Understanding Cox's Regression Model: A Martingale Approach

Richard D. Gill

Journal of the American Statistical Association, Volume 79, Issue 386 (Jun., 1984), 441-447.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28198406%2979%3A386%3C441%3AUCRMAM%3E2.0.CO%3B2-9

Journal of the American Statistical Association

Understanding Cox's Regression Model: A Martingale Approach

RICHARD D. GILL*

An informal discussion is given of how martingale techniques can be used to extend Cox's regression model and to derive its large sample properties.

KEY WORDS: Censoring; Survival data; Partial likelihood; Counting processes; Asymptotic theory.

SUMMARY

Cox's (1972) regression model for analyzing censored survival data, allowing for covariates, has enjoyed an enormous success among applied statisticians. It elegantly combines the advantages of both parametric and nonparametric approaches to statistical inference, and it is beautifully adapted to the kind of data one will obtain in clinical cancer trials and other sources of survival data and life-testing data. By incorporating time-varying, random covariates, it becomes a highly flexible tool for model building.

Despite this its mathematical basis so far is somewhat heuristic. Just to provide motivation for the estimators used, Cox (1975) had to introduce a new principle for inference, based on the concept of partial likelihood. Many papers contain asymptotic results on the estimators (Liu and Crowley 1978; Tsiatis 1978a,b,1981a,b; Link 1979; Bailey 1983; Naes 1981,1982; and Sen 1981) confirming Cox's conjectures, but all are restricted to very special cases. Moreover, in all cases derivations are highly complex and technical. For instance, simple formulas for limiting variances appear as if by surprise after lengthy computations, in the course of which complicated terms cancel one another out.

The purpose of this article is to discuss recent work by Andersen and Gill (1982) that shows how a firm mathematical basis can be given to the model (in its fullest generality) from which asymptotic properties can be derived in a completely natural way. The mathematics is based on the statistical theory of counting processes developed by Aalen (1976,1978). In brief, the idea is as follows. The original hazard rate definition of Cox's model can be directly interpreted as specifying the stochastic intensity of a multivariate counting process (counting occurrences of the event "death" for each of the individuals under observation). This connects immediately with modern martingale and stochastic integral theory, very powerful and deep mathematical tools that are, on the other hand, often no more than a mathematical formulation of many of the intuitive ideas one has, for instance, concerning what kinds of censoring may be allowed and what kinds of covariates. Naes (1982) and Sen (1981) use discrete time martingale theory in an iid setup. However, we feel that continuous time methods are much more appropriate.

After sketching this theory on an intuitive level, we indicate how asymptotic properties of the estimator simply follow from this formulation of the model (Andersen and Gill 1982).

1. INTRODUCTION

It is possible to read this article at several different levels. At the most obvious level, the article summarizes some problems concerning Cox's regression model and indicates solutions to these problems that are further developed in Andersen and Gill (1982).

At the same time, the present article gives just a hint of how Cox's regression model can be extended in many useful ways. Also, taking Cox's model as an example, the article contains an introduction on a very intuitive level to the statistical theory of counting processes that is currently being used, following the work of Aalen (1976), to unify and extend many branches of nonparametric survival analysis. Finally, we hope the article will encourage those analyzing censored survival data to make use of the model. Even if a clinical cancer trial is designed to answer a simple yes/no question on the relative benefits of two treatments, there is no reason why after the trial, the data should not also be analyzed in a more exploratory fashion to look for variables or combinations of variables of prognostic importance and to quantify their simultaneous effects, or to look more closely at how a particular treatment influences survival (perhaps it only improves the hazard rate during the course of treatment and has no lasting effect).

Though the mathematics may at first sight seem formidable, we want to emphasize the fact that the methods are a natural formalization of the heuristic derivations of, for instance, Mantel (1966, p. 169) or Cox (1975, p. 274). This is in contrast to the classical approach to survival analysis, which has been to solve the problem using tools derived from classical nonparametric theory. To oversimplify, this has forced us to direct attention to situations

* Richard D. Gill is Chief of the Department of Mathematical Statistics, Centre for Mathematics and Computer Science, Postbus 4079, 1009 AB Amsterdam, Netherlands.

June 1984, Volume 79, Number 386
Theory and Methods Section

with iid observations and to special censoring models (random censorship, for instance) and away from methods based on hazard rates and the development of a process as time moves forward.

A second point we want to emphasize is that although the mathematical presentation in this article is entirely informal, everything we say can be made rigorous.

We next describe briefly the structure of the article. In Section 2 we give a specification of Cox's regression model in quite restrictive terms, just as it was first introduced. We also summarize the statistical procedures related to the model and give an indication of the controversy that has surrounded them. In Section 3 we give an equivalent reformulation of the model in terms of the intensities of counting processes, and in Section 4 we describe the martingale theory, which will solve many of our problems. In Section 5 we show how this theory can be used to derive asymptotic properties of the statistical procedures appropriate to the model. In Section 6 we discuss some open questions.

2. FIRST SPECIFICATION OF THE MODEL

We specify the model as follows. Let T_i, $i = 1, \ldots, n$, be independent continuously distributed positive random variables representing the times of death of n individuals, each of whom can only be observed on a fixed time interval $[0, c_i]$ for certain censoring times c_i, $i = 1, \ldots, n$. Suppose that individual i has hazard rate

$$\lambda_i(t) = \lim_{h \downarrow 0} \frac{1}{h} P[T_i \leq t + h \mid T_i \geq t] \tag{2.1}$$

of the special form

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_0' z_i(t)), \tag{2.2}$$

where β_0' is the transpose of a (column) vector β_0 of p unknown coefficients, z_i is a column vector of p possibly time-varying covariates, and λ_0 is a fixed unknown baseline hazard rate for an individual with $z \equiv 0$. The observations for the ith individual consist of $T_i \wedge c_i$, $\delta_i = I\{T_i \leq c_i\}$, and $z_i(t)$, $t \in [0, T_i \wedge c_i]$. Here $\wedge$ denotes minimum and $I\{\cdot\}$ is the indicator random variable for the specified event. We are interested in estimation of, or hypothesis testing on, the parameter β_0, while λ_0 assumes the status of an infinite dimensional nuisance parameter. The model can thus be termed semiparametric.

For the interpretation of the model and for examples of how covariates z_i can be chosen, we refer to Cox (1972), Miller et al. (1980), Andersen (1982), and Kalbfleisch and Prentice (1980).

Let

$$\mathcal{R}(t) = \{i: T_i \geq t \quad \text{and} \quad c_i \geq t\}$$

denote the risk set at time t, that is to say, the set of individuals i under observation at time t. Given $\mathcal{R}(t)$ and that at time t one individual in $\mathcal{R}(t)$ is observed to die, the probability that it is precisely individual i can be calculated as

$$\exp(\beta_0' z_i(t)) \Big/ \sum_{j \in \mathcal{R}(t)} \exp(\beta_0' z_j(t));$$

a factor $\lambda_0(t)$ has canceled out in numerator and denominator. Because λ_0 is completely arbitrary, it seems reasonable that what is observed in the intervals of time between observed deaths does not contain any information on β_0. Cox therefore proposed that statistical inference on β_0 could be carried out by considering

$$\mathcal{L}(\beta) = \prod_{i: T_i \leq c_i} \left(\frac{\exp(\beta' z_i(T_i))}{\sum_{j \in \mathcal{R}(T_i)} \exp(\beta' z_j(T_j))} \right) \tag{2.3}$$

as a likelihood function for β, to which standard large-sample maximum likelihood theory could be applied. Each term in this product is the probability that at the time T_i of an observed death, it is precisely individual i who is observed to die.

Whether $\mathcal{L}(\beta)$ is some sort of likelihood function has given rise to much discussion in the literature. It certainly is not a conditional likelihood, that is, a likelihood function for β based on the conditional distribution of the data given some statistic. Nor is it generally a marginal likelihood, that is, a likelihood based on the marginal distribution of some reduction of the data. Cox (1975) introduced the notion of partial likelihood to remedy this defect and showed that $\mathcal{L}(\beta)$ is one (to date, the most important) example of partial likelihood.

Whatever sort of likelihood $\mathcal{L}(\beta)$ may be, it is still not clear that standard large-sample maximum likelihood theory will lead to valid asymptotic (i.e., in practice, approximate) results for inference on β_0. Much effort has been spent in deriving rigorously the required asymptotics: All of the work so far (using classical methods) is very complicated and restricted in scope but does give the hoped-for results. In his partial likelihood paper, Cox gave a very brief sketch of how asymptotic results might be derived. Though it is not recognized as such, there is the germ of a martingale argument in this sketch, a fact that will be of great significance.

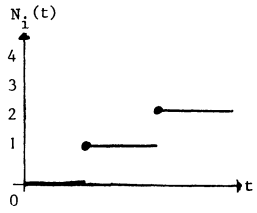
Before taking this point further, let us mention a related class of problems concerning possible extensions to the model. Can we allow other types of censoring than the fixed censoring specified above? Can we allow covariates to be random processes Z_i rather than fixed functions? In this context it is fascinating that by very curious choices of random covariates, one can derive all of the well-known nonparametric k-sample tests for censored survival data as score tests based on $\mathcal{L}(\beta)$ for the hypothesis $\beta_0 = 0$ (see Oakes 1981 and Lustbader 1980). Can we model more complicated situations with repeated events or events of different types (rather than the single event "death") in the life of any individual? In all cases it is easy to write down analogs to $\mathcal{L}(\beta)$, but it is not obvious that they will still have the same properties.

3. SECOND SPECIFICATION OF THE MODEL

We will reformulate Cox's regression model as a model for the random intensity of a multivariate counting process. Let us first discuss the meaning of these terms. A multivariate counting process

$$\tilde{N} = \{N_i(t): 0 \leq t < \infty; \quad i = 1, \ldots, n\}$$

is a stochastic process with n components that can be thought of as counting the occurrences (as time t proceeds) of n different types of event (or the same event for n different individuals). We suppose these events occur singly. The realizations of each component $N_i(\cdot)$, seen as functions of t, are integer-valued step functions, zero at time zero, with jumps of size $+1$ only. We also suppose them to be right-continuous, so that $N_i(t)$ is the (random) number of events of type i in the time interval $[0, t]$. No two components jump at the same time.



Under regularity conditions, which need not concern us, the process $\tilde{N}$ has an intensity process

$$\tilde{\Lambda} = \{\Lambda_i(t): 0 \leq t < \infty; \quad i = 1, \ldots, n\}$$

defined by

$$\Lambda_i(t)dt = P[N_i \text{ jumps in a time interval of length } dt \text{ around time } t \mid \mathcal{F}_{t-}], \tag{3.1}$$

where $\mathcal{F}_{t-}$ denotes the past up to the beginning of the small time interval dt, that is, everything that has happended until just before time t. Here we include a complete specification of the paths $N_j(\cdot)$, $j = 1, \ldots, n$, on $[0, t)$, as well as all other events implicitly or explicitly included in the model that can be thought of as having occurred before time t.

Let us take as an example a very simple multivariate counting process, each component of which jumps at most once. In Cox's model in Section 3, we define

$$N_i(t) = I\{T_i \leq t, T_i \leq c_i\}.$$

So N_i jumps once, if at all, at time $T_i \leq c_i$ of individual i's observed death. What can be said about Λ_i in this case? Given what has happened before the time interval dt, we know that individual i died at the observed time T_i less than t and less than the censoring time c_i, or that individual i was censored at time $c_i < t$, or that individual i is still alive and uncensored. In the first two cases, we know that N_i either has made its only jump or will never jump, so the probability of a jump in the interval dt is zero. In the last case, we know that $T_i \in dt$ or $T_i \geq t$, so by (2.1) the probability of a jump in the interval dt is $\lambda_i(t)dt$. Thus, defining

$$\begin{aligned} Y_i(t) &= I\{T_i \geq t, \quad c_i \geq t\} \\ &= 1 \text{ if individual } i \text{ is under observation just before time } t \\ &= 0 \text{ otherwise,} \end{aligned} \tag{3.2}$$

we have by (2.2) and (3.1)

$$\Lambda_i(t)dt = Y_i(t)\lambda_0(t)\exp\{\beta_0' z_i(t)\}dt.$$

Note that given the past up to (but not including) time t, $Y_i(t)$ and $\Lambda_i(t)$ are fixed or nonrandom. We say in such a case that Y_i and Λ_i are predictable.

An obvious extension of Cox's regression model is now: $\tilde{N}$ is a multivariate counting process with intensity process $\tilde{\Lambda}$ satisfying

$$\Lambda_i(t)dt = Y_i(t)\lambda_0(t)\exp\{\beta_0' Z_i(t)\}dt. \tag{3.3}$$

Here we have replaced the fixed covariate $z_i(t)$ by the random covariate $Z_i(t)$. We no longer require each N_i to make at most one jump, nor do we require Y_i to be of the special form given in (3.2). All we require is that N_i, Y_i, and Z_i are processes that can be observed and that Y_i and Z_i are predictable ($Y_i(t)$ and $Z_i(t)$ are fixed given what has happened before time t). This condition is forced on us by the meaning of $\Lambda_i(t)$ as the intensity or rate with which N_i jumps given the past. This also restricts Y_i to being nonnegative.

Consider an example in which we wish to model the effects of a drug that is given to the treatment group over a possibly varying length of time; there is also a control group. We might want to investigate whether the drug has a different effect during treatment from its effect after treatment has ended. To this end we could define two components of Z_i, say the first two, as follows:

$$\begin{aligned} Z_{i1}(t) &= 1 \text{ during the treatment of a patient in the treatment group} \\ &= 0 \text{ otherwise;} \\ Z_{i2}(t) &= 1 \text{ after treatment of a patient in the treatment group} \\ &= 0 \text{ otherwise.} \end{aligned}$$

If the two corresponding components of β_0 are negative, the treatment is effective; if, moreover, the first component is significantly larger in absolute value than the second, then the effect of the treatment apparently has declined after treatment has stopped. Many variations on this kind of model are possible and sensible. Note that we do not require the treatment period for each patient to be fixed beforehand; it may be adapted or curtailed by, say, the occurrence of side effects. One might even include the occurrence of side effects as yet another 0–1 component of Z_i. The only restriction is that $Z_i(t)$ must indicate the status of the ith patient just before time t.

For an example in which the processes N_i may have

several jumps, see Andersen and Gill (1982). As to the almost arbitrary nature of the process Y_i, note that we may now have patients, for instance, entering observation at times t larger than the start time 0 (representing time of diagnosis, randomization, or operation) or returning to the study after a period during which they were lost to observation. For further discussion of the implications and interpretation of the model, see Self and Prentice (1982).

Finally, we rewrite (2.3) in the new notation. Our proposal is still to estimate β_0 by treating

$$L(\beta) = \prod_{t\geq 0}\prod_{i=1}^{n}\left(\frac{Y_i(t)\exp(\beta' Z_i(t))}{\sum_{j=1}^{n} Y_j(t)\exp(\beta' Z_j(t))}\right)^{dN_i(t)} \tag{3.4}$$

as an ordinary likelihood function for β_0 and derive confidence intervals, significance tests, and so on, using standard large-sample likelihood theory. In formula (3.4), $dN_i(t)$ is the increment of N_i over a small interval dt around the time t and the product over t is a product over disjoint intervals. So (3.4) reduces to a finite product over all i and t for which N_i jumps at time t ($dN_i(t) = 1$); elsewhere $dN_i(t) = 0$. Let $\hat{\beta}$ be the value of β maximizing $L(\beta)$, and also define $L(\beta, u)$ as the likelihood function based on the observations on the time interval $[0, u]$, in which the product over $t \geq 0$ in (3.4) is replaced by a product over $t \geq 0$, $t \leq u$.

4. SOME MARTINGALE THEORY

A martingale $M = \{M(t): t \geq 0\}$ is a stochastic process whose increment over an interval $(u, v]$, given the past up to and including time u, has expectation zero. In symbols, we have

$$\mathcal{E}[M(v) - M(u) \mid \mathcal{F}_u] = 0 \tag{4.1}$$

for all $0 \leq u < v < \infty$. Given $\mathcal{F}_u$, $M(u)$ is fixed. A great deal is known about martingales. There are, for instance, martingale transform theorems, which state that integrating a predictable process with respect to a martingale yields a new martingale, and there are martingale central limit theorems, which give conditions under which the whole process M is approximately normally distributed, with independent increments (so the process looks like a time-transformed Brownian motion).

We will shortly sketch the ideas behind these two topics. First, though, we rewrite the defining property (4.1) by taking the time instants u and v to be just before and just after the time instant t, giving

$$\mathcal{E}[dM(t) \mid \mathcal{F}_{t-}] = 0. \tag{4.2}$$

Let us relate this to the defining property (3.1) of the intensity of a counting process. Note that in a small time interval dt, N_i either jumps once or does not jump at all. So the probability of a jump in that interval is close to the expected number of jumps in the interval. Thus (3.1) states $\Lambda_i(t)dt = \mathcal{E}[dN_i(t) \mid \mathcal{F}_{t-}]$ or, defining $dM_i(t) = dN_i(t) - \Lambda_i(t)dt$, $\mathcal{E}[dM_i(t) \mid \mathcal{F}_{t-}] = 0$. So (3.1) is equivalent to the assertion that M_i, defined by

$$M_i(t) = N_i(t) - \int_0^t \Lambda_i(s)ds, \tag{4.3}$$

is a martingale.

We need one more concept, that of the predictable variation process of a martingale M. That is a process $\langle M\rangle = \{\langle M\rangle(t): t \geq 0\}$ defined by

$$d\langle M\rangle(t) = \mathcal{E}[dM(t))^2 \mid \mathcal{F}_{t-}] = \operatorname{var}[dM(t) \mid \mathcal{F}_{t-}].$$

It is predictable and nondecreasing and can be thought of as the sum of conditional variances of the increments of M over small time intervals partitioning $[0, t]$, each conditional variance being taken given what has happened up to the beginning of the corresponding interval. One can similarly define the predictable covariation process of two martingales, M and M' say, denoted by $\langle M, M'\rangle$.

We illustrate this concept with the counting process martingales M_i, $i = 1, \ldots, n$, of (4.3). Given the past up to the beginning of an interval dt, $dN_i(t)$ is a zero–one variable. Its conditional expectation is $\Lambda_i(t)dt$, and hence its conditional variance is $\Lambda_i(t)dt\,(1 - \Lambda_i(t)dt) \approx \Lambda_i(t)dt$. Thus we expect (and this turns out to be true) that

$$\langle M_i\rangle(t) = \int_0^t \Lambda_i(s)ds.$$

As to the predictable covariance between M_i and M_j, $i \neq j$, recall that we supposed that N_i and N_j never jump simultaneously. Thus $dN_i(t)dN_j(t)$ is always zero, and hence the conditional covariance between $dN_i(t)$ and $dN_j(t)$ is $-\Lambda_i(t)dt.\Lambda_j(t)dt \approx 0$. Indeed, it is the case that

$$\langle M_i, M_j\rangle(t) = 0, \quad \text{for all } t \text{ and } i \neq j.$$

We now can discuss the results mentioned at the beginning of Section 5. Suppose M is a martingale and H is a predictable process. Define a process $M' = \{M'(t): t \geq 0\}$ by

$$M'(t) = \int_0^t H(s)dM(s)$$

or, equivalently, $dM'(t) = H(t)dM(t)$. Then M' is also a martingale, since

$$\begin{aligned}
&\mathcal{E}[dM'(t) \mid \mathcal{F}_{t-}] \\
&\quad = \mathcal{E}[H(t)dM(t) \mid \mathcal{F}_{t-}] \\
&\quad = H(t)\mathcal{E}[dM(t) \mid \mathcal{F}_{t-}] \quad \text{(because } H \text{ is predictable)} \\
&\quad = 0 \qquad\qquad\qquad\quad \text{(because } M \text{ is a martingale).}
\end{aligned}$$

Furthermore, $\langle M'\rangle(t) = \int_0^t H(s)^2 d\langle M\rangle(s)$; this follows because

$$\begin{aligned}
\operatorname{var}[dM'(t) \mid \mathcal{F}_{t-}] &= \operatorname{var}[H(t)dM(t) \mid \mathcal{F}_{t-}] \\
&= H(t)^2 \operatorname{var}[dM(t) \mid \mathcal{F}_{t-}] \\
&= H(t)^2 d\langle M\rangle(t).
\end{aligned}$$

A similar result holds for the predictable covariation process of the integrals of two predictable processes with respect to two martingales.

Second, we must mention martingale central limit theorems. A time-transformed Brownian motion $W = \{W(t): t \geq 0\}$ is a process with the following properties. The realizations $W(\cdot)$ are continuous functions, zero at time zero. For any $t_1, \ldots, t_n$, $W(t_1), \ldots, W(t_n)$ is multivariate normally distributed with zero means and independent increments; thus for $s < t$, $W(t) - W(s)$ is independent of $W(s)$ (and in fact of $W(u)$ for all $u \leq s$).

By independence of increments, the conditional variance of $dW(t)$ given the path of W on $[0, t)$ does not depend on the past. Also the conditional expectation is zero. Thus W is a continuous martingale with predictable variation process $\langle W \rangle$ equal to some deterministic function A, say.

In fact, these properties characterize the distribution of W (Gaussian). So it is not surprising that if a sequence of martingales $M^{(n)}$, $n = 1, 2, \ldots$, is such that (a) the jumps of $M^{(n)}$ get smaller as $n \to \infty$ ($M^{(n)}$ becomes more nearly continuous) and (b) the predictable variation process of $M^{(n)}$ becomes deterministic, that is, $\langle M^{(n)} \rangle(t) \to A(t)$ in probability as $n \to \infty$, where A is a fixed function, then $M^{(n)}$ converges in distribution to W as $n \to \infty$; in particular, $M^{(n)}(t)$ is asymptotically normally distributed with mean zero and variance $A(t)$ and the increments of $M^{(n)}$ are asymptotically independent.

A complete account of martingale and stochastic integral theory can be found in Meyer (1976). The links to counting processes are made in Bremaud and Jacod (1977). The central limit theorem we have sketched above can be found in Rebolledo (1980); still more sophisticated theorems can be found in Liptser and Shiryayev (1980). See also Shiryayev's survey (1981). For surveys aimed at applications in statistics see Aalen (1976,1978) or Gill (1980).

5. LARGE-SAMPLE PROPERTIES OF $\hat{\beta}$

It should be recalled that classically, asymptotic normality of a consistent maximum likelihood estimator can be derived via a Taylor expansion of the first derivative of the log likelihood about the true value $\beta = \beta_0$, evaluated at $\beta = \hat{\beta}$. When writing $D \log L(\beta)$ for the vector of partial derivatives $(\partial/\partial\beta_i) \log L(\beta)$ evaluated at β, the key step is to show that $n^{-1/2} D \log L(\beta_0)$ is asymptotically multivariate normally distributed, with mean zero and covariance matrix equal to the average Fisher information. In a classical setup with iid observations from a density $f(\cdot; \beta_0)$, this result follows from the central limit theorem, since $n^{-1/2} D \log L(\beta_0)$ turns out to be $n^{-1/2}$ times the sum of n random vectors, iid, with means zero and covariance matrices equal to the Fisher information matrix.

We will show that the same approach works here if we simply use a martingale central limit theorem instead of a classical central limit theorem. This means considering $D \log L(\beta_0)$ as a sum (or integral) over time instants t rather than over individuals i. (We shall briefly discuss the problem of proving consistency of $\hat{\beta}$ later.) Recall that $L(\beta, u)$ is the likelihood for β based on observation of N_i, Y_i, and Z_i, $i = 1, \ldots, n$, on the time interval $[0, u]$, and define

$$E_0(t) = \frac{\frac{1}{n}\sum_{j=1}^{n} Y_j(t) Z_j(t) \exp(\beta_0' Z_j(t))}{\frac{1}{n}\sum_{j=1}^{n} Y_j(t) \exp(\beta_0' Z_j(t))}.$$

Then we have, from (3.4),

$$\begin{aligned}
& n^{-1/2} D \log L(\beta_0, u) \\
&= n^{-1/2} \sum_{i=1}^{n} \sum_{t \leq u} \left(Z_i(t) - \frac{\sum_{j=1}^{n} Y_j(t) Z_j(t) \exp(\beta_0' Z_j(t))}{\sum_{j=1}^{n} Y_j(t) \exp(\beta_0' Z_j(t))} \right) dN_i(t) \\
&= \sum_{i=1}^{n} \int_{t=0}^{u} n^{-1/2} (Z_i(t) - E_0(t)) dN_i(t) \\
&= \sum_{i=1}^{n} \int_{t=0}^{u} n^{-1/2} (Z_i(t) - E_0(t)) dM_i(t), \qquad (5.1)
\end{aligned}$$

since $dM_i(t) = dN_i(t) - \Lambda_i(t) dt$, and

$$\begin{aligned}
& \sum_{i=1}^{n} (Z_i(t) - E_0(t)) \Lambda_i(t) \\
&= \sum_{i=1}^{n} Z_i(t) Y_i(t) \lambda_0(t) \exp(\beta_0' Z_i(t)) - E_0(t) \\
&\qquad \times \sum_{i=1}^{n} Y_i(t) \lambda_0(t) \exp(\beta_0' Z_i(t)) \\
&= 0.
\end{aligned}$$

Now $n^{-1/2}(Z_i(t) - E_0(t))$ is a vector of predictable processes (it only depends on the fixed parameter β_0 and the predictable processes Y_j, Z_j, $j = 1, \ldots, n$), so we see by the martingale transform theorem of Section 4 that $n^{-1/2} D \log L(\beta_0, t)$, considered as a stochastic process in t, is the sum of n (vector) martingales, hence also a martingale. It now remains to verify the conditions (1) and (2) of the martingale central limit theorem of Section 4 to show that $M^{(n)}(t) = n^{-1/2} D \log L(\beta_0, t)$ is asymptotically normally distributed.

In fact, we need a vector version of that theorem (which does exist) unless the vectors β and $Z_i(t)$ are scalars. But for simplicity, let us from now on suppose that this is the case. We did not state very precisely what we meant by the jumps of $M^{(n)}$ getting smaller as $n \to \infty$. Let us consider then, a special case in which it is clear that there will be no difficulties—that in which $|Z_i(t)| \leq C < \infty$ for all i and t for some constant C. (This condition is not necessary for our final result.) In that case it is easily seen that the integrand $Z_i(t) - E_0(t)$ in (5.1) is also

bounded by C. Each M_i only has jumps of size $+1$, coinciding with the jumps of N_i. Since there are no multiple jumps, the jumps of $M^{(n)}$ are bounded by $n^{-1/2}C$, which tends to zero as $n \to \infty$. This deals with condition (1).

As for condition (2), we must evaluate the process $\langle M^{(n)} \rangle$. It is easy, using the results of Section 4 and some simple algebra, to show that

$$\langle M^{(n)} \rangle(t) = \int_0^t \frac{1}{n} \sum_{i=1}^{n} (Z_i(s) - E_0(s))^2 \Lambda_i(s) ds$$

$$= \int_0^t \left(\frac{1}{n} \sum_{i=1}^{n} Z_i(s)^2 Y_i(s) \exp(\beta_0' Z_i(s)) - \frac{\left(\frac{1}{n} \sum_{i=1}^{n} Z_i(s) Y_i(s) \exp(\beta_0' Z_i(s)) \right)^2}{\frac{1}{n} \sum_{i=1}^{n} Y_i(s) \exp(\beta_0' Z_i(s))} \right) \lambda_0(s) ds.$$

Thus $\langle M^{(n)} \rangle(t)$ can be expressed in terms of simple averages of $Y_i(s)Z_i(s)^r \exp(\beta_0' Z_i(s))$, $r = 0$, 1, and 2. We would expect to be able to show that $\langle M^{(n)} \rangle(t)$ converges in probability to some constant if these averages converge in probability. This turns out to be the case; moreover, all of the other parts of the classical proof of asymptotic normality of $\hat{\beta}$ also go through under the same conditions (sometimes with β_0 replaced by β close to β_0). In particular this applies to proving consistency of $\hat{\beta}$, which needs to be established before the above arguments can be applied. Martingale theory and concavity of the log likelihood function together yield a simple proof of this; see Andersen and Gill (1982). In conclusion, it turns out that large-sample maximum likelihood theory is valid for $\hat{\beta}$ if n is large enough that the averages

$$\frac{1}{n} \sum_{i=1}^{n} Y_i(t) Z_i(t)^r \exp(\beta' Z_i(t)), \quad r = 0, 1, \text{ and } 2,$$

are almost nonrandom for all t and for β close to β_0.

The martingale property of $M^{(n)}$ is implied in Cox's (1975) definition of partial likelihood (see p. 274). There it is shown that each term in $D \log L(\beta_0)$ has expectation zero given the preceding terms. So it does appear, more generally, that the definition of partial likelihood contains enough structure to ensure that the large-sample properties of maximum likelihood estimation hold for it, too (under similar regularity conditions). For instance, Prentice and Self (1983) show that similar results hold when the effect of covariates on survival is modeled by replacing $\exp(\beta_0' Z_i(t))$ by any other function of $\beta_0' Z_i(t)$ in the hazard rate for the ith individual (see also Thomas 1982).

6. CONCLUDING REMARKS

It was the aim of the previous sections to show that the counting process and martingale approach to Cox's regression model fits both practical and theoretical aspects of the model; that is, it gives a framework in which one can go about constructing practically realistic models, and it supplies the mathematical tools for deriving the statistical properties of the model. We claim that this is true not only for the Cox model but also for many other techniques in survival analysis. In particular we refer to the papers of Aalen and Johansen (1978), Aalen et al. (1980), Andersen et al. (1982), Andersen (1983), Ramlau-Hansen (1983), Sellke and Siegmund (1983), Harrington and Fleming (1982), Gill (1983), and Wei and Gail (1983).

One problem has not been resolved. Large-sample properties of $\hat{\beta}$ are easy to derive because of the martingale property of the derivative of the log (partial) likelihood. Thus the concept of partial likelihood is important and useful. However, Johnansen (1983), using counting process theory, shows that $\hat{\beta}$ can be motivated as a proper maximum likelihood estimator, obtained by maximizing a full likelihood over β and $\lambda(\cdot)$ simultaneously. (Of course these terms have to be defined carefully in non- and semiparametric situations.) Begun et al. (1983) showed that $\hat{\beta}$ and the corresponding maximum likelihood estimator of λ_0 have the kind of asymptotic efficiency properties one would expect of maximum likelihood estimators. These two results have not yet been related.

A similar coincidence arises in connection with the curious fact mentioned in Section 2 that not only the log–rank test but also all other well known k-sample censored-data linear rank tests in survival analysis can be derived as score tests (i.e., tests based on $D \log L(\beta)\,|_{\beta=0}$) when covariates are specified appropriately in the Cox model. This fact can be partially explained as follows. Suppose we assume that in k groups, we have censored observations from survival distributions with densities $f(t; \theta_i)$, $i = 1, \ldots, k$. Thus we have k hazard rates $\lambda(t; \theta_i)$, and by a Taylor expansion, we can write $\log \lambda(t; \theta_i) \approx \log \lambda(t; \theta_k) + (\theta_i - \theta_k) g(t; \theta_k)$ for some function g. Therefore we have, close to the null hypothesis $\theta_1 = \cdots = \theta_k = \theta_0$,

$$\lambda(t; \theta_i) \approx \lambda_0(t) \exp((\theta_i - \theta_k) z(t)), \tag{6.1}$$

where $\lambda_0(t) = \lambda(t; \theta_0)$ and $z(t) = g(t; \theta_0)$. Such a parametric model is close to the Cox model with a vector of $k - 1$ covariates, such that for an individual in group i, the ith component of the covariate at time t equals $z(t)$ and the other components are zero. The choice of covariates in the Cox model, which gives as score test the log-rank test or any other linear rank test, is precisely of this form, except that $z(t)$ is replaced by some predictable process $Z(t)$. For large sample sizes, however, $Z(t)$ is close to some nonrandom function $z(t)$. Corresponding to this function, one can generate parametric families of hazard rates via the relation

$$(\partial/\partial\theta) \log \lambda(t; \theta)\,|_{\theta=\theta_0} = z(t).$$

It turns out (Gill 1980) that each censored-data linear rank

test is asymptotically optimal, when testing against exactly those alternatives implied through (6.1) by its implicit choice of z. For some linear rank tests (e.g., those of Efron 1967 or Gehan 1965), z depends on the censoring distributions in the k samples, so the parametric alternatives generated in this way are not very interesting ones. However, for others, such as Prentice's (1978) test statistics, including the log-rank test, z only depends on the underlying (null-hypothesis) survival function. The parametric alternatives generated in this way are, in this example, precisely the alternatives for which Prentice (1978) designed his test statistics to have high power.

These coincidences of maximum likelihood and efficiency properties cry out for explanation. One may hope that an asymptotic theory of nonparametric maximum likelihood estimation will eventually be developed that will cast some light on these phenomena.

[*Received October 1982. Revised December 1983.*]

REFERENCES

AALEN, O.O. (1976), "Statistical Theory for a Family of Counting Processes," Institute of Mathematical Statistics, University of Copenhagen.

——— (1978), "Nonparametric Inference for a Family of Counting Processes," *Annals of Statistics*, 6, 701–726.

AALEN, O.O., and JOHANSEN, S. (1978), "An Empirical Transition Matrix for Nonhomogeneous Markov Chains Based on Censored Observations," *Scandinavian Journal of Statistics*, 5, 141–150.

AALEN, O.O., BORGAN, Ø., KEIDING, N., and THORMAN, J. (1980), "Interaction Between Life History Events: Nonparametric Analysis of Prospective and Retrospective Data in the Presence of Censoring," *Scandinavian Journal of Statistics*, 7, 161–171.

ANDERSEN, P.K. (1982), "Testing Goodness of Fit of Cox's Regression and Life Model," *Biometrics*, 38, 67–77.

——— (1983), "Comparing Survival Distributions via Hazard Ratio Estimates," *Scandinavian Journal of Statistics*, 10, 77–86.

ANDERSEN, P.K., BORGAN, Ø., GILL, R.D., and KEIDING, N. (1982), "Linear Nonparametric Tests for Comparison of Counting Processes, With Applications to Censored Survival Data" (with discussion), *International Statistical Review*, 50, 219–258.

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

BAILEY, K.R. (1983), "The Asymptotic Joint Distribution of Regression and Survival Parameter Estimates in the Cox Regression Model," *Annals of Statistics*, 11, 39–48.

BEGUN, J.M., HALL, W.J., HUANG, W.M., and WELLNER, J.A. (1983), "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *Annals of Statistics*, 11, 432–452.

BREMAUD, P., and JACOD, J. (1977), "Processes Ponctuels et Martingales: Résultats Récents sur la Modelisation et le Filtrage," *Advances in Applied Probability*, 9, 362–416.

COX, D.R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187–200.

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

EFRON, B. (1967), "The Two-Sample Problem With Censored Data," *Proceedings of the 5th Berkeley Symposium*, 4, 831–854.

GEHAN, E.A. (1965), "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples," *Biometrika*, 52, 203–223.

GILL, R.D. (1980), "Censoring and Stochastic Integrals," *Mathematical Centre Tracts*, No. 124, Amsterdam: The Mathematical Centre.

——— (1983), "Large Sample Behavior of the Product-Limit Estimator on the Whole Line," *Annals of Statistics*, 11, 49–58.

HARRINGTON, D.P., and FLEMING, T.R. (1982), "A Class of Rank Test Procedures for Censored Survival Data," *Biometrika*, 69, 533–546.

JOHANSEN, S. (1983), "An Extension of Cox's Regression Model," *International Statistical Review*, 51, 165–174.

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

LINK, C.L. (1979), "Confidence Intervals for the Survival Function Using Cox's Proportional Hazard Model With Covariates," Technical Report No. 45, Stanford University, Dept. of Biostatistics.

LIPTSER, R.S., and SHIRYAYEV, A.N. (1980), "A Functional Central Limit Theorem for Semimartingales," *Theory of Probability and Its Applications*, 25, 667–688.

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, University of Wisconsin–Madison, Wisconsin Clinical Cancer Center, Dept. of Biostatistics.

LUSTBADER, E.D. (1980), "Time Dependent Covariates in Survival Analysis," *Biometrika*, 67, 697–698.

MANTEL, N. (1966), "Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration," *Cancer Chemotherapy Reports*, 50, 163–170.

MEYER, P.A. (1976), "Un Cours sur les Intégrales Stochastiques," Séminaire de Probabilités X, *Lecture Notes in Mathematics*, 258, 245–400.

MILLER, R.G., Jr., EFRON, B., BROWN, B.W., and MOSES, L.E. (1980), *Biostatistics Casebook*, New York: John Wiley.

NAES, T. (1981), "Estimation of the Non-Regression Part of the Hazard Rate in Cox's Regression Model," Research Report, Norwegian Food Research Institute, NLH-Aas.

——— (1982), "The Asymptotic Distribution of the Estimator for the Regression Parameter in Cox's Regression Model," *Scandinavian Journal of Statistics*, 9, 107–115.

OAKES, D. (1981), "Survival Times: Aspects of Partial Likelihood" (with discussion), *International Statistical Review*, 49, 235–264.

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

PRENTICE, R.L., and SELF, S.G. (1983), "Asymptotic Distribution Theory for Cox-Type Regression Models With General Relative Risk Form," *Annals of Statistics*, 11, 804–813.

RAMLAU-HANSEN, H. (1983), "Smoothing Counting Process Intensities by Means of Kernel Functions," *Annals of Statistics*, 11, 433–466.

REBOLLEDO, R. (1980), "Central Limit Theorems for Local Martingales," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebieten*, 51, 269–286.

SELF, S.G., and PRENTICE, R.L. (1982), "Commentary on Andersen and Gill's 'Cox's Regression Model for Counting Processes: A Large Sample Study,'" *Annals of Statistics*, 10, 1121–1124.

SELLKE, T., and SIEGMUND, D. (1983), "Sequential Analysis of the Proportional Hazards Model," *Biometrika*, 70, 315–326.

SEN, P.K. (1981), "The Cox Regression Model, Invariance Principles for Some Induced Quantile Processes and Some Repeated Significance Tests," *Annals of Statistics*, 9, 109–121.

SHIRYAYEV, A.N. (1981), "Martingales: Recent Developments, Results and Applications," *International Statistical Review*, 49, 199–233.

THOMAS, D.C. (1982), "General Relative-Risk Models for Survival Time and Matched Case-Control Analysis," *Biometrics*, 37, 673–686.

TSIATIS, A.A. (1978a), "A Heuristic Estimate of the Asymptotic Variance of the Survival Probability in Cox's Regression Model," Technical Report No. 524, University of Wisconsin–Madison, Dept. of Statistics.

——— (1978b), "A Large Sample Study of the Estimate for the Integrated Hazard Function in Cox's Regression Model for Survival Data," Technical Report No. 526, University of Wisconsin–Madison, Dept. of Statistics.

——— (1981a), "A Large Sample Study of Cox's Regression Model," *Annals of Statistics*, 9, 93–108.

——— (1981b), "The Asymptotic Distribution of the Efficient Scores Test for the Proportional Hazards Model Calculated Over Time," *Biometrika*, 68, 311–315.

WEI, L.J., and GAIL, M.H. (1983), "Nonparametric Estimation for a Scale-Change With Censored Observations," *Journal of the American Statistical Association*, 78, 382–388.