q be the observed value of Q and let $\hat{Q} = \sup_{t_1 \leqslant t \leqslant t_2} |\hat{B}(t)|$. Then the P-value, $\Pr(Q > q)$, can be approximated by $\Pr(\hat{Q} > q)$, the latter probability being estimated through simulation. If the baseline hazard function is incorrectly parameterized but the remaining specification in model (2.1) is valid, then $\tilde{\Lambda}_0(t; \tilde{\beta})$ converges to the true $\Lambda_0(t)$ whereas $\Lambda_0(t; \hat{\gamma})$ converges to a limit which is different from $\Lambda_0(t)$ at least for some t. Therefore, the supremum test is consistent against any mis-specification of the baseline hazard function.

Different choices of π and ϕ produce tests that are sensitive to different alternatives. The supremum test based on the original process $W(\cdot)$ has the disadvantage of tending to place too much emphasis on the right tail of the distribution, where, as is clear from (3.5), the process is most variable. In this article, we choose $\pi(\cdot) = 1$ and $\phi(x) = \exp(-x)$ so that the resulting process $B(t) = n^{1/2}\{\exp(-\Lambda_0(t; \hat{\gamma})) - \exp(\tilde{\Lambda}_0(t; \tilde{\beta}))\}$ compares the parametric and semi-parametric survival function estimators. In this case, $\hat{B}(t) = -\exp(-\hat{\Lambda}_0(t; \tilde{\beta}))\hat{W}(t)$. Since $\Lambda_0(t; \hat{\gamma})$ is monotone in t and since $\tilde{\Lambda}_0(t; \tilde{\beta})$ and the last two terms on the right-hand side of (3.6) are step functions which jump only at uncensored failure times, say d_j $(j = 1, \ldots, n_d)$, one needs to evaluate $B(t)$ and $\hat{B}(t)$ only at d_j and $d_j -$ $(j = 1, \ldots, n_d)$ when searching for the maxima. (If $\partial\Lambda_0(t; \hat{\gamma})/\partial\gamma$ is not monotone between d_j and d_{j+1}, then more elaborate search is needed for that interval.) Due to the instability on the right tail, we shall restrict all calculations up to the last uncensored failure time.

For each realization of $\{G_i;\ i = 1, \ldots, n\}$, the calculation of $\sup_t |\hat{B}(t)|$ is only of order n. To see why this is the case, note that the third term on the right-hand side of (3.6) is $\hat{W}_3(t) = n^{-1/2}\sum_{i=1}^{n} I(X_i \leqslant t)\Delta_i G_i / S^{(0)}(\tilde{\beta}, X_i)$; therefore, $\hat{W}_3(X_k) = \hat{W}_3(X_{k-1}) + n^{-1/2}\Delta_k G_k / S^{(0)}(\tilde{\beta}, X_k)$ $(k = 2, \ldots, n)$ if $X_k (k = 1, \ldots, n)$ are distinct and are arranged in ascending order. Making use of this recursive relationship and a similar one for $\hat{h}(t)$, we may calculate $\hat{B}(X_k)$ $(k = 1, \ldots, n)$, $\hat{B}(X_{k-})$ $(k = 1, \ldots, n)$, $\sup_{1 \leqslant k < n} |\hat{B}(X_k)|$ and $\sup_{1 \leqslant k < n} |\hat{B}(X_k -)|$ simultaneously in a single loop. (If $X_{k+1} = X_k$, one simply skips X_k in the maximization step.)

3.2. AFT models

Following up the development in section 2.3.2, an Aalen–Breslow type estimator for $\Lambda_\varepsilon(u) = \int_{-\infty}^{u} \lambda_\varepsilon(v)\, dv$ is

$$\Lambda_\varepsilon(u; \tilde{\beta}_R) = \sum_{i=1}^{n} I(e_i(\tilde{\beta}_R) \leqslant u)\, \Delta_i \Big/ \sum_{j=1}^{n} I(e_j(\tilde{\beta}_R) \geqslant e_i(\tilde{\beta}_R)). \tag{3.7}$$

Let $\lambda_\varepsilon(u; \gamma_0)$ be the parametric hazard function of ε. Then the maximum likelihood estimator for $\Lambda_\varepsilon(u; \gamma_0) = \int_{-\infty}^{u} \lambda_\varepsilon(v; \gamma_0)\, dv$ is $\Lambda_\varepsilon(u; \tilde{\gamma}) = \int_{-\infty}^{u} \lambda_\varepsilon(v; \tilde{\gamma})\, dv$, where $\tilde{\gamma}$ is the maximum likelihood estimator for γ_0. The goodness-of-fit process for testing the distributional assumption is $W_\varepsilon(u) = n^{1/2}\{\Lambda_\varepsilon(u; \tilde{\gamma}) - \hat{\Lambda}_\varepsilon(u; \tilde{\beta}_R)\}$. Analogous to (3.4), uniformly in u,

$$W_\varepsilon(u) = g_\varepsilon'(u)\Omega^{-1}n^{-1/2}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\partial\log\lambda_\varepsilon(v;\gamma_0)/\partial\theta\, dM_i(v;\beta_0)$$
$$-h_\varepsilon'(u)A^{-1}n^{-1/2}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\{Z_i-\bar{Z}_R(\beta_0,v)\}\,dM_i(v;\beta_0)$$
$$-n^{-1/2}\sum_{i=1}^{n}\int_{-\infty}^{u}\frac{dM_i(v;\beta_0)}{n^{-1}\sum_{j=1}^{n}I(e_j(\beta_0)\geq v)}+o_p(1), \tag{3.8}$$

where

$$g_\varepsilon(u)=\begin{bmatrix}0\\ \partial\Lambda_\varepsilon(u;\gamma_0)/\partial\gamma\end{bmatrix},\quad h_\varepsilon(u)=\int_{-\infty}^{u}\bar{z}_R(\beta_0,v)\,d\lambda(v;\gamma_0)$$

and $\bar{z}_R(\beta_0, v)$ is the limit of $\bar{Z}_R(\beta_0, v)$. In deriving (3.8), one cannot use the Taylor series expansion to approximate the distribution of $n^{1/2}\{\hat{\Lambda}_\varepsilon(u;\hat{\beta}_R)-\hat{\Lambda}_\varepsilon(u;\beta_0)\}$ because $\hat{\Lambda}_\varepsilon(u;\beta)$ is a discrete function of β. Instead, we used the following asymptotic linear approximation

$$n^{1/2}\{\hat{\Lambda}_\varepsilon(u;\hat{\beta}_R)-\hat{\Lambda}_\varepsilon(u;\beta_0)\}=h_\varepsilon'(u)n^{1/2}(\hat{\beta}_R-\beta_0)+o_p(1),$$

which follows from the kind of arguments for establishing the asymptotic linearity of the rank estimating function $U_R(\beta)$; see Tsiatis (1990, pp. 358–359). As in (3.6), we approximate the null distribution of $W_\varepsilon(u)$ by

$$\hat{W}_\varepsilon(u) = \hat{g}_\varepsilon'(u)\hat{\mathscr{I}}^{-1}(\hat{\theta})n^{-1/2}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\partial\log\lambda_\varepsilon(v;\hat{\gamma})/\partial\theta G_i\, dN_i(v;\hat{\beta})$$
$$-\hat{h}_\varepsilon'(u)\hat{A}^{-1}n^{-1/2}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\{Z_i-\bar{Z}_R(\hat{\beta},v)\}G_i\,dN_i(v;\hat{\beta})$$
$$-n^{-1/2}\sum_{i=1}^{n}\int_{-\infty}^{u}\frac{G_i\,dN_i(v;\hat{\beta})}{n^{-1}\sum_{j=1}^{n}I(e_j(\hat{\beta})\geq v)}, \tag{3.9}$$

where

$$\hat{g}_\varepsilon(u)=\begin{bmatrix}0\\ \partial\Lambda_\varepsilon(u;\hat{\gamma})/\partial\gamma\end{bmatrix},\quad \hat{h}_\varepsilon(u)=\int_{-\infty}^{u}\bar{Z}_R(\hat{\beta},v)\,d\lambda_\varepsilon(v;\hat{\gamma})\quad\text{and}$$

$$\hat{A}=n^{-1}\sum_{i=1}^{n}\int_{-\infty}^{\infty}I(e_i(\hat{\beta})\geq u)\{Z_i-\bar{Z}_R(\hat{\beta},u)\}^{\otimes 2}\,d\lambda_\varepsilon(u;\gamma).$$

One may then follow the ideas given in the latter part of section 3.1 to develop goodness-of-fit tests for the distributional assumption on ε.

It is somewhat cumbersome to calculate the rank estimator $\hat{\beta}_R$. Simpler goodness-of-fit procedures may be constructed by replacing $\hat{\beta}_R$ in (3.7) with the maximum likekihood estimator $\hat{\beta}$ and thereby considering the process $W_\varepsilon^*(u)=n^{1/2}\{\Lambda_\varepsilon(u;\hat{\gamma})-\hat{\Lambda}_\varepsilon(u;\hat{\beta})\}$. It follows from the above arguments that, uniformly in u,

$$W_\varepsilon^*(u) = \kappa'(u)\Omega^{-1}n^{-1/2}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\partial\log\lambda_\varepsilon(v;\gamma_0)/\partial\theta\, dM_i(v;\beta_0)$$
$$-n^{-1/2}\sum_{i=1}^{n}\int_{-\infty}^{u}\frac{dM_i(v;\beta_0)}{n^{-1}\sum_{j=1}^{n}I(e_j(\beta_0)\geq v)}+o_p(1),$$

where

$$\kappa(u) = \begin{bmatrix} -\int_{-\infty}^{u} \bar{z}_R(\beta_0, v)\, d\lambda_\varepsilon(v; \gamma_0) \\ \partial \Lambda_\varepsilon(u; \gamma_0)/\partial \gamma \end{bmatrix}.$$

Thus, we approximate the distribution of $W_\varepsilon^*(u)$ by

$$\hat{\kappa}'(u)\mathcal{I}^{-1}(\hat{\theta})n^{-1/2}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\partial \log \lambda_\varepsilon(v;\hat{\gamma})/\partial\theta G_i\, dN_i(v;\hat{\beta}) - n^{-1/2}\sum_{i=1}^{n}\int_{-\infty}^{u}\frac{G_i\, dN_i(v;\hat{\beta})}{n^{-1}\sum_{j=1}^{n} I(e_j(\hat{\beta}) \geqslant v)},$$

where

$$\hat{\kappa}(u) = \begin{bmatrix} -\int_{-\infty}^{u} \bar{Z}_R(\hat{\beta}, v)\, d\lambda_\varepsilon(v; \hat{\gamma}) \\ \partial \Lambda_\varepsilon(u; \hat{\gamma})/\partial \gamma \end{bmatrix}.$$

The price to pay for using W_ε^* instead of W_ε is that one can no longer claim the consistency of the corresponding supremum tests against all violations of the distributional assumption. The reason is that, unlike $\hat{\Lambda}_\varepsilon(u; \hat{\beta}_R)$, the estimator $\hat{\Lambda}_\varepsilon(u; \hat{\beta})$ may converge to the same limit as $\Lambda_\varepsilon(u; \hat{\gamma})$ does for every u under certain alternatives, though such scenarios are unlikely to occur in applications. Incidentally, Pierce & Kopecky (1979) and Loynes (1980) studied processes similar to W_ε^* for non-censored data.

We have restricted Z to be time-invariant in model (2.2). As shown by Cox & Oakes (1984, pp. 64–68), time-dependent covariates may be incorporated into the AFT model in a natural fashion. Their formula (5.10) provides the essential ingredients for the fully parametric analysis. Recently, a rank-type estimator with time-dependent covariates analogous to $\hat{\beta}_R$ has been studied by Robins & Tsiatis (1992) and Lin & Ying (1995). It is straightforward to extend the above goodness-of-fit results to the case of time-dependent covariates.

4. Martingale residuals

The parametric survival model $\lambda(t \mid Z, \theta_0)$ may be regarded as a (possibly time non-homogeneous) Poisson model. In view of this connection, it is natural to assess the fit of the model by examining the martingale residuals

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(v)\lambda(v \mid Z_i, \hat{\theta})\, dv \quad i = 1, \ldots, n.$$

Note that $\hat{M}_i(t)$ is the observed number of failures by time t on the ith subject minus its (estimated) expectation under the assumed model. Similar residuals were previously studied by Barlow & Prentice (1988), Therneau *et al.* (1990) and Lin *et al.* (1993) for the semi-parametric PH model.

As in the setting of ordinary residuals, one may plot the martingale residuals $\hat{M}_i = \hat{M}_i(\infty)$ $(i = 1, \ldots, n)$ against a covariate to (informally) check its functional form. However, it is often difficult to conclude whether a seemingly abnormal residual pattern reflects functional form mis-specification or is a phenomenon that has a high probability of occurring even when the model holds. To make the inspection more objective, we shall ascertain the null distribution of the cumulative sum of the martingale residuals over all possible values of the covariate. (In this paper, the null hypothesis always means that the whole model is correctly specified.) In fact, we shall develop a number of analytical and graphical methods by

considering various special cases of the following two classes of multi-parameter stochastic processes:

$$\mathscr{W}_z(t, z) = n^{-1/2} \sum_{i=1}^{n} f(Z_i) I(Z_i \leqslant z) \hat{M}_i(t)$$

and

$$\mathscr{W}_r(t, r) = n^{-1/2} \sum_{i=1}^{n} f(Z_i) I(\hat{\beta}' Z_i \leqslant r) \hat{M}_i(t),$$

where f is a known smooth function, $z = (z_1, \ldots, z_p)' \in R^p$, and $I(Z_i \leqslant z) = I(Z_{1i} \leqslant z_1, \ldots, Z_{pi} \leqslant z_p)$. We shall restrict to fixed covariates in this section.

4.1. Null distributions of $\mathscr{W}_z$ and $\mathscr{W}_r$

By Taylor series expansions and some probabilistic arguments, $\mathscr{W}_z(t, z) = \tilde{\mathscr{W}}_z(t, z) + o_p(1)$ uniformly in t and z, where

$$\tilde{\mathscr{W}}_z(t, z) = n^{-1/2} \sum_{i=1}^{n} f(Z_i) I(Z_i \leqslant z) M_i(t) - J(t, z; \theta_0) \mathscr{I}^{-1}(\theta_0) n^{-1/2} \sum_{i=1}^{n} \int_0^{\infty} \partial \log \lambda(v \mid Z_i, \theta_0)/\partial\theta \, dM_i(v), \tag{4.1}$$

and

$$J(t, z; \theta) = n^{-1} \sum_{i=1}^{n} f(Z_i) I(Z_i \leqslant z) \int_0^t Y_i(v) \, \partial\lambda(v \mid Z_i, \theta)/\partial\theta' \, dv.$$

Similar to the approximations used in section 3, the null distribution of $\mathscr{W}_z(t, z)$ will be approximated by

$$\hat{\mathscr{W}}_z(t, z) = n^{-1/2} \sum_{i=1}^{n} f(Z_i) I(Z_i \leqslant z) G_i N_i(t) - J(t, z; \hat{\theta}) \mathscr{I}^{-1}(\hat{\theta}) n^{-1/2} \sum_{i=1}^{n} \int_0^{\infty} \partial \log \lambda(v \mid Z_i, \hat{\theta})/\partial\theta G_i \, dN_i(v), \tag{4.2}$$

which is obtained from (4.1) by replacing $\{M_i(\cdot)\}$ with $\{N_i(\cdot)G_i\}$ $(i = 1, \ldots, n)$ and θ_0 with $\hat{\theta}$.

Let $H_i(v, t, z, \theta) = I(v \leqslant t) f(Z_i) I(Z_i \leqslant z) - J(t, z; \theta) \mathscr{I}^{-1}(\theta) \, \partial \log \lambda(v \mid Z_i, \theta)/\partial\theta$, and let $\tilde{H}_i(v, t, z, \theta)$ by the same expression except that $J(t, z; \theta)$ and $\mathscr{I}(\theta)$ are replaced by their limits. Then $\tilde{\mathscr{W}}_z(t, z) = n^{-1/2} \Sigma_{i=1}^n \int_0^{\infty} H_i(v, t, z, \theta_0) \, dM_i(v) = n^{-1/2} \Sigma_{i=1}^n \int_0^{\infty} \tilde{H}_i(v, t, z, \theta_0) \, dM_i(v) + o_p(1)$ and $\hat{\mathscr{W}}_z(t, z) = n^{-1/2} \Sigma_{i=1}^n \int_0^{\infty} H_i(v, t, z, \hat{\theta}) G_i \, dN_i(v)$. Conditional on $\{N_i(\cdot), Y_i(\cdot), Z_i(\cdot)\}$ $(i = 1, \ldots, n)$, the covariance function for $\hat{\mathscr{W}}_z(\cdot, \cdot)$ between (t, z) and $(t^\dagger, z^\dagger)$ is

$$n^{-1} \sum_{i=1}^{n} \int_0^{\infty} H_i(v, t, z, \hat{\theta}) H_i(v, t^\dagger, z^\dagger, \hat{\theta})' \, dN_i(v). \tag{4.3}$$

Due to the strong consistency of $\hat{\theta}$ and the strong law of large numbers, (4.3) converges almost surely to

$$E\left\{\int_0^{\infty} \tilde{H}_1(v, t, z, \theta_0) \tilde{H}_1(v, t^\dagger, z^\dagger, \theta_0)' \, dN_1(v)\right\},$$

which is the asymptotic covariance function for $\tilde{\mathscr{W}}_z(\cdot, \cdot)$. By the Lindeberg–Feller theorem and the Cramér–Wold device, the finite-dimensional distributions of $\tilde{\mathscr{W}}_z$ and $\hat{\mathscr{W}}_z$ are both

asymptotically normal. The tightness of $\tilde{\mathcal{W}}_z$ follows from lem. 1 of Lin *et al.* (1993) while that of $\hat{\mathcal{W}}_z$ can be verified by extending the arguments given in the Appendix of this paper. Therefore, the processes $\tilde{\mathcal{W}}_z$ and $\hat{\mathcal{W}}_z$ have the same limiting Gaussian distribution.

Likewise, the null distribution of $\mathcal{W}_r(t, r)$ can be approximated by that of $\hat{\mathcal{W}}_r(t, r)$, where $\hat{\mathcal{W}}_r(t, r)$ is obtained from (4.2) by substituting $I(\hat{\beta}'Z_i \leqslant r)$ for $I(Z_i \leqslant z)$ $(i = 1, \ldots, n)$. The finite-dimensional asymptotic normality and the asymptotic equivalence of the covariance functions for $\mathcal{W}_r$ and $\hat{\mathcal{W}}_r$ follows from the arguments given in the preceding paragraph. We refer the interested reader to app. B of Lin *et al.* (1992) for the techniques to prove the tightness of $\mathcal{W}_r$ and $\hat{\mathcal{W}}_r$.

4.2. *Omnibus tests*

Since $\mathcal{W}_z(\cdot, \cdot)$ is a process defined on the entire product space of T and Z, whose null distribution is centred around zero, it is natural to construct a global goodness-of-fit test based on the statistic $\mathcal{Q} = \sup_{t,z} |\mathcal{W}_z(t, z)|$. An extreme observed value of $\mathcal{Q}$ would indicate model mis-specification. The P-value can be estimated through simulation as in section 3.

We now show that the $\mathcal{Q}$ test is consistent against any departures from the assumed parametric model $\lambda(t \mid Z, \theta_0)$. Under mis-specified models, $\hat{\theta}$ converges to some constant θ^* which is assumed to be the unique root for the limiting score function $u(\theta) = \lim n^{-1}U(\theta)$ (Hjort, 1992). Let L be the distribution function of Z. It can be shown that $n^{-1/2}\mathcal{W}_z(t, z)$ converges in probability to

$$\int_{x=-\infty}^{z} f(x) \int_{v=0}^{t} E\{Y(v) \mid x\}\{\lambda(v \mid x) - \lambda(v \mid x, \theta^*)\}\, dv\, dL(x), \tag{4.4}$$

which will be non-zero for some t and/or z if $\lambda(t \mid Z, \theta_0)$ is not the true model for $\lambda(t \mid Z)$. This establishes our claim. In the sequel, we shall take $f(\cdot) = 1$ for the omnibus test, denoting the corresponding process by $\mathcal{W}_o(\cdot, \cdot)$.

In computing the maxima, one needs to evaluate $\mathcal{W}_o(t, z)$ and $\hat{\mathcal{W}}_o(t, z)$ at X_k and $X_k - (k = 1, \ldots, n)$ as well as at all distinct combinations of the observed covariate values. By using recursive relationships in t similar to that of $\hat{W}_3(t)$ given in the last paragraph of section 3.1, one may avoid any calculations of orders higher than nn^* for $\sup_{t,z} |\hat{\mathcal{W}}_o(t, z)|$, where n^* is the total number of observed covariate patterns. The computation will be intensive if n^* is very large, in which case one may have to define the test statistic to be the supremum over the follow-up time and over some representative values in the covariate space.

Since non-censored data may be regarded as a special case of censored data, we in fact provide an omnibus test for any parametric regression model. To our knowledge, such tests have not been available.

For the one-sample problem, Andersen *et al.* (1993, pp. 456–471) considered the goodness-of-fit process $\mathcal{Y}(t) = n^{-1/2} \sum_{i=1}^{n} \int_0^t K(v; \hat{\gamma})\, d\hat{M}_i(v)$, where K is a weight process. Realizing that $\mathcal{Y}(\cdot)$ does not have an independent increment structure asymptotically (except for the one-parameter case with $K(t; \gamma) = \partial \log \lambda_0(t; \gamma)/\partial\gamma$), these authors suggested that $\mathcal{Y}(\cdot)$ be substracted from its compensator to achieve a Brownian motion limit. As an alternative approach, one may use the compensator to achieve a Brownian motion limit. As an alternative approach, one may use the ideas presented in section 4.1 to simulate the null distribution of $\mathcal{Y}(\cdot)$ directly. In fact, $\mathcal{W}_o$ reduces to $\mathcal{Y}$ in the one-sample case under $K = 1$.

4.3. Checking specific model components

In this subsection, we demonstrate how certain one-dimensional cumulative sums of the martingale residuals may be used to detect specific departures from the assumed model. The general results presented in section 4.1 will enable us to assess the goodness of fit in a more formal fashion than the conventional residual analysis.

It is natural to plot the cumulative sum of $\hat{M}_i$ $(i = 1, \ldots, n)$ against a covariate to check its functional form. The partial-sum process $\mathscr{W}^{(j)}(x) = n^{-1/2} \Sigma_{i=1}^{n} I(Z_{ji} \leqslant x)\hat{M}_i$ for the jth covariate component is a special case of $\mathscr{W}_z(t, z)$ with $f(\cdot) = 1$, $t = \infty$ and $z_k = \infty$ for all $k \neq j$. According to the results of section 4.1, the null distribution of $\mathscr{W}^{(j)}$ can be approximated through simulating the corresponding zero-mean Gaussian process $\hat{\mathscr{W}}^{(j)}$. To assess how unusual the observed residual pattern is under the assumed model, one may plot it along with a few, say 20, realizations from the distribution of the $\hat{\mathscr{W}}^{(j)}$ process. To further enhance the objectivity of the new graphical technique, one may complement the residual plot with the P-value for the supremum test $\sup_x |\mathscr{W}^{(j)}(x)|$.

A simple way of checking the link function (e.g. the exponential regression form of model (2.1) or the linearity of model (2.2)) is to plot the residuals against the fitted value $\hat{\beta}' Z_i$ $(i = 1, \ldots, n)$, or more formally to consider the following special case of the $\mathscr{W}_r(\cdot, \cdot)$ process,

$$\mathscr{W}_l(r) = n^{-1/2} \sum_{i=1}^{n} I(\hat{\beta}' Z_i \leqslant r)\hat{M}_i.$$

The null distribution of this process can be approximated by the zero-mean Gaussian process $\hat{\mathscr{W}}_l(\cdot)$. Again, one may plot the observed process along with a few realizations of $\hat{\mathscr{W}}_l$, and supplement the graphical display with an estimated P-value for the supremum test $\sup_r |\mathscr{W}_l(r)|$.

To provide some insights into the aforementioned procedures, we consider the PH model $\lambda(t \mid Z, \theta_0) = \lambda_0(t; \gamma_0) \exp(\beta_0 Z)$, where Z is one-dimensional. Suppose that the true regression function is $\psi(Z)$ rather than $\exp(\beta_0 Z)$. It then follows from (4.4) that $n^{-1/2}\mathscr{W}(x)$ converges to

$$\int_{x=-\infty}^{z} \int_{v=0}^{\infty} E\{Y(v) \mid x\}\{\lambda_0(v; \gamma_0)\psi(x) - \lambda_0(v; \gamma^*) \exp(\beta^* x)\}\, dv\, dL(x). \tag{4.5}$$

Apart from the discrepancy between γ^* and γ_0, (4.5) compares the true and assumed regression functions at each covariate value. Note that (4.5) will be non-zero for some z unless

$$\psi(x) = \exp(\beta^* x) \frac{\int_0^{\infty} E\{Y(v) \mid x\}\lambda_0(v; \gamma^*)\, dv}{\int_0^{\infty} E\{Y(v) \mid x\}\lambda_0(v; \gamma_0)\, dv} \tag{4.6}$$

for all x. If λ_0 is exponential, then the right-hand side of (4.6) is $\exp(\beta^* x)(\gamma^*/\gamma_0)$, which clearly cannot be equal to $\psi(x)$ for all x.

The partial-likelihood score process has been known to be informative about the proportional hazards assumption under the Cox model. Here we show that a similar process may serve the same purpose in the parametric case. Under model (2.1), the score function for β_0 evaluated at time t and at the true parameter value θ_0 is $\mathscr{U}(\theta_0, t) = n^{-1/2} \Sigma_{i=1}^{n} Z_i M_i(t)$. Clearly, $\mathscr{U}(\hat{\theta}, t)$ is a special case of $\mathscr{W}_z(t, z)$ with $z = \infty$ and $f(x) = x$. Thus the results described in section 4.1 can be used to simulate the null distribution of $\mathscr{U}(\hat{\theta}, t)$. One may then

plot the p standardized score components $\{\mathscr{I}^{-1}(\hat{\theta})_{jj}\}^{1/2}\mathscr{U}_j(\hat{\theta}, t)$ $(j = 1, \ldots, p)$, and use the supremum test statistics $\sup_t |\mathscr{U}_j(\hat{\theta}, t)|$ $(j = 1, \ldots, p)$ and $\sup_t \Sigma_{j=1}^{p} \{\mathscr{I}^{-1}(\hat{\theta})_{jj}\}^{1/2}|\mathscr{U}_j(\hat{\theta}, t)|$.

To see why the $\mathscr{U}$ process is informative about the hazard ratio, let us assume that the true model is $\lambda(t \mid Z) = \lambda_0(t; \gamma_0) \exp(\zeta(t)'Z)$ where $\zeta(t)$ is not time-invariant. It is easy to see that $n^{-1/2}\mathscr{U}(\hat{\theta}, t)$ converges to

$$\int_{v=0}^{t} \int_{x=-\infty}^{\infty} xE\{Y(v) \mid x\}\{\lambda_0(v; \gamma_0) \exp(\zeta(v)'x) - \lambda_0(v; \gamma^*) \exp(\beta'^*x)\}\, dL(x)\, dv. \tag{4.7}$$

Expression (4.7) characterizes the (cumulative) difference between the true and assumed hazard ratios at time t albeit that γ^* may be different from γ_0. By the defining property of θ^*, (4.7) is zero at $t = \infty$. It is highly unlikely that (4.7) is zero for every t. This is particularly clear if λ_0 is exponential, in which case $\lambda_0(t; \gamma_0)/\lambda_0(t; \gamma^*)$ is a constant.

All of the procedures described in this subsection are easy to implement because the maximization of a one-dimensional process is only an n^2 operation. Although these marginal residual plots are helpful in model building, they should not be interpreted without caution since mis-specification of one model component may be reflected in the residual plots for other components. Strictly speaking, a significant result for a (one-dimensional) supremum test implies that there is mis-specification for some aspect(s) of the model, not necessarily the component that is being checked.

5. Simulation studies

Extensive Monte Carlo studies were carried out to investigate the finite-sample behaviour of the supremum tests described in section 3. The proposed approach was used to test the exponentiality under the PH model with a standard normal covariate. For power assessement, failure times were generated from the Weibull distribution with survival function $S(t) = \exp(-t^{\gamma})$ and the log-normal distribution with survival function $S(t) = 1 - \Phi(\sigma^{-1} \log t)$, where Φ is the standard normal distribution function. Censorship was imposed by generating independent uniform random variables on the interval $(0, c)$, where c was a suitably chosen real number such that observations in each simulation sample had a desired probability of being censored.

For comparisons, we also evaluated Hjort's test. Hjort (1990) considered a goodness-of-fit process similar to our $W(\cdot)$, but he replaced the maximum partial likelihood estimator $\tilde{\beta}$ in the Breslow estimator (2.6) by the maximum likelihood estimator $\hat{\beta}$ and discretized the process to form the chi-squared test. He proposed a class of tests with various choices of a weight function $K_n(\cdot)$. In our studies, we let $K_n(\cdot) = 1$. When estimating the variance–covariance matrix for the chi-squared test, we estimated the baseline hazard function by the Breslow-type estimator, though one might use the maximum likelihood estimator instead. (The variance estimator may be negative no matter how λ_0 is estimated). To perform Hjort's test, one needs to partition the time axis into several cells, say $I_j = (a_{j-1}, a_i]$ $(j = 1, \ldots, m)$, where $a_0 = 0$. Due to the instability at the right tail, we let a_m be the last uncensored failure time, which turned out to be a much better choice than the last observation time. It is less clear how one should choose $a_1, \ldots, a_{m-1}$. A common practice for this type of problem is to choose a partition such that there are (roughly) equal numbers of observed failures among the cells. Our simulations indicated that such data-dependent partitions would not yield proper tests. Following Hollander & Peña (1992), we divided the time axis in such a way that the cells have the same expected number of failures under the null hypothesis. The chi-squared test is expected to have the best control of the type I error under this partition, though such a partition would be hard to obtain in practice since it requires prior knowledge

Table 1. *Monte Carlo estimates for the sizes/powers of the supremum and chi-squared tests at the 0.05 level for testing the exponentiality under the model* $\lambda(t \mid Z) = \lambda_0(t) \exp(0.3Z)$

True distributions of $\lambda_0(\cdot)$	Tests	$n = 50$			$n = 100$			$n = 200$		
		CP = 0.25	0.5	0.75	CP = 0.25	0.5	0.75	CP = 0.25	0.5	0.75
Unit exponential	sup	0.06	0.06	0.04	0.06	0.07	0.04	0.06	0.07	0.05
	χ^2	0.07	0.09	0.15	0.05	0.06	0.07	0.04	0.04	0.04
Weibull ($\gamma = 0.5$)	sup	0.96	0.73	0.36	1.00	0.92	0.64	1.00	0.97	0.87
	χ^2	0.87	0.26	0.20	0.99	0.51	0.12	1.00	0.84	0.38
Weibull ($\gamma = 2$)	sup	0.98	0.84	0.35	1.00	0.99	0.72	1.00	1.00	0.94
	χ^2	0.97	0.87	0.46	1.00	0.99	0.63	1.00	1.00	0.82
log-normal($\sigma^2 = 0.5$)	sup	0.84	0.65	0.36	0.99	0.88	0.69	1.00	0.98	0.91
	χ^2	0.92	0.87	0.51	0.99	0.98	0.64	1.00	1.00	0.81
log-normal ($\sigma^2 = 3$)	sup	0.75	0.30	0.07	0.97	0.48	0.10	1.00	0.74	0.12
	χ^2	0.62	0.29	0.20	0.93	0.39	0.08	1.00	0.69	0.02

Note: Z is a standard normal variable. The supremum and chi-squared tests are denoted by sup and χ^2, respectively. CP stands for censoring probability. The estimates for the sizes and powers are based on 5000 and 1000 replicates of data, respectively. For each replicate of data, 1000 realizations of the $\hat{W}$ process are generated to calculate the P-value for the supremum test. The rejection proportions are based on the simulated P-values. (The null hypothesis is rejected if the simulated P-value is less than 0.05.) The chi-squared test uses the 95% percentile of the χ^2_1 distribution. The data sets for which the variance estimates of the chi-squared test are negative are excluded from its size/power estimation. The random number generator of Wichmann & Hill (1982) is used.

of unknown parameters. The choise for the number of cells m is not an easy task, either. In our studies, we let $m = 3$, 2 and 1 under 25%, 50% and 75% censoring probabilities, respectively. Finer partitioning would result in severe anti-conservativeness for the sample sizes considered here.

The main results from the aforementioned studies are summarized in Table 1. The supremum test maintains its size near the nominal level, even for sample size of 50 with heavy censoring. The chi-squared test does not have proper size in small samples with substantial censoring; the distortion is much worse for other partitions not shown in the table. The supremum test tends to have higher power than the chi-squared test although the latter appears to be more powerful in some small samples. (Of course, one needs to be cautious when comparing the powers of two tests with different sizes.)

We also conducted extensive numerical studies on the martingale-residuals methods developed in section 4. The results showed that the proposed supremum tests have proper sizes for practical samples and are sensitive to model mis-specification. Here we report some Monte Carlo estimates for $n = 50$, 25% (uniform) censoring and significance level of 0.05. (The estimates for the sizes and powers were based on 5000 and 1000 replicates of data, respectively.) Under the exponential model considered in Table 1, the sizes of the omnibus test $\sup_{t,z} |\mathcal{W}_o(t, z)|$, the functional form test $\sup_x |\mathcal{W}^{(1)}(x)|$ (or equivalently the link function test $\sup_r |\mathcal{W}_l(r)|$) and the PH test $\sup_t |\mathcal{U}(\hat{\theta}, t)|$ were estimated at 0.06, 0.04 and 0.05, respectively. If the true distributions are Weibull rather than exponential, the omnibus test has estimated powers of 0.96 and 0.99 for $\gamma = 0.5$ and 2, respectively. When Z^2 is omitted from the true model $\lambda(t \mid Z) = \exp(0.3Z - 0.6Z^2)$, the estimated powers for the functional form test and the omnibus test are 0.74 and 0.47, respectively. Suppose now that Z takes the values 0 and 1 with equal probability and that T is unit exponential under $Z = 0$ and Weibull with $\gamma = 2$ under $Z = 1$. Then the estimated powers for the PH and omnibus tests are 0.88

and 0.60, respectively. In the last two cases, the estimated powers for Hjort's tests are only 0.15 and 0.46, respectively.

6. Real examples

We now illustrate the proposed methods with the well-known Stanford heart transplant data reported in Miller & Halpern (1982). The data set contains the survival times of 184 patients who received heart transplants along with their ages at the time of the first transplant and T5 mismatch scores. As in Miller & Halpern (1982), the 27 patients who did not have T5 mismatch scores are excluded from our analysis. Out of the remaining 157 patients, 55 were censored as of the date of data listings.

One of the final models obtained by Miller & Halpern (1982) is the PH model with two covariates, age and age^2. We are interested in assessing whether this age quadratic model adequately characterizes the dependence of survival time on age and in ascertaining the form of the failure time distribution. Table 2 presents the parameter estimates under the exponential and Weibull distributions as well as the Cox semi-parametric model. We centre age at its sample mean 41.7 for each interpretation. (A zero survival time is set to 0.5.) The Weibull model takes the form $\lambda(t \mid Z, \beta_0, \alpha_0, \rho_0) = \alpha_0 \rho_0 t^{\rho_0 - 1} \exp(\beta_0' Z)$, reducing to the exponential model under $\rho_0 = 1$.

Figure 1 shows the Breslow estimate for the baseline survival function and the maximum likelihood estimate under the exponential and Weibull models. Under the former model, the observed value for the supremum test statistic Q is $(157)^{1/2}0.18$ with P-value of 0.0015, whereas the observed statistic is $(157)^{1/2}0.07$ with P-value of 0.228 under the latter. The

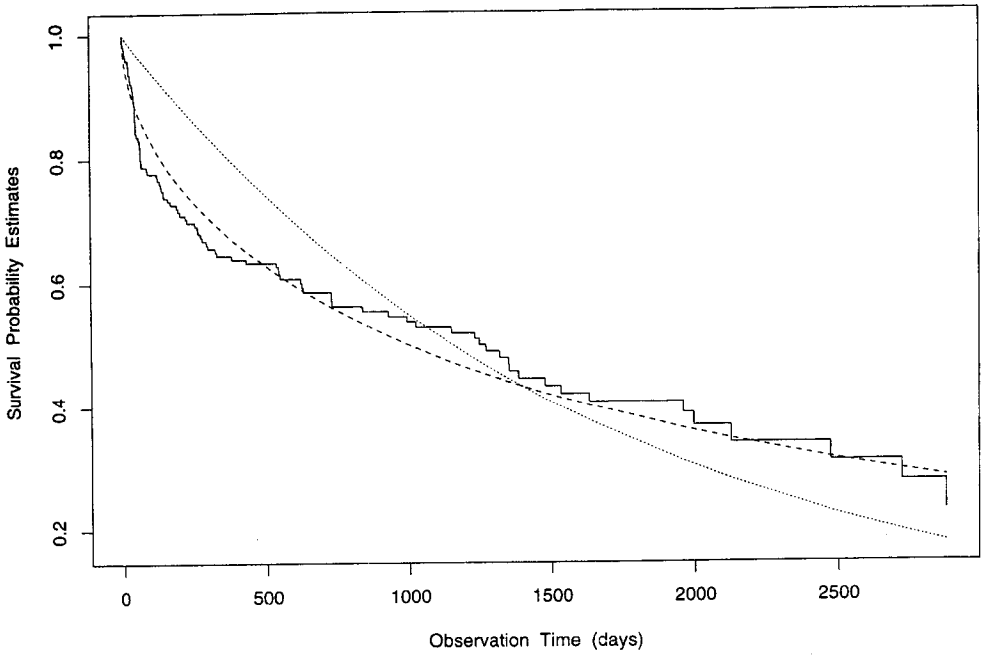

Fig. 1. Estimates of the baseline survival function under age quadratic PH models for the Stanford heart transplant data. The estimates under the Cox, exponential and Weibull models are shown by the solid, dotted and dashed curves, respectively.

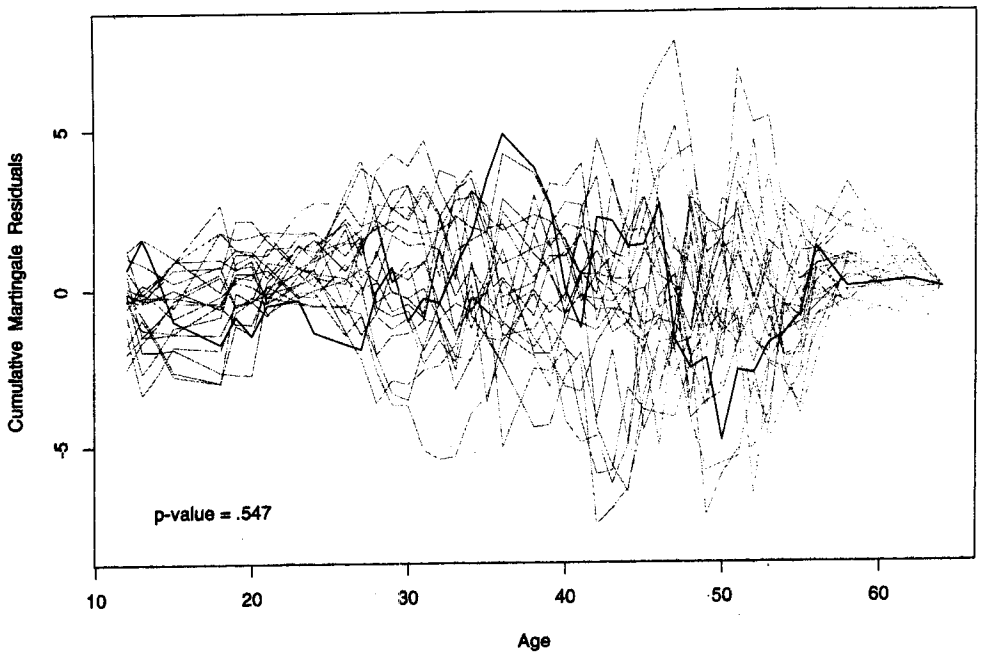

Fig. 2. Plot of (non-normalized) cumulative martingale residuals vs age under the age quadratic Weibull model for the Stanford heart transplant data. The observed process is shown by the solid curve, and 20 simulated processes are shown by light dotted curves. The estimated P-value for the supremum test of the functional form is based on 10 000 realizations.

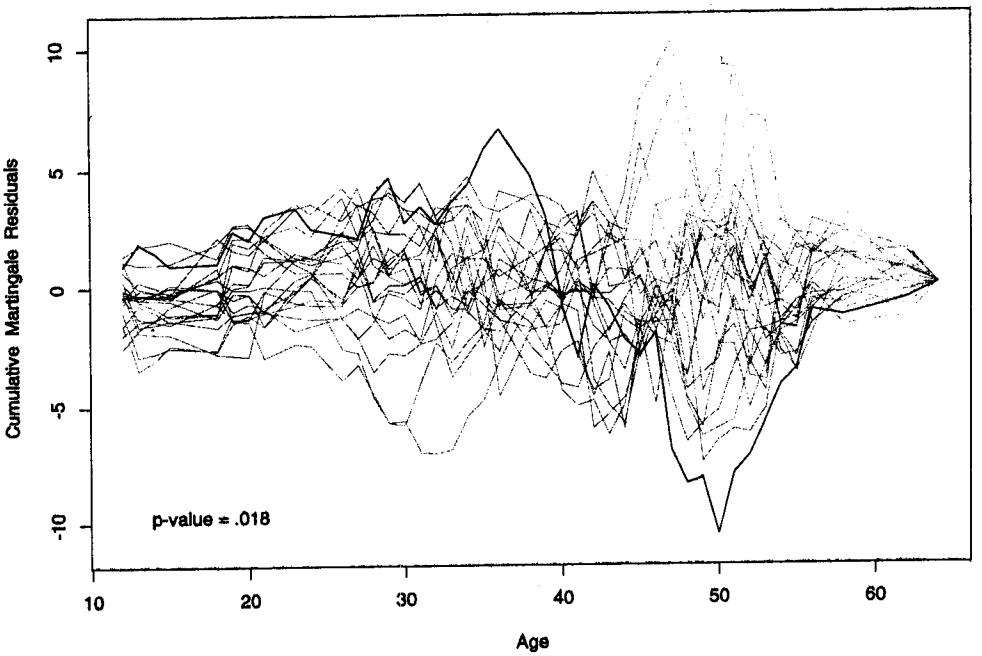

Fig. 3. Plot of (non-normalized) cumulative martingale residuals vs age under the age linear Weibull model for the Stanford heart transplant data. The observed process is shown by the solid curve, and 20 simulated processes are shown by light dotted curves. The estimated P-value for the supremum test of the functional form is based on 10 000 realizations.

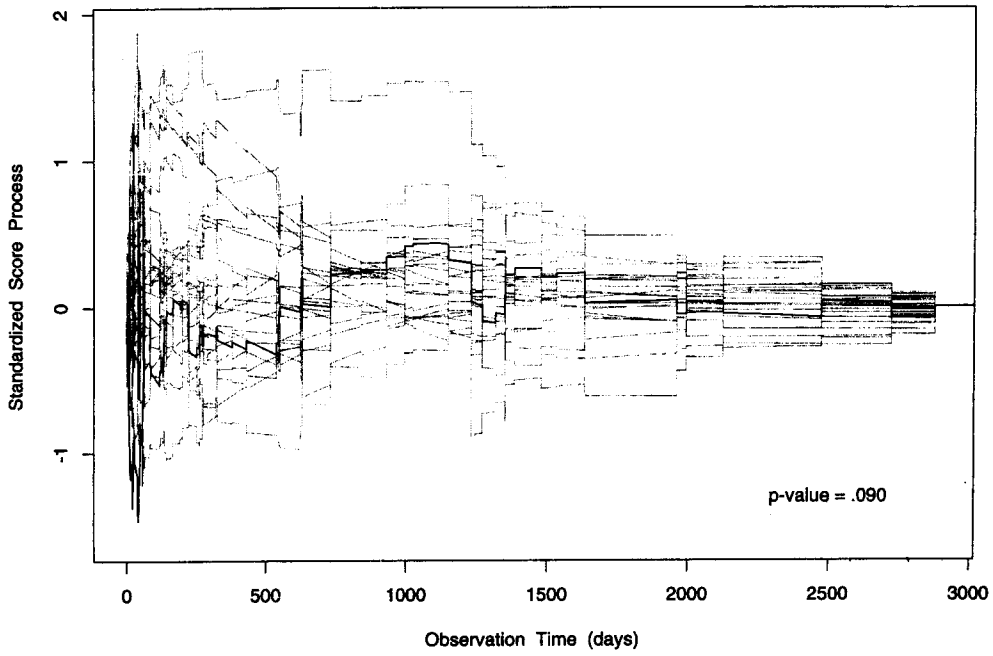

Fig. 4. Plot of standardized score process versus time with respect to age under the age quadratic Weibull model for the Stanford heart transplant data. The observed process is shown by the solid curve, and 20 simulated processes are shown by light dotted curves. The estimated P-value for the supremum test of the PH assumption is based on 10 000 realizations.

omnibus tests yield P-values of <0.0001 and 0.041 for the exponential and Weibull models, respectively. (The p-values reported in this section are based on 10 000 replications, though 1000 replications are sufficient for practical purposes.) These results suggest that the Weibull baseline distribution is adequate but there is some mis-specification for the deterministic part of the model.

Figure 2 plots the (non-normalized) cumulative sum of the martingale residuals against age, i.e. $\Sigma_{i=1}^{n} I(Z_{1i} \leqslant \cdot)\hat{M}_i$, in the age quadratic Weibull model, which indicates that the functional form for age is appropriate. The link function test yields a P-value of 0.318. Thus, the exponential regression function with age and age^2 seems satisfactory.

Figure 3 displays the residual plot for age when age^2 is deliberately omitted from the model. The age linear model vastly overestimates the hazard rates for ages between 40 and 50, implying that the age effect is non-linear.

Since the omnibus test indicated that the age quadratic Weibull model is imperfect while the tests for the distributional form, the functional form and the link function did not reveal any problems, one suspects that the proportional hazards assumption may fail. Indeed, the P-values for the proportional hazard tests are 0.090 and 0.041 for age and age^2, respectively. As shown in Fig. 4, the lack of proportionality for age lies in the first 50 days. A similar plot reveals lack of proportionality for age^2 in the first 100 days. In conclusion, the age quadratic Weibull model given in Table 2 provides a fairly good discription of the data, though the proportional hazards assumption is not satisfactory in the early follow-up.

We now briefly mention another example. The Cox model shown in Table 3 is the Mayo Clinic model for primary biliary cirrhosis (PBC) studied extensively by Fleming & Harrington (1991) and Lin *et al.* (1993). This model has been extremely useful in counselling patients

Table 2. *Parameter estimates for the regression analysis of the Stanford heart transplant data*

	Models		
Parameters	Exponential	Weibull	Cox
age—41.7	0.0626	0.0472	0.0446
	(0.0114)	(0.0109)	(0.0109)
$(\text{age}-41.7)^2$	0.0032	0.0023	0.0022
	(0.0007)	(0.0007)	(0.0007)
$\log \alpha_0$	−7.4214	−4.3052	—
	(0.1386)	(0.3618)	—
$\log \rho_0$	—	−0.5628	—
	—	(0.0836)	—

Note: The standard error estimates are shown in parentheses. The logarithmic transformations are taken on α_0 and ρ_0 to remove the range restrictions.

Table 3. *Parameter estimates for the regression analysis of the Mayo PBC Data*

	Models	
Parameters	Cox	Weibull
age	0.0394	0.0389
	(0.0077)	(0.0076)
log (albumin)	−2.5328	−2.4526
	(0.6482)	(0.6392)
log (bilirunin)	0.8707	0.8476
	(0.0826)	(0.0799)
edema	0.8592	0.9258
	(0.2711)	(0.2671)
log (protime)	2.3797	2.5822
	(0.7666)	(0.7436)
$\log \alpha_0$	—	−12.3298
	—	(0.7279)
$\log \rho_0$	—	0.3846
	—	(0.0632)

Note: see Note of Table 2.

and in understanding the course of PBC for untreated patients. We would like to characterize the baseline distribution with a parametric function, which would deepen our understanding of the natural history and ease the applications of the model (especially in predicting survival experience for future patients). Again, we centre the covariates at their respective sample means for better interpretability. The P-values for testing the exponential and Weibull assumptions are 0.0001 and 0.061, respectively. Figure 5 indicates that there are considerable discrepancies between the Breslow and parametric estimates on the right tail. When restricted to the time interval between 0 and 9.8 years, the P-value for the Weibull assumption

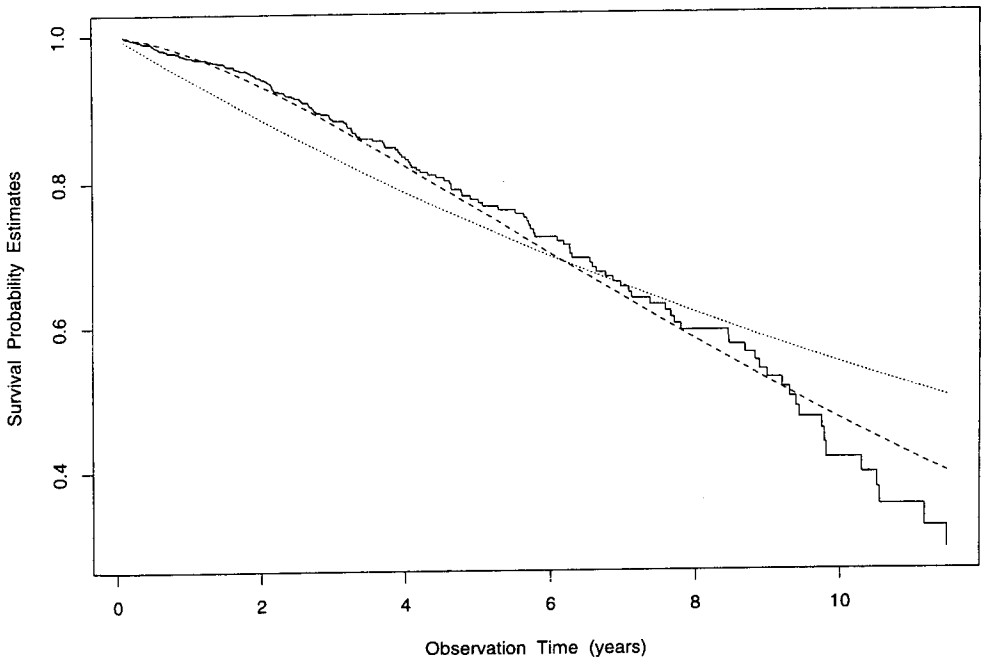

Fig. 5. Estimates of the baseline survival function for the Mayo PBC data. The estimates under the Cox, exponential and Weibull models are shown by the solid, dotted and dashed curves, respectively.

jumps to 0.61. As shown in Table 3, the Weibull and Cox models provide similar estimates for the covariate effects, though the standard error estimates (for the regression parameters) are smaller under the Weibull model. A practical advantage of the Weibull model over the Cox model is that it can be easily reproduced. (To reproduce the Cox model, one has to know the Breslow estimates at all 257 uncensored failure time points.)

References

Akritas, M. G. (1988). Pearson-type goodness-of-fit tests: the univariate case. *J. Amer. Statist. Assoc.* **83,** 222–230.

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

Barlow, W. E. & Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika* **75,** 65–74.

Breslow, N. (1972). Contribution to the discussion of the paper by D. R. Cox. *J. Roy. Statist. Soc. Ser. B* **34,** 216–217.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34,** 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62,** 269–276.

Cox, D. R. & Oakes, D. (1984). *Analysis of survival data.* Chapman & Hall, London.

Fleming, T. R. & Harrington, D. P. (1991). *Counting processes and survival analysis.* Wiley, New York.

Gray, R. J. & Pierce, D. A. (1985). Goodness-of-fit tests for censored survival data. *Ann. Statist.* **13,** 552–563.

Habib, M. G. & Thomas, D. R. (1986). Chi-squared goodness-of-fit tests for randomly censored data. *Ann. Statist.* **14,** 759–765.

Hjort, N. L. (1990). Goodness of fit tests in models for life history data based on cumulative hazard rates. *Ann. Statist.* **18,** 1221–1258.

Hjort, N. L. (1992). On inference in parametric survival data models. *Int. Statist. Rev.* **60,** 355–387.

Hollander, M. & Peña, E. A. (1992). A chi-squared goodness-of-fit test for randomly censored data. *J. Amer. Statist. Assoc.* **87,** 458–463.

Kalbfleisch, J. D. & Prentice, R. L. (1980). *The statistical analysis of failure time data.* Wiley, New York.
Lawless, J. F. (1982). *Statistical models and methods for lifetime data.* Wiley, New York.
Li, G. & Doss, H. (1993). Generalized Pearson–Fisher chi-squared goodness-of-fit tests, with applications to models with life history data. *Ann. Statist.* **21,** 772–797.
Lin, D. Y. & Geyer, C. J. (1992). Computational methods for semiparametric linear regression with censored data. *J. Computat. Graphical Statist.* **1,** 77–90.
Lin, D. Y. & Ying, Z. (1995). Semiparametric inference for the accelerated life model with time-dependent covariates. *J. Statist. Planning Inference*, in press.
Lin, D. Y., Wei, L. J. & Ying, Z. (1992). Checking the Cox model with cumulative sums of martingale-based residuals. Technical Report 111, University of Washington, Department of Biostatistics.
Lin, D. Y., Wei, L. J. & Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80,** 557–572.
Loynes, R. M. (1980). The empirical distribution function of residuals from generalised regression. *Ann. Statist.* **8,** 285–298.
Miller, R. & Halpern, J. (1982). Regression with censored data. *Biometrika* **69,** 521–531.
Pierce, D. A. & Kopecky, K. J. (1979). Testing goodness of fit for the distribution of errors in regression models. *Biometrika* **66,** 1–5.
Robins, J. M. & Tsiatis, A. A. (1992). Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika* **79,** 311–319.
Shorak, G. R. & Wellner, J. A. (1986). *Empirical processes with applications to statistics.* Wiley, New York.
Therneau, T. M., Grambsch, P. M. & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* **77,** 147–160.
Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18,** 354–372.
Wei, L. J., Ying, Z. & Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77,** 845–851.
Wichmann, B. A. & Hill, I. D. (1982). An efficient and portable pseudo-random number generator. *Appl. Statist.* **31,** 188–190.
Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21,** 76–99.

Received August 1994, in final form April 1995

D. Y. Lin, Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195, USA

Appendix. Weak convergence of $\hat{W}(\cdot)$

We shall prove that the conditional distribution of the process $\hat{W}(\cdot)$ given $\{Y_i(\cdot), N_i(\cdot), Z_i(\cdot)\}$ $(i = 1, \ldots, n)$ is the same in the limit as the (unconditional) distribution of $W(\cdot)$. Conditional on $\{Y_i(\cdot), N_i(\cdot), Z_i(\cdot)\}$ $(i = 1, \ldots, n)$, $\hat{W}(t)$ is a sum of n independent zero-mean random variables for every t. By the Lindeberg–Feller theorem, the finite-dimensional distribution of $W(\cdot)$ is asymptotically zero-mean normal under the regularity conditions stated in sections 2.1 and 2.2. Let us write $\hat{W}(t) = \hat{W}_1(t) - \hat{W}_2(t) - \hat{W}_3(t)$, where $\hat{W}_1(t)$, $\hat{W}_2(t)$ and $\hat{W}_3(t)$ are the three terms on the right side of (3.6). For notational simplicity, assume for the moment that γ is one-dimensional, in which case the last component of the vector $\mathscr{I}^{-1}(\hat{\theta})n^{-1/2}$ $\Sigma_{i=1}^n \int_0^\infty \partial \log \lambda(v \mid Z_i, \hat{\theta})/\partial\theta \, dN_i(v)G_i$ can be expressed as $n^{-1/2}\Sigma_{i=1}^n c_i G_i$, where the c_i are bounded scalars. Then for $0 \leqslant t_1 \leqslant t \leqslant t_2 \leqslant \tau$,

$$
\begin{aligned}
&E[|\{\hat{W}_1(t) - \hat{W}_1(t_1)\}\{\hat{W}_1(t_2) - \hat{W}_1(t)\}|] \\
&\quad = E\left\{\left|\int_{t_1}^{t} \partial\lambda_0(v;\hat{\gamma})/\partial\gamma \, dv \int_{t}^{t_2} \partial\lambda_0(v;\hat{\gamma})/\partial\gamma \, dv\left(n^{-1/2}\sum_{i=1}^{n} c_i G_i\right)^2\right|\right\} \\
&\quad = \left|\int_{t_1}^{t} \partial\lambda_0(v;\hat{\gamma})/\partial\gamma \, dv\right| \left|\int_{t}^{t_2} \partial\lambda_0(v;\hat{\gamma})/\partial\gamma \, dv\right| n^{-1}\sum_{i=1}^{n} c_i^2 \leqslant \mathscr{K}(t - t_1)(t_2 - t),
\end{aligned}
$$

where $0 < \tilde{\mathscr{K}} < \infty$ is some constant. Similar moment inequalities can be established for $\hat{W}_2(\cdot)$ and $\hat{W}_3(\cdot)$. Therefore, $\hat{W}(t)$ $(0 \leqslant t \leqslant \tau)$ converges weakly to a zero-mean Gaussian process (Shorack & Wellner, 1986, th. 2.3.6). Now, the conditional expectation

$$E\{\hat{W}_1(t)\hat{W}_1(t^\dagger)\} = \hat{g}'(t)\mathscr{I}^{-1}(\hat{\theta})\left[n^{-1}\sum_{i=1}^{n}\int_0^\infty \{\partial \log \lambda(v \mid Z_i, \hat{\theta})/\partial\theta\}^{\otimes 2}\, dN_i(v)\right]\mathscr{I}^{-1}(\hat{\theta})\hat{g}(t^\dagger).$$

By the strong law of large numbers, $\mathscr{I}(\theta_0) \xrightarrow{\text{a.s.}} \Omega$ and $n^{-1}\Sigma_{i=1}^{n}\int_0^\infty \{\partial \log \lambda(v \mid Z_i, \theta_0)/\partial\theta\}^{\otimes 2}\, dN_i(v) \xrightarrow{\text{a.s.}} \Omega$. It then follows from the consistency of $\hat{\theta}$ and the regularity of $\lambda(v \mid Z, \theta)$ that $E\{\hat{W}_1(t)\hat{W}_1(t^\dagger)\} \xrightarrow{\text{a.s.}} g'(t)\Omega^{-1}g(t^\dagger)$. Continuing this line of calculations for the other terms in $E[\{\hat{W}_1(t) - \hat{W}_2(t) - \hat{W}_3(t)\}\{\hat{W}_1(t^\dagger) - \hat{W}_2(t^\dagger) - \hat{W}_3(t^\dagger)\}]$, we find that the conditional expectation $E\{\hat{W}(t)\hat{W}(t^\dagger)\} \xrightarrow{\text{a.s.}} \xi(t, t^\dagger)$, where $\xi(t, t^\dagger)$ is the limiting covariance function of $W(\cdot)$ given in (3.5). This completes our proof.