stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SW 81/82 FEBRUARI
(DEPARTMENT OF MATHEMATICAL STATISTICS)

R.D. GILL

UNDERSTANDING COX'S REGRESSION MODEL: A MARTINGALE APPROACH

Preprint

kruislaan 413 1098 SJ amsterdam

BIBLIOTHEEK MATHEMATISCH CENTRUM
AMSTERDAM

Printed at the Mathematical Centre, 413 Kruislaan, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

1980 Mathematics subject classification: 62P10, 62F12

Understanding Cox's regression model: a martingale approach [*)]

by

R.D. Gill

ABSTRACT

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 & PHRASES: *censoring, survival data, Cox's regression model, partial likelihood, martingale, counting processes, asymptotic theory*

---

*) This report will be submitted for publication elsewhere.

0. 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 advantage of both parametric and nonparametric approaches to statistical inference, and is beautifully adapted to the kind of data one will obtain in clinical cancer trials and other sources of survival data and lifetesting data. By incorporating time-varying, random, covariates it becomes a highly flexible tool for model building.

Despite this its mathematical basis so far is almost entirely heuristic. Even just to intuitively motivate 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 & CROWLEY (1978), TSIATIS (1978a, 1978b, 1981a, 1981b), LINK (1979), BAILEY (1979), NAES (1981a, 1981b)), all confirming Cox's original conjectures, but all restricted to very special cases. Moreover, in all cases derivations are highly complex and technical. For instance, simple formulae 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 paper is to discuss recent work by JOHANSEN (1981) and ANDERSEN & GILL (1981) which shows how a firm mathematical basis can be given to the model (in its fullest generality) from which for instance 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 the model of Cox 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 up immediately with modern martingale and stochastic integral theory, very powerful and deep mathematical tools which 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, what kinds of covariates, etc.

After sketching this theory on an intuitive level, we indicate how it

can be used to derive Cox's estimator as an *ordinary maximum likelihood* estimator (JOHANSEN, 1981), and how asymptotic properties of the estimator also follow simply from this formulation of the model (ANDERSEN & GILL, 1981).

1. INTRODUCTION

Hopefully it will be possible to read this paper at several different levels. At the most obvious level, the paper summarizes some outstanding problems concerning Cox's regression model and indicates solutions to these problems which are further developed in ANDERSEN & GILL (1981) and JOHANSEN (1981). At the same time, the paper 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 paper contains an introduction at a very intuitive level to the statistical theory of counting processes which is currently being used, following the work of AALEN (1976), to unify and extend many branches of nonparametric survival analysis. Finally, we hope the paper will encourage those analysing 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 be also analysed 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 used are a natural formalization of the heuristic derivations of, for instance, MANTEL (1966) page 169, or COX (1975) page 274. This is in contrast to the classical approach to survival analysis, which has been to solve its problems using classical tools derived for instance from classical nonparametric theory. To oversimplify, this has forced us to restrict attention to situations with independent and identically distributed 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 though the mathematical presentation in this paper is entirely informal, everything we say can be made rigorous.

We next briefly describe the structure of the paper. In the next section 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 which 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 and 6 we show how this theory can be used to derive asymptotic properties of the statistical procedures appropriate to the model, and we show that these procedures can be motivated by the classical maximum likelihood method without reference to partial likelihood. The last section contains some concluding remarks. We suggest that the less mathematically inclined reader should skip Sections 4 to 6. The statistician who wants to understand the general counting process approach used in such papers as AALEN (1978), AALEN, BORGAN, KEIDING & THORMANN (1980) or ANDERSEN, BORGAN, GILL & KEIDING (1981) could skip Sections 5 and 6 which are specific to the Cox model.

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

$$(2.1)\qquad \lambda_i(t) = \lim_{h\downarrow 0} \frac{1}{h} P\,[T_i \le t+h \mid T_i \ge t]$$

of the special form

$$(2.2)\qquad \lambda_i(t) = \lambda_0(t)\exp(\beta_0' z_i(t))$$

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 "base-line" hazard rate for an individual with $z \equiv 0$. The observations for the i-th individual consist of

$$T_i \wedge c_i,$$

$$\delta_i = I\{T_i \le c_i\}, \text{ and}$$

$$z_i(t),\ t \in [0, T_i \wedge c_i].$$

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 semi-parametric.

For the interpretation of the model and for examples of how covariates z_i can be chosen, we refer to COX (1972), ANDERSEN & GILL (1981), MILLER, EFRON, BROWN & MOSES (1980), ANDERSEN (1981,1982), and KALBFLEISCH & PRENTICE (1980).

Let

$$\mathcal{R}(t) = \{i\colon T_i \ge t \text{ and } c_i \ge t\}$$

denote the risk set at time t, that is to say, the set of individuals i who are under observation at time t. Given 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 to be

$$\frac{\exp(\beta_0' z_i(t))}{\sum_{j \in \mathcal{R}(t)} \exp(\beta_0' z_j(t))},$$

a factor $\lambda_0(t)$ has cancelled 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 proposed therefore that statistical inference on β_0 could be carried out by considering

$$(2.3) \qquad L(\beta) = \Pi_{i:T_i \le c_i} \left(\frac{\exp(\beta' z_i(T_i))}{\sum_{j \in R(T_i)} \exp(\beta' z_j(T_j)} \right)$$

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 or not $L(\beta)$ is some sort of likelihood function has given rise to much discussion in the literature. It certainly is not a *conditional* likelihood: i.e. 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 to say, 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 $L(\beta)$ is one (to date, the most important example of partial likelihood).

Whatever sort of likelihood $L(\beta)$ 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 rigorously deriving the required asymptotics: all 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 which will turn out to be of great significance.

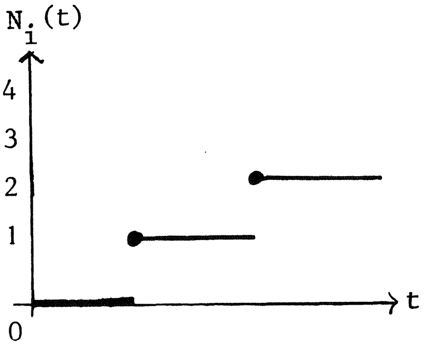
Before taking this point further, let us mention a related class of problems concerning possible extensions to the model. Can we allow other types of censoring than the "fixed censoring" specified above? Can we allow covariates to be random processes Z_i rather than fixed functions? (In this context it is fascinating that, by very curious choices of random covariates, one can derive all the well known non-parametric k-sample tests for censored survival data as score test based on $L(\beta)$ 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 analogues to $L(\beta)$, but not obvious that it will still have the same properties.

3. SECOND SPECIFICATION OF THE MODEL

We are going to reformulate Cox's regression model as a model for the random intensity of a multivariate counting process. So let us first discuss the meaning of these terms. A multivariate counting process

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

is a stochastic process with n components which can be thought of as counting the occurrences (as time t proceeds) of n different types of event. 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

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

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; i.e. everything that has happened till 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 which 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 of the previous section, define

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

So N_i jumps once, if at all, at the time $T_i \le c_i$ 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 either know that individual i has died at the observed time T_i less that 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 that the probability of a jump in the interval dt is zero. In the last case, we know that $T_i \in dt$ or $T_i \ge t$ so that by (2.1) the probability of a jump in the interval dt is $\lambda_i(t)dt$. Thus defining

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

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) the time t, $Y_i(t)$ and $\Lambda_i(t)$ are fixed or non-random. 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

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

Here we have replaced the fixed covariate $z_i(t)$ by the random covariate $Z_i(t)$. We do not any 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 which can be observed; 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 which 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:

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

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

Hopefully the two corresponding components of β_0 are negative; 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 i'th patient *just before* time t.

For an example in which the processes N_i may have several jumps, see ANDERSEN & GILL (1981). 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 patients may *return* to the study after a period during which they were lost to observation.

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

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

as an ordinary likelihood function for β_0, and derive confidence intervals, significance tests, etc., using standard large sample likelihood theory. In formula (3.4), we mean by $dN_i(t)$ 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

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

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

$$(4.2) \qquad E[dM(t) \mid F_{t-}] = 0$$

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

$$\Lambda_i(t)dt = E[dN_i(t) \mid F_{t-}]$$

or, defining $dM_i(t) = dN_i(t) - \Lambda_i(t)dt$

$$E[dM_i(t) \mid F_{t-}] = 0.$$

So (3.1) is equivalent to the assertion that M_i defined by

$$(4.3) \qquad M_i(t) = N_i(t) - \int_0^t \Lambda_i(s)ds$$

is a martingale.

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

$$d<M>(t) = E[dM(t))^2 \mid F_{t-}] = \mathrm{var}[dM(t) \mid F_{t-}].$$

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; it is denoted by <M,M'>.

We illustrate this concept with the counting process martingales M_i, i = 1,...,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 $\Lambda_i(t)dt$, 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

$$<M_i>(t) = \int_0^t \Lambda_i(s)ds.$$

As to the predictable covariance between M_i and M_j, $i \neq j$, recall that we have 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

$$<M_i,M_j>(t) = 0 \qquad \text{for all } t \text{ and } i \neq j.$$

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

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

or equivalently $dM'(t) = H(t)dM(t)$. Then M' is also a martingale; for we have

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

Furthermore $<M'>(t) = \int_0^t H(s)^2dM(s)$; this follows because

$$\begin{aligned} \text{var}[dM'(t)|F_{t-}] &= \text{var}[H(t)dM(t)|F_{t-}] \\ &= H(t)^2\text{var}[dM(t)|F_{t-}] \\ &= H(t)^2d<M>(t). \end{aligned}$$

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

Secondly 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

(1) the jumps of $M^{(n)}$ get smaller as $n \to \infty$ ($M^{(n)}$ becomes more nearly continuous), and

(2) the predictable variation process of $M^{(n)}$ becomes deterministic, i.e. $\langle M^{(n)}\rangle(t) \to A(t)$ 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 & JACOD (1977). The central limit theorem we have sketched above can be found in REBOLLEDO (1980); more sophisticated theorems still can be found in LIPTSER & SHIRYAYEV (1980). 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 maximum likelihood estimator can be derived via a Taylor expansion of the first derivative of the log likelihood, evaluated at $\beta = \hat{\beta}$, about the true value $\beta = \beta_0$. Writing $D\log L(\beta)$ for the vector of partial derivatives

$(\partial/\partial\beta_i)\log L(\beta)$ evaluated at β, then the key step is to show that $n^{-\frac{1}{2}}D\log L(\beta_0)$ is asymptotically multivariate normally distributed, with mean zero, and covariance matrix equal to the average Fisher information. In a classical set-up with independent and identically distributed observations from a density $f(\cdot;\beta_0)$, this result follows from the central limit theorem, for $n^{-\frac{1}{2}}D\log L(\beta_0)$ turns out to be $n^{-\frac{1}{2}}$ times the sum of n random vectors, independent and identically distributed, with means zero and covariance matrices equal to the Fisher information matrix.

We shall show that the same approach works here, if we simply use a martingale central limit theorem instead of a classical central limit theorem. 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)

$$(5.1)\qquad n^{-\frac{1}{2}}D\log L(\beta_0,u)$$

$$= n^{-\frac{1}{2}}\sum_{i=1}^n \sum_{t\le u}\left(Z_i(t) - \frac{\sum_{j=1}^n Y_j(t)Z_j(t)\exp(\beta_0'Z_i(t))}{\sum_{j=1}^n Y_j(t)\exp(\beta_0'Z_j(t))}\right)dN_i(t)$$

$$= \sum_{i=1}^n \int_{t=0}^u n^{-\frac{1}{2}}(Z_i(t) - E_0(t))dN_i(t)$$

$$= \sum_{i=1}^n \int_{t=0}^u n^{-\frac{1}{2}}(Z_i(t) - E_0(t))dM_i(t)$$

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

$$\sum_{i=1}^n (Z_i(t) - E_0(t))\Lambda_i(t)$$

$$= \sum_{i=1}^n Z_i(t)Y_i(t)\lambda_0(t)\exp(\beta_0'Z_i(t)) - E_0(t)\sum_{i=1}^n Y_i(t)\lambda_0(t)\exp(\beta_0'Z_i(t))$$

$$= 0.$$

Now $n^{-\frac{1}{2}}(Z_i(t) - E_0(t))$ 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 $n^{-\frac{1}{2}}D\log L(\beta_0,t)$, 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 $M^{(n)}(t) = n^{-\frac{1}{2}}D \log L(\beta_0,t)$ 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. Also 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 $n^{-\frac{1}{2}}C$, 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

$$\begin{aligned}\langle M^{(n)}\rangle(t) &= \int_0^t \frac{1}{n}\sum_{i=1}^n (Z_i(s) - E_0(s))^2\Lambda_i(s)ds \\ &= \int_0^t \left(\frac{1}{n}\sum_{i=1}^n Z_i(s)^2Y_i(s)\exp(\beta_0'Z_i(s)) \right. \\ &\qquad \left. - \frac{(\frac{1}{n}\sum_{i=1}^n Z_i(s)Y_i(s)\exp(\beta_0'Z_i(s)))^2}{\frac{1}{n}\sum_{i=1}^n Y_i(s)\exp(\beta_0'Z_i(s))}\right)\lambda_0(s)ds\end{aligned}$$

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 the other parts of the classical proof of asymptotic normality of $\hat{\beta}$ turn out also to go through under the same conditions (sometimes with β_0 replaced by β close to β_0). In conclusion, it turns out that large sample maximum likelihood theory is valid for $\hat{\beta}$ if n is large enough that the averages $\frac{1}{n} \sum_{i=1}^{n} Y_i(t)Z_i(t)^r \exp(\beta' Z_i(t))$, $r = 0,1$ and 2, are almost non-random 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 page 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 in more generality 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).

6. $\hat{\beta}$ AS A MAXIMUM LIKELIHOOD ESTIMATOR

The result of Section 5 shows that $\hat{\beta}$ has the expected large sample properties, whatever sort of likelihood $L(\beta)$ may be. These properties, and other statistical efficiency properties of this estimator which are beginning to appear in the literature, all point to a very close connection to classical likelihood theory. In this Section we sketch JOHANSEN'S (1981) proof that $\hat{\beta}$ is an *ordinary* maximum likelihood estimator for β_0, obtained by maximizing a joint likelihood for β_0 and λ_0. This proof depends on a result from martingale theory which we did not mention in Section 4, (to be found in BREMAUD & JACOD, 1977) which shows how in general a likelihood function can be written down based on observing a multivariate counting process. Here we need to make two assumptions. Firstly, $Z_i(t)$ and $Y_i(t)$ are only random through dependence on $N_j(s)$, $j = 1,\ldots,n$, $s < t$ and through dependence on other events which can be considered as taking place at time $t = 0$. Secondly, the distribution of these time zero events does not depend on the parameters β_0 and λ_0.

Thus in computing a likelihood based on all that is observed $(N_i, Y_i, Z_i;\ i = 1,\ldots,n)$ we may condition on what happens at time zero and then look at the distribution of $N_i;\ i = 1,\ldots,n$ only; the rest of what is observed is now determined. It turns out that the likelihood tion may now be determined exactly as one would expect: commute the distribution of $dN_i(t)$, $i = 1,\dots,n$, given the past up to time t; write down the corresponding density functions and multiply t in order to obtain an unconditional density for $\tilde{N}$.

Now given the past, $dN_i(t)$, $i = 1,\dots,n$, are approximately distributed as independent zero-one variables with expectations $\Lambda_i(t)dt$. Equally well we can say that they are approximately distributed as Poisson variables with expectations $\Lambda_i(t)dt$. Their joint probability density can be written down as a product of the distribution of the sum $d\bar{N}(t) = \sum_{i=1}^n dN_i(t)$ (Poisson with expectation $\bar{\Lambda}(t)dt = \sum_{i=1}^n \Lambda_i(t)dt$) and the conditional distribution given the sum, which is multinomial with parameters $d\bar{N}(t)$ and $\frac{\Lambda_i(t)dt}{\bar{\Lambda}(t)dt}$, $i = 1,\dots,n$. Thus we obtain the following joint likelihood for λ and β:

$$L(\beta,\lambda) = \prod_t \left\{ \frac{(\bar{\Lambda}(t)dt)^{d\bar{N}(t)} \exp(-\bar{\Lambda}(t)dt)}{(d\bar{N}(t))!} \, . \binom{d\bar{N}(t)}{dN_1(t) \ \dots \ dN_n(t)} \prod_{i=1}^n \left(\frac{Y_i(t)\exp(\beta' Z_i(t))}{\sum_{j=1}^n Y_j(t)\exp(\beta' Z_j(t)} \right)^{dN_i(t)} \right\}$$

which is proportional to

$$\prod_t (\bar{\Lambda}(t)dt)^{d\bar{N}(t)} \exp(-\bar{\Lambda}(t)dt) . L(\beta)$$

in which of course $\bar{\Lambda}(t)dt = \lambda(t)dt \sum_{i=1}^n Y_i(t)\exp(\beta' Z_i(t))$ is considered as a function of λ and β, whose true values are λ_0 and β_0.

Let us for the moment try to maximize this function unthinkingly over the parameters β and $\lambda(t)dt$, $t \geq 0$. We will consider the sense of the conclusion afterwards. For any fixed β, maximization over $\lambda(t)dt$ gives the equation

$$\bar{\Lambda}(t)dt = \lambda(t)dt \sum_{i=1}^n Y_i(t)\exp(\beta' Z_i(t)) = d\bar{N}(t)$$

hence

$$\lambda(t)dt = \frac{d\bar{N}(t)}{\sum_{i=1}^n Y_i(t)\exp(\beta' Z_i(t))} .$$

Denoting this λ by $\hat{\lambda}|_\beta$ we obtain

$$L(\beta,\hat{\lambda}|_\beta) = \prod_t (d\bar{N}(t))^{d\bar{N}(t)} L(\beta).$$

Thus Cox's partial likelihood is in fact a partially maximized likelihood: $L(\beta) = \max_\lambda L(\beta,\lambda)$ (up to a constant of proportionality). The overall maximum likelihood estimator of λ_0 is then given by

$$\hat{\lambda}(t)|_{\hat{\beta}} dt = \frac{d\bar{N}(t)}{\sum_{i=1}^n Y_i(t)\exp(\hat{\beta}'Z_i(t))}$$

Define $H_0(t) = \int_0^t \lambda_0(s)ds$. Equivalently we can say that the maximum likelihood estimator of H_0 is $\hat{H}$ defined by

$$\hat{H}(t) = \int_0^t \frac{d\bar{N}(t)}{\sum_{i=1}^n Y_i(t)\exp(\hat{\beta}'Z_i(t))}$$

This takes us outside our original model in which H_0 must be a *continuous* function. This is not surprising; by letting λ peak more and more at jump times of $N_1,\ldots,N_n$ and be zero elsewhere, we make the probability of the observations all the larger. In fact maximum likelihood estimators with $\hat{H}$ continuous do not exist. We can better reformulate our problem and look for the maximum likelihood estimators of β_0 and H_0 without any restrictions on $\hat{H}$. But what is the extended model which is implied by this problem? Looking back at the Poisson approximation we used, we see that the extended model is: for t at which H_0 is continuous, nothing is changed (i.e. the model is the intensity specification (3.3) (with $\lambda_0(t)dt$ replaced by $dH_0(t)$). However for t at which H_0 jumps, we are assuming that (given the past) $dN_i(t)$, $i = 1,\ldots,n$, are independent Poisson with parameters $Y_i(t)dH_0(t)\exp(\beta_0'Z_i(t))$. This extended model allows multiple jumps and may not seem very realistic. However this fact is not important as we are *not* proposing that it should be used for truly discrete data (for which one might have different N_i's jumping at the same time, but hardly ever arbitrarily sized jumps of one N_i). Rather we consider the extended method as a pure mathematical construct (though a very natural one) which allows us to consider $\hat{\beta}$ and $\hat{H}$ as joint maximum likelihood estimators of β_0 and H_0. Hopefully in the future a large-sample theory of nonparametric maximum likelihood estimation will be developed which will then cover this model too.

Finally we should come back to the curious fact mentioned in Section 2 that not only the log rank test but all other well known k-sample 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 explained as follows. Suppose we assume that in k groups we have survival distributions with densities $f(t;\theta_i)$, $i = 1,\dots,k$. 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 $\theta_1 =\dots= \theta_k$,

$$(6.1) \qquad \lambda(t;\theta_i) \approx \lambda_0(t)\exp((\theta_i-\theta_k)z(t))$$

where $\lambda_0(t) = \lambda(t;\theta_k)$ and $z(t) = g(t;\theta_k)$. So 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 equals z(t), and the other components are zero. However z(t) is not known, but in such models it can be consistently estimated. It turns out (GILL, 1980) that each censored data linear rank test is asymptotically optimal when testing exactly those parametric alternatives implied through (6.1) by its implicit choice of z_i. This supplies yet another example of how treating $L(\beta)$ as a likelihood function is not misleading since the resulting tests also enjoy the expected optimality properties.

7. CONCLUSIONS

It was the aim of the previous sections to show that the counting process and martingale approach to Cox's regression model is one which fits both practical and theoretical aspects of the model: i.e. 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 not only true for the Cox model but also for many other techniques in survival analysis.

One problem has not been resolved. Large sample properties of $\hat{\beta}$ (Section 5) 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, despite the fact that in Section 6 we saw that the concept was not needed to motivate the estimator $\hat{\beta}$. This fact also shows that other models too will be tractable with martingale techniques. The particular choice of $\exp(\beta' Z_i(t))$ as a way of parametrizing the effect of covariates on survival has mathematically speaking many advantages. However it is practically speaking an arbitrary choice, and one might want to fit other forms of dependence.

REFERENCES

AALEN, O.O. (1976), *Statistical theory for a family of counting processes,* Inst. Math. Stat., Univ. of Copenhagen.

AALEN, O.O. (1978), *Nonparametric inference for a family of counting processes,* Ann. Statist. 6 701-726.

AALEN, O.O., BORGAN, Ø., KEIDING, N. & THORMANN, J. (1980), *Interactions between life history events: nonparametric analysis for prospective and retrospective data in the presence of censoring,* Scand. J. Statist. 7 161-171.

ANDERSEN, P.K. (1981), *Measuring and evaluating prognosis using the proportional hazards model,* Research report 81/8, Statistical Research Univ, Copenhagen.

ANDERSEN, P.K. (1982), *Testing goodness of fit of Cox's regression and life model,* Biometrics 38 (to appear).

ANDERSEN, P.K. & GILL, R.D. (1981), *Cox's regression model for counting processes: a large sample study,* Report SW 73/81, Mathematical Centre, Amsterdam.

ANDERSEN, P.K., BORGAN, Ø., GILL, R.D. & KEIDING, N. (1982), *Linear nonparametric tests for comparison of counting processes, with applications to censored survival data,* Research report 81/4, Statistical Research Univ, Copenhagen.(To appear in Int. Stat Rev.)

BAILEY, K.R. (1979), *The general maximum likelihood approach to the Cox regression model,* Ph. D. dissertation, University of Chicago, Illinois.

BREMAND, P. & JACOD, J. (1977), *Processus ponctuels et martingales: résultats récents sur la modelisation et le filtrage*, Adv. Appl. Prob. 9 362-416.

COX, D.R. (1972), *Regression models and life tables*, J. Roy. Statist. Soc. (B) 34 187-200 (with discussion).

COX, D.R. (1975), *Partial likelihood*, Biometrika 62 269-276.

GILL, R.D. (1980), *Censoring and stochastic integrals*, M.C. Tract 124, Mathematical Centre, Amsterdam.

JOHANSEN, S. (1981), *An extension of Cox's regression model*, Preprint from Inst. of Math. Stat., Univ. of Copenhagen.

KALBFLEISCH, J.G. & PRENTICE, R.L. (1980), *The statistical analysis of failure time data*, Wiley, New York.

LINK, C.L. (1979), *Confidence intervals for the survival function using Cox's proprotional hazard model with covariates*, Tech. Rep. No. 45, Division of Biostatistics, Stanford University.

LIPTSER, R.S. & SHIRYAYEV, A.N. (1980), *A functional central limit theorem for semimartingales*, Th. Prob. Appl. 25 667-688.

LIU, P.Y. & CROWLEY, J. (1978), *Large sample theory of the m.l.e. based on Cox's regression model for survival data*, Tech. Rep. No. 1, Wisconsin Clinical Cancer Centre, Biostatistics, Univ. of Wisconsin-Madison.

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 Chemother. Rep. 50 163-170.

MEYER, P.A. (1976), *Un cours sur les intégrales stochastiques, Séminaire de Probabilités* X 245-400, Lecture Notes in Mathematics 258, Springer, Berlin.

MILLER, R.G. Jr., EFRON, B., BROWN, B.W. Jr. & MOSES, L.E. (1980), *Biostatistics casebook*, Wiley, New York.

NAES, T. (1981b), *Estimation of the non-regression part of the hazard-rate in Cox's regression model*, Research Report from the Norwegian Food Research Institute, NLH-As.

OAKES, D. (1981), *Survival times: aspects of partial likelihood*, Int. Stat. Rev. 49 235-264 (with discussion).

REBOLLEDO, R. (1980), *Central limit theorems for local martingales*, Z. Wahrsch. u. verw. Geb. 51 269-286.

TSIATIS, A.A. (1978a), *A heuristic estimate of the asymptotic variance of the survival probability in Cox's regression model*, Tech. Rep. No. 524, Dept. of Stat., University of Wisconsin-Madison.

TSIATIS, A.A. (1978b), *A large sample study of the estimate for the integrated hazard function in Cox's regression model for survival data*, Tech. Rep. No. 526, Dept. of Stat., University of Wisconsin-Madison.

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

TSIATIS, A.A. (1981b), *The asymptotic distribution of the efficient scores test for the proportional hazards model calculated over time*, Biometrika 68 311-315.

ONTVANGEN 2 2 FEB. 1982