Checking the Cox model with cumulative sums of martingale-based residuals

By D. Y. LIN

Department of Biostatistics, SC-32, University of Washington, Seattle, Washington 98195, U.S.A.

L. J. WEI

Department of Biostatistics, Harvard University, 677 Huntington Ave., Boston, Massachusetts 02115, U.S.A.

AND Z. YING

Department of Statistics, 101 Illini Hall, University of Illinois, Champaign, Illinois 61820, U.S.A.

Summary

This paper presents a new class of graphical and numerical methods for checking the adequacy of the Cox regression model. The procedures are derived from cumulative sums of martingale-based residuals over follow-up time and/or covariate values. The distributions of these stochastic processes under the assumed model can be approximated by zero-mean Gaussian processes. Each observed process can then be compared, both visually and analytically, with a number of simulated realizations from the approximate null distribution. These comparisons enable the data analyst to assess objectively how unusual the observed residual patterns are. Special attention is given to checking the functional form of a covariate, the form of the link function, and the validity of the proportional hazards assumption. An omnibus test, consistent against any model misspecification, is also studied. The proposed techniques are illustrated with two real data sets.

Some key words: Censoring; Goodness of fit; Link function; Omnibus test; Proportional hazards; Regression diagnostic; Residual plot; Survival data.

1. Introduction

The proportional hazards model (Cox, 1972) with the partial likelihood principle (Cox, 1975) has become exceedingly popular for the analysis of failure time observations. This model specifies that the hazard function for the failure time T associated with a $p \times 1$ vector of covariates Z takes the form of

$$\lambda(t; Z) = \lambda_0(t) \exp(\beta_0' Z), \tag{1·1}$$

where $\lambda_0(.)$ is an unspecified baseline hazard function, and β_0 is a $p \times 1$ vector of unknown regression parameters.

Let C denote the censoring time. Assume that Z is bounded and that T and C are independent conditional on Z. Suppose that the data consist of n independent replicates of (X, Δ, Z), where $X = \min(T, C)$, $\Delta = I(T \leqslant C)$, and $I(.)$ is the indicator

function. Then the partial likelihood score function for β_0 is

$$U(\beta) = \sum_{i=1}^{n} \Delta_i \{Z_i - \bar{Z}(\beta, X_i)\}, \tag{1·2}$$

where

$$\bar{Z}(\beta, t) = \sum_{l=1}^{n} Y_l(t) \exp(\beta' Z_l) Z_l \Big/ \sum_{l=1}^{n} Y_l(t) \exp(\beta' Z_l), \quad Y_l(t) = I(X_l \geqslant t).$$

For future reference, we denote the denominator of $\bar{Z}(\beta, t)$ by $S^{(0)}(\beta, t)$. The maximum partial likelihood estimator $\hat{\beta}$ is the solution to the estimating equation $U(\beta) = 0$. Under some mild regularity conditions (Andersen & Gill, 1982), the random vector $\mathscr{I}^{\frac{1}{2}}(\hat{\beta})(\hat{\beta} - \beta_0)$ is asymptotically zero-mean normal with an identity covariance matrix, where $\mathscr{I}(\beta)$ is minus the derivative matrix of $U(\beta)$.

Model (1·1) may fail in three ways: (i) the proportional hazards assumption, viz. the time invariance of the hazard ratio $\lambda(t; Z)/\lambda_0(t)$, does not hold; (ii) the functional forms of individual covariates in the exponent of the model are misspecified; (iii) the link function, viz. the exponential form for the hazard ratio, is inappropriate. The model misspecification can have detrimental effects on the validity and efficiency of the partial likelihood inference (Lagakos & Schoenfeld, 1984; Struthers & Kalbfleisch, 1986; Lagakos, 1988; Lin & Wei, 1989).

Numerous graphical and analytical methods have been suggested for checking the adequacy of Cox models. A partial review of the contributions in this area was given by Lin & Wei (1991). Many of the existing methods are related to the so-called martingale residuals.

To describe the martingale residuals, we define the counting processes $N_i(t) = \Delta_i I(X_i \leqslant t)$ $(i = 1, \ldots, n)$. These processes have the intensity functions $Y_i(t)\lambda_0(t) \exp(\beta_0' Z_i)$ $(i = 1, \ldots, n)$. The differences between the counting processes and their respective integrated intensity functions,

$$M_i(t) = N_i(t) - \int_0^t Y_i(u) \exp(\beta_0' Z_i) \lambda_0(u)\, du \quad (i = 1, \ldots, n),$$

are martingales. The martingale residuals are defined as

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(u) \exp(\hat{\beta}' Z_i)\, d\hat{\Lambda}_0(u) \quad (i = 1, \ldots, n),$$

where

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{l=1}^{n} dN_l(u)}{S^{(0)}(\hat{\beta}, u)}.$$

For convenience, we denote $\hat{M}_i(\infty)$ by $\hat{M}_i$. The martingale residual $\hat{M}_i(t)$ can be interpreted as the difference at time t between the observed and expected numbers of events for the ith subject. These residuals have some properties reminiscent of ordinary residuals in linear models. Most notably, for any t, $\sum \hat{M}_i(t) = 0$, where the summation is over the range $i = 1, \ldots, n$, and

$$E\{\hat{M}_i(t)\} \doteq \operatorname{cov}\{\hat{M}_i(t), \hat{M}_j(t)\} \doteq 0 \quad (i \neq j)$$

in large samples.

The score function (1·2) can be written as $U(\beta, \infty)$, where

$$U(\beta, t) = \sum_{i=1}^{n} \int_0^t \{Z_i - \bar{Z}(\beta, s)\}\, dN_i(s).$$

Note that $U(\hat{\beta}, t) = \sum Z_i \hat{M}_i(t)$, where the summation is over $i = 1, \ldots, n$, which is a

function of the martingale residuals. We call $U(\hat{\beta}, t) = (U_1(\hat{\beta}, t), \ldots, U_p(\hat{\beta}, t))'$ the empirical score process. The increments in this score process are the well-known partial residuals introduced by Schoenfeld (1982).

The martingale residuals and their transforms can be used to detect model departures (Barlow & Prentice, 1988; Therneau, Grambsch & Fleming, 1990). For example, plots of the martingale residuals against covariate components provide useful clues on appropriate functional forms of covariates in the exponent of the model. Also, plotting the score process versus follow-up time may reveal violation of the proportional hazards assumption. Interpreting the results from such residual plots, however, can be quite challenging. It is often difficult to conclude whether a trend exhibited in a residual plot reflects model misspecification or is a phenomenon that is likely to occur even when the model is correctly specified.

In the present paper, we offer an objective graphical solution to model checking. Our approach is based on various partial-sum processes of the martingale residuals and their transforms. Such processes include

$$U_j(\hat{\beta}, t) \quad (t > 0;\ j = 1, \ldots, p), \quad \sum_{i=1}^{n} I(Z_{ji} \leqslant x)\hat{M}_i \quad (-\infty < x < \infty;\ j = 1, \ldots, p),$$

where Z_{ji} is the jth covariate component for the ith subject. Under the null hypothesis of no model misspecification, the distributions of these processes can be easily approximated through simulating certain Gaussian processes. Each observed process can then be plotted along with a number of realizations from the corresponding Gaussian process. The plots enable the data analyst to assess visually how unusual the observed patterns are. To make the graphical inspection even more objective, some numerical measures for the lack of fit may be attached.

The aforementioned graphical procedures are useful during the stage of model building. In some applications, however, such marginal plots may not be highly informative due to the nonlinear nature of the model and nonorthogonality of covariates. Thus, there is a need for omnibus lack-of-fit tests, which are consistent against any departures from model (1·1). In this paper, we develop such a test based on the martingale residuals.

The new techniques for model checking are described in § 2. The underlying theoretical developments and computational methods are relegated to the appendices. In § 3, we provide illustrations with the Mayo liver disease data (Fleming & Harrington, 1991, pp. 359–75) and the Stanford heart transplant data (Miller & Halpern, 1982). In § 4, generalizations to the settings of time-dependent covariates and other relative risk models are discussed.

2. Model checking techniques

2·1. *General*

It is easier to approximate the distribution of a summary statistic from a group of residuals than those of individual residuals. In this section, we derive diagnostic methods for the Cox model by grouping the martingale-based residuals cumulatively with respect to follow-up time and/or covariate values. These partial sums of residuals are special cases of the following two classes of multi-parameter stochastic processes:

$$W_z(t, z) = \sum_{i=1}^{n} f(Z_i) I(Z_i \leqslant z)\hat{M}_i(t), \tag{2·1}$$

$$W_r(t,r) = \sum_{i=1}^{n} f(Z_i) I(\hat{\beta}' Z_i \leqslant r) \hat{M}_i(t), \tag{2·2}$$

where $f(.)$ is a known smooth function, $z = (z_1, \ldots, z_p)' \in R^p$, and the event $\{Z_i \leqslant z\}$ means that all the p components of Z_i are no larger than the respective components of z. If model (1·1) holds, these processes will fluctuate randomly around zero. In §2·2, we show how to approximate their null distributions.

2·2. *Null distributions of W_z and W_r*

The process $W_z(t,z)$ is a smooth function of $\hat{\beta}$. By the Taylor series expansions of $W_z(t,z)$ and $U(\hat{\beta})$ at β_0 and some simple probabilistic arguments, the process $n^{-\frac{1}{2}} W_z(t,z)$ is seen to be asymptotically equivalent to the process $n^{-\frac{1}{2}} \tilde{W}_z(t,z)$, where

$$\begin{aligned} \tilde{W}_z(t,z) = & \sum_{l=1}^{n} \int_0^t \{ f(Z_l) I(Z_l \leqslant z) - \tilde{g}(\beta_0, u, z) \} \, dM_l(u) \\ & - \sum_{k=1}^{n} \int_0^t Y_k(s) \exp(\beta_0' Z_k) f(Z_k) I(Z_k \leqslant z) \{ Z_k - \tilde{Z}(\beta_0, s) \}' \lambda_0(s) \, ds \\ & \times \mathscr{I}^{-1}(\beta_0) \sum_{l=1}^{n} \int_0^{\infty} \{ Z_l - \tilde{Z}(\beta_0, u) \} \, dM_l(u). \end{aligned} \tag{2·3}$$

In (2·3), $\tilde{Z}(\beta,t)$ is the limit of $\bar{Z}(\beta,t)$ and $\tilde{g}(\beta,t,z)$ is the limit of

$$g(\beta,t,z) = \frac{\sum_{k=1}^{n} Y_k(t) \exp(\beta' Z_k) f(Z_k) I(Z_k \leqslant z)}{\sum_{k=1}^{n} Y_k(t) \exp(\beta' Z_k)}.$$

It is proved in Appendix 1 that $n^{-\frac{1}{2}} \tilde{W}_z(t,z)$ converges to a zero-mean Gaussian process. We now show how to approximate the limiting distribution through Monte Carlo simulations. If we knew the stochastic structure of the martingale process $M_l(u)$, we could easily simulate $\tilde{W}_z$ after replacing the unknown quantities in (2·3) by their respective consistent estimates. The distributional form of $M_l(u)$, however, is unknown. One way to tackle this problem is to replace $M_l(u)$ in (2·3) by a similar process, say $\tilde{M}_l(u)$, which has a known distribution. Note that the variance function of $M_l(u)$ is $E\{N_l(u)\}$ (Fleming & Harrington, 1991, Lemma 2.3.2, Theorem 2.5.3). Thus a natural candidate for $\tilde{M}_l(u)$ is $N_l(u) G_l$, where $N_l(u)$ is the observed counting process and $\{G_l; l = 1, \ldots, n\}$ denotes a random sample of standard normal variables. Upon replacing β_0, $\lambda_0(s)\,ds$, $\tilde{Z}$, $\tilde{g}$ and $\{M_l\}(.)$ in (2·3) by $\hat{\beta}$, $d\hat{\Lambda}_0(s)$, $\bar{Z}$, g and $\{N_l(.) G_l\}$, respectively, we obtain

$$\begin{aligned} \hat{W}_z(t,z) = & \sum_{l=1}^{n} I(X_l \leqslant t) \Delta_l \{ f(Z_l) I(Z_l \leqslant z) - g(\hat{\beta}, X_l, z) \} G_l \\ & - \sum_{k=1}^{n} \int_0^t Y_k(s) \exp(\hat{\beta}' Z_k) f(Z_k) I(Z_k \leqslant z) \{ Z_k - \bar{Z}(\hat{\beta}, s) \}' \, d\hat{\Lambda}_0(s) \\ & \times \mathscr{I}^{-1}(\hat{\beta}) \sum_{l=1}^{n} \Delta_l \{ Z_l - \bar{Z}(\hat{\beta}, X_l) \} G_l. \end{aligned} \tag{2·4}$$

Although individual $M_l(u)$ may not be Gaussian, we show in Appendix 1 that the conditional distribution of $n^{-\frac{1}{2}} \hat{W}_z(t,z)$ given the observed data $\{X_i, \Delta_i, Z_i\}$ is the same in the limit as the unconditional distribution of $n^{-\frac{1}{2}} \tilde{W}_z(t,z)$. In the sequel, $\{X_i, \Delta_i, Z_i\}$ are

regarded as fixed for $\hat{W}_z$. To approximate the distribution of W_z, we simulate a number of realizations from $\hat{W}_z$ by repeatedly generating normal random samples $\{G_l\}$ while holding the observed data $\{X_i, \Delta_i, Z_i\}$ fixed.

Similarly, it is shown in Appendix 2 that in large samples the distribution of $W_r(t, r)$ can be approximated by that of $\hat{W}_r(t, r)$, where $\hat{W}_r(t, r)$ is obtained from (2·4) by substituting $I(\hat{\beta}' Z_l \leqslant r)$ for $I(Z_l \leqslant z)$ $(l = 1, \ldots, n)$. Again, one may approximate the distribution of W_r through simulations.

In §2·3–2·6, we develop model checking techniques by considering some special cases of the W_z and W_r processes. The following notation will be used: a capital letter (e.g. W_z or S_z) refers to an original process or statistic, a small-case letter (e.g. w_z or s_z) to its observed value, and the corresponding quantities under the Gaussian approximations are indicated by '^' (e.g. $\hat{W}_z$, $\hat{w}_z$, $\hat{S}_z$ and $\hat{s}_z$).

2·3. *Checking the functional form of a covariate*

Therneau et al. (1990) showed that a smoothed plot of the $\hat{M}_i$ versus a covariate omitted from the fitted model provides approximately the correct functional form to be placed in the exponent of the Cox model if the omitted covariate is uncorrelated with the covariates in the model. Unfortunately, it is not clear how much confidence one can have in such a scatterplot smoother. In fact, different smoothing techniques or even the same technique with varying values of the smoothing parameter may result in quite distinct smoothers.

Here, we suggest a less subjective approach. Instead of plotting the raw martingale residuals, one plots the partial-sum processes of the $\hat{M}_i$,

$$W_j(x) = \sum_{i=1}^{n} I(Z_{ji} \leqslant x)\hat{M}_i \quad (j = 1, \ldots, p).$$

Note that $W_j(x)$ is a special case of $W_z(t, z)$ with $f(.) = 1$, $t = \infty$ and $z_k = \infty$ (for all $k \neq j$). According to the general results presented in §2·2, the null distribution of $W_j(.)$ can be approximated through simulating the corresponding zero-mean Gaussian process $\hat{W}_j(.)$. To assess how unusual the observed process $w_j(.)$ is under model (1·1), one may plot it along with a few, say 20, realizations from the $\hat{W}_j(.)$ process (see examples in §3).

To further enhance the objectivity of the new graphical technique, one may complement the residual plot with some numerical values which measure the extremity of $w_j(.)$. Since $W_j(.)$ fluctuates randomly around zero under the null hypothesis, a natural numerical measure is $s_j = \sup_x |w_j(x)|$. An unusually large value of s_j would suggest that the functional form for Z_j may be inappropriate. The p-value, $\text{pr}(S_j \geqslant s_j)$, can be approximated by $\text{pr}(\hat{S}_j \geqslant s_j)$, where $\hat{S}_j = \sup_x |\hat{W}_j(x)|$. Note that the calculation of $\text{pr}(\hat{S}_j \geqslant s_j)$ is conditional on $\{X_i, \Delta_i, Z_i\}$. The results in Appendix 1 indicate that $\text{pr}(\hat{S}_j \geqslant s_j)$ converges almost surely to $\text{pr}(S_j \geqslant s_j)$ as $n \to \infty$. In turn, $\text{pr}(\hat{S}_j \geqslant s_j)$ can be estimated by generating a large number of normal samples $\{G_l\}$. As justified in Appendix 3, S_j is a reasonable test because it is consistent against incorrect functional forms for Z_j if there is no additional model misspecification and if Z_j is independent of all other covariates.

When the foregoing analysis shows that $w_j(.)$ is too extreme, an appropriate functional form for Z_j may be identified from the observed pattern of $w_j(.)$ or from the scatterplot smoother of the raw martingale residuals. This point will be further elaborated in §3.

2·4. *Checking the link function*

To check the exponential link function, we consider the following special case of the $W_r(.,.)$ process,

$$W_l(x) = \sum_{i=1}^{n} I(\hat{\beta}' Z_i \leqslant x)\hat{M}_i .$$

The null distribution of this process can be approximated by the zero-mean Gaussian process $\hat{W}_l(x)$. As in §2·3, one may plot the observed process $w_l(.)$ along with a few realizations of $\hat{W}_l(.)$, and supplement the graphical display with an estimated p-value for $\sup_x |w_l(x)|$. Note that the w_l plot resembles the residual plot against fitted values for checking the linearity in the classical linear model. An unusual pattern of $w_l(.)$ would suggest an alternative link function. As shown in Appendix 3, the $\sup_x |W_l(x)|$ test is consistent against incorrect link functions in the form of $g(\beta^{*\prime} Z)$, where g is not exponential and β^* is the limit of $\hat{\beta}$.

2·5. *Checking the proportional hazards assumption*

If Z is a dichotomous variable, the standardized score process $\mathscr{I}^{-\frac{1}{2}}(\hat{\beta})U(\hat{\beta}, t)$ is asymptotically equivalent to the Brownian bridge B^0, and the corresponding supremum test is consistent against nonproportional hazards alternatives (Wei, 1984). For $p \geqslant 1$, each of the proportional hazards test statistics,

$$\sup_t \{\mathscr{I}^{-1}(\hat{\beta})_{jj}\}^{\frac{1}{2}} |\, U_j(\hat{\beta}, t)\,| \quad (j = 1, \ldots, p),$$

has the distribution of $\sup_{0 \leqslant u \leqslant 1} |B^0(u)|$ asymptotically if $\{V(t)\}_{jk} = 0$ $(j \neq k)$ for all t, where $V(.)$ is the limiting covariance matrix for $n^{-\frac{1}{2}}U(\beta_0, .)$ (Therneau et al., 1990). This general result is of limited practical use, however, because the assumption on $V(.)$, which essentially requires the independence of covariates, usually fails.

Note that $U(\hat{\beta}, t)$ is a special case of $W_z(t, z)$ with $z = \infty$ and $f(x) = x$. Thus the results described in §2·2 can be used to simulate the distributions of

$$\sup_t \{\mathscr{I}^{-1}(\hat{\beta})_{jj}\}^{\frac{1}{2}} |\, U_j(\hat{\beta}, t)\,| \quad (j = 1, \ldots, p).$$

The resulting p-values are valid asymptotically regardless of the covariance structure $V(.)$. One can also conduct graphical inspections of the proportional hazards assumption by comparing the observed score processes with the simulated ones.

For assessing the overall proportionality, it is natural to consider the test statistic

$$\sup_t \| U(\hat{\beta}, t) \| \quad \text{or} \quad \sup_t \sum_{j=1}^{p} \{\mathscr{I}^{-1}(\hat{\beta})_{jj}\}^{\frac{1}{2}} |\, U_j(\hat{\beta}, t)\,|.$$

As shown in Appendix 3, such tests are consistent against the nonproportional hazards alternative: $\lambda(t; Z) = \lambda_0(t) \exp\{\theta(t)' Z\}$, where $\theta(t)$ is not time-invariant. The powers tend to be high if $\theta(.)$ is monotone.

2·6. *An omnibus test*

With $f(.) = 1$, the $W_z(.,.)$ process becomes

$$W_o(t, z) = \sum_{i=1}^{n} I(Z_i \leqslant z)\hat{M}_i(t).$$

Plotting the $W_o(t, z)$ process versus t and z simultaneously would permit a global

assessment of the model adequacy; however, high-dimensional graphics is still in its infancy. Since the null distribution of the $W_o(.,.)$ process is centred around zero, it is natural to construct a lack-of-fit test based on the statistic $S_o = \sup_{t,z} | W_o(t,z) |$. An extreme value of s_o would indicate model misspecification. The p-value, $\mathrm{pr}(S_o \geqslant s_o)$, can again be estimated through simulations. Because the supremum is taken over the entire product space of the follow-up time and covariates, the S_o test is consistent against any departures from model (1·1), as is proven in Appendix 3.

Schoenfeld (1980) proposed a chi-squared goodness-of-fit test by comparing the observed and expected numbers of events in cells arising from a partition of the Cartesian product of the range of covariates and the time axis. A key criticism of this approach has been its arbitrariness in partitioning. The relationship of our omnibus test with Schoenfeld's is analogous to that of Kolmogorov's supremum test versus Pearson's chi-squared test for a hypothesized continuous distribution.

2·7. *Simulation experiments*

From our numerical studies, we have found the aforementioned Gaussian approximations to be satisfactory for practical sample sizes. In one key experiment, we assumed the Cox model $\lambda(t;h) = \exp(\beta_0 h)$, where $h = 0, 1, \ldots, 9$ with equal proportions, and generated censoring times from uniform $(0, \tau)$. For $\beta_0 = 0{\cdot}2$, $\tau = 3$, $n = 50$ and significance level of 0·05, the sizes of three supremum tests, $\sup_{t,z} | W_o(t,z) |$, $\sup_x | W_1(x) |$ and $\sup_t | U(\hat{\beta}, t) |$, were estimated at 0·04, 0·04 and 0·05, respectively. In this and all other studies, we used 1000 realizations of the Gaussian process with 1000 replications of the data. To evaluate the $\sup_x | W_1(x) |$ test, we adopted the same set of experimental parameters except that $\lambda(t;h) = \exp(-0{\cdot}2h + 0{\cdot}1h^2)$. The estimated size was 0·04. Additional experiments confirmed that the supremum tests did indeed preserve the size well.

Our numerical studies have also indicated that the proposed supremum tests are sensitive to model misspecification. For example, when h^2 is omitted from the true model $\lambda(t;h) = \exp(0{\cdot}5h - 0{\cdot}1h^2)$, the estimated powers for the tests $\sup_x | W_1(x) |$, or equivalently $\sup_x | W_l(x) |$, and $\sup_{t,z} | W_o(t,z) |$ with the 0·05 significance level were, respectively, 0·85 and 0·79, for $n = 50$ and 25% uniform censoring. Schoenfeld's test which partitions the time axis into two intervals (0, 0·5) and (0·5, ∞) and h into three subsets (0, 1, 2), (3, 4, 5) and (6, 7, 8, 9) had power of about 0·63. Further partitioning would lead to unacceptably small cell counts, while splitting h into two categories would render the test completely insensitive to the quadratic trend. The optimal test in this case is the partial likelihood score test for testing no h^2 effect. Its power was estimated at 0·96. Obviously, this is an unfair competitor since the score tests are designed for specific and nested alternatives. To study nonproportional hazards alternatives, we generated failure times from the Weibull model with density $\lambda(t;h) = (\gamma h) t^{\gamma h - 1}$. For $\gamma = 0{\cdot}2$, $\tau = 5$, $n = 50$ and the 0·05 significance level, the estimated powers of the $\sup_t | U(\hat{\beta}, t) |$ and $\sup_{t,z} | W_o(t,z) |$ tests were 0·90 and 0·56, respectively. Schoenfeld's test which partitions the time axis into two intervals (0, 0·8) and (0·8, ∞) and h into three groups (0, 1, 2), (3, 4, 5) and (6, 7, 8, 9) had estimated power of about 0·65.

3. Worked examples

3·1. *General*

We now apply the proposed techniques to two familiar data sets. In our illus-

trations, the p-value for the supremum-type test is always based on 10 000 realizations, though 1000 are recommended for general use. In each graphical display, the observed process is indicated by a solid curve and 20 simulated processes are plotted in dotted curves. The p-value for the supremum test is also shown on the graph. The dotted curves are unavoidably crowded, though distinguishable when plotted successively on an X-window.

3·2. *Mayo liver disease data*

The Mayo Clinic developed a database for 418 patients with primary biliary cirrhosis (PBC), a fatal chronic liver disease. These data are tabulated in Appendix D.1 of Fleming & Harrington (1991). As of the date of data listings, 161 patients had died. The PBC data were used by Dickson et al. (1989) to build a Cox model for the natural history of the disease with five covariates, log(bilirubin), log(protime), log(albumin), age and oedema. The covariates are mildly correlated, all correlation estimates being smaller than 0·35. The parameter estimates for the five covariates are, respectively, 0·871, 2·380, −2·533, 0·039 and 0·859, the respective estimated standard errors being 0·083, 0·767, 0·648, 0·008 and 0·271. The Mayo PBC Model has played an extremely important role in the liver disease research.

In Figure 4.6.5 of Fleming & Harrington (1991, p. 183), raw martingale residuals from a model with the discrete covariate oedema and three of the four continuous variables, log(bilirubin), log(protime), log(albumin) and age, are plotted against the omitted variable. Approximate linearity of each of the four scatterplot smoothers provides support for the selected transformations, but departures from the linear fit in the right-hand tail are noticeable for log(protime) and age due to some outlying covariate values (Fleming & Harrington, 1991, p. 184). It is difficult to make an objective conclusion from this figure regarding the functional forms.

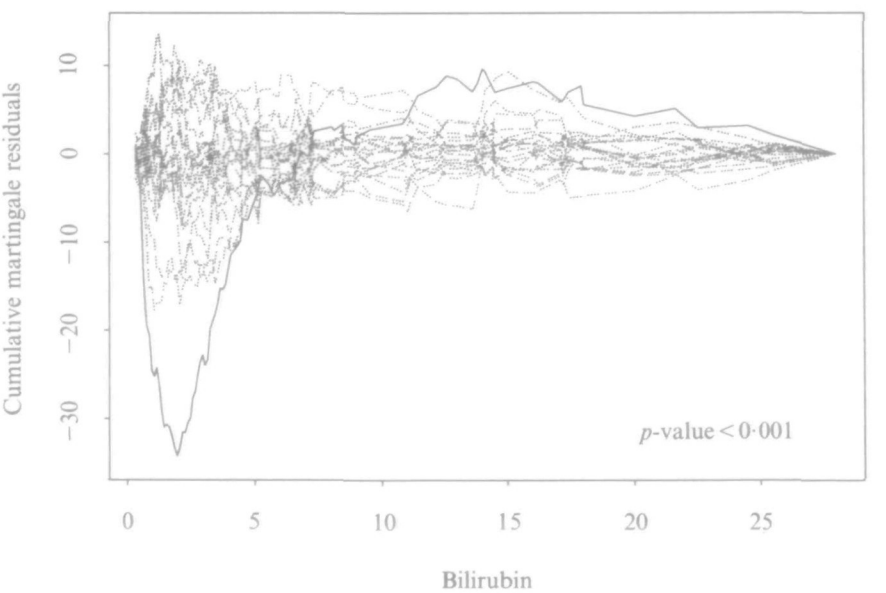

Fig. 1. Plot of cumulative martingale residuals versus bilirubin in the Cox model with bilirubin, log(protime), log(albumin), age and oedema for the Mayo PBC data.

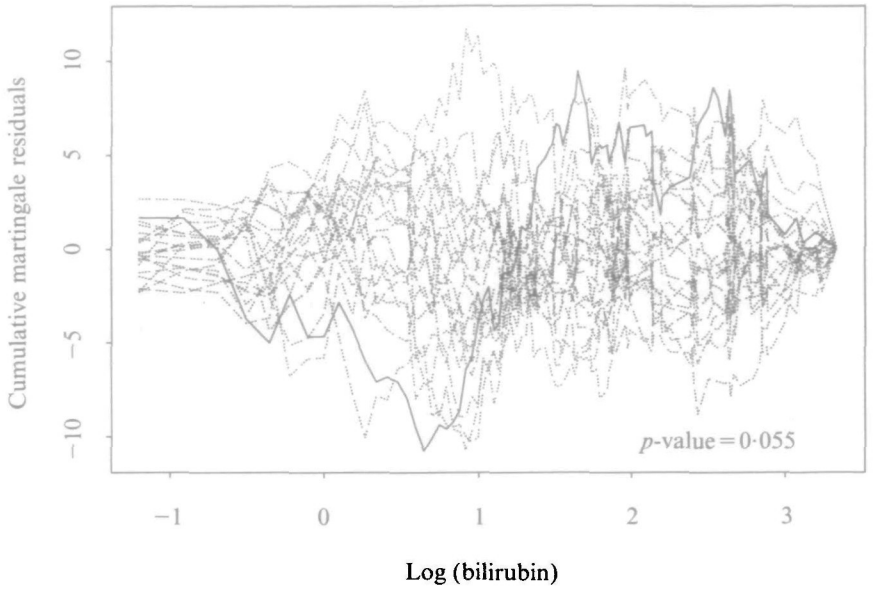

Fig. 2. Plot of cumulative martingale residuals versus log(bilirubin) in the Mayo PBC Model with log(bilirubin), log(protime), log(albumin), age and oedema.

Figure 1 of the present paper plots the cumulative martingale residuals against bilirubin in the Cox model with bilirubin, log(protime), log(albumin), age and oedema. The deliberate use of the untransformed bilirubin is clearly inappropriate. The fitted model vastly overestimates the hazards for the very low end of the bilirubin values and underestimates the hazards for most of the remaining bilirubin values. This pattern suggests a logarithmic transformation. As shown in Fig. 2, log(bilirubin) is a much better functional form, though by no means perfect. Additional analyses indicate that

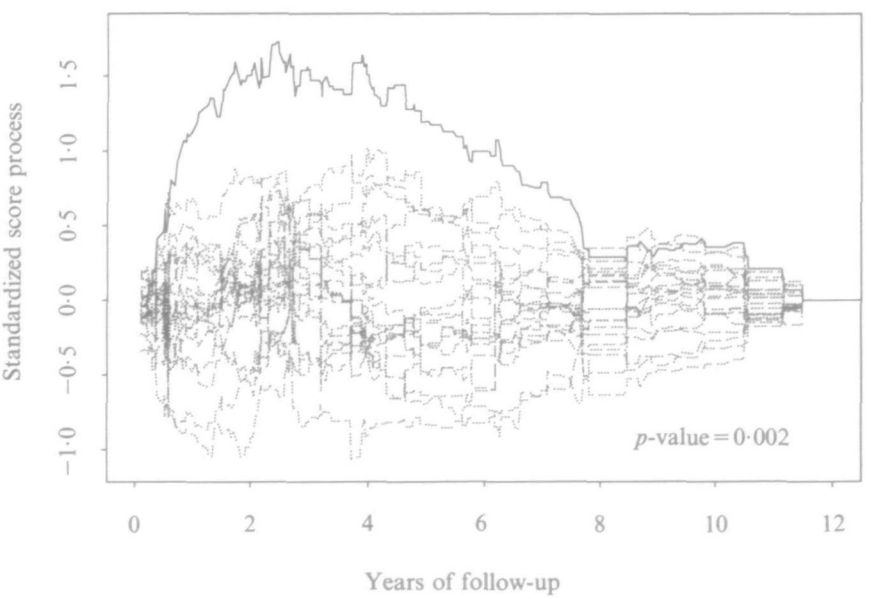

Fig. 3. Plot of standardized score process versus time for log(protime) in the Mayo PBC Model.

the functional forms for the remaining covariates are satisfactory, the p-values of the supremum tests all being greater than 0·30. Furthermore, the plot of the cumulative martingale residuals against the risk index $\hat{\beta}'Z$ suggests that the exponential link function is reasonable, the p-value of the supremum test being 0·272.

Figure 3 displays the score process for log(protime), revealing violation of the proportional hazards assumption. This finding confirms an earlier conclusion reached by Therneau et al. (1990), who used the critical values of $\sup_{0\leqslant u\leqslant 1} |B^0(u)|$ without adjusting for the dependence of covariates. The same computer run gave p-values of 0·114, 0·448, 0·473 and 0·031 for the proportional hazards tests with respect to log(bilirubin), log(albumin), age and oedema, respectively, indicating nonproportional hazards for oedema. The overall test $\sup_t \Sigma\{\mathscr{I}^{-1}(\hat{\beta})_{jj}\}^{\frac{1}{2}} |U_j(\hat{\beta}, t)|$, where the summation is over the range $j = 1, \ldots, 5$, yielded a p-value of 0·009. The nonproportionality may be corrected by introducing time-varying covariates or by stratifications, which we shall not pursue here.

3·3. *Stanford heart transplant data*

The Stanford heart transplant data as of February 1980 were described by Miller & Halpern (1982). Out of the 157 patients who are included in our analysis, 55 were censored as of the date of data listings.

We first fit a Cox model with covariate age only. The parameter estimate is 0·030 with an estimated standard error of 0·011. The omnibus test yields a p-value of 0·045, discrediting the assumed model. The proportional hazards test turns out nonsignificant (p-value = 0·244). As shown in Fig. 4, the model misspecification lies in the functional form of the covariate. The observed pattern of the cumulative martingale residuals is 'opposite' to that shown in Fig. 1 and calls for addition of the squared term. Note that checking the link function is equivalent to checking the function form of the covariate when there is only one covariate.

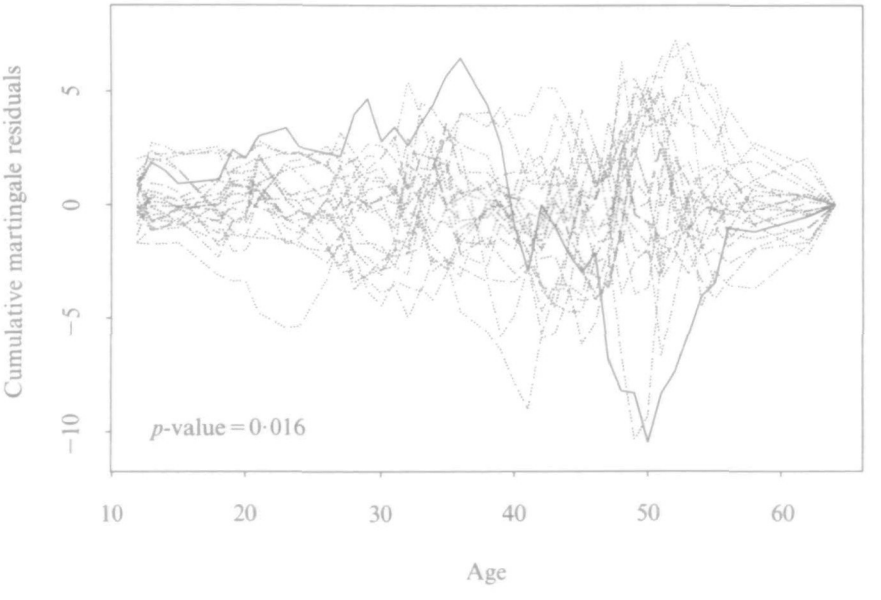

Fig. 4. Plot of cumulative martingale residuals versus age in the Cox model with age only for the Stanford heart transplant data.

When age^2 is added to the model, the p-value for the omnibus test jumps from 0·045 to 0·313 and that of the supremum test for the functional form of age from 0·016 to 0·499. The supremum test for the link function is not significant (p-value = 0·322). Individual proportional hazards tests give p-values of 0·134 and 0·108 for age and age^2, respectively, and the p-value for the overall test is 0·118. Due to a high correlation between age and age^2, the observed supremum values of the standardized score processes for the two covariates are both over 6·0. The use of the critical values from $\sup_{0 \leqslant u \leqslant 1} | B^0(u) |$, for example 1·628 for the 0·01 significant level, would result in very misleading conclusions.

4. Remarks

We have confined our attention to time-independent covariates. Allowing Z to vary over time not only enables one to study time-varying risk factors, but also provides a flexible way of adjusting for nonproportional hazards. The score process in the presence of time-dependent covariates can be expressed as

$$U(\hat{\beta}, t) = \sum_{i=1}^{n} \int_0^t Z_i(s)\, d\hat{M}_i(s).$$

One can again check the proportional hazards assumption by examining the score process. The Gaussian process for simulations is obtained from (2·4) with the replacement of the Z_l by the $Z_l(.)$. In theory, one may also extend the techniques described in § 2·3 and 2·4 to the setting of time-dependent covariates. The resulting procedures are of little practical use, however, because one cannot plot the partial-sum process against a time-varying covariate or risk index in a two-dimensional graphical display.

The ideas presented in this paper can be applied to relative risk regression models with other link functions and also to parametric survival models. The martingale residuals in those two settings were considered by Barlow & Prentice (1988) and Therneau et al. (1990). For noncensored data, Su & Wei (1991) used cumulative sums of ordinary residuals to check the generalized linear model.

Recently, McKeague & Utikal (1991) developed goodness-of-fit methods for the Cox model by comparing an estimator of a doubly cumulative hazard function under the assumed model with a fully nonparametric estimator of the same function. They derived a Schoenfeld-type test. It would be valuable to construct a supremum test by simulating their null process, as we have done here.

Acknowledgements

This work was supported by the National Institutes of Health for D. Y. Lin and L. J. Wei, and by the National Science Foundation and the National Security Agency for Z. Ying. The authors thank the reviewers and Dr Barbara McKnight for helpful comments.

Appendix 1

Weak convergence of $n^{-\frac{1}{2}} W_z$ and $n^{-\frac{1}{2}} \hat{W}_z$

We begin with a tightness lemma.

Lemma 1. *Let*

$$\xi_n(z, t) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^t q(Z_i, s)\, dM_i(s) I(Z_i \leqslant z),$$

where q and Z_i are bounded, without loss of generality, by 1. *Then ξ_n is tight in $\mathscr{D}([-1,1]^p \times [0,\infty))$.*

Proof. For simplicity, assume $p = 1$. We first show that ξ_n is tight in $\mathscr{D}([-1,1] \times [0,\tau_0])$ for any τ_0 such that $\Lambda_0(\tau_0) < \infty$, by applying Theorems 1 and 3 and the remark on p. 1665 of Bickel & Wichura (1971). To do so, it suffices to verify that, for $t_1 < t < t_2$ and $z_1 < z < z_2$ with $\text{pr}\{Z \in [z_1, z)\} \geqslant 1/n$ and $\text{pr}\{Z \in [z, z_2)\} \geqslant 1/n$, the following two inequalities hold:

$$E[\{\psi_n(z_1, z; t_1, t)\}^2 \{\psi_n(z_1, z; t, t_2)\}^2] \leqslant K(t_2 - t)(t - t_1)\,\text{pr}^2\{Z \in [z_1, z)\}, \tag{A1·1}$$

$$E[\{\psi_n(z_1, z; t_1, t_2)\}^2 \{\psi_n(z, z_2; t_1, t_2)\}^2] \leqslant K(t_2 - t_1)^2\,\text{pr}\{Z \in [z_1, z)\}\,\text{pr}\{Z \in [z, z_2)\}, \tag{A1·2}$$

where $K > 0$ is some constant and

$$\psi_n(z_1, z_2; t_1, t_2) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_{t_1}^{t_2} q(Z_i, s)\, dM_i(s) I(Z_i \in [z_1, z_2)).$$

The proofs for (A1·1) and (A1·2) were given in Technical Report #111 from Department of Biostatistics, University of Washington, and will not be shown here due to pressure on space.

It remains to show tightness at the endpoint, i.e., for any $\varepsilon > 0$, there exist n_0 and τ_0 such that

$$\text{pr}\{\sup_{z,s \geqslant t} |\xi_n(z, s) - \xi_n(z, t)| \geqslant \varepsilon\} \leqslant \varepsilon, \tag{A1·3}$$

for all $n \geqslant n_0$ and $t \geqslant \tau_0$. Rearrange $\{Z_i\}$ to make it increasing in i. Then

$$\sup_{z, t \geqslant \tau_0} |\xi_n(z, t) - \xi_n(z, \tau_0)| \leqslant \sup_{k, t \geqslant \tau_0} \left| n^{-\frac{1}{2}} \sum_{i=1}^{k} \int_{\tau_0}^{t} q(Z_i, s)\, dM_i(s) \right|.$$

By a similar argument as in the proof of Wichura's inequality (Shorack & Wellner, 1986, pp. 876–7), we can show that, for sufficiently large n,

$$\text{pr}\left\{ \sup_{k, t \geqslant \tau_0} \left| n^{-\frac{1}{2}} \sum_{i=1}^{k} \int_{\tau_0}^{t} q(Z_i, s)\, dM_i(s) \right| \geqslant \varepsilon \right\} \leqslant \frac{8}{\varepsilon^2 n} \sum_{i=1}^{n} E\left\{ \int_{\tau_0}^{\infty} q^2(Z_i, s)\, d\langle M_i \rangle(s) \right\},$$

which can be made arbitrarily small by choosing τ_0 large enough. This completes the proof. □

We now use Lemma 1 to show that $n^{-\frac{1}{2}} W_z$ is tight. Since $n^{-\frac{1}{2}} W_z$ and $n^{-\frac{1}{2}} \tilde{W}_z$ are asymptotically equivalent, it suffices to show the tightness of $n^{-\frac{1}{2}} \tilde{W}_z$. Let

$$A_n(t, z) = \mathscr{I}^{-1}(\beta_0) \sum_{k=1}^{n} \int_0^t Y_k(s) \exp(\beta_0' Z_k) f(Z_k) I(Z_k \leqslant z) \{Z_k - \tilde{Z}(\beta_0, s)\} \lambda_0(s)\, ds.$$

Then

$$\begin{aligned} n^{-\frac{1}{2}} \tilde{W}_z(t, z) = {} & n^{-\frac{1}{2}} \sum_{l=1}^{n} \int_0^t \{f(Z_l) I(Z_l \leqslant z) - \tilde{g}(\beta_0, u, z)\}\, dM_l(u) \\ & - A_n'(t, z) n^{-\frac{1}{2}} \sum_{l=1}^{n} \int_0^{\infty} \{Z_l - \tilde{Z}(\beta_0, u)\}\, dM_l(u). \end{aligned} \tag{A1·4}$$

From Lemma 1, the first term is tight. By the law of large numbers, A_n converges to some non-random function. Thus the second term is also tight since

$$n^{-\frac{1}{2}} \sum_{l=1}^{n} \int_0^{\infty} \{Z_l - \tilde{Z}(\beta_0, u)\}\, dM_l(u)$$

converges in distribution.

Conditional on $\{X_i, \Delta_i, Z_i\}$, the only random components in $\hat{W}_z$ are the independent standard normal variables $\{G_l\}$. Thus, it is easy to get moment inequalities similar to (A1·1) and (A1·2), from which the tightness of $n^{-\frac{1}{2}} \hat{W}_z$ follows.

For fixed t and z, $\tilde{W}_z(t,z)$ is essentially a sum of independent zero-mean random vectors. It then follows from the multivariate central limit theorem and the above tightness result that the process $n^{-\frac{1}{2}}\tilde{W}_z(.,.)$ converges to a zero-mean Gaussian random field. Furthermore, conditional on $\{X_i, \Delta_i, Z_i\}$, the process $n^{-\frac{1}{2}}\hat{W}_z$ is zero-mean Gaussian with a covariance function that will be shown next to converge to the same limit as that of $n^{-\frac{1}{2}}\tilde{W}_z$.

Let us rewrite $\tilde{W}_z(t,z)$ as

$$\tilde{W}_z(t,z) = \sum_{l=1}^{n} \int_0^\infty h_l(\beta_0, t, z, u)\, dM_l(u),$$

where

$$h_l(\beta, t, z, u) = I(u \leqslant t)\{f(Z_l)I(Z_l \leqslant z) - \tilde{g}(\beta, u, z)\} - A_n'(t,z)\{Z_l - \tilde{Z}(\beta, u)\}.$$

It then becomes clear that the covariance function for $n^{-\frac{1}{2}}\tilde{W}_z$ is

$$E\{n^{-1}\tilde{W}_z(t_1,z_1)\tilde{W}_z(t_2,z_2)\} = E\left\{\int_0^\infty h_l(\beta_0,t_1,z_1,u)h_l(\beta_0,t_2,z_2,u)\,Y_l(u)\exp(\beta_0'Z_l)\lambda_0(u)\,du\right\}. \tag{A1·5}$$

Now due to the strong consistency of $\hat{\beta}$ and $\hat{\Lambda}_0(.)$ (Tsiatis, 1981; Shorack & Wellner, 1986, p. 304),

$$E[n^{-1}\{\hat{W}_z(t,z) - \hat{W}_z^*(t,z)\}^2 | \{X_i, \Delta_i, Z_i\}] \to 0$$

almost surely, where

$$\hat{W}_z^*(t,z) = \sum_{l=1}^{n} \int_0^\infty h_l(\beta_0, t, z, u)\, dN_l(u) G_l.$$

Therefore, conditional on $\{X_i, \Delta_i, Z_i\}$, the asymptotic covariance function for $n^{-\frac{1}{2}}\hat{W}_z$ is

$$n^{-1}\sum_{l=1}^{n}\int_0^\infty h_l(\beta_0,t_1,z_1,u)\,dN_l(u)\int_0^\infty h_l(\beta_0,t_2,z_2,u)\,dN_l(u)$$

$$= n^{-1}\sum_{l=1}^{n}\int_0^\infty h_l(\beta_0,t_1,z_1,u)h_l(\beta_0,t_2,z_2,u)\,dN_l(u),$$

which converges almost surely to (A1·5) by the law of large numbers since $Y_l(u)\exp(\beta_0'Z_l)\lambda_0(u)$ is the intensity function of $N_l(u)$.

Appendix 2

Weak convergence of $n^{-\frac{1}{2}}W_r$ and $n^{-\frac{1}{2}}\hat{W}_r$

Proving the tightness of $n^{-\frac{1}{2}}W_r$ is similar to but considerably more complicated than proving the tightness of $n^{-\frac{1}{2}}W_z$. We again refer the interested reader to our Technical Report for details. On the other hand, exactly the same arguments used in Appendix 1 for getting the tightness of $n^{-\frac{1}{2}}\hat{W}_z$ can be applied to $n^{-\frac{1}{2}}\hat{W}_r$. The weak convergence of $n^{-\frac{1}{2}}W_r$ and $n^{-\frac{1}{2}}\hat{W}_r$ to the same limiting Gaussian process can then be verified by showing that their finite-dimensional distributions are asymptotically the same.

Appendix 3

Consistency of supremum tests

Consistency of $\sup_{t,z}|W_z(t,z)|$. We claim that the omnibus test based on $\sup_{t,z}|W_z(t,z)|$ is consistent against the general alternative that there do not exist a constant vector β_0 and a function $\lambda_0(.)$ such that $\lambda(t;z) = \lambda_0(t)e^{\beta_0'z}$ for almost all $t > 0$ and z generated by the random vector Z. Let H be the distribution function of Z. Under the alternative hypothesis, $\hat{\beta} \to \beta^*$

and $\hat{\Lambda}_0(t) \to \int \lambda^*(u)\,du$, where the integral is over $(0,t)$, and where β^* is some constant vector and $\lambda^*(.)$ is a deterministic function (Lin & Wei, 1989). To prove that the omnibus test is consistent, it suffices to show that asymptotically $\sup_{t,z} |n^{-1}W_z(t,z)|$ is nonzero under the alternative hypothesis. By some simple arguments, $n^{-1}W_z(t,z)$ converges almost surely to

$$\int_{x\leqslant z, u\leqslant t} f(x)E\{Y(u)\,|\,x\}\{\lambda(u;x) - e^{\beta^{*\prime}x}\lambda^*(u)\}\,dH(x)\,du,$$

which will be nonzero for some t and z under the alternative. This establishes our claim.

Consistency of $\sup_z |\sum_i f(Z_i)I(Z_i \leqslant z)\hat{M}_i|$ *and related tests.* For simplicity, assume $f \equiv 1$. Consider the following alternative: $\lambda(t;z) = \lambda_0(t)g(z)$ but there does not exist a β such that $g(z)/e^{\beta' z}$ is a constant for all the z in the support of H. To prove the consistency of the test based on $\sup_z |\sum I(Z_i \leqslant z)\hat{M}_i|$, where the summation is over $i = 1,\ldots,n$, it suffices to show that asymptotically

$$\sup_z \left| n^{-1}\sum_{i=1}^{n} I(Z_i \leqslant z)\hat{M}_i \right|$$

is nonzero under the alternative hypothesis. Note that $n^{-1}\sum I(Z_i \leqslant z)\hat{M}_i$ converges almost surely to

$$\int_{x\leqslant z, 0<t<\infty} e^{\beta^{*\prime}x}E\{Y(t)\,|\,x\}\left\{\frac{g(x)}{e^{\beta^{*\prime}x}} - J(t)\right\}\lambda_0(t)\,dH(x)\,dt,$$

where

$$J(t) = \frac{\int g(x)E\{Y(t)\,|\,x\}\,dH(x)}{\int e^{\beta^{*\prime}x}E\{Y(t)\,|\,x\}\,dH(x)} = \frac{\int \{g(x)/e^{\beta^{*\prime}x}\}e^{\beta^{*\prime}x}E\{Y(t)\,|\,x\}\,dH(x)}{\int e^{\beta^{*\prime}x}E\{Y(t)\,|\,x\}\,dH(x)}.$$

Let x^* be the maximizer of $g(x)/e^{\beta^{*\prime}x}$. Then, under the alternative hypothesis,

$$g(x^*)/e^{\beta^{*\prime}x} - J(t) > 0$$

(Hardy, Littlewood & Polya, 1934, p. 136). Thus, the test is consistent.

As a by-product of the foregoing result, the $\sup_x |W_j(x)|$ test is consistent against misspecification of the functional form for Z_j provided that the components of $\hat{\beta}$ for the remainder of Z converge to the true values. Arguments similar to those of Struthers & Kalbfleisch (1986) indicate that the asymptotic bias of $(\hat{\beta}_1,\ldots,\hat{\beta}_{j-1},\hat{\beta}_{j+1},\ldots,\hat{\beta}_p)$ is generally small if there is no additional model misspecification and if Z_j is independent of all other covariates. It also follows from the arguments of the previous paragraph that the $\sup_x |W_j(x)|$ is consistent against misspecification of the link function in the form of $g(\beta^{*\prime}Z)$, where g is not exponential.

Consistency of $\sup_t \| U(\hat{\beta},t) \|$ *and related tests.* We claim that the $\sup_t \| U(\hat{\beta},t) \|$, or

$$\sup_t \sum_j \{\mathscr{I}^{-1}(\hat{\beta})_{jj}\}^{\frac{1}{2}} |\, U_j(\hat{\beta},t)\,|,$$

test is consistent against the nonproportional hazards alternative: $\lambda(t;z) = \lambda_0(t)\,e^{\theta(t)'z}$, where $\theta(t)$ is not time-invariant. It is straightforward to show that $n^{-1}U(\hat{\beta},t) \to h(\beta^*,t)$ under this alternative, where

$$h(\beta,t) = \int_0^t \left[\frac{E\{Y(s)\,e^{\theta(s)'Z}Z\}}{E\{Y(s)\,e^{\theta(s)'Z}\}} - \frac{E\{Y(s)\,e^{\beta'Z}Z\}}{E\{Y(s)\,e^{\beta'Z}\}}\right] E\{Y(s)\,e^{\theta(s)'Z}\}\lambda_0(s)\,ds.$$

If $h(\beta^*,t) = 0$ for all t, then $\mu(\theta(t),t) = \mu(\beta^*,t)$ for all t, where

$$\mu(\eta,t) = E\{Y(t)\,e^{\eta'Z}Z\}/E\{Y(t)\,e^{\eta'Z}\}.$$

Since $\partial\mu(\eta,t)/\partial\eta$ is positive definite, $\mu(\theta(t),t) = \mu(\beta^*,t)$ implies $\theta(t) = \beta^*$. Thus, our claim is true.

Appendix 4

Computational methods

We now discuss numerical issues in implementing the omnibus test S_o, which is computationally the most complicated procedure. For simplicity of presentation, assume that Z is univariate. Denote the distinct values of $\{Z_1, \ldots, Z_n\}$ by $\{Z_1^*, \ldots, Z_{n'}^*\}$. Note that $\sup_{t,x} |w_o(t,x)| = \max_{i,j} |w_o(X_j, Z_i^*)|$. Thus, it is straightforward to calculate the supremum. When computing $\max_{i,j} |\hat{w}_o(X_j, Z_i^*)|$ for each realization of $\hat{W}_o(.,.)$, one can avoid any calculations of orders higher than $n'n$ by using the following algorithm.

If the data are sorted in the ascending order of the failure times and if there are no ties, then

$$\hat{w}_o(X_j, Z_i^*) = \hat{w}_o^{(1)}(X_j, Z_i^*) - Q(X_j, Z_i^*)\mathscr{I}^{-1}(\hat{\beta})\hat{w}_o^{(2)},$$

where

$$\hat{w}_o^{(1)}(X_j, Z_i^*) = \sum_{l=1}^{j} \Delta_l \{I(Z_l \leqslant Z_i^*) - g(\hat{\beta}, X_l, Z_i^*)\} G_l,$$

$$Q(X_j, Z_i^*) = \sum_{l=1}^{j} \sum_{k=1}^{n} \Delta_l Y_k(X_l) \exp(\hat{\beta}' Z_k) I(Z_k \leqslant Z_i^*) \{Z_k - \bar{Z}(\hat{\beta}, X_l)\} / S^{(0)}(\hat{\beta}, X_l),$$

$$\hat{w}_o^{(2)} = \sum_{l=1}^{n} \Delta_l \{Z_l - \bar{Z}(\hat{\beta}, X_l)\} G_l.$$

Since g and Q do not involve $\{G_l\}$, they only need to be evaluated outside the simulations. Computing g is trivial. Note that

$$Q(X_j, Z_i^*) = Q(X_{j-1}, Z_i^*) + \sum_{k=1}^{n} \Delta_j Y_k(X_j) \exp(\hat{\beta}' Z_k) I(Z_k \leqslant Z_i^*) \{Z_k - \bar{Z}(\hat{\beta}, X_j)\} / S^{(0)}(\hat{\beta}, X_j).$$

This recursive relationship enables one to evaluate $\{Q(X_j, Z_i^*); i = 1, \ldots, n'; j = 1, \ldots, n\}$ efficiently.

Note now that $\hat{w}_o^{(2)}$ does not involve X_j and Z_i^* and can be calculated before the maximization. Also note that

$$\hat{w}_o^{(1)}(X_j, Z_i^*) = \hat{w}_o^{(1)}(X_{j-1}, Z_i^*) + \Delta_j \{I(Z_j \leqslant Z_i^*) - g(\hat{\beta}, X_j, Z_i^*)\} G_j.$$

With the use of this recursive relationship, evaluating $\{\hat{w}_o^{(1)}(X_j, Z_i^*); i = 1, \ldots, n'; j = 1, \ldots, n\}$ is of order $n'n$. Thus, computing $\max_{i,j} |\hat{w}_o(X_j, Z_i^*)|$ is an $n'n$ process given the input of g and Q.

The aforementioned formulae can also be used when there are tied failure times; however, one should skip the maximization step for those X_j's that are equal to the X_{j+1}'s. The extension to the multiple covariate setting is straightforward. Computation can become quite extensive if there exists a very large number of distinct covariate patterns.

Calculations of the p-values for the supremum tests described in §§2·3–2·5 are considerably simpler than that for the omnibus test. Since one is always dealing with one-parameter processes, all those graphical and numerical procedures can be implemented in a short period of time.

Computer software implementing the proposed methods is available from D. Y. Lin.

References

Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100–20.

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

Bickel, P. J. & Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42**, 1656–70.

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

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–76.

DICKSON, E. R., GRAMBSCH, P. M., FLEMING, T. R., FISHER, L. D. & LANGWORTHY, A. (1989). Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology* **10**, 1–7.

FLEMING, T. R. & HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

HARDY, G. H., LITTLEWOOD, J. E. & POLYA, G. (1934). *Inequalities*. Cambridge University Press.

LAGAKOS, S. W. (1988). The loss in efficiency from misspecifying covariates in proportional hazard regression models. *Biometrika* **75**, 156–60.

LAGAKOS, S. W. & SCHOENFELD, D. A. (1984). Properties of proportional-hazards score tests under misspecified regression models. *Biometrics* **40**, 1037–48.

LIN, D. Y. & WEI, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Am. Statist. Assoc.* **84**, 1074–8.

LIN, D. Y. & WEI, L. J. (1991). Goodness-of-fit tests for the general Cox regression model. *Statistica Sinica* **1**, 1–17.

MCKEAGUE, I. W. & UTIKAL, K. J. (1991). Goodness-of-fit tests for additive hazards and proportional hazards models. *Scand. J. Statist.* **18**, 117–95.

MILLER, R. & HALPERN, J. (1982). Regression with censored data. *Biometrika* **69**, 521–31.

SCHOENFELD, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* **67**, 145–53.

SCHOENFELD, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239–41.

SHORACK, G. R. & WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. New York: Wiley.

STRUTHERS, C. A. & KALBFLEISCH, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73**, 363–9.

SU, J. Q. & WEI, L. J. (1991). A lack-of-fit test for the mean function in a generalized linear model. *J. Am. Statist. Assoc.* **86**, 420–6.

THERNEAU, T. M., GRAMBSCH, P. M. & FLEMING, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147–60.

TSIATIS, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9**, 93–108.

WEI, L. J. (1984). Testing goodness-of-fit for proportional hazards model with censored observations. *J. Am. Statist. Assoc.* **79**, 649–52.

[*Received January* 1992. *Revised November* 1992]