Statistica Sinica 1(1991), 1-17

GOODNESS-OF-FIT TESTS FOR THE GENERAL COX REGRESSION MODEL

D. Y. Lin and L. J. Wei

University of Washington and University of Wisconsin

Abstract: In this article, we extend the information matrix tests proposed by White (1982) for detecting parametric model misspecification to the partial likelihood setting with particular interest in the Cox semi-parametric regression model. First we identify two model-based consistent estimators for the inverse of the asymptotic covariance matrix of the maximum partial likelihood estimator in the Cox model. We then show that under the assumed model the difference between these two estimators is asymptotically normal with mean zero and with a covariance matrix which can be consistently estimated. Goodness-of-fit tests for the Cox model are constructed based on these asymptotic results. Extensive Monte Carlo studies indicate that the large-sample approximation is appropriate for practical use. In addition, we demonstrate that the proposed tests tend to be more powerful than other numerical methods in the literature. Two examples are provided for illustrations.

Key words and phrases: Information matrix, martingale, model misspecification, partial likelihood, proportional hazards, survival data.

1. Introduction

The Cox (1972) regression model has become the most widely used statistical tool for analyzing censored failure time data due to its flexibility and versatility. The model specifies that the hazard function $h(t) = \lim_{d\downarrow 0} d^{-1} \Pr[T \leq t+d | T \geq t]$ for the failure time T of an individual with a possibly time-varying p-vector of covariates Z has the following form

$$h\{t|Z(t)\} = \lambda_0(t) \exp\{\beta_0' Z(t)\}, \tag{1.1}$$

where β_0 is a p-vector of unknown regression coefficients, and $\lambda_0(t)$ is an unspecified baseline hazard function.

Let $X_1, \ldots, X_n$ be n possibly right-censored failure times and $Z_1, \ldots, Z_n$ be the corresponding covariate vectors. Then the parameter vector β_0 is usually estimated by $\hat{\beta}$ which maximizes the partial likelihood function (Cox (1972, 1975))

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\{\beta' Z_i(X_i)\}}{\sum_{j \in \mathcal{R}_i} \exp\{\beta' Z_j(X_i)\}} \right]^{\Delta_i}, \tag{1.2}$$

where $\mathcal{R}_i$ is the set of labels attached to the individuals at risk at time X_i^-, and $\Delta_i = 1$ if X_i is an observed failure time and $\Delta_i = 0$ otherwise.

Model (1.1) assumes that (i) all relevant covariates are included; (ii) the regression form of the hazard function on covariates is exponential; and (iii) the relationship between the baseline hazard function and the regression function of covariates is multiplicative. The violation of these assumptions may have adverse effects on the statistical inference. For instance, when an independent covariate is omitted from a proportional hazards model, $\hat{\beta}$ as an estimator of the regression parameter in the true model is asymptotically biased toward zero (Struthers and Kalbfleisch (1986)). In addition, model misspecification can lead to distortion of the size and reduction of the power of the partial likelihood score test (Lagakos and Schoenfeld (1984), Lagakos (1988), Lin and Wei (1989)).

Various graphical techniques have been proposed to check the aforementioned assumptions (e.g., Crowley and Hu (1977), Kay (1977), Cox (1979), Kalbfleisch and Prentice (1980, pp. 87–98), Lagakos (1980), Andersen (1982), Schoenfeld (1982), Crowley and Storer (1983), Arjas (1988)). The diagnostic plots can be quite informative. The difficulty with these procedures is that they are rather subjective.

Numerical tests for the fit of the Cox model have also been studied. In his original paper, Cox (1972) proposed a way of model checking by introducing a 'dummy' time-varying covariate. This method is restricted to testing against a specific alternative. Schoenfeld (1980) compared the observed and the expected numbers of deaths in the cells arising from a partition of the Cartesian product of the range of covariates and the times axis. Similar approaches were taken by Moreau, O'Quigley and Mesbah (1985), and Moreau, O'Quigley and Lellouch (1986). However, the partitions of time-axis and covariates are often arbitrary. In addition, different partitions might lead to conflicting results (see Section 3). Wei (1984) proposed an omnibus test for the two-sample problem. The extension of his method to the general one-parameter Cox model is straightforward (see Wei (1984), Haara (1987)). Gill and Schumacher (1987) constructed simple tests for detecting monotone departures from the constant hazard ratio assumption by comparing different generalized rank estimators of the relative risk. Other numerical procedures appeared in Andersen (1982), Breslow, Edler and Berger (1984), Ciampi and Etezadi-Amoli (1985), Nagelkerke, Oosting and Hart (1984), and O'Quigley and Pessione (1989). Again, they are only applicable to specific problems or require arbitrary decisions by the user. Therefore, it is desirable to develop global goodness-of-fit tests for the general Cox model without the above constraints.

In the ordinary likelihood setting, White (1982) exploited the properties of the information matrix to yield several useful tests for model misspecification. The idea of White is as follows. Let $X_1, \ldots, X_n$ be a random sample (without censoring) from a distribution $F(X)$. Suppose that one is interested in testing whether the density function of $F(X)$ is given by $f_0(X;\theta_0)$, where θ_0 is a vector of unknown parameters. Let $\hat{\theta}$ denote the maximum likelihood estimator of θ_0 from the log-likelihood function $l(\theta) = \sum l_i(\theta)$, where $l_i(\theta) = \log f_0(X_i;\theta)$. Then, under the assumed model, the Fisher's information matrix can be consistently estimated by either the score derivative matrix $A_n(\hat{\theta}) = -n^{-1}\partial^2 l(\theta)/\partial\theta^2|_{\theta=\hat{\theta}}$ or the squared score matrix $B_n(\hat{\theta}) = n^{-1}\sum\{\partial l_i(\theta)/\partial\theta\}\{\partial l_i(\theta)/\partial\theta\}'|_{\theta=\hat{\theta}}$, provided that some regularity conditions are satisfied. A significant discrepancy between $A_n(\hat{\theta})$ and $B_n(\hat{\theta})$ indicates that the model $f_0(X;\theta_0)$ is misspecified.

In Section 2, we apply White's idea to the partial likelihood setting with particular interest in the Cox regression model. First, we identify two model-based consistent estimators $A_n(\hat{\beta})$ and $B_n(\hat{\beta})$, which are similar to $A_n(\hat{\theta})$ and $B_n(\hat{\theta})$ defined above, for the inverse of the asymptotic covariance matrix of $n^{1/2}\hat{\beta}$. Secondly, it is shown that under model (1.1) the statistic $n^{1/2}(A_n(\hat{\beta}) - B_n(\hat{\beta}))$ is asymptotically normal with mean zero and with a covariance matrix for which a consistent estimator is proposed. Goodness-of-fit tests are then constructed based on this statistic. Two real-life examples are provided in Section 3 for illustrations. The finite-sample properties of the proposed tests are investigated in Section 4.

2. Construction of Test Statistics

For $i = 1, \ldots, n$, let $N_i(t) = I(X_i \le t, \Delta_i = 1)$ and $Y_i(t) = I(X_i \ge t)$, where $I(\cdot)$ is the indicator function. We assume that (N_i, Y_i, Z_i) $(i = 1, \ldots, n)$ are independent and identically distributed and that covariates are bounded.

The logarithm of the partial likelihood function (1.2) can be expressed as $l(\beta) = \sum l_i(\beta)$, where

$$l_i(\beta) = \int_0^\infty \{\beta' Z_i(u) - \log[S^{(0)}(\beta, u)]\} dN_i(u) \tag{2.1}$$

with $S^{(0)}(\beta, u) = n^{-1}\sum Y_j(u)\exp\{\beta' Z_j(u)\}$.

The asymptotic normality of the maximum partial likelihood estimator $\hat{\beta}$ for the Cox model has been established by Liu and Crowley (1978), Tsiatis (1981), Andersen and Gill (1982), Naes (1982), and Bailey (1983) among others. Let Ω denote the inverse of the asymptotic covariance matrix of $n^{1/2}(\hat{\beta} - \beta_0)$ under the assumed model (1.1). Then Ω can be consistently estimated by the score derivative matrix $A_n(\hat{\beta}) = -n^{-1}\partial^2 l(\beta)/\partial\beta^2|_{\beta=\hat{\beta}}$, i.e.,

$$A_n(\hat{\beta}) = n^{-1}\int_0^\infty \left\{\frac{S^{(2)}(\hat{\beta},u)}{S^{(0)}(\hat{\beta},u)} - \frac{S^{(1)}(\hat{\beta},u)^{\otimes 2}}{S^{(0)}(\hat{\beta},u)^2}\right\} d\bar{N}(u), \tag{2.2}$$

where

$$S^{(1)}(\beta,t) = n^{-1}\sum Y_j(t)\exp\{\beta' Z_j(t)\}Z_j(t),$$
$$S^{(2)}(\beta,t) = n^{-1}\sum Y_j(t)\exp\{\beta' Z_j(t)\}Z_j(t)^{\otimes 2},$$

$\bar{N} = \sum N_j$, and $a^{\otimes 2}$ denotes the matrix aa' for a column vector a (see Andersen and Gill (1982), Theorem 4.2). On the other hand, we show in the theorem given below that the estimator in the outer-product form

$$B_n(\hat{\beta}) = n^{-1}\sum_{i=1}^n \{\partial l_i(\beta)/\partial\beta\}\{\partial l_i(\beta)/\partial\beta\}'|_{\beta=\hat{\beta}},$$

i.e.,

$$B_n(\hat{\beta}) = n^{-1}\sum_{i=1}^n \left\{\int_0^\infty \left[Z_i(u) - \frac{S^{(1)}(\hat{\beta},u)}{S^{(0)}(\hat{\beta},u)}\right] dN_i(u)\right\}^{\otimes 2} \tag{2.3}$$

is also consistent for Ω. A significant difference between $A_n(\hat{\beta})$ and $B_n(\hat{\beta})$ suggests that model (1.1) is incorrect.

Let $D_n(\hat{\beta}) = A_n(\hat{\beta}) - B_n(\hat{\beta})$, and let vec$(H)$ denote the mn-vector whose $((i-1)m+j)$th component is the (i,j)th element of an $m \times n$ matrix H. The asymptotic null distribution of $n^{1/2}\text{vec}(D_n(\hat{\beta}))$ is given in the following theorem.

Theorem 2.1. *Under the assumed Cox model* (1.1), $n^{1/2}\text{vec}(D_n(\hat{\beta}))$ *is asymptotically normal with zero mean and with covariance matrix* $Q(\beta_0)$, *where the* $p^2 \times p^2$ *matrix* $Q(\beta_0)$ *can be consistently estimated by* $Q_n(\hat{\beta})$ *given in* (A.5) *of Appendix* A.

The proof of Theorem 2.1 is given in Appendix A.

Let the $p(p+1)/2$-vector $d_n(\hat{\beta})$ consist of the upper triangular elements of $D_n(\hat{\beta})$. The covariance matrix of $d_n(\hat{\beta})$, denoted by $\tilde{Q}_n(\hat{\beta})$, is a subset of $Q_n(\hat{\beta})$. A class of goodness-of-fit tests can be constructed based on the asymptotic distribution of $d_n(\hat{\beta})$. For example, let T_n be the maximum absolute value of the $p(p+1)/2$ standardized components of $d_n(\hat{\beta})$. Then an extremely large value of T_n is an indication of model misspecification. The p-value for this maximum test T_n can be easily obtained by either numerical integration or simulation. Alternatively, one can form the Wald statistic $W_n = nd_n(\hat{\beta})'\tilde{Q}_n(\hat{\beta})^{-1}d_n(\hat{\beta})$. It follows from Theorem 2.1 that W_n converges weakly to a central chi-square variable with $p(p+1)/2$ degrees of freedom under model (1.1). Notice that the maximum test T_n and the Wald test W_n are equivalent for the one-parameter Cox model. The consistency of the proposed tests is discussed in Appendix B.

In the ordinary likelihood situation, some components of d_n may be linear combinations of others (White (1982)), leading to singularity of the covariance matrix $\tilde{Q}_n$. It is not clear whether this is likely to occur in the Cox model. The singularity of $\tilde{Q}_n(\hat{\beta})$ would not be an obstacle to the test T_n, but would require appropriate adjustment for the test W_n. Specifically, $\tilde{Q}_n(\hat{\beta})^{-1}$ would be replaced by a reflexive g-inverse of $\tilde{Q}_n(\hat{\beta})$ (Rao (1973), p. 26), and the degrees of freedom $p(p+1)/2$ by the rank of $\tilde{Q}_n(\hat{\beta})$. The singularity of $\tilde{Q}_n(\hat{\beta})$ can be deleted empirically by the so-called condition number, which is defined as the ratio of the largest over the smallest eigenvalue (see Kennedy and Gentle (1980), p. 279).

A FORTRAN-77 program for implementing the above two tests is available from the first author.

3. Examples

We now apply the proposed goodness-of-fit tests to two familiar data sets. The first one is taken from Freireich et al. (see Cox (1972)). As shown in Table 1, the data set consists of the times to remission for two groups of leukemia patients. The only covariate is the group indicator, which is coded as 0 or 1. The Cox estimator $\hat{\beta}$ is 1.5092. Let $SE_1(\hat{\beta})$ and $SE_2(\hat{\beta})$ denote the estimated standard errors of $\hat{\beta}$ based on $A_n(\hat{\beta})$ and $B_n(\hat{\beta})$, respectively. In this example, $SE_1(\hat{\beta}) = 0.4096$ and $SE_2(\hat{\beta}) = 0.4227$. The z-score for T_n or W_n equals 0.7049, which provides no evidence against the assumed model. This result confirms the findings of Cox (1972), Nagelkerke, Oosting and Hart (1984), Wei (1984), and Gill and Schumacher (1987). Schoenfeld's test for this data set yields a p-value of 0.43 when the time axis is divided at 11 weeks (Schoenfeld (1980)) versus a p-value of 0.08 when the time axis is divided at 5 weeks (Schoenfeld (1982)).

As another example, we consider the Stanford heart transplant data as of February 1980. This data set contains the survival times of 184 heart-transplanted patients along with their ages at the time of the first transplant and $T5$ mismatch scores. 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 February 1980. Data listings and further details can be found in Miller and Halpern (1982).

Three proportional hazards models are fitted to this data set and the results summarized in Table 2. For computational reasons and ease of interpretation, the variable age is centered around 41.7, the approximate sample mean of the patients' ages. The condition numbers for the three covariance matrices $\tilde{Q}_n(\hat{\beta})$ are 2.23, 7.90 and 6.27, respectively, which indicates that none of them are near-singular.

The first model $\beta_1(\text{age}) + \beta_2(T5)$, which is labeled Model 1 in Table 2, includes only the linear effects. Notice that $SE_1(\hat{\beta}_1)$ is about 20% larger than $SE_2(\hat{\beta}_1)$. The goodness-of-fit tests provide strong evidence to discredit this model.

The more general model $\beta_1(\text{age}) + \beta_2(T5) + \beta_3(\text{age})^2$, which is denoted by Model 2 in Table 2, fits the data quite well. Here, $SE_1(\hat{\beta}_1)$ is almost identical to $SE_2(\hat{\beta}_1)$. $T5$ is clearly insignificant and could be deleted.

The final age quadratic model $\beta_1(\text{age}) + \beta_3(\text{age})^2$, which is Model 3 in Table 2, provides a satisfactory description of the data. As in the previous two models, the p-value based on T_n is rather similar to that of W_n in this model. The two sets of standard error estimates agree extremely well. Both linear and quadratic age effects are highly significant. Incidently, the plot of generalized residuals by Miller and Halpern (1982) also indicated the appropriateness of this age quadratic model.

4. Finite-Sample Properties of Test Statistics

Extensive Monte Carlo studies were carried out to assess the performance of the new goodness-of-fit tests for practical sample sizes. For comparisons, the tests proposed by Cox (1972), Schoenfeld (1980), Wei (1984), and Gill and Schumacher (1987) were also evaluated. The fitted models were proportional hazards models with a dichotomous or continuous covariate. The Cox goodness-of-fit test was defined as the partial likelihood score test for testing the hypothesis that the regression coefficient associated with the 'dummy' time-varying covariate tZ is 0. The log-rank and Prentice's Wilcoxon weight functions were used in Gill-Schumacher's test. For Schoenfeld's test, the time axis was always partitioned at the sample median of survival times; each value of a dichotomous covariate was used as a partition in two-sample cases; and a continuous covariate was split at its mean. For larger samples, finer partitions might have been more appropriate but for simplicity were not chosen in our studies.

The survival distributions selected in the studies included exponential with density function $\rho\exp(-\rho x)$, Weibull with density $\tau\rho(\rho x)^{\tau-1}\exp\{-(\rho x)^{\tau}\}$, gamma with density $x^{k-1}e^{-x}/\Gamma(k)$, and log-normal with density $(2\pi)^{-1/2}(\sigma x)^{-1}$ $\exp\{-(\log x-\mu)^2/(2\sigma^2)\}$. Censorship was imposed by the generation of independent uniform random variables U on the interval $(0,c)$, where c was a suitably chosen integer so that observations in each simulation sample had about 25%

chance of being censored. Uniform random numbers were generated through an algorithm provided by Press, Flannery, Teukolsky and Vetterling (1986, pp. 196-197).

The key results from these Monte Carlo studies are summarized in Tables 3 through 7. Shown in the headings to the tables are the true survival model, censoring distribution, empirical proportion of censored observations, sample size and nominal level α. The empirical sizes or powers of the goodness-of-fit tests being compared are presented in the same order as the order of their appearance in the title of the table. Each entry in the table was calculated from 1,000 replications, so the observed tail probabilities have standard errors of about 0.003 for the first percentile and about 0.007 for the fifth percentile.

Table 3 displays the empirical Type I error probabilities of the five goodness-of-fit tests for the one-parameter Cox model with a normal or dichotomous covariate. The empirical error probability of the new test (T_n or W_n) may slightly exceed the nominal level in some smaller samples especially at the 1% level, whereas Wei's test seems rather conservative. The empirical sizes of the other tests are fairly close to the nominal level.

Table 4 presents the empirical powers of the five tests for detecting the violation of the constant hazard ratio assumption in the two-sample case. The new test has greater power than the other four tests in case (A) but is not as powerful as the others in case (B). Cox's and Gill-Schumacher's tests are expected to perform well in case (B), where the true hazard ratio is $3t$, since these two tests were designed to detect such a monotone departure from proportionality. As one referee pointed out, the new test may not be very sensitive to monotone deviations because the two model-based variance estimators contain no direct information about the relationship between residuals (Schoenfeld (1982)) and failure times.

Table 5 demonstrates the superior performance of the new test for detecting the omission of relevant covariates. In both cases (A) and (B), the omission of covariates induces non-monotone departures from the proportional hazards assumption, which Cox's and Gill-Schumacher's tests have little power in detecting. In addition, one would not expect Schoenfeld's test with two partitions on the time axis and the covariate space to be sensitive to a non-monotone departure from proportionality or to a quadratic alternative to a linear fit. Case (A) is more interesting than case (B) because the size of the partial likelihood score test for testing the effect of Z being zero is distorted in the former but not in the latter (see Lin and Wei (1989)). In case (A), the partial likelihood score test for testing no effect of Z^2 with 0.05 Type I error has power of 0.75, 0.95 and 1.00 for $n = 50, 100$ and 200, respectively, without censoring, and 0.68, 0.92 and 0.99 for $n = 50, 100$ and 200, respectively, in the presence of censoring. The score test, which is optimal for this specific alternative, is more powerful than the proposed goodness-of-fit test, but the difference becomes marginal as the sample size increases.

In cases (A) and (B) of Table 6, the true regression forms of the hazard function on the normal covariate are, respectively, linear and natural-logarithmic instead of exponential. Again, only the new test has adequate power.

Table 7 consists of two models which are not proportional hazards models. Case (A) is an accelerated failure time model with a normal error term, which specifies that the effect of the covariate is multiplicative on the failure time T rather than on the hazard function. Case (B) is similar to case (A) except that the error term acts on T additively rather than multiplicatively. Cox's test is somewhat more powerful than the other three tests in case (A) whereas the new test performs the best in case (B).

In summary, the new test maintains its size near the nominal level, which reflects the appropriateness of the normal approximation for practical use. Furthermore, it has adequate power for detecting violation of the three assumptions in the Cox model. The tests proposed by Cox (1972), Schoenfeld (1980), Wei (1984), and Gill and Schumacher (1987) are sensitive to monotone departures, but are not powerful for testing other alternatives. Since the proposed test is designed to detect general departures from Model (1.1), it can not be expected to have optimal power properties against specific alternatives that can be tested by specific tests. To identify the direction of model misspecification after the proposed test rejects the null hypothesis, the user is encouraged to use some specific tests and diagnostic plots.

Acknowledgements

This research was supported by a Population Council Fellowship (for Lin) and the National Institutes of Health Grant RO1–CA45544 (for Wei). The authors are grateful to two referees, an editorial board member, Morton B. Brown, Charles S. Davis and Robert Wolfe for their helpful comments.

Appendix A: Proof of Theorem 2.1

The Cox model (1.1) can be re-expressed in terms of the multivariate counting process $N = (N_1, \ldots, N_n)$ with the intensity process

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta_0' Z_i(t)\}, \quad i = 1, \ldots, n \tag{A.1}$$

(see Andersen and Gill (1982)). We will repeatedly use the assumptions that (N_i, Y_i, Z_i) $(i = 1, \ldots, n)$ are independent and identically distributed and that the covariates are bounded without specifically referring to them in the proof. We also assume that Ω is positive definite and that $\Pr(Y_i(t) = 1, \text{ for all } t \leq \tau) > 0$ for each $\tau < \infty$ and all i.

Taylor series expansion of $\text{vec}(D_n(\hat{\beta}))$ around β_0 gives

$$\text{vec}(D_n(\hat{\beta})) = \text{vec}(D_n(\beta_0)) + \nabla\text{vec}(D_n(\tilde{\beta}))(\hat{\beta} - \beta_0),$$

where $\nabla\text{vec}(D_n(\beta))$ is the $p^2 \times p$ Jacobian matrix of $\text{vec}(D_n(\beta))$ with respect to β, and $\tilde{\beta}$ is on the line segment between $\hat{\beta}$ and β_0. Using the arguments similar to those given in the proof of the consistency of $A_n(\hat{\beta})$ for Ω on pp. 1107–1108 of Andersen and Gill (1982), we can show that $\nabla\text{vec}(D_n(\beta))$ converges in probability to a deterministic matrix $J(\beta)$ uniformly in β around β_0.

Now, let

$$M_i(t) = N_i(t) - \int_0^t \lambda_i(u)du, \quad i = 1, \ldots, n. \tag{A.2}$$

Then $n^{1/2}(\hat{\beta} - \beta_0)$ is asymptotically equivalent to

$$n^{-1/2}\Omega^{-1} \sum_{i=1}^{n} \int_0^{\infty} \{Z_i(u) - E(\beta_0, u)\} dM_i(u),$$

where $E(\beta, u) = S^{(1)}(\beta, u)/S^{(0)}(\beta, u)$. The matrix $B_n(\beta)$ can be rewritten as $n^{-1} \sum \int_0^{\infty} \{Z_i(u) - E(\beta, u)\}^{\otimes 2} dN_i(u)$ due to the fact that $\int_0^{\infty} G(\beta, u) dN_i(u) = \Delta_i G(\beta, X_i)$ for any predictable matrix $G(\beta, u)$. Thus,

$$n^{1/2}\text{vec}(D_n(\beta_0)) = n^{-1/2} \sum_{i=1}^{n} \int_0^{\infty} \text{vec}(R_i(\beta_0, u)) dN_i(u), \tag{A.3}$$

where $R_i(\beta, u) = S^{(2)}(\beta, u)/S^{(0)}(\beta, u) - E(\beta, u)^{\otimes 2} - \{Z_i(u) - E(\beta, u)\}^{\otimes 2}$. It is easy to show that the right hand side of (A.3) equals $n^{-1/2} \sum \int_0^{\infty} \text{vec}(R_i(\beta_0, u))$ $dM_i(u)$ under the assumed model (A.1). It follows that $n^{1/2}\text{vec}(D_n(\hat{\beta}))$ is asymptotically equivalent to

$$n^{-1/2} \sum_{i=1}^{n} \int_0^{\infty} \Big\{ \text{vec}(R_i(\beta_0, u)) + J(\beta_0)\Omega^{-1}\left[Z_i(u) - E(\beta_0, u)\right] \Big\} dM_i(u). \tag{A.4}$$

By Rebolledo's Central Limit Theorem for local square integrable martingales (see Theorem I.2 of Andersen and Gill (1982) and Theorems 4.2.1 and 4.3.1 of Gill (1980)), expression (A.4) converges weakly to a normal variable with mean 0 and with covariance matrix $Q(\beta_0)$, which can be consistently estimated by

$$Q_n(\hat{\beta}) = n^{-1} \sum_{i=1}^{n} \int_0^{\infty} \Big\{ \text{vec}(R_i(\hat{\beta}, u)) + \nabla \text{vec}(D_n(\hat{\beta})) A_n(\hat{\beta})^{-1} \left[Z_i(u) - E(\hat{\beta}, u)\right] \Big\}^{\otimes 2} dN_i(u). \tag{A.5}$$

Note that $\nabla\text{vec}(D_n(\hat\beta)) = n^{-1}\sum\int_0^\infty \nabla\text{vec}(R_i(\hat\beta,u))dN_i(u)$. Again, the consistency of $Q_n(\hat\beta)$ for $Q(\beta_0)$ can be established by the arguments on pp. 1107–1108 of Andersen and Gill (1982), which completes the proof.

Appendix B: Consistency of Test Statistics

Let $h_i(t)$ denote the true hazard function of the ith individual. It is convenient to introduce the following notation

$$S^{(r)}(t) = n^{-1}\sum_{i=1}^n Y_i(t)h_i(t)Z_i(t)^{\otimes r},$$
$$s^{(r)}(t) = \mathcal{E}\{S^{(r)}(t)\}, \qquad s^{(r)}(\beta,t) = \mathcal{E}\{S^{(r)}(\beta,t)\}$$

for $r = 0,1,2$, where the expectations are taken with respect to the true model of (X_i,Δ_i,Z_i) $(i = 1,\dots,n)$. Also, let $E(\beta,t) = S^{(1)}(\beta,t)/S^{(0)}(\beta,t)$, $e(\beta,t) = s^{(1)}(\beta,t)/s^{(0)}(\beta,t)$, $V(\beta,t) = S^{(2)}(\beta,t)/S^{(0)}(\beta,t) - E(\beta,t)^{\otimes 2}$, and $v(\beta,t) = s^{(2)}(\beta,t)/s^{(0)}(\beta,t) - e(\beta,t)^{\otimes 2}$. The quantities $E(t), e(t), V(t)$, and $v(t)$ are defined similarly.

Under a possibly misspecified Cox model, the maximum partial likelihood estimator $\hat\beta$ converges in probability to a p-vector of constants β^* which is the unique solution to the system of p equations

$$\int_0^\infty \{e(t) - e(\beta,t)\}s^{(0)}(t)dt = 0 \tag{B.1}$$

provided that the $p\times p$ matrix $\int_0^\infty v(\beta^*,t)s^{(0)}(t)dt$ is positive definite (Struthers and Kalbfleisch (1986), Lin and Wei (1989)).

It follows from (A.2) and (A.3) that

$$\begin{aligned} D_n(\beta) &= n^{-1}\sum_{i=1}^n\int_0^\infty \{V(\beta,t) - [Z_i(t) - E(\beta,t)]^{\otimes 2}\}dN_i(t) \\ &= n^{-1}\sum_{i=1}^n\int_0^\infty \{V(\beta,t) - [Z_i(t) - E(\beta,t)]^{\otimes 2}\}dM_i(t) \\ &\quad - \int_0^\infty \{V(t) - V(\beta,t) + [E(t) - E(\beta,t)]^{\otimes 2}\}S^{(0)}(t)dt. \end{aligned}$$

By standard counting process techniques, we can show that $D_n(\hat\beta)$ converges in probability to $D(\beta^*)$, where

$$D(\beta) = -\int_0^\infty \{v(t) - v(\beta,t) + [e(t) - e(\beta,t)]^{\otimes 2}\}s^{(0)}(t)dt.$$

In addition, the covariance matrix estimator $Q_n(\hat{\beta})$ given in (A.5) converges in probability to a positive semidefinite matrix. Therefore, the goodness-of-fit tests based on $D_n(\hat{\beta})$ are consistent against any model misspecification under which $D(\beta^*)$ is nonzero.

When model (1.1) is incorrect, for any fixed β, $e(t) \neq e(\beta, t)$ and $v(t) \neq v(\beta, t)$ in some time interval of t. The p-vector β^* determined by (B.1) entails $\{e(t) - e(\beta^*, t)\}$ to vary over t in such a way that the integration of $\{e(t) - e(\beta^*, t)\}s^{(0)}(t)$ is zero. This value of β^* generally does not satisfy $D(\beta^*) = 0$.

We now examine $D(\beta^*)$ in detail for the two-sample problem. Let Z be the treatment indicator with $\Pr(Z = 1) = \rho$. Also, let $h_1(t)$ and $h_0(t)$ denote the hazard functions of the treated and the control groups, respectively. By the basic properties of conditional expectations, we have

$$s^{(1)}(\beta, t) = \rho e^{\beta} \Pr\{Y(t) = 1|Z = 1\}$$

$$s^{(0)}(\beta, t) = \rho e^{\beta} \Pr\{Y(t) = 1|Z = 1\} + (1 - \rho) \Pr\{Y(t) = 1|Z = 0\}$$

$$s^{(1)}(t) = \rho h_1(t) \Pr\{Y(t) = 1|Z = 1\}$$

$$s^{(0)}(t) = \rho h_1(t) \Pr\{Y(t) = 1|Z = 1\} + (1 - \rho) h_0(t) \Pr\{Y(t) = 1|Z = 0\}.$$

Thus,

$$e(\beta, t) = \left\{1 + \frac{(1 - \rho) \Pr\{Y(t) = 1|Z = 0\}}{\rho \Pr\{Y(t) = 1|Z = 1\} e^{\beta}}\right\}^{-1}, \tag{B.2}$$

$$e(t) = \left\{1 + \frac{(1 - \rho) \Pr\{Y(t) = 1|Z = 0\} h_0(t)}{\rho \Pr\{Y(t) = 1|Z = 1\} h_1(t)}\right\}^{-1}. \tag{B.3}$$

By noting that $s^{(2)}(\beta, t) = s^{(1)}(\beta, t)$ and $s^{(2)}(t) = s^{(1)}(t)$ and that β^* satisfies (B.1), we can show that

$$D(\beta^*) = 2 \int_0^{\infty} e(\beta^*, t)\{e(t) - e(\beta^*, t)\} s^{(0)}(t) dt.$$

Next suppose that the hazard ratio $h_1(t)/h_0(t)$ is monotone. Then according to (B.1)–(B.3) there exists a time point $0 < t^* < \infty$ such that $\{e(t) - e(\beta^*, t)\}$ takes opposite signs between $(0, t^*)$ and (t^*, ∞). Hence, $D(\beta^*)$ will be nonzero if, for example, $e(\beta^*, t)$ is monotone in t, which is true if the curves of $h_1(t)$ and $h_0(t)$ do not intersect and the censoring distributions are identical between the two groups.

Table 1. Times to remission (weeks) for leukemia patients

Sample 0	6*, 6, 6, 6, 7, 9*, 10*, 10, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
Sample 1	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

NOTE: * indicates censored.

Table 2. Reanalysis of Stanford heart transplant data as of February 1980 by Cox regression models

	Model		
	1	2	3
age			
$\hat{\beta}_1$	0.02955	0.04471	0.04478
$SE_1(\hat{\beta}_1)$	0.01135	0.01095	0.01089
$SE_2(\hat{\beta}_1)$	0.00949	0.01071	0.01049
$T5$			
$\hat{\beta}_2$	0.16956	0.17506	—
$SE_1(\hat{\beta}_2)$	0.18312	0.18306	—
$SE_2(\hat{\beta}_2)$	0.16730	0.16908	—
age^2			
$\hat{\beta}_3$	—	0.00223	0.00221
$SE_1(\hat{\beta}_3)$	—	0.00070	0.00069
$SE_2(\hat{\beta}_3)$	—	0.00072	0.00071
Maximum test			
T_n	2.466	1.578	1.001
p-value	0.041	0.478	0.631
Wald test			
W_n	9.356	6.999	1.587
d.f.	3	6	3
p-value	0.025	0.321	0.662

Table 3. Empirical sizes of five goodness-of-fit tests: (i) new test, (ii) Cox's test, (iii) Schoenfeld's test, (iv) Wei's test and (v) Gill-Schumacher's test for one-parameter Cox models

Sample Size	(A) $h(t\|Z) = \exp(0.2Z)$[a]				(B) Weibull($\rho = 1, \tau = 2$) vs. Weibull($\rho = 2, \tau = 2$)[b]			
	No Censoring		25% Censoring[c]		No Censoring		22% Censoring[d]	
	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$
50	0.015	0.053	0.023	0.053	0.021	0.070	0.022	0.075
	0.018	0.059	0.016	0.064	0.014	0.055	0.013	0.052
	0.013	0.056	0.011	0.048	0.013	0.047	0.011	0.062
	0.002	0.040	0.005	0.041	0.003	0.030	0.006	0.029
	—	—	—	—	0.019	0.059	0.027	0.070
100	0.012	0.047	0.012	0.058	0.009	0.049	0.019	0.068
	0.009	0.054	0.015	0.061	0.013	0.054	0.009	0.052
	0.009	0.059	0.017	0.061	0.009	0.045	0.007	0.043
	0.007	0.044	0.009	0.036	0.004	0.027	0.008	0.033
	—	—	—	—	0.010	0.037	0.013	0.050
200	0.010	0.048	0.009	0.044	0.012	0.048	0.011	0.049
	0.012	0.045	0.014	0.059	0.009	0.040	0.014	0.052
	0.018	0.066	0.009	0.045	0.008	0.041	0.009	0.050
	0.009	0.047	0.008	0.043	0.004	0.033	0.005	0.040
	—	—	—	—	0.007	0.041	0.015	0.052

a. Z is a standard normal variable truncated at ± 5.
b. Subjects are allocated to two groups with equal probability.
c. $U(0, 4)$ censoring distribution.
d. $U(0, 3)$ censoring distribution.

Table 4. Empirical powers of five goodness-of-fit tests: (i) new test, (ii) Cox's test, (iii) Schoenfeld's test, (iv) Wei's test and (v) Gill-Schumacher's test for detecting nonconstant hazard ratio in the two-sample case

Sample Size	(A) Gamma($k = 2$) vs. Gamma($k = 3$)				(B) Weibull($\rho = 1, \tau = 0.5$) vs. Weibull($\rho = 1, \tau = 1.5$)			
	No Censoring		25% Censoring[a]		No Censoring		26% Censoring[b]	
	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$
50	0.169	0.370	0.124	0.313	0.009	0.120	0.044	0.129
	0.062	0.162	0.038	0.126	0.866	0.950	0.709	0.892
	0.044	0.135	0.037	0.127	0.673	0.853	0.520	0.772
	0.027	0.089	0.022	0.080	0.760	0.901	0.555	0.815
	0.077	0.170	0.061	0.168	0.933	0.982	0.852	0.942
100	0.297	0.521	0.214	0.423	0.238	0.719	0.064	0.296
	0.129	0.299	0.112	0.258	1.000	1.000	0.977	0.994
	0.083	0.211	0.059	0.188	0.968	0.991	0.908	0.966
	0.085	0.226	0.058	0.167	0.992	1.000	0.953	0.988
	0.161	0.326	0.097	0.241	1.000	1.000	0.988	0.996
200	0.527	0.726	0.391	0.570	0.948	0.991	0.428	0.729
	0.296	0.533	0.232	0.492	1.000	1.000	1.000	1.000
	0.220	0.452	0.163	0.380	1.000	1.000	1.000	1.000
	0.248	0.503	0.157	0.384	1.000	1.000	1.000	1.000
	0.397	0.649	0.269	0.497	1.000	1.000	1.000	1.000

NOTE: Subjects are allocated to two groups with equal probability. Cox models with indicator covariate are fitted.
a. $U(0, 10)$ censoring distribution.
b. $U(0, 4)$ censoring distribution.

Table 5. Empirical powers of five goodness-of-fit tests: (i) new test, (ii) Cox's test, (iii) Schoenfeld's test, (iv) Wei's test and (v) Gill-Schumacher's test for detecting proportional hazards models with omitted covariates

Sample	(A) $h(t\|Z) = \exp(0.2Z + 0.3Z^2)$[a]				(B) $h(t\|Z, Z^*) = \exp(0.5Z + 2Z^*)$[b]			
Size	No Censoring		26% Censoring[c]		No Censoring		28% Censoring[d]	
	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$
50	0.025	0.297	0.024	0.224	0.074	0.220	0.060	0.192
	0.023	0.094	0.024	0.090	0.021	0.098	0.010	0.068
	0.013	0.057	0.011	0.057	0.009	0.058	0.007	0.049
	0.045	0.135	0.034	0.103	0.007	0.027	0.006	0.030
	—	—	—	—	0.023	0.060	0.016	0.061
100	0.403	0.790	0.313	0.750	0.078	0.244	0.095	0.221
	0.019	0.086	0.029	0.105	0.027	0.106	0.018	0.074
	0.011	0.060	0.016	0.054	0.022	0.066	0.013	0.061
	0.057	0.172	0.048	0.158	0.017	0.061	0.011	0.052
	—	—	—	—	0.033	0.090	0.032	0.098
200	0.946	0.996	0.878	0.983	0.133	0.307	0.099	0.254
	0.034	0.111	0.052	0.135	0.028	0.100	0.020	0.093
	0.010	0.045	0.012	0.060	0.040	0.121	0.024	0.089
	0.075	0.204	0.085	0.222	0.034	0.115	0.018	0.076
	—	—	—	—	0.056	0.146	0.035	0.104

NOTE: One-parameter Cox models with covariate Z only are fitted.

a. Z is a standard normal variable truncated at ± 5.

b. $\Pr(Z = -1) = \Pr(Z = 1) = 1/2$. Z^* is an independent standard normal variable truncated at ± 5.

c. $U(0, 3)$ censoring distribution.

d. $U(0, 7)$ censoring distribution.

Table 6. Empirical powers of four goodness-of-fit tests: (i) new test, (ii) Cox's test, (iii) Schoenfeld's test and (iv) Wei's test for detecting proportional hazards models with non-exponential regression forms

Sample	(A) $h(t\|Z) = 1 + 0.5Z$				(B) $h(t\|Z) = \log(2 + 0.5Z)$			
Size	No Censoring		24% Censoring[a]		No Censoring		24% Censoring[b]	
	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$
50	0.114	0.224	0.088	0.192	0.118	0.246	0.082	0.207
	0.018	0.121	0.012	0.080	0.020	0.109	0.011	0.058
	0.020	0.055	0.008	0.067	0.018	0.063	0.009	0.062
	0.010	0.035	0.003	0.019	0.007	0.039	0.001	0.026
100	0.152	0.321	0.149	0.290	0.151	0.325	0.139	0.300
	0.073	0.208	0.020	0.085	0.060	0.160	0.015	0.068
	0.013	0.070	0.013	0.062	0.012	0.059	0.016	0.049
	0.018	0.066	0.006	0.034	0.003	0.053	0.005	0.035
200	0.271	0.480	0.258	0.447	0.303	0.529	0.242	0.440
	0.179	0.400	0.022	0.125	0.130	0.280	0.024	0.086
	0.008	0.058	0.013	0.056	0.011	0.074	0.005	0.040
	0.003	0.114	0.009	0.049	0.027	0.102	0.011	0.043

NOTE: Z is a standard normal variable truncated at ± 1.96. One-parameter Cox models (with exponential regression form) are fitted.

a. $U(0, 5)$ censoring distribution.

b. $U(0, 7)$ censoring distribution.

Table 7. Empirical powers of four goodness-of-fit tests: (i) new test, (ii) Cox's test, (iii) Schoenfeld's test and (iv) Wei's test for detecting nonproportional hazards models

Sample Size		(A) $\log T = -0.5Z + 0.5\epsilon^{a}$				(B) $T = \exp(-0.5Z) + \psi^{b}$			
		No Censoring		25% Censoring[c]		No Censoring		25% Censoring[d]	
		$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$
50	(i)	0.075	0.180	0.086	0.177	0.461	0.625	0.384	0.558
	(ii)	0.120	0.281	0.119	0.264	0.286	0.465	0.199	0.374
	(iii)	0.041	0.130	0.031	0.125	0.201	0.395	0.167	0.372
	(iv)	0.048	0.164	0.032	0.118	0.112	0.266	0.085	0.192
100	(i)	0.136	0.272	0.116	0.249	0.745	0.855	0.654	0.788
	(ii)	0.301	0.499	0.223	0.433	0.606	0.780	0.470	0.661
	(iii)	0.107	0.258	0.096	0.221	0.514	0.710	0.471	0.701
	(iv)	0.185	0.371	0.131	0.288	0.358	0.597	0.266	0.464
200	(i)	0.254	0.404	0.203	0.348	0.934	0.972	0.919	0.957
	(ii)	0.594	0.790	0.501	0.729	0.919	0.968	0.816	0.922
	(iii)	0.269	0.491	0.224	0.425	0.900	0.966	0.865	0.955
	(iv)	0.521	0.748	0.363	0.598	0.822	0.934	0.624	0.817

NOTE: Z is a standard normal variable truncated at ± 5. One-parameter Cox models (with exponential regression form) are fitted.

a. ϵ is an independent standard normal variable.

b. ψ is an independent log-normal variable with $\mu = 0$ and $\sigma = 0.5$.

c. $U(0, 5)$ censoring distribution.

d. $U(0, 9)$ censoring distribution.

References

Andersen, P. K. (1982). Testing goodness of fit of Cox's regression and life model. *Biometrics* **38**, 67–77.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100–1120.

Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *J. Amer. Statist. Assoc.* **83**, 204–212.

Bailey, K. R. (1983). The asymptotic joint distribution of regression and survival parameter estimates in the Cox regression model. *Ann. Statist.* **11**, 39–48.

Breslow, N. E., Edler, L. and Berger, J. (1984). A two-sample censored-data rank test for acceleration. *Biometrics* **40**, 1049–1062.

Ciampi, A. and Etezadi-Amoli, J. (1985). A general model for testing the proportional hazards and the accelerated failure time hypotheses in the analysis of censored survival data with covariates. *Comm. Statist. Ser. A* **14**, 651–667.

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. (1979). A note on the graphical analysis of survival data. *Biometrika* **66**, 188–190.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *J. Amer. Statist. Assoc.* **72**, 27–36.

Crowley, J. and Storer, B. E. (1983). Comment (on "A reanalysis of the Stanford heart transplant data" by M. Aitkin, N. Laird and B. Francis). *J. Amer. Statist. Assoc.* **78**, 277–281.

Gill, R. D. (1980). *Censoring and Stochastic Integrals* (Mathematical Centre Tracts, No. 124). The Mathematisch Centrum, Amsterdam.

Gill, R. D. and Schumacher, M. (1987). A simple test of the proportional hazards assumption. *Biometrika* **74**, 289–300.

Haara, P. (1987). A note on the asymptotic behaviour of the empirical score in Cox's regression model for counting processes. *Proceedings of the 1st World Congress of the Bernoulli Society*, 139–142. VNU Science Press, Tashkent.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* John Wiley, New York.

Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Appl. Statist.* **26**, 227–237.

Kennedy, W. J. and Gentle, J. E. (1980). *Statistical Computing.* Marcel Dekker, New York.

Lagakos, S. W. (1980). The graphical evaluation of explanatory variables in proportional hazard regression models. *Biometrika* **68**, 93–98.

Lagakos, S. W. (1988). The loss in efficiency from misspecifying covariates in proportional hazards regression models. *Biometrika* **75**, 156-160.

Lagakos, S. W. and Schoenfeld, D. A. (1984). Properties of proportional-hazards score tests under misspecified regression models. *Biometrics* **40**, 1037–1048.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Amer. Statist. Assoc.* **84**, 1074–1078.

Liu, P. Y. and Crowley, J. (1978). Large sample theory of the MLE based on Cox's regression model for survival data. Technical Report No. 1, Biostatistics Center, University of Wisconsin–Madison.

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika* **69**, 521–531.

Moreau, T., O'Quigley, J. and Lellouch, J. (1986). On D. Schoenfeld's approach for testing the proportional hazards assumption. *Biometrika* **73**, 513–515.

Moreau, T., O'Quigley, J. and Mesbah, M. (1985). A global goodness-of-fit statistic for the proportional hazards model. *Appl. Statist.* **34**, 212–218.

Naes, T. (1982). The asymptotic distribution of the estimator for the regression parameter in Cox's regression model. *Scand. J. Statist.* **9**, 107–115.

Nagelkerke, N. J. D., Oosting, J. and Hart, A. A. M. (1984). A simple test for goodness of fit of Cox's proportional hazards model. *Biometrics* **40**, 483–486.

O'Quigley, J. and Pessione, F. (1989). Score tests for homogeneity of regression effect in the proportional hazards model. *Biometrics* **45**, 135–144.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1986). *Numerical Recipes: the Art of Scientific Computing.* Cambridge University Press, Cambridge.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd edition. John Wiley, New York.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* **67**, 145–153.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239–241.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73**, 363–369.

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. Amer. Statist. Assoc.* **79**, 649–652.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195, U.S.A.
Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706, U.S.A.

(Received August 1989; accepted June 1990)