Biometrika (1994), **81**, 3, *pp.* 515–26
Printed in Great Britain

Proportional hazards tests and diagnostics based on weighted residuals

BY PATRICIA M. GRAMBSCH

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.

AND TERRY M. THERNEAU

Department of Health Science Research, Mayo Clinic, Rochester, Minnesota 55905, U.S.A.

SUMMARY

Nonproportional hazards can often be expressed by extending the Cox model to include time varying coefficients; e.g., for a single covariate, the hazard function for subject i is modelled as $\exp\{\beta(t)Z_i(t)\}$. A common example is a treatment effect that decreases with time. We show that the function $\beta(t)$ can be directly visualized by smoothing an appropriate residual plot. Also, many tests of proportional hazards, including those of Cox (1972), Gill & Schumacher (1987), Harrell (1986), Lin (1991), Moreau, O'Quigley & Mesbah (1985), Nagelkerke, Oosting & Hart (1984), O'Quigley & Pessione (1989), Schoenfeld (1980) and Wei (1984) are related to time-weighted score tests of the proportional hazards hypothesis, and can be visualized as a weighted least-squares line fitted to the residual plot.

Some key words: Cox model; Loess; Schoenfeld residuals; Weighted regression.

1. INTRODUCTION AND NOTATION

Using the approach of Fleming & Harrington (1991), consider each subject to be an independent counting process $\{N_i(t), t \geqslant 0, i = 1, \ldots, n\}$ with intensity function given by

$$Y_i(t) \exp\{\beta' Z_i(t)\}\, d\Lambda_0(t), \tag{1}$$

where $Y_i(t)$ is a 0–1 process which indicates whether the ith subject is at risk at time t, β is a p vector of regression parameters, Z_i is a p vector of covariate processes and $d\Lambda_0(t)$ is an unspecified hazard function. One can estimate β by maximizing the log partial likelihood (Cox, 1972):

$$\sum_{i=1}^{n} \int_0^\infty \left[Y_i(t)\beta' Z_i(t) - \log\left\{ \sum_{j=1}^{n} Y_j(t) e^{\beta' Z_j(t)} \right\} \right] dN_i(t). \tag{2}$$

Define

$$S^{(r)}(\beta, t) = \sum_{i=1}^{n} Y_i(t) \exp\{\beta' Z_i(t)\} Z_i(t)^{\otimes r}$$

for $r = 0, 1, 2$, where, for a column vector a, $a^{\otimes 2}$ denotes the outer product aa', $a^{\otimes 1}$ denotes the vector a, and $a^{\otimes 0}$ denotes the scalar 1. Then the conditional weighted mean and

variance of the covariate vector at time t are

$$M(\beta, t) = S^{(1)}(\beta, t)/S^{(0)}(\beta, t),$$

$$V(\beta, t) = \frac{S^{(2)}(\beta, t)}{S^{(0)}(\beta, t)} - \left\{\frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)}\right\}^{\otimes 2}.$$

We use i $(i = 1, \ldots, n)$ to index subjects, j $(j = 1, \ldots, p)$ to index covariates and k $(k = 1, \ldots, d)$ to index event times. Let $Z_{(k)}$ be the covariate vector of the subject with an event at time t_k. The total number of events d may be less than n due to censoring or greater than n in data sets with multiple events per subject.

An alternative to proportional hazards is time-varying coefficients. Let

$$\beta_j(t) \equiv \beta_j + \theta_j g_j(t),$$

where $g_j(t)$ is a predictable process, and suppose the intensity function is

$$Y_i(t) \exp\{\textstyle\sum_j \beta_j(t) Z_{ij}(t)\}\, d\Lambda_0(t).$$

For identifiability we assume that g varies about 0, but this is not essential. The Schoenfeld residuals are defined as

$$r_k(\beta) = Z_{(k)} - M(\beta, t_k).$$

Section 2 shows that $E\{r_k(\beta)\} \doteq V(\beta, t_k) G(t_k)\theta$, where $G(t)$ is a $p \times p$ diagonal matrix with $G_{jj}(t) = g_j(t)$, and that the score test for $H_0: \beta(t) = \beta$ is equivalent to a generalized least squares test on the Schoenfeld residuals. Many published goodness-of-fit tests are shown to be special cases. Using Monte Carlo simulations and actual data sets, § 3 shows that smoothed plots of the standardized Schoenfeld residuals can reveal the form of $g_j(t)$. Section 4 discusses some related work by Pettitt & Bin Daud (1990) and Gray (1990).

2. A REGRESSION APPROACH TO TESTS FOR NONPROPORTIONALITY

Let $\{F_t, t \geqslant 0\}$ be the right-continuous filtration, specifying the process history:

$$F_t = \sigma\{Z_i(u), N_i(u), Y_i(u+): 0 \leqslant u \leqslant t, i = 1, \ldots, n\},$$

and let $g_j(t)$ $(j = 1, \ldots, p)$ be a vector of predictable processes with respect to F_t. Suppose one wishes to test the null hypothesis of proportional hazards as in (1) versus the alternative of time-varying coefficients with intensity function given by

$$\begin{aligned} h_i(t)\, dt &= Y_i(t) \exp\{\beta'(t) Z_i(t)\}\, d\Lambda_0(t) \\ &= Y_i(t) \exp[\{\beta + G(t)\theta\}' Z_i(t)]\, d\Lambda_0(t). \end{aligned} \tag{3}$$

We generalize Schoenfeld's approach (1982), which considered a nonproportional departure in one covariate only.

To begin, assume (3) and known β. Then

$$\begin{aligned} r_k(\beta) &= Z_{(k)} - M(\beta, t_k) \\ &= [Z_{(k)} - M\{\beta(t_k), t_k\}] + [M\{\beta(t_k), t_k\} - M(\beta, t_k)], \end{aligned}$$

which is the sum of the Schoenfeld residual from the true model, a mean 0 random variable, and the difference between the weighted covariate means under the true and null models. Expand $M\{\beta(t_k), t_k\}$ in a one-term Taylor's expansion about $\beta(t_k) = \beta$ in the

second summand to obtain

$$E\{r_k(\beta)\,|\,F_{t_k}\} \doteq V(\beta, t_k)G_k\theta, \tag{4}$$

where $G_k = G(t_k)$ is a diagonal matrix with jj element $g_j(t_k)$.

Let $r_k^* = r_k^*(\beta) = V^{-1}(\beta, t_k)r_k(\beta)$ be the scaled Schoenfeld residual. Then

$$E(r_k^*\,|\,F_{t_k}) \doteq G_k\theta, \tag{5}$$

$$\text{var}\,(r_k^*\,|\,F_{t_k}) = V^{-1}(\beta, t_k)V\{\beta(t_k), t_k\}V^{-1}(\beta, t_k) \doteq V^{-1}(\beta, t_k). \tag{6}$$

Further, the r_k's are uncorrelated; see Cox (1975) or use martingale arguments. Equations (5) and (6) suggest a standard linear model for r_k^*. With $V_k \equiv V(\beta, t_k)$ generalized least squares gives

$$\hat{\theta} = (\sum G_k V_k G_k)^{-1} \sum G_k r_k. \tag{7}$$

This leads to an asymptotic χ^2 test statistic on p degrees of freedom:

$$(\sum G_k r_k)'(\sum G_k V_k G_k)^{-1}(\sum G_k r_k) \tag{8}$$

to test $H_0: \theta = 0$.

Now assume β is unknown and let $\hat{\beta}$ be the maximum partial likelihood estimate under H_0. Let $\hat{V}_k = V(\hat{\beta}, t_k)$ and $\hat{r}_k = r_k(\hat{\beta})$. As above, we have $E(\hat{V}_k^{-1}\hat{r}_k) \doteq G_k\theta$. However, since $\sum \hat{r}_k = 0$, the residuals are correlated with $\text{cov}\,(\hat{r}_l, \hat{r}_m)$ consistently estimated by $\delta_{lm}\hat{V}_l - \hat{V}_l(\sum \hat{V}_k)^{-1}\hat{V}_m$ asymptotically under H_0 (Schoenfeld, 1982). Thus $\text{var}(\hat{V}_k^{-1}\hat{r}_k) \doteq \hat{V}_k^{-1} - (\sum_l \hat{V}_l)^{-1}$. Generalized least squares gives

$$\hat{\theta} = D^{-1} \sum G_k \hat{r}_k, \tag{9}$$

with

$$D = \sum G_k \hat{V}_k G_k' - (\sum G_k \hat{V}_k)(\sum \hat{V}_k)^{-1}(\sum G_k \hat{V}_k)'. \tag{10}$$

Under H_0 the asymptotic variance of $n^{-\frac{1}{2}} \sum G_k \hat{r}_k$ can be consistently estimated by $n^{-1}D$, leading to an asymptotic χ^2 test statistic on p degrees of freedom:

$$T(G) = (\sum G_k \hat{r}_k)' D^{-1} (\sum G_k \hat{r}_k). \tag{11}$$

The estimator (9) and test statistic (11) are familiar from another context. They are respectively a one-step Newton–Raphson algorithm estimator of θ and the Rao score test of $H_0: (\beta, \theta) = (\hat{\beta}, 0)$, based on the partial likelihood. A formal proof of the asymptotic null hypothesis distribution of $T(G)$ therefore follows from standard results for score processes from partial likelihoods, using martingale theory. The fully-iterated maximum partial likelihood estimator of θ_j could be obtained by continuing with additional Newton–Raphson steps.

For any single variate, these derivations suggest a plot with $g(t_k)$ on the horizontal axis and $\hat{V}_k^{-1}\hat{r}_k$ on the vertical axis, along with a test for linear association. An ordinary least squares slope estimate and the corresponding test for zero slope will have a different form from (9) and (11) since least squares does not incorporate the constraint that the $\hat{r}_k$'s sum to zero, but the effect is slight.

Different choices for G result in different tests for model misspecification. In all of the special cases below, G is diagonal, so for simplicity we will refer to a univariate function $g(t)$.

Test 1. If $g(t)$ is a specified function of time, then $T(G)$ is a score test for the addition

of the time-dependent variable $g(t)Z$ to the model, a test initially suggested by Cox (1972). Chappell (1992) describes the relationship between this test and the test of Gill & Schumacher (1987).

Test 2. If g is piecewise constant on non-overlapping time intervals with the intervals and constants chosen in advance, $T(G)$ is the score test proposed by O'Quigley & Pessione (1989), which generalizes and extends goodness-of-fit tests proposed by Schoenfeld (1980) and Moreau et al. (1985). As the authors point out, this test has the disadvantage that the investigator must choose a partition of the time axis, but they suggest guidelines for doing so.

Test 3. If $g(t) = \bar{N}(t-)$ then $T(G)$ is the covariance between the scaled Schoenfeld residual and the rank of the event times. The resulting test is similar to one proposed by Harrell (1986), who uses the correlation between the unscaled residuals and rank of the event times.

Test 4. Lin (1991) suggests comparing $\hat{\beta}$ to the solution $\hat{\beta}_g$ of the weighted estimating equation

$$\sum G_k r_k(\beta) = 0$$

with $g(t)$ one of the scalar weight functions commonly chosen for weighted log rank tests, such as the left-continuous version of the Kaplan–Meier estimator. He showed that asymptotically $\hat{\beta} - \hat{\beta}_g$ is multivariate normal with mean 0 and a variance matrix derived from martingale counting process theory. If the estimator $\hat{\beta}_g$ were based on a one-step Newton–Raphson algorithm starting from $\hat{\beta}$, his test would be identical to $T(G)$.

Test 5. Let $g_j(t_1) = 0$ and $g_j(t_{k+1}) = a_j^2 \hat{r}_{jk}$, where $j = 1, \ldots, p$. This gives the test statistic of Nagelkerke, Oosting & Hart (1984), who suggest using the serial correlation of the Schoenfeld residuals for a univariate predictor, or, for multivariate covariates, the correlation of a weighted sum, $a'\hat{r}_k$. The authors standardize by using a permutational approach to estimate the variance rather than (10). They suggest $a = \hat{\beta}$ as a natural choice for the weights, followed by examination of individual covariates if the test is significant.

The key point is that each of the above tests can be directly visualized as a simple trend test applied to the plot of $g(t)$ versus $\hat{r}^*$. An alternative group of tests is based on the plot of cumulative residuals, as suggested by Wei (1984). These can also be explored using T, but we have not found them to be as useful in practice.

3. GRAPHICAL DIAGNOSTICS

The test statistics imply a prespecified form for departures from proportionality as given by $G(t)$, but when the investigator has no hypotheses about the nature of the non-proportionality graphical displays let the data speak for themselves. Schoenfeld (1982) and Lin (1991) recommended plotting the elements of the Schoenfeld residuals against failure times, and Wei (1984) and Therneau, Grambsch & Fleming (1990) recommended plotting the cumulative sums. Equations (3) and (5) suggest a smoothed scatter plot of the quantities $\hat{V}_k^{-1}\hat{r}_k + \hat{\beta}$ versus t_k will reveal the functional form of $\beta(t)$.

The computation of $\hat{V}_k$ at each death time can frequently be avoided. For most data sets the variance matrix of $Z(t)$ varies slowly, and is quite stable until the last few death times. One can substitute the average value $\bar{V} = \mathcal{I}/d$, where $\mathcal{I}^{-1}$ is the covariance matrix

of $\hat{\beta}$. This estimate may even be preferable, as the last few $\hat{V}_k$'s may be based on a very small number of subjects each, and $\hat{V}_k$ will be singular if the number at risk is less than the number of covariates. A slight modification of this approach may be needed for binary covariates and light censoring: if only one group remains at risk, both $\hat{r}_k$ and $\hat{V}_k$ will be zero. An alternative is $\mathscr{I}/d^*$, where d^* is the number of events where individuals from both groups remain at risk. The computations for $\bar{V}$ and $\hat{r}_k$'s are unchanged if there are tied events, but the usual caveats about biased estimates apply if the proportion of ties is large.

To explore the properties of these plots, we consider Monte Carlo simulations, beginning with a single binary covariate. Although the linear model for the Schoenfeld residuals was derived from a Taylor expansion assuming small amplitude variations in G, simulations suggest that the plots are useful even when G has large scale variations.

We simulated two nonproportional hazards models. For both, one group had an exponential failure time with a hazard rate of 2. In simulation 1, the second group had

$$\beta_1(t) = \begin{cases} -\log(4) & (0 \leqslant t < 0{\cdot}5), \\ \log(4) & (0{\cdot}5 \leqslant t), \end{cases}$$

so the group had a piecewise exponential hazard with rate of 0·5 until time = 0·5 and then an abrupt change to a rate of 8. In simulation 2, $\beta_2(t) = 5t$ and so the second group had a hazard rate log linear in time. To generate the failure times for this group, let

$$T = \theta^{-1} \log(1 + \theta T_0/\lambda),$$

where T_0 is a standard exponential pseudo-random variable. In each simulation, the sample size was 80, 40 per group. The censoring distribution was independent of the failure time distribution. In simulation 1, the censoring distribution was uniform on [0·5, 0·8] resulting in a censoring rate of 29%, and in simulation 2 it was uniform on [0·3, 0·6] with a

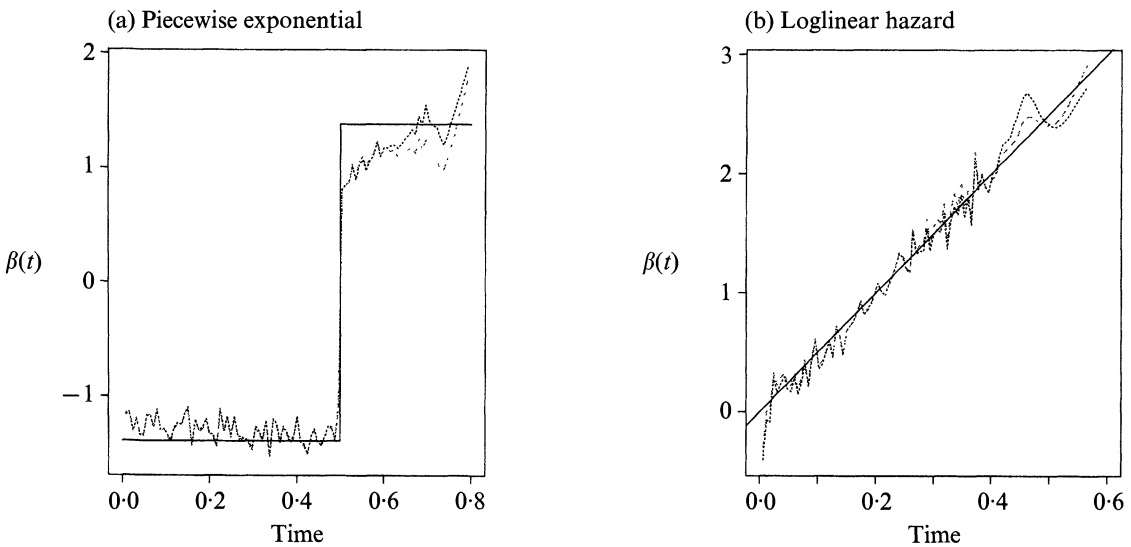

Fig. 1. Smoothed (loess, span = 0·01) standardized Schoenfeld residuals + mean $\hat{\beta}$ plotted against event times from two Monte Carlo simulations for nonproportional hazards models with a single binary covariate described in the text. Two means of standardization are shown: dotted line shows event specific variance; dashed line, average variance. Solid line shows the true functional form of the interaction between time and covariate.

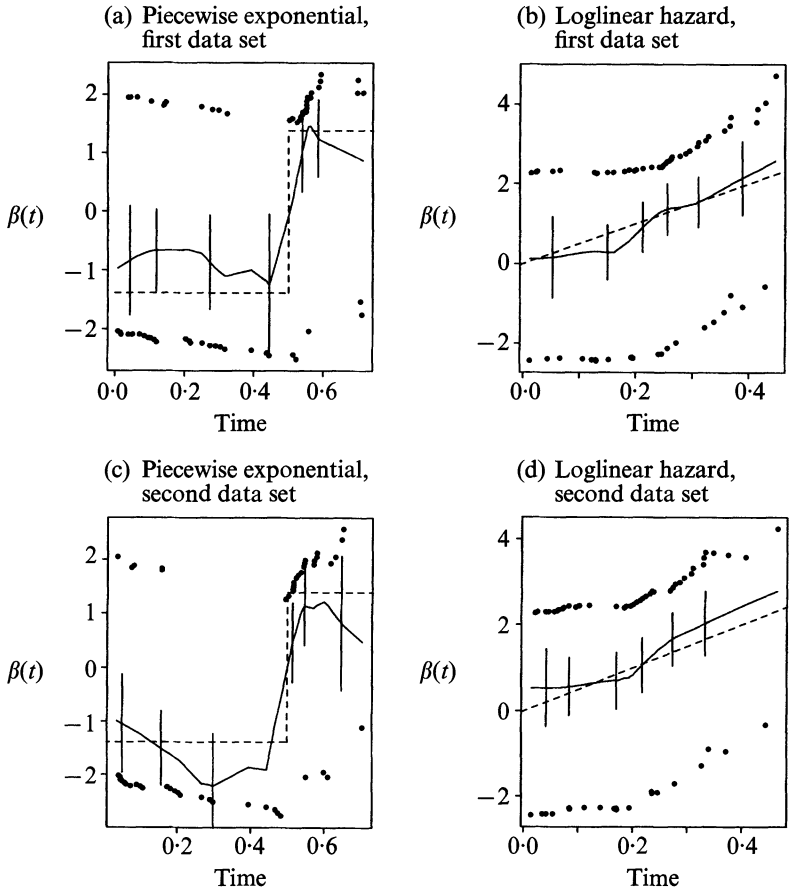

Fig. 2. Standardized Schoenfeld residuals $+\hat{\beta}$ plotted against event time for two data sets from each simulation (compare Fig. 1). A loess smooth (span = $\frac{1}{2}$ and degrees of freedom = 4 for piecewise exponential, and span = $\frac{3}{4}$ and degrees of freedom = 3 for the loglinear hazard) is shown on each plot. Average variance standardization is used with 90% confidence intervals at the 6th, 15th, 24th, 33rd, 42nd and 52nd events. Dashed line shows $\beta(t)$.

censoring rate of 24%. One thousand replicates were generated. A proportional hazards Cox model with Z_i, a binary 0–1 group indicator, as covariate was fitted for each replicate and Schoenfeld residuals computed.

Plots of the average scaled Schoenfeld residuals revealed the functional form of $\beta_j(t)$ quite nicely. Figure 1 shows the average from all 1000 simulations, computed by applying a low bandwidth smoother to the set of all scaled Schoenfeld residuals. The two methods of standardizing, event specific variance $\hat{V}_k^{-1}\hat{r}_k$ and average variance $\mathscr{I}^{-1}\hat{r}_k/d^*$, are nearly indistinguishable. The true function $\beta_j(t)$ is superimposed. Even in the face of large amplitude nonproportionality, the mean scaled Schoenfeld residuals track the nonproportionality closely, although there appears to be some bias in the piecewise exponential model immediately following the change in hazard.

Figure 2 shows scaled Schoenfeld residuals plotted against time for two individual data sets from each simulation. Standardization by the average variance is presented; standardization by event-time specific variance, not shown, is virtually identical. The true function $\beta(t)$ has been superimposed on the scatter plot, as well as a loess smooth: see

below. The importance of smoothing is obvious in these plots. As is typical for a binary covariate, the residuals fall into two horizontal bands with no apparent structure. The smooths show the departure from proportionality; Fig. 2(a), (c) suggest a substantial increase in β at about $t = 0{\cdot}5$, and Fig. 2(b), (d) show β to be roughly monotonically increasing. Pointwise 90% confidence intervals, based on the null distribution of the smooths as discussed in the Appendix, are given at six event times. They are fairly broad; any single realization, based on fewer than 60 events, does not contain enough information to resolve the details of the functional form.

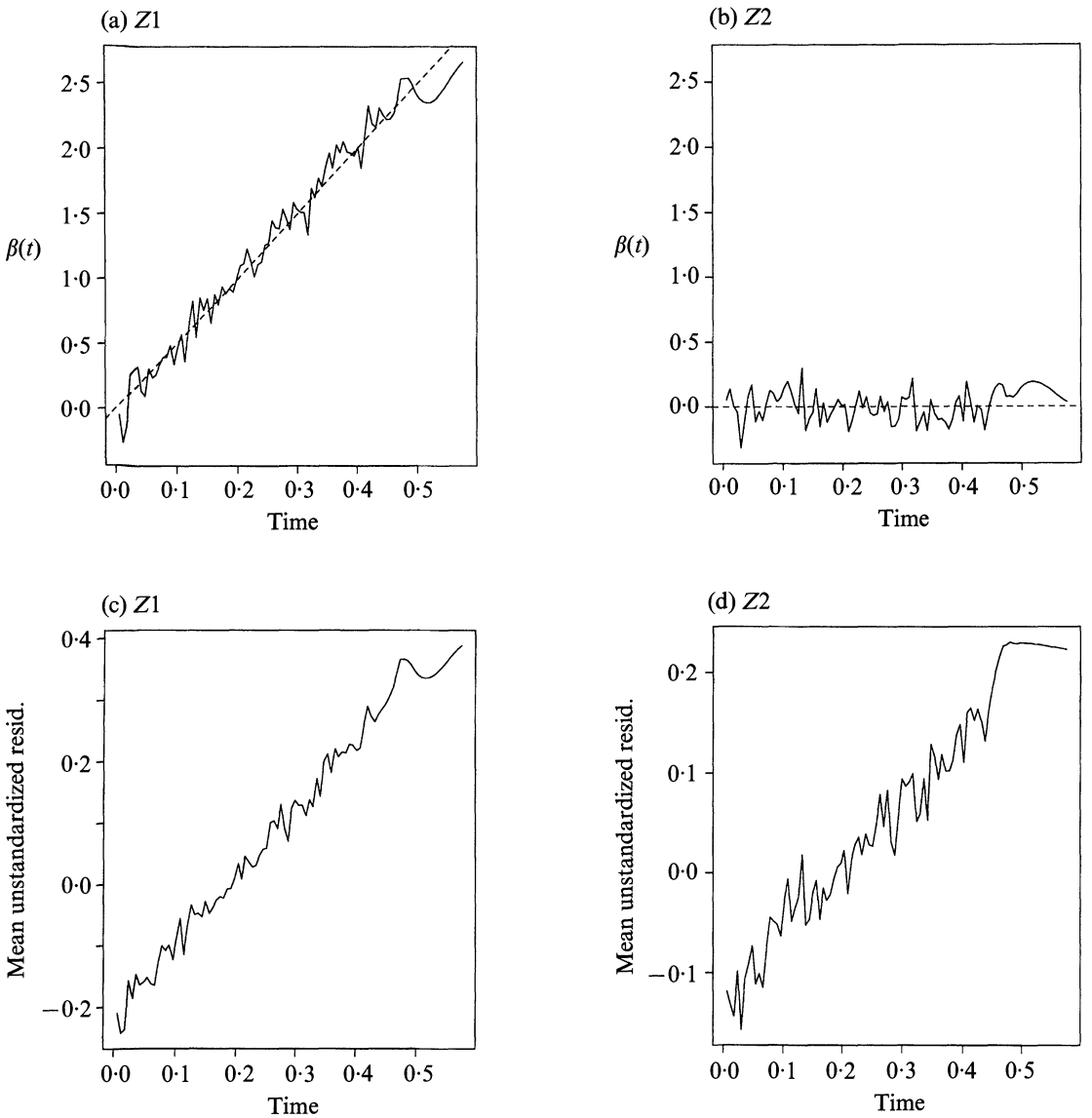

Fig. 3(a), (b). Smoothed (loess, span = 0·01) standardized Schoenfeld residuals + mean $\hat{\beta}$, using average variance standardization plotted against event time for a simulation of a nonproportional hazards model with two correlated binary covariates, described in text. Solid line shows the true functional form of the interaction between time and covariate. (c), (d). Smoothed (loess, span = 0·01) unstandardized Schoenfeld residuals for the same simulation as in (a), (b). Both components show a linear time trend, although only $Z1$ enters the model in a nonproportional hazards fashion.

Chambers et al. (1983) discuss scatterplot smoothing in detail, and show that even for continuous data such smooths are an essential aid to our visual system. This observation is borne out by nearly all the plots shown here. The choice of smoothing technique is usually not very important, as long as the smoother is sensitive to local rather than global features of the data set and has an appropriate number of degrees of freedom (Hastie & Tibshirani, 1990, Ch. 3, 4). The smooths displayed here are based on the loess algorithm (Cleveland & Devlin, 1988; Cleveland, Grosse & Shyu, 1992). The value of the smooth at any time point t^* is computed as the fitted value from a weighted linear least squares regression of the nearby scaled Schoenfeld residuals on event times. The proportion of event times with nonzero weight is called the span and the weights on points falling within the span decrease as a tricube function of the distance to t^*. The number of degrees of freedom is inversely proportional to the span. The smooths displayed here have between 3 and 4 degrees of freedom. Other smooths, such as smoothing or regression splines with similar degrees of freedom, give visually similar results.

Another simulation examined the case of two correlated binary covariates. Each realization had a sample size of 100. There were 40 each with covariate configuration $Z1 = Z2 = 0$ and $Z1 = Z2 = 1$, and 10 each with $Z1 = 0$, $Z2 = 1$ and $Z1 = 1$, $Z2 = 0$. The hazard was the same as in simulation 2:

$$h_i(t) = 2 \exp(5tZ_{1i}).$$

The same censoring distribution was used and there were 1000 replicates. A main effects Cox regression model with covariates Z_1 and Z_2 was fitted to each replicate.

Figure 3(a), (b) shows a smooth of the scaled Schoenfeld residuals. The first component shows a linear trend with slope 5 and the second component gives a straight line at 0·0, both in agreement with the true model. Figure 3(c), (d) plots the smoothed unstandardized Schoenfeld residuals against event times separately for each covariate. The plot for $Z1$ correctly suggests a linear departure from proportionality. The plot for $Z2$ also suggests a linear departure, even though the hazard does not depend on $Z2$. This phenomenon is due to the strong correlation between $Z1$ and $Z2$.

A real data application is the Veterans Administration lung cancer data (Kalbfleisch & Prentice, 1980, pp. 223–4), from a clinical trial of 137 male patients with advanced inoperable lung cancer. The end point was time to death and there were six covariates measured at randomization: cell type (squamous cell, large cell, small cell and adenocarcinoma), Karnofsky performance status, time in months from diagnosis, age in years, prior therapy (yes/no) and therapy (test chemotherapy versus standard). Lin's test (1991) comparing the Cox model $\hat{\beta}$ to a weighted estimate with the Peto–Prentice weight function found a

Table 1. *$T(G)$ tests for the Veterans Administration data with $g = \log(time)$*

Covariate	Test	Degrees of freedom	p
All (global test)	32·45	8	0·0001
Cell type	14·17	3	0·0027
Karnofsky score	9·93	1	0·0016
Months since diagnosis	0·31	1	0·5748
Age	3·25	1	0·0716
Prior therapy	2·69	1	0·1010
Treatment	0·24	1	0·6225

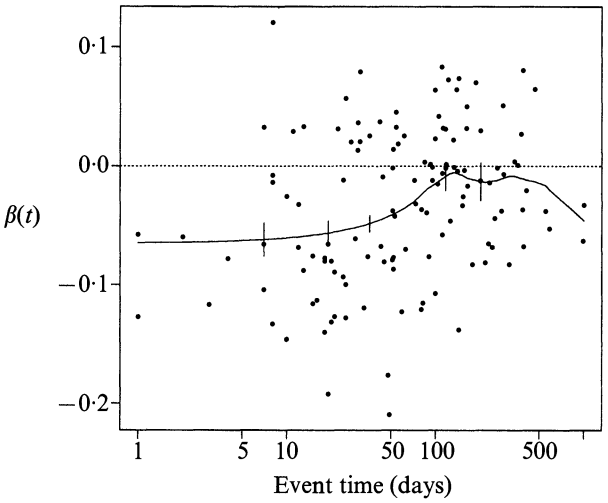

Fig. 4. Scaled Schoenfeld residuals $+\hat{\beta}$ for Karnofsky score plotted against event time from a main effects Cox model fitted to the Veterans data. Average variance standardization is used. A loess smooth (span $= \frac{3}{4}$ and degrees of freedom $= 3{\cdot}7$) is superimposed with 90% pointwise confidence intervals at the 6th, 26th, 46th, 86th and 106th events.

highly significant difference ($p = 0{\cdot}00002$), suggestive of nonproportionality. Table 1 summarizes test statistics, using (11) with log of event times for $g(t)$. Karnofsky score and cell-type were the only predictors with significant coefficients in the initial Cox model or with significant nonproportionality. Figure 4 shows the smooth scatter plot for Karnofsky score. Because the survival times have a long-tailed distribution, $\log(t_k)$ is used for the x-axis. Use of the Kaplan–Meier values for the x-axis (Peto–Prentice scores) gives a similar figure. The impact of Karnofsky score clearly changes with time. Early, a low score is protective. However, the effect diminishes over time and is effectively zero by 100 days. Another way of interpreting this would be that a 3–4 month old Karnofsky score is no longer medically useful. The downturn at the right end of the plot is probably an artifact of small numbers and disappears if the last four points are excluded.

4. Discussion

We have shown that many of the popular tests for proportional hazards are essentially tests for nonzero slopes in a generalized linear regression of the rescaled residuals on chosen function(s) of time. This regression framework clarifies the relationship among these tests; for example, because $N(t-)$ and $\hat{F}(t-)$ are proportional if there is no censoring, the tests of Harrell (1986) and Lin (1991) are equivalent in that case. More importantly, these tests are linked to an interpretable parameter $\beta(t)$, which is estimated by the smoothed scatterplot, and which helps a user understand the effect and importance of any nonproportionality which may be discovered by one of the formal tests. As seen in Fig. 2, the estimate of $\beta(t)$ may not be very accurate, however, when the study is small. Other explanations of nonproportional hazards besides a time varying coefficient are also possible. Andersen et al. (1993, Examples 7.3.1 and 7.3.4) achieve a good fit to a decaying treatment effect by introducing a frailty parameter.

Many of the individual ideas in this paper are not new: the expansion of $\beta(t)$ found in equation (4) is due to Schoenfeld (1982); time varying coefficient models are considered in general by Hastie & Tibshirani (1993), and for the Cox model in particular by Gamerman (1991) and by Zucker & Karr (1990). In each of these the focus is on estimating $\beta(t)$ directly by including it in the likelihood function. A test statistic T of the form of (11) is used by Jones & Crowley (1990) although not to test proportional hazards. A set of similar plots is given by Gray (1990), who discusses kernel-based smoothing of standard cumulative hazard estimates, including several methods of constructing partial residual plots. The derivation and formulae of his approach are quite different from those presented here. Henderson & Milner (1991) suggest plotting quantities that are nearly the cumulative sum of the scaled Schoenfeld residuals against event time. The largest concordance is with the work of Pettitt & Bin Daud (1990) who consider one-dimensional departures from proportional hazards. They also use equation (4) as a motivation to create standardized Schoenfeld residuals, and suggest smoothed scatter plots to explore the form of the departure. However, their standardization uses only the diagonal elements of $\hat{V}_k$ and is equivalent to ours only if the covariates are uncorrelated at each time point. When the covariates are correlated, the Pettitt & Bin Daud standardization may be misleading, as shown in Fig. 3, because time-dependencies in one covariate may show up in the plot of another covariate. This partial standardization also precludes equivalence of their method to existing tests of proportional hazards. Pettitt & Bin Daud discuss a variety of smoothing methods, and develop a pointwise confidence interval for the smooths. The width of these intervals may be inflated, because their methodology ignores the negative covariance between residuals as discussed in the Appendix.

A strength of our approach is its ease of application. To apply the average variance standardization, one needs only the Schoenfeld residuals, along with the coefficients and variance matrix from a standard time independent Cox model fit. These are readily available in standard statistical software, e.g., S-Plus (StatSci Inc., 1991). Virtually any of the commonly available scatter-plot smoothers will be adequate to reveal nonproportionality. If the smoother is a linear operator, such as loess or regression splines, standard linear model calculations can be used to create confidence intervals.

The plots should be treated as guides, with due caution, since they are computed using a one-step update to the fitted proportional hazards model. Many studies are designed with barely enough power to test $H_0: \beta = 0$ versus a practically significant alternative. It is then unrealistic to expect our plots to reveal fine details of the form of $\beta(t)$, and the scatterplot smoother would normally be adjusted to have only a few degrees of freedom. The confidence intervals, in particular, are based on a series of approximations, although our simulations showed them to be accurate for some simple cases.

If more refined estimates of $\beta(t)$ or its confidence intervals are required, the method given here could be iterated or one could model $\beta(t)$ directly by including appropriate time-dependent terms in the Z matrix. Another approach, suggested by a referee, replaces the individual variable scatterplot smooths with a global method that incorporates the between-variable correlations. However, the gain in accuracy is likely to be small, especially when compared to the increase in complexity and the difficulty of implementing these ideas.

APPENDIX

Confidence intervals

Most of the widely-used scatterplot smoothers are linear in the dependent variable y: for any x and y the fit is $\hat{y} = Ly$, where the smoothing matrix L depends on the spacing of the x values, a

prespecified smoothing parameter and any prior weights specified by the user, but not on y. Then we can write $\hat{Y} = L\hat{R}^* + \hat{\beta}$, where L is a $d \times d$ matrix and $\hat{R}^*$ has kth row $\hat{r}_k^* = \hat{V}_k^{-1}\hat{r}_k$. Under the null hypothesis of proportional hazards, each column j of $\hat{R}^*$ is asymptotically multivariate normal with mean 0 and variance-covariance S_j, say. Note that conditioning on the event times and risk sets var $(\hat{r}_k^*)$ can be estimated consistently by $\hat{V}_k^{-1} - \mathscr{I}^{-1}$ and cov $(\hat{r}_k^*, \hat{r}_l^*)$ by $-\mathscr{I}^{-1}$. Therefore, we estimate S_j by $A - \mathscr{I}_{j,j}^{-1}J + \mathscr{I}_{j,j}^{-1}I$, where A is a $d \times d$ diagonal matrix whose kth diagonal element is $\hat{V}_{k,j,j}^{-1}$, J is a $d \times d$ matrix of 1's, and I is the identity matrix. The third term is the variance of $\hat{\beta}_j$. Due to the optimality of the partial likelihood score equation under H_0 (Chang & Hsiung, 1990), $\hat{\beta}$ and the Schoenfeld residuals are asymptotically uncorrelated. If they were not, a more efficient estimator could be constructed from a combination of the two.

Conditioning on the observed failure times so that we can treat L as deterministic, the jth column of $\hat{Y}$ will be asymptotically normal with mean 0 and variance LS_jL'. Confidence intervals can be formed by standard linear model calculations, e.g. Scheffé intervals using the rank of LS_jL' for simultaneous confidence bands or simple z-intervals for pointwise estimates.

As a simplification discussed in § 4, one might consider using $\bar{V} = d^{-1}\mathscr{I}$ in place of $\hat{V}_k$. Then S_j simplifies to $\mathscr{I}_{jj}^{-1}\{(d+1)I - J\}$. For smoothers based on linear regression against a basis matrix X, such as splines, the calculation is very similar to that for the ordinary regression hat matrix $H = X(X'X)^{-1}X'$, in that a $d \times d$ matrix need not ever be explicitly constructed. If the covariate term J were ignored the result is exactly $(d+1)\mathscr{I}_{jj}^{-1}H$. The problem is computationally more complex for the loess smoother (Cleveland, Devlin & Grosse, 1988). However, one typically wants pointwise confidence intervals at only a few time points and therefore only the rows of L corresponding to those points are needed. As a practical matter, we have rarely found the simplified approach to lead to different conclusions from using the $\hat{V}_k$'s. The following simulation results give an example of how well it performs under the null hypothesis. We considered two groups with exponential failure times and rates 1 and 2 respectively. There were 30 per group, no censoring and 1000 replications. For each replication, we fitted the Cox proportional hazards model with group indicator as covariate, computed the standardized Schoenfeld residuals with average variance standardization and calculated a loess smooth for a span of 0·40 (5·1 equivalent degrees of freedom) and 0·75 (3·1 equivalent degrees of freedom). Table A1 compares the true, i.e. simulation, standard error of the smooth with the mean estimated standard error and shows the proportion of nominal 90% confidence intervals that covered the true value for the 10th, 20th, 30th and 40th event times. The empirical results are close to the theoretical in all cases.

Table A1. *Simulation results*

	Span = 0·75				Span = 0·40			
	Event rank				Event rank			
	10	20	30	40	10	20	30	40
True standard error	0·352	0·230	0·323	0·349	0·438	0·505	0·525	0·513
Mean estimated standard error:								
using theoretical covariance	0·352	0·223	0·302	0·344	0·438	0·486	0·490	0·506
assuming 0 covariance	0·450	0·358	0·413	0·444	0·521	0·562	0·565	0·579
Coverage percent for nominal 90% confidence interval:								
using theoretical covariance	90·6	90·7	88·5	90·1	90·3	88·8	87·5	91·1
assuming 0 covariance	96·9	98·8	96·5	97·6	96·4	93·8	92·0	95·1

Although the covariance is roughly $-1/d$ times the variance under H_0, it cannot be neglected in computing standard errors for the smooth. The rows of Table A1 labelled 'assuming 0 covariance' show the results of ignoring the covariance (J) term. The estimated standard errors are inflated. The 90% confidence intervals are too wide and have coverage probabilities closer to 95%.

References

Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer-Verlag.

CHAMBERS, J. M., CLEVELAND, W. S., KLEINER, B. & TUKEY, P. A. (1983). *Graphical Methods for Data Analysis.* Belmont, CA: Wadsworth.

CHANG, I. S. & HSIUNG, C. A. (1990). Finite sample optimality of maximum partial likelihood estimation in Cox's model for counting processes. *J. Statist. Plan. Inf.* **25**, 35–42.

CHAPPELL, R. (1992). A note on linear rank tests and Gill & Schumacher's tests of proportionality. *Biometrika* **79**, 199–201.

CLEVELAND, W. S. & DEVLIN, S. J. (1988). Locally weighted regression: and approach to regression analysis by local fitting. *J. Am. Statist. Assoc.* **83**, 596–610.

CLEVELAND, W. S., DEVLIN, S. J. & GROSSE, E. (1988). Regression by local fitting: methods, properties, and computational algorithms. *J. Econometrics* **37**, 87–114.

CLEVELAND, W. S., GROSSE, E. & SHYU, W. M. (1992). Local regression models. In *Statistical Models in S*, Ed. J. M. Chambers and J. J. Hastie, pp. 309–76. Pacific Grove, CA: Wadsworth & Brooks.

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.

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

GAMERMAN, D. (1991). Dynamic Bayesian models for survival data. *Appl. Statist.* **40**, 63–79.

GILL, R. & SCHUMACHER, M. (1987). A simple test of the proportional hazards assumption. *Biometrika* **74**, 289–300.

GRAY, R. J. (1990). Some diagnostic methods for Cox regression models through hazard smoothing. *Biometrika* **46**, 93–102.

HARRELL, F. E. (1986). *The PHGLM Procedure, SAS Supplemented Library User's Guide*, Version 5, Cary, NC: SAS Institute Inc.

HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models.* London: Chapman and Hall.

HASTIE, T. & TIBSHIRANI, R. (1993). Varying-coefficient models (with discussion). *J. R. Statist. Soc.* B **55**, 757–96.

HENDERSON, R. & MILNER, A. (1991). Aalen plots under proportional hazards. *Appl. Statist.* **40**, 401–9.

JONES, M. P. & CROWLEY, J. (1990). Asymptotic properties of a general class of nonparametric tests for survival analysis. *Ann. Statist.* **18**, 1203–20.

KALBFLEISCH, J. D. & PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: John Wiley.

LIN, D. Y. (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *J. Am. Statist. Assoc.* **86**, 725–8.

MOREAU, T., O'QUIGLEY, J. & MESBAH, M. (1985). A global goodness-of-fit statistic for the proportional hazards model. *Appl. Statist.* **34**, 212–8.

NAGELKERKE, N. J. D., OOSTING, J. & HART, A. A. M. (1984). A simple test for goodness of fit of Cox's proportional hazards model. *Biometrics* **40**, 483–6.

O'QUIGLEY, J. & PESSIONE, F. (1989). Score tests for homogeneity of regression effects in the proportional hazards model. *Biometrics* **45**, 135–44.

PETTITT, A. N. & BIN DAUD, I. (1990). Investigating time dependence in Cox's proportional hazards model. *Appl. Statist.* **39**, 313–29.

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

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

STATSCI INC. (1991). *S-Plus User's Manual.* Seattle, WA: Statistical Sciences Inc.

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

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

ZUCKER, D. M. & KARR, A. F. (1990). Non-parametric survival analysis with time-dependent covariate effects: a penalized likelihood approach. *Ann. Statist.* **18**, 329–53.

[*Received February* 1993. *Revised January* 1994]