Goodness-of-Fit Analysis for the Cox Regression Model Based on a Class of Parameter Estimators

D. Y. LIN*

In this article we propose a class of estimation functions for the vector of regression parameters in the Cox proportional hazards model with possibly time-dependent covariates by incorporating the weight functions commonly used in weighted log-rank tests into the partial likelihood score function. The resulting estimators behave much like the conventional maximum partial likelihood estimator in that they are consistent and asymptotically normal. When the Cox model is inappropriate, however, the estimators with different weight functions generally converge to nonidentical constant vectors. For example, the magnitude of the parameter estimator using the Kaplan–Meier survival estimator as the weight function will be stochastically larger than that of the maximum partial likelihood estimator if covariate effects diminish over time. Such facts motivate us to develop goodness-of-fit methods for the Cox regression model by comparing parameter estimators with different weight functions. Under the assumed model, the normalized difference between the maximum partial likelihood estimator and a weighted parameter estimator is shown to converge weakly to a multivariate normal with mean zero and with a covariance matrix for which a consistent estimator is proposed. The asymptotic properties of the weighted parameter estimators and those of the related goodness-of-fit tests under misspecified Cox models are also investigated. In particular, it is demonstrated that a goodness-of-fit test with a monotone weight function is consistent against monotone departures from the proportional hazards assumption. Versatile testing procedures with broad sensitivities can be developed based on simultaneous use of several weight functions. Three examples using real data are presented.

KEY WORDS: Censoring; Lack of fit; Martingale; Model misspecification; Proportional hazards; Survival data.

1. INTRODUCTION

The Cox (1972) proportional hazards model specifies that the hazard function for the failure time T of an individual with a $p \times 1$ vector of possibly time-varying covariates Z has the following form

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

where β_0 is a $p \times 1$ vector of unknown regression parameters and $\lambda_0(t)$ is an unspecified baseline hazard function. Let $X = \min(T, G)$, where G is the censoring time variable, and let $\Delta = 1$ if $X = T$ and $\Delta = 0$ otherwise. Assume that T and G are independent conditional on Z and that (X_i, Δ_i, Z_i) $(i = 1, \ldots, n)$ are independent and identically distributed replicates of (X, Δ, Z). Then the parameter vector β_0 is commonly estimated by $\hat{\beta}$ which maximizes the partial likelihood function (Cox 1972; Cox 1975)

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

where $Y_j(t) = 1$ if $X_j \geq t$ and $Y_j(t) = 0$ otherwise. The maximum partial likelihood estimator $\hat{\beta}$ possesses the usual asymptotic properties of ordinary maximum likelihood estimators (Andersen and Gill 1982).

The immense popularity of the Cox regression model has made the issue of model checking extremely important. Several procedures have been suggested to verify the proportional hazards assumption; see Lin (1989) for a review. Virtually all of the existing goodness-of-fit tests for the *general* Cox model, however, are based on some arbitrary partition of the time axis and/or are difficult to compute. Consequently, goodness-of-fit analysis has rarely been performed by the users of the Cox regression model despite increasing awareness of the adverse effects of model misspecification on the statistical inference.

Gill and Schumacher (1987) presented a simple test of the proportional hazards assumption for two-sample censored data. The rationale behind their procedure is that two rank estimators of the hazard ratio, one of which assigns relatively more weights to early failures than the other, should give different answers in nonproportional hazards situations, especially when the hazard ratio is monotone.

This article extends the idea of Gill and Schumacher (1987) to the general Cox regression model with possibly time-varying covariates. In the next section, we introduce a class of estimation functions for β_0 by incorporating the weight functions commonly used in weighted log-rank tests into the partial likelihood score function. The resulting weighted parameter estimators are consistent for β_0. In Section 3, we show that under model (1.1) the difference between the maximum partial likelihood estimator and a weighted parameter estimator is asymptotically normal with mean zero and with a covariance matrix that can be consistently estimated. These results are used to develop goodness-of-fit methods for the Cox model. The properties of the proposed procedures under misspecified models are carefully investigated. Three examples using real data are provided in Section 4 as illustrations.

2. WEIGHTED PARAMETER ESTIMATORS

Define $S^{(r)}(\beta, t) = n^{-1} \sum_{j=1}^{n} Y_j(t) \exp\{\beta' Z_j(t)\} Z_j(t)^{\otimes r}$ for $r = 0, 1, 2$, where for a column vector a, $a^{\otimes 2}$ refers to the matrix aa', $a^{\otimes 1}$ to the vector a and $a^{\otimes 0}$ to the scalar 1. The

* D. Y. Lin is Assistant Professor, Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195. This research was supported by the National Institutes of Health Grants 5 R01 AI 24643 and 5 R01 AI 29168. The author thanks David Schoenfeld and Robert Gray for their helpful comments and is grateful to the associate editor and referees for their reviews of the manuscript.

Journal of the American Statistical Association
September 1991, Vol. 86, No. 415, Theory and Methods

logarithm of the partial likelihood function (1.2) can be expressed as

$$l(\beta) = \sum_{i=1}^{n} \Delta_i[\beta' Z_i(X_i) - \log\{S^{(0)}(\beta, X_i)\}].$$

The corresponding score function is

$$U(\beta) = \sum_{i=1}^{n} \Delta_i\{Z_i(X_i) - E(\beta, X_i)\}, \qquad (2.1)$$

where $E(\beta, t) = S^{(1)}(\beta, t)/S^{(0)}(\beta, t)$. The maximum partial likelihood estimator $\hat{\beta}$ is the solution to the system of partial likelihood equations $U(\beta) = 0$. It has been shown that the random vector $n^{1/2}(\hat{\beta} - \beta_0)$ is asymptotically normal with mean zero and with a covariance matrix that can be consistently estimated by $C(\hat{\beta}) = \{n^{-1}\Sigma_{i=1}^{n} \Delta_i V(\hat{\beta}, X_i)\}^{-1}$, where $V(\beta, t) = S^{(2)}(\beta, t)/S^{(0)}(\beta, t) - E(\beta, t)^{\otimes 2}$ (Andersen and Gill 1982).

The score function (2.1) is the sum over the failure times of the difference between the observed value of $Z_i(X_i)$ and the conditional mean of $Z_j(X_i)$ on $\{j: X_j \geq X_i\}$ with respect to a probability distribution proportional to $\exp\{\beta' Z_j(X_i)\}$. In addition, all failures receive the same weight. Alternatively, we may assign unequal weights to different failures according to the times of their occurrences as in the case of weighted log-rank tests (Tarone and Ware 1977). Specifically, let us define a class of weighted score functions in the form

$$U_w(\beta) = \sum_{i=1}^{n} \Delta_i W(X_i)\{Z_i(X_i) - E(\beta, X_i)\} \qquad (2.2)$$

and define the weighted parameter estimator $\hat{\beta}_w$ as the solution to the system of estimation equations $U_w(\beta) = 0$. The random weight function $W(t)$ in (2.2) is a predictable stochastic process that converges in probability to a nonnegative bounded function $w(t)$ uniformly in $t \in [0, \infty)$. The weight functions commonly chosen for weighted log-rank tests can be used here. Note that $U_w(\beta)$ reduces to the partial likelihood score function $U(\beta)$ if $W(t) = 1$ and that $U_w(\beta)$ is related to the Peto–Prentice generalization of the Wilcoxon statistic (Peto and Peto 1972; Prentice 1978) if $W(t) = \hat{F}(t)$, where $\hat{F}(\cdot)$ is the left-continuous version of the Kaplan–Meier estimator for the survival function of T computed from the pooled sample. The asymptotic results for the weighted parameter estimator $\hat{\beta}_w$ are established in the following theorem.

Theorem 2.1. If model (1.1) holds, the random vector $n^{1/2}(\hat{\beta}_w - \beta_0)$ is asymptotically normal with zero mean and with a covariance matrix that can be consistently estimated by $C_w(\hat{\beta}) = A_w(\hat{\beta})^{-1}B_w(\hat{\beta})A_w(\hat{\beta})^{-1}$, where $A_w(\beta) = n^{-1} \times \Sigma_{i=1}^{n} \Delta_i W(X_i)V(\beta, X_i)$ and $B_w(\beta) = n^{-1}\Sigma_{i=1}^{n} \Delta_i W(X_i)^2 V(\beta, X_i)$.

The proof of Theorem 2.1 is similar to the development of the asymptotic theory for the maximum partial likelihood estimator $\hat{\beta}$ given in Andersen and Gill (1982) and is therefore omitted.

Andersen (1983) proposed a class of explicitly computable two-sample rank estimators of the hazard ratio. The estimators in that class generally differ from our estimators $\exp(\hat{\beta}_w)$'s for the same weight functions. The Cox estimator $\exp(\hat{\beta})$ always has a smaller asymptotic variance than any member of Andersen's family (Andersen 1983).

3. GOODNESS-OF-FIT ANALYSIS

When the assumed model (1.1) is valid, a weighted estimator $\hat{\beta}_w$ with a nonconstant weight function and the unweighted estimator $\hat{\beta}$ should be close to each other for a given data set since both estimators are consistent. On the other hand, the two estimators tend to differ under misspecified models. For instance, if the effects of covariates diminish over time, then $|\hat{\beta}_w|$ with the Peto–Prentice weight function $\hat{F}(\cdot)$ is stochastically larger than $|\hat{\beta}|$. These facts motivate us to develop goodness-of-fit methods based on the difference between $\hat{\beta}_w$ and $\hat{\beta}$.

Theorem 3.1. If model (1.1) holds, the random vector $n^{1/2}(\hat{\beta}_w - \hat{\beta})$ is asymptotically normal with zero mean and with a covariance matrix that can be consistently estimated by $D_w(\hat{\beta}) = C_w(\hat{\beta}) - C(\hat{\beta})$.

A crucial step in proving Theorem 3.1 is to show that the random vectors $n^{-1/2}U_w(\beta_0)$ and $n^{-1/2}U(\beta_0)$ are asymptotically joint normal with a covariance matrix that can be consistently estimated by $A_w(\hat{\beta})$. This can be accomplished by the application of standard counting process techniques (Andersen and Gill 1982) to the vector of local square-integrable martingales $n^{-1/2}\{U_w(\beta_0, t)', U(\beta_0, t)'\}$, where $U(\beta, t)$ and $U_w(\beta, t)$ are, respectively, the score function and weighted score function evaluated at time t.

Goodness-of-fit procedures for the Cox regression model can be derived from Theorem 3.1. In particular, the quadratic form $Q_w = n(\hat{\beta}_w - \hat{\beta})'D_w(\hat{\beta})^{-1}(\hat{\beta}_w - \hat{\beta})$ has an asymptotic central χ^2 distribution on p degrees of freedom when model (1.1) holds. The sources of model misspecification could be identified by component-wise comparisons of $\hat{\beta}$ and $\hat{\beta}_w$.

In the two-sample case, the goodness-of-fit test Q_w does not reduce to the test of Gill and Schumacher (1987), the reason being that the latter employs the rank estimators of Andersen (1983). Unlike the variance estimator used by Gill and Schumacher (1987), the variance estimator for the proposed test is always nonnegative.

In order to study the asymptotic behavior of the weighted parameter estimator $\hat{\beta}_w$ and the goodness-of-fit test statistic Q_w under model misspecification, we define $s^{(r)}(t) = \mathscr{E}\{Y(t)\lambda(t; Z)Z(t)^{\otimes r}\}$ and $s^{(r)}(\beta, t) = \mathscr{E}\{S^{(r)}(\beta, t)\}$ for $r = 0, 1, 2$, where $\mathscr{E}$ denotes the expectation with respect to the true model of (X, Δ, Z). In addition, let

$$h_w(\beta) = \int_0^\infty w(t)\{e(t) - e(\beta, t)\}s^{(0)}(t)\,dt$$

and

$$A_w^*(\beta) = \int_0^\infty w(t)\{s^{(2)}(\beta, t)/s^{(0)}(\beta, t) - e(\beta, t)^{\otimes 2}\}s^{(0)}(t)\,dt,$$

where $e(t) = s^{(1)}(t)/s^{(0)}(t)$ and $e(\beta, t) = s^{(1)}(\beta, t)/s^{(0)}(\beta, t)$.

Theorem 3.2. Under a possibly misspecified Cox model, the weighted parameter estimator $\hat{\beta}_w$ converges in probability to a $p \times 1$ vector of constants β_w^*, where β_w^* is the unique solution to the system of equations $h_w(\beta) = 0$ if the matrix $A_w^*(\beta_w^*)$ is positive definite.

The proof of Theorem 3.2 is similar to the proofs of lemma 3.1 in Andersen and Gill (1982) and of theorem 2.1 in Struthers and Kalbfleisch (1986) and is not shown here. By the arguments used in the proofs of theorems 3.2 and 4.2 of Andersen and Gill (1982), we can show that $D_w(\hat{\beta})$ converges in probability to a finite quantity. Hence, the following result is a direct outcome of Theorem 3.2.

Corollary 3.3. The goodness-of-fit test Q_w is consistent against any model misspecification under which $\beta_w^* \neq \beta^*$ or $h_w(\beta^*) \neq 0$, where β^* is the probability limit of $\hat{\beta}$.

Suppose now that we intend to test H_0: $\lambda(t; Z) = \lambda_0(t) \exp\{\beta_0 Z(t)\}$ against H_1: $\lambda(t; Z) = \lambda_0(t) \exp\{\theta(t)Z(t)\}$, where Z is a scalar and $\theta(t)$ is an unspecified monotone function of t. Note that H_1 corresponds to monotone departures from the proportional hazards assumption. It is easy to see that, under H_1, the function $\{e(t) - e(\beta^*, t)\}$ crosses the time axis only at the point where $\theta(t) = \beta^*$ because $e(\beta, t)$ is a monotonically increasing function of β for any fixed t. Thus $h_w(\beta^*) \neq 0$ if $w(t)$ is monotone in t, which indicates that Q_w with a monotone weight function is consistent against H_1.

Since monotone departures from the proportionality are the most common and most important forms of model misspecification in real applications, we suggest that Q_w with a monotone weight function such as $\hat{F}(\cdot)$ be used. Other weight functions may be supplemented to increase power against nonmonotone alternatives. For example, the test with $W(\cdot) = \hat{F}(\cdot)\{1 - \hat{F}(\cdot)\}$ is sensitive to quadratic trends. To preserve a given overall type I error probability when several weight functions are used simultaneously, a global test based on the joint distribution of the differences between the maximum partial likelihood estimator and the weighted estimators of interest can be used.

4. EXAMPLES

A double-blind, placebo-controlled trial on the efficacy of AZT (azidothymidine) for treating AIDS patients was conducted in 1986 (Fischl, Richman, Grieco et al. 1987). Two hundred eighty-one patients were enrolled in the experiment, among whom 144 were assigned to AZT and 137 to placebo. By the end of the study, 25 patients in the AZT group and 51 patients in the placebo group had developed at least one opportunistic infection. As an illustration, the Cox model is used to regress the time from entering the trial to the first infection on the indicator function of AZT and the logarithm of the *current* CD4 count, the latter covariate being time-dependent. Such a model is useful in evaluating the role of CD4 change as a surrogate endpoint in AIDS clinical trials. The maximum partial likelihood estimate $\hat{\beta}$ is $(-.6029, -.7592)'$, and the weighted estimate $\hat{\beta}_w$ with the Peto–Prentice weight function is $(-.5495, -.7507)'$. The covariance matrix estimate $D_w(\hat{\beta})$ is

$$\begin{bmatrix} .2446 & -.0318 \\ -.0318 & .0645 \end{bmatrix}.$$

Thus, the observed value of Q_w is 4.396 on 2 df, yielding a p value of .111. The low p value is attributed to the considerable difference between the weighted and unweighted estimates of the AZT effect. Incidentally, partial likelihood score tests indicate that both covariates are highly significant.

Two other examples will be mentioned briefly. The first one of these uses data on the time to remission for two groups of leukemia patients presented by Cox (1972). The maximum partial likelihood estimate and the weighted estimate with the Peto–Prentice weight function are 1.5092 and 1.5382, respectively. The observed value of Q_w is .034 with a p value of .85. The Gill–Schumacher test using the same weight function gives a p value of .72 (Gill and Schumacher 1987). The p value for the Q_w test with $\hat{F}(\cdot)\{1 - \hat{F}(\cdot)\}$ as the weight function is .32. Nearly all of the other existing goodness-of-fit tests have also concluded that the Cox model fits this data set well.

The last example uses the Veterans Administration lung cancer data (Kalbfleisch and Prentice 1980, pp. 223–224). Kalbfleisch and Prentice (1980, pp. 89–90) fit a proportional hazards model with seven covariates to these data. For that model, the maximum partial likelihood estimate and weighted estimates give strikingly different results. The p value for the Q_w test with the Peto–Prentice weight function is .00002, which strongly discredits the assumed model.

5. REMARKS

We have described a simple approach to analyzing the fit of the Cox regression model. The results of the analysis can be interpreted in a very natural way. The new procedures can be easily incorporated into an existing computer program for fitting the Cox model, and software is available from the author.

The sample-size requirement for the proposed methods is similar to that of the standard Cox regression analysis. The structure of the test statistic indicates that Q_w with an appropriate weight function is rather sensitive to model misspecification, which has been confirmed by our empirical findings. For this reason, we recommend that users should assess the practical importance of model misspecification by examining the magnitude of the difference between the unweighted and weighted estimators rather than relying on the p value of the Q_w test alone.

The proposed goodness-of-fit methods formalize a graphical technique suggested by Schoenfeld (1982). Schoenfeld (1982) defined partial residuals for the proportional hazards model as the difference between the observed covariate value $Z_i(X_i)$ and its estimated conditional expectation $E(\hat{\beta}, X_i)$. These residuals can be plotted against the time axis to visually examine the fit of the model and detect outlying covariate values.

[*Received May 1990. Revised February 1991.*]

REFERENCES

Andersen, P. K. (1983), "Comparing Survival Distributions Via Hazard Ratio Estimates," *Scandinavian Journal of Statistics*, 10, 77–85.

Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics,* 10, 1100–1120.

Cox, D. R. (1972), "Regression Models and Life-Tables" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 34, 187–220.

——— (1975), "Partial Likelihood," *Biometrika,* 62, 269–276.

Fischl, M. A., Richman, D. D., Grieco, M. H., et al. (1987), "The Efficacy of Azidothymidine (AZT) in the Treatment of Patients With AIDS and AIDS-Related Complex," *The New England Journal of Medicine,* 317, 185–191.

Gill, R. D., and Schumacher, M. (1987), "A Simple Test of the Proportional Hazards Assumption," *Biometrika,* 74, 289–300.

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

Lin, D. Y. (1989), *Goodness-of-Fit Tests and Robust Inference for the Cox Proportional Hazards Model,* unpublished Ph.D. dissertation, University of Michigan, Dept. of Biostatistics.

Peto, R., and Peto, J. (1972), "Asymptotically Efficient Rank Invariant Test Procedures" (with discussion), *Journal of the Royal Statistical Society,* Ser. A, 135, 185–206.

Prentice, R. L. (1978), "Linear Rank Tests With Right Censored Data," *Biometrika,* 65, 167–179.

Schoenfeld, D. (1982), "Partial Residuals for the Proportional Hazard Regression Model," *Biometrika,* 69, 239–241.

Struthers, C. A., and Kalbfleisch, J. D. (1986), "Misspecified Proportional Hazard Models," *Biometrika,* 73, 363–369.

Tarone, R. E., and Ware, J. (1977), "On Distribution-Free Tests for Equality of Survival Distributions," *Biometrika,* 64, 156–160.