The Efficiency of Cox's Likelihood Function for Censored Data

BRADLEY EFRON*

D.R. Cox has suggested a simple method for the regression analysis of censored data. We carry out an information calculation which shows that Cox's method has full asymptotic efficiency under conditions which are likely to be satisfied in many realistic situations. The connection of Cox's method with the Kaplan-Meier estimator of a survival curve is made explicit.

KEY WORDS: Censored data; Cox likelihood; Survival curves.

1. INTRODUCTION

A recent California study investigated the survival times of residents at a senior citizens' facility. New arrivals joined the facility at various ages past 65, sometimes moved out of the facility, and of course not all had died by the end of the study. Complicated data-censoring patterns such as this are common in studies involving human beings. In a heavily censored situation standard regression techniques are inappropriate for analyzing the effects of covariates (such as race, sex, and blood pressure in the example above) on survival time.

D.R. Cox (1972) has suggested a regression analysis for survival data which cleverly finesses censoring difficulties. Cox's model assumes that the ith subject has hazard rate

$$h_i(t) = \theta_i(t, \boldsymbol{\beta})h(t, \boldsymbol{\gamma}) \ , \qquad (1.1)$$

where the unobserved vector $\boldsymbol{\beta}$, which parameterizes the regression of survival time on the observed covariates, is the main object of interest. Cox uses the parameterization $\theta_i(t, \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}\cdot\mathbf{z}_i(t))$, where $\mathbf{z}_i(t)$ is the possibly time-varying vector of observed covariates, but this particular form does not play a crucial role in the analysis. The unknown nuisance function $h(t, \boldsymbol{\gamma})$ modifies all the individual hazard rates equally, depending for its form on another unobserved vector $\boldsymbol{\gamma}$ of parameters.

In order to visualize (1.1) more concretely, it helps to imagine the time axis divided into infinitesimal intervals of length ϵ. We have a collection of (time-varying) coins indexed by $i = 1, 2, \ldots, n$ corresponding to all the subjects ever observed in the study. During time interval $(t, t + \epsilon)$ a subset $\mathcal{R}(t)$ of these coins, called the "risk set at time t," are each flipped once, with the probability of heads (death, in the senior citizens' study) equal to $h_i(t)\epsilon$ for the ith coin, independently of all other coins. This process proceeds sequentially in time. Once a head is achieved that coin is removed from subsequent flippings. Coins may be removed from the risk set for reasons other than death, and new coins may come on risk, i.e., join $\mathcal{R}(t)$ as t increases.

Cox's analysis proceeds as follows: let $t_1 < t_2 < \ldots < t_J$ be the observed failure times, assuming no ties, say for items $i_1, i_2, \ldots, i_J$, respectively, and let $\mathcal{R}(t_j)$ be the risk set of items on test just before the jth failure. Given $\mathcal{R}(t_j)$ and the fact that one item failed at time t_j, the conditional probability that item i_j failed is

$$\theta_{i_j}(t_j, \boldsymbol{\beta}) / \sum_{i \in \mathcal{R}(t_j)} \theta_i(t_j, \boldsymbol{\beta}) \ .$$

Simply multiplying these factors together gives the "partial likelihood function"

$$\prod_{j=1}^{J} \{\theta_{i_j}(t_j, \boldsymbol{\beta}) / \sum_{i \in \mathcal{R}(t_j)} \theta_i(t_j, \boldsymbol{\beta})\} \ . \qquad (1.2)$$

The coin-tossing model in the preceding paragraph clarifies the derivation of (1.2).

Cox treats (1.2) as an ordinary likelihood function for the purposes of inference on $\boldsymbol{\beta}$. Maximum likelihood estimates, hypothesis tests, and asymptotic confidence intervals are then derived in the usual way. Cox's analysis relates to earlier work by many authors, in particular Mantel and Haenzel (1959) and Peto and Peto (1972). A "major outstanding problem," which is the main topic of this paper, is the efficiency of inferences about $\boldsymbol{\beta}$ based on (1.2) (Cox 1972).

There are three very attractive features of Cox's approach: (1) The nuisance function $h(t, \boldsymbol{\gamma})$ is completely removed from the inference process on $\boldsymbol{\beta}$; (2) Covariate information on the different items is easily incorporated into (1.1), for example in the form $\theta_i(t, \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}\cdot\mathbf{z}_i(t))$ suggested by Cox; and (3) Data censoring patterns often encountered in life tests, such as those in the senior citizens study, do not affect (1.2).

Qualms about (1.2) were expressed in the discussion following Cox's paper. It is not really a likelihood function since it ignores a factor in the likelihood, essentially that relating to the "nonfailure intervals," $t_1, t_2 - t_1, t_3 - t_2, \ldots, t_J - t_{J-1}$, nor is it a conditional or marginal likelihood, except in very special cases. (See Kalbfleisch

* Bradley Efron is Professor, Department of Statistics and Biostatistics, Stanford University, Stanford, CA 94305. Research was supported in part by National Science Foundation Grant MPS74-21416 and Public Health Service Grant 5 R01 GM1215-02.

© Journal of the American Statistical Association
September 1977, Volume 72, Number 359
Theory and Methods Section

and Prentice 1973 and also Remark E, Section 6.) Cox's (1975) theory of *partial likelihood* shows among other things that (1.2) produces inferences similar to ordinary likelihood procedures. We use his results in Section 3.

In this article, the meaning of (1.2) is set in context by considering the complete likelihood function of all the observed data. The heuristic argument of Section 3 shows that if the class of nuisance functions $h(t, \gamma)$ is moderately large, then inferences about β based on (1.2) are asymptotically equivalent to those based on all the data. In a rough sense this solves Cox's "outstanding problem."

In practice, $h(t, \gamma)$ may be an important quantity in its own right rather than a nuisance. The connection between (1.2) and inferences about $h(t, \gamma)$ is considered briefly in Section 5, particularly as it concerns the Kaplan-Meier estimator. This analysis is closely related to that in Breslow (1974). There is also considerable overlap with Breslow and Crowley (1974), and the work of Aalen (1975) which concerns the efficiency of (1.2) for testing purposes.

We begin in Section 2 with the case of many identical items on test, to which the Kaplan-Meier estimator refers. The main result is in Section 3 with the proof deferred until Section 7. Section 4 illustrates the general theory in the special case of the two-sample problem. Section 6 consists of several brief remarks on Cox's likelihood and the Kaplan-Meier estimator.

2. IDENTICAL ITEMS ON TEST

Suppose several identical items are on test, each obeying the same hazard function $h(t)$. A typical item has lifetime T, a continuous positive random variable with

$$\text{Prob}\,\{T > t_2 \mid T > t_1\} = \exp\left\{-\int_{t_1}^{t_2} h(t)dt\right\} . \quad (2.1)$$

We wish to infer h from the observed failure times $t_1 < t_2 < \ldots < t_J$; h is assumed to belong to some parametric family, which for the moment won't be indicated in the notation; and the nonparametric case is the limit when the family is allowed to include all hazard functions. Let

$$n(t) \equiv \text{number of items on test just before time } t \ . \quad (2.2)$$

In what follows, $n(t)$ is assumed to be a step function continuous from the left, changing value (due to losses, failures, and introduction of new items) only finitely often in any finite interval. The likelihood of the observed data, considered as a function of the unknown hazard rate h, is

$$f_h(\text{data}) = \exp\left\{-\int_0^\infty n(t)h(t)dt\right\} \prod_{j=1}^{J} n(t_j)h(t_j) \ . \quad (2.3)$$

This is derived from standard Poisson process arguments by noting that the probability of no event between t_{j-1} and t_j is

$$\exp\left\{-\int_{t_{j-1}}^{t_j} n(t)h(t)dt\right\} ,$$

while the probability of the single event, "one out of $n(t_j)$ items fails at time t_j," is proportional to $n(t_j)h(t_j)$. A more careful derivation is obtained by dividing the time axis into infinitesimal discrete units as in the introduction, see also Aalen (1975). Formula (2.3) assumes that $h(t)$ is continuous at the failure times t_j.

It can be shown that in the nonparametric case, the unrestricted maximizer of (2.3), say $h^*(t)$, satisfies

$$\exp\left\{-\int_{t_j^-}^{t_j^+} h^*(t)dt\right\} = 1 - \frac{1}{n(t_j)}$$
$$j = 1, 2, \ldots, J \ . \quad (2.4)$$

This leads to the familiar "Kaplan-Meier estimate" of the survival function (1958).

$$\text{Prob}^*\,\{T > t\} = \prod_{t_j \leq t} \left[1 - \frac{1}{n(t_j)}\right]. \quad (2.5)$$

There are some minor technical difficulties in deriving (2.4) from (2.3) because $h^*(t)$ does not refer to a continuous distribution for T. The discretization argument mentioned above avoids this difficulty.

3. COX'S PARTIAL LIKELIHOOD FUNCTION

We return to the situation where the different items on test have different hazard rates,

$$h_i(t) = \theta_i(t)h(t) \quad i = 1, 2, \ldots, n \ . \quad (3.1)$$

Here n is the number of items ever on test during the course of the experiment. The parameterization of the unknown functions θ_i and h introduced below is slightly different from (1.1); for the moment it will not be indicated in the notation.

The likelihood function of the observed data is now

$$f_{\theta,h}(\text{data}) = \exp\left\{-\int_0^\infty (\sum_{i\in\mathcal{R}(t)} \theta_i(t))h(t)dt\right\}$$
$$\cdot \prod_{j=1}^{J} \theta_{i_j}(t_j)h(t_j) \ , \quad (3.2)$$

where as before t_j is the jth ordered failure time, i_j the index of the failed item, and $\mathcal{R}(t)$ the risk set of items on test just before time t. This is derived in the same way as (2.3). Aalen (1975) gives a rigorous derivation. Equation (3.2) assumes that $h_{i_j}(t)$ is continuous at t_j and that the risk sets are continuous from the left and change only finitely often in any finite interval.

We will rewrite (3.2) to emphasize its relation to the partial likelihood (1.2) and the likelihood (2.3) for the identical items situation. Define

$$H(t) \equiv (\sum_{i=1}^{n} \theta_i(t)/n)h(t) \ , \quad (3.3)$$

the average hazard rate if all n items were on test at time t, and also

$$N(t) \equiv n\{\sum_{i\in\mathcal{R}(t)} \theta_i(t)/\sum_{i=1}^{n} \theta_i(t)\} \ . \quad (3.4)$$

If all the items are identical, i.e., if $\theta_i(t)$ doesn't depend on i, then $N(t) = n(t)$, the number at risk at time t. In general $N(t)/n$ is the proportion of the total possible hazard on test at time t. To put it another way, $N(t)$ identical items each with hazard rate $H(t)$ would have the same total hazard as the items actually in $\mathcal{R}(t)$.

The likelihood function (3.2) can now be written as

$$f_{\theta,h}(\text{data}) = \{\prod_{j=1}^{J} [\theta_{i_j}(t_j)/\sum_{\mathcal{R}(t)} \theta_i(t_j)]\}$$
$$\cdot\left\{\left[\exp - \int_0^\infty N(t)H(t)dt\right] \prod_{j=1}^{J} N(t_j)H(t_j)\right\} . \quad (3.5)$$

The first factor is the Cox likelihood, while the second factor is similar to (2.3).

The parameterization we will use assumes that the relative value of $\theta_i(t)$ and $\theta_{i'}(t)$, for any two indices i and i', is

$$\frac{\theta_i(t)}{\theta_{i'}(t)} = \frac{\exp\{\boldsymbol{\beta}\mathbf{z}_i(t)\}}{\exp\{\boldsymbol{\beta}\mathbf{z}_{i'}(t)\}} , \quad (3.6)$$

where $\boldsymbol{\beta}$ is a $1 \times B$ unknown parameter vector, and $\mathbf{z}_i(t)$ is a $B \times 1$ possibly time-varying vector of observed covariates. This parameterization makes (1.2) equal to

$$\prod_{j=1}^{J} [\exp\{\boldsymbol{\beta}\mathbf{z}_{i_j}(t_j)\}/ \sum_{i\in\mathcal{R}(t_j)} \exp\{\boldsymbol{\beta}\mathbf{z}_i(t_j)\}] , \quad (3.7)$$

as in Cox (1972); (3.4) becomes

$$N(t) \equiv N(t, \boldsymbol{\beta})$$
$$= n \sum_{\mathcal{R}(t)} \exp\{\boldsymbol{\beta}\mathbf{z}_i(t)\}/\sum_{1}^{n} \exp\{\boldsymbol{\beta}\mathbf{z}_i(t)\} . \quad (3.8)$$

Notice that (3.6) is weaker than the assumption $\theta_i(t) = \exp\{\boldsymbol{\beta}\mathbf{z}_i(t)\}$ mentioned in Section 1. We will work directly with (3.7) and (3.8), obviating the need to explicitly parameterize the functions $\theta_i(t)$.

The function $H(t)$ is assumed to be of the form

$$H(t, \boldsymbol{\gamma}) = \exp\{\boldsymbol{\gamma}\mathbf{w}(t)\} , \quad (3.9)$$

where $\boldsymbol{\gamma}$ is a $1 \times C$ unknown parameter vector functionally independent of $\boldsymbol{\beta}$, and $\mathbf{w}(t)$ is another time-varying $C \times 1$ vector of observed covariates. Substituting (3.7)–(3.9) into (3.5) gives the likelihood expression,

$$f_{\beta,\gamma}(\text{data}) = \left\{\prod_{j=1}^{J} \frac{\exp\{\boldsymbol{\beta}\mathbf{z}_{i_j}(t_j)\}}{\sum_{\mathcal{R}(t_j)} \exp\{\boldsymbol{\beta}\mathbf{z}_i(t_j)\}}\right\}$$
$$\cdot\left\{\exp\left\{-\int_0^\infty N(t, \boldsymbol{\beta})H(t, \boldsymbol{\gamma})dt\right\} \prod_{j=1}^{J} N(t_j, \boldsymbol{\beta})H(t_j, \boldsymbol{\gamma})\right\} . \quad (3.10)$$

(See Remark H, Section 6.)

Cox (1975) shows that the first factor can be treated as an ordinary likelihood function for the purpose of large-sample inference. In particular, the "maximum likelihood estimator" of $\boldsymbol{\beta}$ obtained by maximizing (3.7) will asymptotically have mean $\boldsymbol{\beta}$ and a covariance matrix which is the inverse of the "Fisher information matrix," the covariance matrix of the partial derivatives of the log of (3.7) with respect to the components of $\boldsymbol{\beta}$. The quotation marks used here serve as a reminder that (3.7) is not really a likelihood function. (For example it it *not* in general the likelihood of the reduced data set $(\mathcal{R}(t_1), i_1), (\mathcal{R}(t_2), i_2), \ldots, (\mathcal{R}(t_J), i_J)$.)

In what follows we will calculate the actual Fisher information matrix for $\boldsymbol{\beta}$ from (3.10) and give a heuristic demonstration that asymptotically it equals the information matrix based just on (3.7) assuming that the class of hazards $H(t, \boldsymbol{\gamma})$ is moderately large. This equality shows that the maximum likelihood estimate of $\boldsymbol{\beta}$ based on (3.7) must be asymptotically equivalent to that based on all the data. Similar statements hold true for asymptotic testing and confidence procedures (see Aalen 1975).

For convenience we consider only the case where β and, therefore, $z_i(t)$ is a scaler rather than a vector. The vector case is discussed briefly in Remark A, Section 6. Define

$$E_\beta\{z \mid \mathcal{R}(t)\} \equiv \sum_{i\in\mathcal{R}(t)} z_i(t) \exp\{\beta z_i(t)\}/ \sum_{i\in\mathcal{R}(t)} \exp\{\beta z_i(t)\} ,$$
$$(3.11)$$
$$E_\beta z \equiv \sum_{i=1}^{n} z_i(t) \exp\{\beta z_i(t)\}/\sum_{i=1}^{n} \exp\{\beta z_i(t)\} ,$$

and

$$\text{var}_\beta\{z \mid \mathcal{R}(t)\} = \sum_{i\in\mathcal{R}(t)} [z_i(t) - E_\beta\{z \mid \mathcal{R}(t)\}]^2$$
$$\cdot \exp\{\beta z_i(t)\}/ \sum_{i\in\mathcal{R}(t)} \exp\{\beta z_i(t)\} . \quad (3.12)$$

$E_\beta\{z \mid \mathcal{R}(t)\}$ and $\text{var}_\beta\{z \mid \mathcal{R}(t)\}$ are the conditional mean and variance of $z_i(t)$ with respect to a probability distribution proportional to $\exp\{\beta z_i(t)\}$ on $i \in \mathcal{R}(t)$. They are functions of β and the random variable $\mathcal{R}(t)$. The following lemma computes the Fisher information in (3.10) for estimating β, i.e., one over the Cramér-Rao lower bound for unbiased estimation.

Lemma: The Fisher information for estimating β in (3.10) is

$$\inf_{\mathbf{g}} \int_0^\infty \mathcal{E}(\{\text{var}_\beta\{z \mid \mathcal{R}(t)\} + [(E_\beta\{z \mid \mathcal{R}(t)\}$$
$$- E_\beta z) - \mathbf{g}\mathbf{w}(t)]^2\} N(t, \beta)H(t, \boldsymbol{\gamma}))dt , \quad (3.13)$$

where the infimum is over all choices of the C dimensional vector $\mathbf{g}$, and $\mathcal{E}$ indicates expectation over the randomness in the risk sets $\mathcal{R}(t)$. The same expression without the term in square brackets is the Fisher information for β based just on Cox's partial likelihood (3.7). (The proof is given in Section 7.)

Recall that if A and B are any two random variables, B is nonnegative, and a is any constant, then

$$E(A - a)^2 B = [\text{var}_B A + (a - E_B A)^2] EB ,$$

where (3.14)

$$E_B A \equiv EAB/EB \text{ and } \text{var}_B A \equiv E(A - E_B A)^2 B/EB .$$

Let $\eta(t, \beta)$ be the expectation, over the randomness in $\mathcal{R}(t)$, of $N(t, \beta)$,

$$\eta(t, \beta) \equiv \mathcal{E}N(t, \beta)$$
$$= n(\mathcal{E} \sum_{\mathcal{R}(t)} \exp\{\beta z_i(t)\}/\sum_{i=1}^{n} \exp\{\beta z_i(t)\}) \ ; \quad (3.15)$$

define

$$B \equiv N(t, \beta) \ , \quad A \equiv E_\beta\{z \mid \mathcal{R}(t)\} - E_\beta z \ ,$$

and $a \equiv \mathbf{gw}(t)$. Using (3.14), the integrand of (3.13) can be expressed as

$$\{\mathcal{E}[N(t, \beta)/\eta(t, \beta)] \operatorname{var}_\beta \{z \mid \mathcal{R}(t)\} + \operatorname{var}_N E_\beta\{z \mid \mathcal{R}(t)\}$$
$$+ [e_\beta(t) - \mathbf{gw}(t)]^2\}\eta(t, \beta)H(t, \gamma) \ , \quad (3.16)$$

where

$$e_\beta(t) \equiv \mathcal{E}_N(E_\beta\{z \mid \mathcal{R}(t)\} - E_\beta z) \ , \quad (3.17)$$

and $\operatorname{var}_N E_\beta\{z \mid \mathcal{R}(t)\}$ indicates a weighted variance, as in (3.14) with the random quantity being $\mathcal{R}(t)$.

A simple calculation shows that if $P_i(t)$ is the probability that item i is in $\mathcal{R}(t)$, then

$$e_\beta(t) = \sum_1^n P_i(t) z_i(t) \exp\{\beta z_i(t)\}/\sum_1^n P_i(t) \exp\{\beta z_i(t)\}$$
$$- \sum_1^n z_i(t) \exp\{\beta z_i(t)\}/\sum_1^n \exp\{\beta z_i(t)\} \ . \quad (3.18)$$

Notice that $P_i(t)$ is also a function of β and γ and possibly other extraneous random factors.

The principle implied by the lemma and (3.16), admittedly in a rough manner, is the following: *If, as the number items tested goes to infinity, the function $e_\beta(t)$ can be approximated arbitrarily well by a linear combination of the functions $w_1(t)$, $w_2(t)$, $\ldots$, $w_C(t)$, then the Cox likelihood is asymptotically fully efficient for the estimation of β.* In other words, the Fisher information for β based on the Cox likelihood has asymptotic ratio unity with that based on all the data. Section 4 illustrates this principle in a particularly simple special case.

Suppose for a moment that $e_\beta(t) = \mathbf{gw}(t)$ for all t for some choice of $\mathbf{g}$. This eliminates the last term in square brackets from (3.16). The additional information for estimating β *not* in the Cox likelihood corresponds to the term $\operatorname{var}_N E_\beta\{z \mid \mathcal{R}(t)\}$. Intuitively this comes from local variations in $N(t, \beta)$ due to random fluctuations in the risk sets, which influence the observed times between failures. These random fluctuations can not be explained away by any possible choice of $H(t, \gamma)$ since this is necessarily a fixed (nonrandom) function of time. However, the magnitude of this term tends to be $0(1/\eta)$ compared to the term $(N/\eta) \operatorname{var}_\beta \{z \mid \mathcal{R}(t)\}$ from the partial likelihood, essentially because $E_\beta\{z \mid \mathcal{R}(t)\}$ is the average of about η random quantities. (See Remark I, Section 6.)

For asymptotic efficiency we don't need $e_\beta(t)$ to actually be in the linear space generated by $w_1, w_2, \ldots, w_C$,

$$\mathcal{L}(\mathrm{w}) \equiv \{\sum_{c=1}^{C} g_c w_c(t)\} \ , \quad (3.19)$$

but only that it be increasingly well approximated by some function in $\mathcal{L}(\mathrm{w})$ as the number of tested items grows large. In other words, we need to be able to ignore the term $[e_B(t) - \mathbf{gw}(t)]^2$ in (3.16).

In order for the partial likelihood to estimate β with reasonable efficiency in finite samples it is necessary for $e_\beta(t)$ to be in or at least near $\mathcal{L}(\mathrm{w})$. Is this a realistic assumption? In many situations the answer is yes. For example, if the z_i are not functions of time, and if there is no censoring, then (3.18) shows that $e_\beta(t)$ is monotonic. For $\beta > 0$, $e_\beta(t)$ will decrease monotonically in time as those items with large values of z_i are selectively removed by earlier failure. Censoring can distort $e_\beta(t)$ but not seriously unless a large proportion of the items have the same fixed censoring time. (See Section 4.) In the absence of firm prior knowledge it may be reasonable to assume that $H(t, \gamma) = \exp\{\gamma \mathrm{w}(t)\}$ can be any smooth monotonic function, which in this case guarantees the asymptotic efficiency of the partial likelihood.

Of course there are situations in which the partial likelihood by itself produces seriously inefficient inferences. For example $\mathcal{L}(\mathrm{w})$ might be known to be the class of linear functions $w_1 + w_2 t$ while $e_\beta(t)$ is some considerably more complicated function. In theory at least, the statistician can always calculate the actual maximum likelihood estimator (MLE) of β from (3.10) in such cases. Kalbfleisch (1974) gives an efficiency calculation in one such case, which reinforces faith in using (3.7) by itself, as do the calculations of Section 4.

4. THE TWO-SAMPLE PROBLEM

The general calculations of Section 3 are more understandable in special cases, the most special of which we consider now: the two-sample problem with exponentially distributed lifetimes. Let

$$e^\beta \equiv \alpha \quad (4.1)$$

be the ratio of expectations for the two samples, and let β be the parameter to be estimated. The two samples are of sizes, say, n_0 and n_1, respectively, $n_0 + n_1 = n$, with sample membership being indicated by the dummy variable

$$z_i = 0 \quad \text{if item } i \text{ is in sample } 0, i = 1, 2, \ldots, n \ ;$$
$$= 1 \text{ if item } i \text{ is in sample } 1, i = 1, 2, \ldots, n \ . \quad (4.2)$$

Also let

$$q \equiv n_0/n \ , \quad p \equiv n_1/n \ , \quad \text{and} \quad D_\alpha \equiv q + p\alpha \ . \quad (4.3)$$

To parameterize this situation as in Section 3 the hazard rates (3.1) are written as

$$h_i(t) = \alpha^{z_i} e^\gamma / D_\alpha \ , \quad i = 1, 2, \ldots, n \ . \quad (4.4)$$

This makes $H(t, \gamma)$ defined at (3.3) equal e^γ, and, as will be apparent, there is no loss of generality in assuming $\gamma = 0$, $H(t, \gamma) \equiv 1$. We see that the probability of item

i's lifetime exceeding t equals

$$P_i(t) = \exp\{-t\alpha^{z_i}/D_\alpha\} \ , \tag{4.5}$$

assuming that there is no censoring.

In the absence of censoring, (3.13) and (3.16) give a simple expression for the asymptotic variance of β^*, which is the MLE based on Cox's partial likelihood function (3.7). We will consider the effects of censoring later. Let

$$n_\ell(t) \equiv \text{number of sample } \ell \text{ members in } \mathcal{R}(t),\ \ell = 0, 1 \ ,$$

and

$$n(t) \equiv n_0(t) + n_1(t) \ , \quad q_t \equiv n_0(t)/n(t) \ , \quad p_t \equiv n_1(t)/n(t) \ . \tag{4.6}$$

Then

$$N(t, \beta) = n(t)(q_t + p_t\alpha)/D_\alpha \ , \tag{4.7}$$
$$\text{var}_\beta \{z \,|\, \mathcal{R}(t)\} = q_t p_t \alpha/(q_t + p_t\alpha)^2 \ ,$$

by substitution in (3.8), (3.12). As n gets large, the random quantities $n(t)/n$, $q(t)$, and $p(t)$ approach constants easily determined from (4.5) giving

$$\lim_{n\to\infty} \frac{1}{n} \mathcal{E}[\text{var}_\beta \{z \,|\, \mathcal{R}(t)\}]N(t, \beta)H(t, \gamma) = \frac{pq\alpha \exp\{-t\alpha/D_\alpha\}}{D_\alpha[q + p\alpha \exp\{-t(\alpha - 1)/D_\alpha\}]} . \tag{4.8}$$

The lemma gives the expression

$$\lim_{n\to\infty} \frac{1}{n \text{ var } \beta^*} = \int_0^1 \frac{pqdu}{q + p\alpha u^{(\alpha-1)/\alpha}} \tag{4.9}$$

for the limiting variance of β^*, where the substitution $u = \exp\{-t\alpha/D_\alpha\}$ has been made in (3.13), (3.16). The limiting variance of β^{**}, the MLE based on all the data, is

$$\lim_{n\to\infty} n \text{ var } \beta^{**} = 1/pq \tag{4.10}$$

under (4.4), as shown by standard Fisher information arguments. The asymptotic relative efficiency (ARE) of the partial likelihood estimate compared to full maximum likelihood is

$$\text{ARE} \equiv \lim_{n\to\infty} \frac{\text{var } \beta^{**}}{\text{var } \beta^*} = \int_0^1 \frac{du}{q + p\alpha u^{(\alpha-1)/\alpha}} . \tag{4.11}$$

The first line of the table tabulates (4.11) for various choices of α with $p = q = \frac{1}{2}$.

The Asymptotic Relative Efficiency of the Partial Likelihood Estimate of β Compared to the Full Maximum Likelihood Estimate, p = q = ½.

$\alpha \equiv e^\beta$	*1*	*2*	*4*	*8*	*16*
ARE from (4.11)	1.000	.901	.705	.502	.334
ARE under model (4.15)–(4.17)					
No censoring	1.000	.982	.959	.914	.819
Censoring pattern 1	.991	.978	.950	.912	.819
Censoring pattern 2	.994	.987	.967	.915	.816

NOTE: The ARE increases under the larger model (4.15)–(4.17) for $H(t,\gamma)$. Censoring has little effect on these calculations. Pattern 1 has half of sample 1 censored at the median of the distribution for sample 0. Pattern 2 has one quarter of sample 1 censored at each quartile of the distribution for sample 0.

Any inefficiency of β^* compared to β^{**} comes from the last term in (3.16), $[e_B(t) - \mathbf{g}\mathbf{w}(t)]^2$. By assuming $H(t, \gamma)$ constant we have restricted $\mathcal{L}(\mathbf{w})$, (3.19), to involve only constant functions. The inefficiency, 1 − ARE, equals

$$(pq)^{-1} \min_g \int_0^\infty [e_\beta(t) - g]^2 \frac{\eta(t, \beta)}{n} dt \tag{4.12}$$

(the factor $(pq)^{-1}$ coming from (4.10)), where

$$\frac{\eta(t, \beta)}{n} = \frac{q \exp\{-t/D_\alpha\} + p\alpha \exp\{-t\alpha/D_\alpha\}}{D_\alpha} \tag{4.13}$$

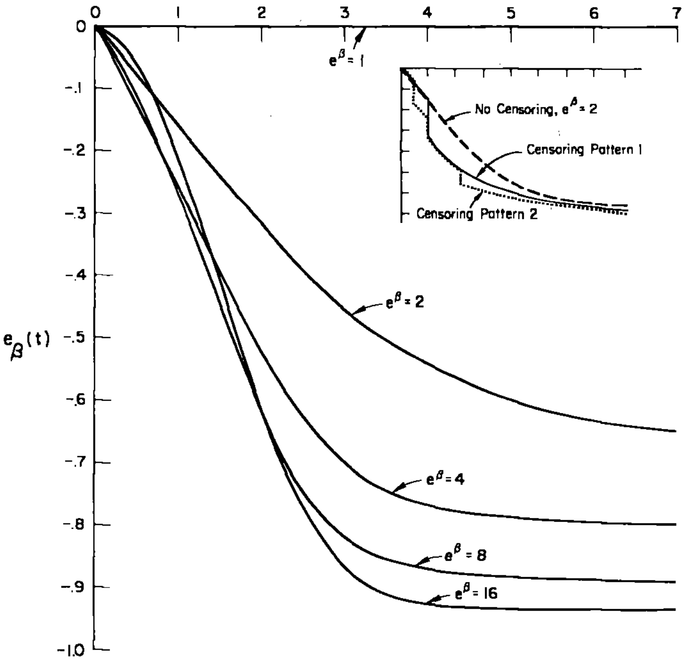
by (3.15) and (4.5).

It is easy to evaluate $e_\beta(t)$ from (3.18),

$$e_\beta(t) = \frac{p\alpha}{D_\alpha}\left[\frac{D_\alpha}{q \exp\{t(\alpha - 1)/D_\alpha\} + p\alpha} - 1\right] . \tag{4.14}$$

The figure graphs $e_\beta(t)$ for $\alpha = e^\beta = 1, 2, 4, 8,$ and 16 in the case $p = q = .5$ showing a smooth monotonic decrease to the asymptote $-p\alpha/D_\alpha$ (for $\alpha > 1$). Notice that in the absence of censoring, $e_1(t) \in \mathcal{L}(\mathbf{w})$ which explains the full asymptotic efficiency in this case.

The Function $e_\beta(t)$, (4.14) for $\alpha = e^\beta = 1,2,4,8,16$ $(p = q = .5)$[a]



[a] The insert shows $e_\beta(t)$, $e^\beta = 2$, for the two censoring patterns mentioned in the text.

Suppose now that we are not willing to assume an exponential model for the lifetimes, but are willing to assume that the relative hazard rate between the two groups is constant, $h_1(t)/h_0(t) = e^\beta$. This type of "Lehmann alternative" is more in the spirit of Cox's article. As a first step, (4.4) can be expanded to

$$h_i(t) = \frac{\alpha^{z_i} \exp\{\gamma_0 + \gamma_1 w_1(t)\}}{D_\alpha} \quad i = 1, 2, \ldots, n \ , \tag{4.15}$$

making

$$H(t, \boldsymbol{\gamma}) = \exp\{\gamma_0 + \gamma_1 w_1(t)\} \ ,$$
$$\mathcal{L}(\mathbf{w}) = \{g_0 + g_1 w_1(t)\} \ . \tag{4.16}$$

Here $w_1(t)$ is some specified function we are willing to use to expand the class of possible hazard rates. For example,

$$w_1(t) = e^{-t} \tag{4.17}$$

allows $h_i(t)$ to vary monotonically by a finite factor as t goes from zero to infinity.

The partial likelihood (3.7) is unaffected by changes in $H(t, \boldsymbol{\gamma})$, so (4.9) remains valid when sampling under $\boldsymbol{\gamma} = (0, 0)$. The full MLE β^{**} now has greater limiting variance, so the ARE of β^* to β^{**} is larger, as shown in the fourth line of the table. The asymptotic inefficiency of β^* relative to β^{**} depends on the magnitude of

$$\min_{g_0, g_1} \int_0^\infty [e_\beta(t) - g_0 - g_1 w_1(t)]^2 \frac{\eta(t, \beta)}{n} dt \ . \tag{4.18}$$

The choice $w_1(t) = e^{-t}$ gives high ARE in this case because it closely matches the shape of $e_\beta(t)$, at least for $e^\beta \leq 8$.

It is easiest to interpret the table in terms of the asymptotic variance of the MLE based on all the data, relative to that of the MLE based just on the partial likelihood (3.7). The numbers also have a testing interpretation as Pitman efficiencies. For example, under Censoring pattern 1, the locally most powerful test of $\alpha = 1$ vs $\alpha > 1$ based on (3.7) has Pitman efficiency .991 compared to that based on all the data. (The test based on (3.7) is a generalization of the Savage rank test, described in Cox (1972) and also in Thomas (1971).) A more general interpretation is that using (3.7) rather than the full likelihood asymptotically wastes nine out of 1,000 observations in this particular situation, for any inferential purpose at all.

Of course there is no real reason behind the choice (4.17). In most practical problems there isn't any obvious choice, beyond perhaps a qualitative preference for monotonic reasonably smooth hazard rates. The functions $e_\beta(t)$ in the figure fit this description. If we take $w_1(t) = e\beta(t)$, the ARE of β^* to β^{**} is one. It is the author's opinion that Cox's method will usually give high efficiency under any reasonably realistic assumptions on the class of possible hazard rates.

Censoring seems to have little effect on the efficiency calculations. The insert to the figure shows $e_\beta(t)$, $e^\beta = 2$, for two censoring patterns: (1) sample 0 uncensored, 50 percent of sample 1 censored at the median of the distribution for sample 0; (2) sample 0 uncensored, 25 percent of sample 1 censored at each quartile of the distribution for sample 0. The discontinuities in $e_\beta(t)$ come from the $P_i(t)$ in (3.18) going suddenly to zero as the fixed censoring times are encountered. Nevertheless, the ARE stays almost constant as the last two lines of the table show.

All of these calculations are asymptotic in nature. In finite samples there is a further loss of efficiency for β^* compared to β^{**} coming from the term $\operatorname{var}_N E_\beta\{z \mid \mathcal{R}(t)\}$ in (3.16). The calculations in Kalbfleisch (1974), in particular his Table 1 and equation (15), suggest an additional efficiency loss of about 10 percent for $n = 10$, 6 percent for $n = 20$, and 5 percent for $n = 40$.

5. ESTIMATING THE HAZARD RATES

Suppose we are willing to rely on the first factor in (3.10), the Cox likelihood, for the estimation of $\boldsymbol{\beta}$. We can treat the estimate obtained in this way, say $\boldsymbol{\beta}^*$, as if it were the true value of $\boldsymbol{\beta}$ and then maximize the second factor in (3.10) to estimate $\boldsymbol{\gamma}$.

Let $\boldsymbol{\gamma}^*$ be the "maximum likelihood" estimator of $\boldsymbol{\gamma}$ obtained in this way, the quotes indicating that $\boldsymbol{\gamma}^*$ is really only the conditional maximizer given the value $\boldsymbol{\beta}^*$ obtained from the Cox likelihood. From (3.1), (3.3), and (3.6), we get

$$h_i(t) = n[\theta_i(t) / \sum_{i'=1}^{n} \theta_{i'}(t)] H(t)$$
$$= n[\exp\{\boldsymbol{\beta}\mathbf{z}_i(t)\} / \sum_{i'=1}^{n} \exp\{\boldsymbol{\beta}\mathbf{z}_{i'}(t)\}] H(t, \boldsymbol{\gamma}) \ ; \tag{5.1}$$

therefore, the corresponding estimate of the hazard rate for item i is

$$h_i^*(t)$$
$$= n[\exp\{\boldsymbol{\beta}^*\mathbf{z}_i(t)\} / \sum_{i'=1}^{n} \exp\{\boldsymbol{\beta}^*\mathbf{z}_{i'}(t)\}] H(t, \boldsymbol{\gamma}^*) \ . \tag{5.2}$$

In the Kaplan-Meier nonparametric situation, $H(t, \boldsymbol{\gamma}^*)$ approaches $H^*(t)$, a sum of delta functions at $t_1, t_2, \ldots, t_J$ satisfying

$$\exp\left\{-\int_{t_j^-}^{t_j^+} H^*(t) dt\right\} = 1 - \frac{1}{N(t_j, \boldsymbol{\beta}^*)} \cdot \tag{5.3}$$

Assuming that the functions $\mathbf{z}_i(t)$, $i \in \mathcal{R}(t_j)$ are continuous at t_j, this gives

$$\exp\left\{-\int_{t_j^-}^{t_j^+} h_i^*(t) dt\right\} = \left[1 - \frac{1}{N(t_j, \boldsymbol{\beta}^*)}\right]^{\phi_{ij}^*} , \tag{5.4}$$

where

$$\phi_{ij}^* \equiv n \exp\{\boldsymbol{\beta}^*\mathbf{z}_i(t_j)\} / \sum_{1}^{n} \exp\{\boldsymbol{\beta}^* z_{i'}(t_j)\} \ .$$

The estimate of the ith cdf is

$$F_i^*(t) = \prod_{t_j \leq t} \left[1 - \frac{1}{N(t_j, \boldsymbol{\beta}^*)}\right]^{\phi_{ij}^*}$$
$$\approx \exp - [\sum_{t_j \leq t} (\exp\{\boldsymbol{\beta}^*\mathbf{z}_i(t_j)\} /$$
$$\sum_{i \in \mathcal{R}(t)} \exp\{\boldsymbol{\beta}^*\mathbf{z}_{i'}(t_j)\})] \ ; \tag{5.5}$$

this last form is essentially the same as that derived in Breslow (1974) and also in Kalbfleisch and Prentice (1973) for the case which is not time-dependent. (See Remark c, Section 6 of this article.)

6. SOME REMARKS

A. The information calculations of Section 3 carry over directly to the case where $\boldsymbol{\beta}$ is a vector. The expression for the information matrix for estimating $\boldsymbol{\beta}$ is the multivariate analog of (3.13),

$$\inf_{\mathbf{G}} \int_0^\infty \mathcal{E}(\{\mathrm{cov}_\beta\{\mathbf{z} \mid \mathcal{R}(t)\} + [(E_\beta\{\mathbf{z} \mid \mathcal{R}(t)\} - E_\beta \mathbf{z}) - \mathbf{G}\mathbf{w}(t)][(E_\beta\{\mathbf{z} \mid \mathcal{R}(t)\} - E_\beta \mathbf{z}) - \mathbf{G}\mathbf{w}(t)]'\} \cdot N(t, \boldsymbol{\beta}) H(t, \boldsymbol{\gamma}))dt \ , \quad (6.1)$$

with the infimum being taken over all $B \times C$ matrices $\mathbf{G}$.

B. There is no particular advantage to the exponential forms $\exp\{\boldsymbol{\beta}\mathbf{z}_i(t)\}$, $\exp\{\boldsymbol{\gamma}\mathbf{w}(t)\}$ used in Section 3. Any other simple positive function serves just as well and may be more natural in some situations. Suppose, e.g., that the event $T < 1$ is hypothesized to follow a linear logistic law in terms of $\boldsymbol{\beta}$ and the (non-time-varying) covariate $\mathbf{z}_i$,

$$\mathrm{Prob}\,\{T_i < 1\} = \exp\{\boldsymbol{\beta}\mathbf{z}_i\}/1 + \exp\{\boldsymbol{\beta}\mathbf{z}_i\} \ . \quad (6.2)$$

This implies

$$\theta_1(\boldsymbol{\beta}) \propto \log\left[1 + \exp\{\boldsymbol{\beta}\mathbf{z}_i\}\right] \quad (6.3)$$

rather than $\theta_i(\boldsymbol{\beta}) \propto \exp\{\boldsymbol{\beta}\mathbf{z}_i\}$.

C. If m is a large positive number then

$$\log(1 - 1/m) = -1/[m - c(m)] \ , \quad (6.4)$$

where $c(m) = \frac{1}{2} - 1/12m + \ldots$. Expression (2.5) for the Kaplan-Meier estimator can be written as

$$\mathrm{Prob}^*\{T > t\} = \exp\{-\sum_{t_j \leq t} 1/[n(t_j) - c(n(t_j))]\} \ . \quad (6.5)$$

Ignoring the correction term $c(n(t_j))$ leads to the last expression in (5.5).

D. The Kaplan-Meier estimator corresponds to the limit of continuous hazard functions putting mass $1/[n(t_j) - c(n(t_j))]$ at t_j, *not* mass $1/n(t_j)$

$$(\text{since } \exp - \{\text{mass at } t_j\} = 1 - 1/n(t_j)) \ .$$

E. The likelihood expressions (3.2), (3.5), and (3.10) assume that the risk sets $\mathcal{R}(t)$ are themselves uninformative for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. It is allowable for $\mathcal{R}(t)$ to depend on all data observed before time t, plus random elements whose distributions don't depend on $\boldsymbol{\beta}$ or $\boldsymbol{\gamma}$. Subject to these restrictions, a malevolent censorer trying to confuse the statistician cannot affect the likelihood function or any Bayesian/likelihood based inferences, though he can affect expectations connected with the likelihood such as the Fisher information.

Kalbfleisch and Prentice (1973) tacitly make a stronger assumption about the censoring mechanism; it in no way depends on the real time axis except through the ordering of the observed events. Otherwise, their marginal likelihood interpretation of Cox's likelihood can easily be contradicted. Take $n = 3$, and suppose that z_1, z_2, z_3 do not depend on time, so that θ_1, θ_2, θ_3 are time independent. Suppose also that no observations are censored if $\min\{T_1, T_2, T_3\} \leq 1.5$, but if the first observation is T_1 and it exceeds 1.5, then further observation on T_2 is immediately censored. An easy calculation gives the probability of observing the partial ordering "T_1 less than $\min\{T_2, T_3\}$" to be

$$\mathrm{Prob}\,\{(1, 2, 3,) \cup (1, 3, 2)\} = \exp\left\{-(\theta_1 + \theta_2 + \theta_3)\int_0^{1.5} h(t)dt\right\} \cdot \theta_1/(\theta_1 + \theta_2 + \theta_3) \ , \quad (6.6)$$

which does not equal the Cox likelihood $\theta_1/(\theta_1 + \theta_2 + \theta_3)$.

F. Another hidden assumption in (3.2) is that once an item leaves the experiment due to censoring it does not return on test at a later time. Suppose an item did drop out at time $t = a$ and returned at $t = b$; then either it will be known to have failed during that interval, multiplying the likelihood function by the ungainly factor $1 - \exp\{-\int_a^b \theta_i(t)h(t)dt\}$, or it will be seen not to have failed during that interval, in which case it really was observed. This point does not arise in the Kaplan-Meier situation of Section 2 unless we add labels to the identical test items in order to make them identifiable.

The two types of allowable changes in the risk sets, aside from failure, are illustrated in the senior citizen study. These are caused by items entering the study late, without any information on those failing before entry (*left truncation*), and items leaving the study before failure (*right censoring*).

G. Real censored data problems are often *discrete*; items are reported to fail during intervals, not by exact times. (In the senior citizen study, e.g., deaths and changes in the risk sets were reported by day but not by minute and second.) Let us add the assumption that the ratio of hazards (3.6) is constant during any one such reporting interval, and that no changes in $\mathcal{R}(t)$ occur within such an interval except those due to failure. Then given the information that the m items i_{j1}, i_{j2}, $\ldots$, i_{jm} failed during the jth reporting interval, we know that the Cox likelihood for the (unobservable) continuous data takes on one of $m!$ possible values, corresponding to the $m!$ possible orderings of i_{j1}, i_{j2}, $\ldots$, i_{jm}, each with equal probability. It is notationally messy to average these $m!$ quantities, but an obvious approximation for the jth factor in the Cox likelihood is

$$\frac{\theta_{i_{j1}}(t_j)\theta_{i_{j2}}(t_j)\ldots\theta_{i_{jm}}(t_j)}{\prod_{\ell=0}^{m-1}\left[\sum_{i \in \mathcal{R}(t_j)} \theta_i(t_j) - \frac{\ell}{m}\sum_{h=1}^{m} \theta_{i_{jh}}(t_j)\right]} \ . \quad (6.7)$$

This is a slightly more accurate approximation than those suggested in the discussion following Cox's 1972 article, but as Peto suggests there, it probably doesn't make much difference.

H. The parameterization, (3.6)-(3.9), which leads to the likelihood expression (3.10) assumes that the relative hazard rates for the different items in the experiment do not functionally determine the total hazard rate. More precisely, the information calculations at, say, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$ require that the possible $\boldsymbol{\gamma}$ vectors corresponding to $\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$ include an open set around $\boldsymbol{\gamma}^{(0)}$.

An alternative parameterization which seems appealing is to let $\bar{h}(t, \boldsymbol{\gamma}) \equiv (\sum_{\mathcal{R}(t)} \theta_i(t, \boldsymbol{\beta})/n(t))h(t)$ be the average hazard rate of those items on test at time t, where $n(t)$ is the number of items in $\mathcal{R}(t)$, and to assume $\theta_i(t, \boldsymbol{\beta}) = \exp\{\boldsymbol{\beta}\mathbf{z}_i(t)\}$, $h(t, \boldsymbol{\gamma}) = \exp\{\boldsymbol{\gamma}\mathbf{w}(t)\}$. This makes the second factor in (5.10) equal to

$$\exp\left\{-\int_0^\infty n(t)\bar{h}(t, \boldsymbol{\gamma})dt\right\} \prod_{j=1}^{J} n(t_j)\bar{h}(t_j, \boldsymbol{\gamma}) \ , \qquad (6.8)$$

which is much simpler since it does not involve $\boldsymbol{\beta}$ at all. However, this parameterization is untenable. The function $\bar{h}$ must depend on $\boldsymbol{\beta}$ in some way, since if $\boldsymbol{\beta}$ is not zero, $\bar{h}$ changes value discontinuously whenever the risk set changes. This is impossible for any function of the form $\exp\{\boldsymbol{\gamma}\mathbf{w}(t)\}$, except in very restricted situations.

I. Suppose that all the items act independently of each other in terms of failures and censoring. Then standard expansion methods would show that the quantity $\text{var}_N\, E_\beta\{z \mid \mathcal{R}(t)\}$, which figures in (3.16), approximately equals

$$(1/\eta)(\sum_{i=1}^{n} P_i Q_i \phi_i^2 (z_i - R)^2/\eta) \ , \qquad (6.9)$$

where t, β, and γ have been dropped from the notation, $Q_i \equiv 1 - P_i$, and

$$\phi_i \equiv n \exp\{\beta z_i\} / \sum_{i'=1}^{n} \exp\{\beta z_{i'}\} \ ,$$

$$R \equiv \sum_{i=1}^{n} P_i z_i \exp\{\beta z_i\} / \sum_{i=1}^{n} P_i \exp\{\beta z_i\} \ . \qquad (6.10)$$

Assuming the $z_i(t)$ are bounded, (6.9) is $0(1/\eta)$ as η goes to infinity.

7. PROOF OF THE LEMMA

To prove the lemma of Section 3, we calculate the score functions for β and $\gamma_1, \gamma_2, \ldots, \gamma_C$ from (3.8)–(3.10),

$$S_\beta \equiv \frac{\partial \log f_{\beta,\gamma}}{\partial \beta} = \sum_{j=1}^{J} [z_{i_j}(t_j) - E_\beta\{z \mid \mathcal{R}(t)\}] + \int_0^\infty [E_\beta\{z \mid \mathcal{R}(t)\} - E_\beta z] \cdot [\mathbf{J}(t) - N(t, \beta)H(t, \gamma)]dt \ , \qquad (7.1)$$

and

$$S_{\gamma_c} \equiv \frac{\partial \log f_{\beta,\gamma}}{\partial \gamma_c} = \int_0^\infty w_c(t)[\mathbf{J}(t) - N(t, \beta)H(t, \gamma)]dt \ ,$$

where $\mathbf{J}(t) = \sum_{j=1}^{J} \delta(t - t_j)$, which is the sum of delta functions at $t_1, t_2, \ldots, t_\delta$. For an arbitrary choice of $\mathbf{g} = (g_{,1}\, g_2, \ldots, g_C)$ we can write

$$S_\beta - \sum_{c=1}^{C} g_c S_{\gamma_c} = \int_0^\infty [U(t) + (D_\beta(t) - \mathbf{gw}(t))]dK(t) \ , \qquad (7.2)$$

where

$$U(t) = z_{i_j}(t_j) - E_\beta\{z \mid \mathcal{R}(t_j)\} \ , \ \text{if} \quad t = t_j$$
$$= \quad 0 \ , \ \text{if} \quad t \notin \{t_1, t_2, \ldots, t_j\} \ ,$$

$$D_\beta(t) = E_\beta\{z \mid \mathcal{R}(t)\} - E_\beta z \ , \qquad (7.3)$$

and

$$dK(t) = [\mathbf{J}(t) - N(t, \beta)H(t, \gamma)]dt \ .$$

Given the observed value of $\mathcal{R}(t)$, $D_\beta(t)$ is a fixed number while $U(t)$ is a random variable with mean zero and variance

$$\text{var}\{U(t) \mid \mathcal{R}(t)\} = \text{var}_\beta\{z \mid \mathcal{R}(t)\} \ , \ \text{if} \quad t = t_j$$
$$= 0 \ , \quad \text{if} \quad t \notin \{t_1, t_2, \ldots, t_j\} \ , \qquad (7.4)$$

with $\text{var}_\beta\{z \mid \mathcal{R}(t)\}$ as defined in (3.12). Also, still assuming $\mathcal{R}(t)$ given,

$$dK(t) = 1 - N(t, \beta)H(t, \gamma)dt \quad \text{with Prob } N(t, \beta)H(t, \gamma)dt$$
$$= -N(t, \beta)H(t, \gamma)dt \quad \text{with Prob } 1 - N(t, \beta)H(t, \gamma)dt \ . \qquad (7.5)$$

Notice that the two cases for $dK(t)$ correspond to the two cases for $U(t)$ given in (7.3). Expressions (7.4)–(7.5) are easier to understand in the coin-tossing formulation of Section 1.

Putting (7.3)–(7.5) together gives

$$E\{([U(t) + (D_\beta(t) - \mathbf{gw}(t))]dK(t))^2 \mid \mathcal{R}(t)\} = \{\text{var}_\beta\{z \mid \mathcal{R}(t)\} + [(E_\beta\{z \mid \mathcal{R}(t)\} - E_\beta z) - \mathbf{gw}(t)]^2\} N(t, \beta)H(t, \gamma)dt \ , \qquad (7.6)$$

and, for $t' < t$,

$$E\{([U(t') + (D_\beta(t') - \mathbf{gw}(t'))]dK(t')) \cdot ([U(t) + (D_\beta(t) - \mathbf{gw}(t))]dK(t)) \mid \mathcal{R}(s) \ , \ 0 \leq s \leq t\} = 0 \ . \qquad (7.7)$$

(In deriving (7.7) we have used $E\{dK(t) \mid \mathcal{R}(t)\} = E\{U(t)dK(t) \mid \mathcal{R}(t)\} = 0$.) Therefore, writing the integral (7.2) as a Reimann sum and conditioning successively on $\mathcal{R}(t)$ as t increases from 0 to ∞ gives the expected value of $(S_\beta - \sum_{c=1}^{C} g_c S_{\gamma_c})^2$ to be the integral in (3.13). But the reciprocal of the Cramér-Rao lower bound for β, by definition the Fisher information for β, is the infimum of the expected value of $(S_\beta - \sum_{c=1}^{C} g_c S_{\gamma_c})^2$ over all choices of $\mathbf{g}$. This proves the first part of the lemma. The second part follows by a similar argument which is made easier by the fact that (3.7) does not involve $\boldsymbol{\gamma}$.

[*Received May 1976. Revised November 1976.*]

REFERENCES

Aalen, Odd O. (1975), "Statistical Inference for a Family of Counting Processes," Ph.D. dissertation, Department of Statistics, University of California, Berkeley.

Breslow, N. (1974), "Covariance Analysis of Censored Data," *Biometrics*, 30, 89–99.

Breslow, N., and Crowley, J. (1974), "A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship," *Annals of Statistics*, 2, 437–53.

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, 269–79.

Kalbfleisch, J., and Prentice, R. (1973), "Marginal Likelihoods Based on Cox's Regression and Life Model," *Biometrika*, 60, 267–79.

———, (1974), "Some Efficiency Calculations for Survival Distributions," *Biometrika*, 61, 31–8.

Kaplan, E.L., and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–81.

Mantel, N., and Haenzel, W. (1959), "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease," *Journal of the National Cancer Institute*, 22, 719–48.

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

Thomas, D.R. (1971), "On the Asymptotic Normality of a Generalized Savage Statistic for Comparing Two Arbitrarily Censored Samples," Technical Report, Department of Statistics, Oregon State University.