THE EFFICIENCY OF COX'S LIKELIHOOD FUNCTION FOR CENSORED DATA

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 77
NOVEMBER 24, 1975

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MPS74-21416

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

THE EFFICIENCY OF COX'S LIKELIHOOD FUNCTION FOR CENSORED DATA

by

BRADLEY EFRON

TECHNICAL REPORT NO. 77
NOVEMBER 24, 1975

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MPS74-21416

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Also issued as Technical Report No. 15 under Contract Public Health Service Grant 1 R01 GM21215-01.

The Efficiency of Cox's Likelihood Function For Censored Data

Bradley Efron

Abstract. 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.

1. Introduction.

Suppose several items are undergoing a life test, the $i^{\underline{th}}$ one of which has hazard rate

$$h_i(t) = \theta_i(t,\underset{\sim}{\beta})h(t,\underset{\sim}{\gamma}) \ , \tag{1.1}$$

where $\underset{\sim}{\beta}$ is an unknown parameter vector we wish to estimate. Here $\underset{\sim}{\gamma}$ is a nuisance parameter indexing a broad class of possible h functions. As a frequently occurring limiting case h is considered completely unknown, except perhaps for continuity assumptions.

Cox (1972) suggested the following analysis: let $t_1 < t_2 < \cdots < t_J$ 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 $j^{\underline{th}}$ 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 $\theta_{i_j}(t_j,\underset{\sim}{\beta})/\sum_{i\in\mathcal{R}(t_j)}\theta_i(t_j,\underset{\sim}{\beta})$. Simply multiplying these factors together gives

$$\prod_{j=1}^{J} \frac{\theta_{i_j}(t_j,\underset{\sim}{\beta})}{\sum_{i\in\mathcal{R}(t_j)} \theta_i(t_j,\underset{\sim}{\beta})}, \tag{1.2}$$

henceforth referred to as <u>Cox's likelihood function</u>. Cox treats this as an ordinary likelihood function for the purposes of inference on $\underset{\sim}{\beta}$. 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).

There are three very attractive features of this approach. 1. The nuisance function $h(t,\underset{\sim}{\gamma})$ is completely removed from the inference process on $\underset{\sim}{\beta}$. 2. Covariance information on the different items is easily incorporated into (1.1). If $\underset{\sim}{z}_i(t)$ is a possibly time-varying vector of observed covariates then the parametrization suggested by Cox is $\theta_i(t,\underset{\sim}{\beta}) = \exp\{\underset{\sim}{\beta}\underset{\sim}{z}_i(t)\}$, taking $\underset{\sim}{\beta}$ a row vector and $\underset{\sim}{z}_i(t)$ a column vector for notational convenience. 3. Data censoring patterns often encountered in life tests do not affect (1.2). For example a recent study conducted in California 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 all had not died by the end of the study. Complicated censoring patterns such as this affect (1.2) only through the risk sets $\mathcal{R}(t_j)$. Analysis of the effect of covariates such as race, sex, and blood pressure proceeds as easily as if there were no censoring.

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 "non-failure 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.

Kalbfleisch and Prentice (1973) pointed out that under certain conditions (1.2) is the marginal likelihood of the pattern of observed risks, but the conditions are rather restrictive; the covariate vectors $\underset{\sim}{z}_i$ are not allowed to be time dependent, and any censoring is assumed to occur immediately following an observation time t_j. (See Remark F, section 5). Recently Cox (1975) proposed a theory of partial likelihood intended to show 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 $h(t,\underset{\sim}{\gamma})$ is moderately large then inferences about $\underset{\sim}{\beta}$ based on (1.2) are asymptotically equivalent to those based on all the data.

In practice $h(t,\underset{\sim}{\gamma})$ may be an important quantity in its own right rather than a nuisance. The connection between the Cox likelihood and inferences about $h(t,\underset{\sim}{\gamma})$ is considered briefly in section 4, particularly as it concerns the Kaplan-Meier estimator. This analysis is closely related to that in Breslow (1974). There is also considerable overlap with the much more carefully justified work of Aalen (1975), and Breslow and Crowley (1974).

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 6. Section 5 consists of several brief remarks on Cox's likelihood and the Kaplan-Meier estimator.

2. The Kaplan-Meier estimator.

Let T be a non-negative continuous random variable with right-sided cumulative distribution function

$$F(t) \equiv \mathrm{Prob}\{T > t\} \equiv e^{-\int_0^t h(s)ds} . \qquad (2.1)$$

Here h(t) is the hazard function; taking logarithms and differentiating (2.1) gives

$$h(t) = \frac{f(t)}{F(t)} , \qquad (2.2)$$

where f(t) is the density of T. For $0 \leq t_1 < t_2$, (2.1) gives

$$\mathrm{Prob}\{T > t_2 | T > t_1\} = e^{-\int_{t_1}^{t_2} h(s)ds} . \qquad (2.3)$$

Suppose several identical items are on test, each independently obeying the same hazard function h(t). We wish to infer h from the observed failure times. Imagine the time axis $[0,\infty)$ divided into ϵ width intervals. Define h_ℓ to be the conditional probability of a single item failing before the end of the $\ell^{\underline{th}}$ interval given that it has survived to the beginning,

$$\begin{aligned} h_\ell &\equiv \mathrm{Prob}\{T \leq \ell\epsilon | T > (\ell-1)\epsilon\} \\ &= 1-e^{-\int_{(\ell-1)\epsilon}^{\ell\epsilon} h(s)ds} . \end{aligned} \qquad (2.4)$$

Let n_ℓ be the number of items on test at the beginning of the $\ell^{\underline{th}}$ interval and suppose single items have been observed to fail at times $t_1 < t_2 < \cdots < t_J$, where t_j exists in the $\ell_j^{\underline{th}}$ interval, $\ell_1 < \ell_2 < \cdots < \ell_J$.

The likelihood of this discretized version of the data is

$$f_h^{(\epsilon)}(\text{data}) = \prod_{\ell \notin \{\ell_1, \ell_2, \ldots, \ell_J\}} (1-h_\ell)^{n_\ell} \prod_{j=1}^{J} n_{\ell_j} h_{\ell_j} (1-h_{\ell_j})^{n_{\ell_j}-1} , \qquad (2.5)$$

thought of as a function of $h = (h_1, h_2, h_3, \ldots)$.

Maximizing (2.5) with respect to the parameters h_ℓ gives the familiar Kaplan-Meier (1958) estimate of the hazard function

$$h_\ell^* = \begin{matrix} 0 & \ell \notin \{\ell_1, \ell_2, \ldots, \ell_J\} \\ \frac{1}{n_{\ell_j}} & \ell = \ell_j, \; j=1,2,\ldots,J \end{matrix} \qquad (2.6)$$

Letting $\epsilon \to 0$ gives the Kaplan-Meier estimate of F,

$$F^*(t) = \prod_{t_j \leq t} \left(1 - \frac{1}{n(t_j)}\right) , \qquad (2.7)$$

where

$$n(t) \equiv \text{number of items on test just before time } t \, . \qquad (2.8)$$

In what follows, $n(t)$ is assumed to be a step function continuous from the left, changing value only finitely often in any finite interval.

Now let $h(t)$ be any hazard function continuous at $t_1, t_2, \ldots, t_J$, and let $\epsilon \to 0$ in (2.5). This gives a limiting version of the likelihood as a function of the unknown hazard rate h,

$$\begin{aligned} f_h(\text{data}) &\equiv \lim_{\epsilon \to 0} \frac{f_h^{(\epsilon)}(\text{data})}{\epsilon^J} \\ &= e^{-\int_0^\infty n(t)h(t)dt} \prod_{j=1}^{J} n(t_j)h(t_j) \; . \end{aligned} \tag{2.9}$$

Letting $\underset{\sim}{J}(t)$ be the sum of delta functions at $t_1, t_2, \ldots, t_J$,

$$\underset{\sim}{J}(t) \equiv \sum_{j=1}^{J} \delta(t-t_j) \; , \tag{2.10}$$

we can write the log likelihood, as in Aalen (1975), to be

$$\log f_h(\text{data}) = \int_0^\infty \{\underset{\sim}{J}(t) \log n(t)h(t) - n(t)h(t)\} dt \; . \tag{2.11}$$

Equations (2.9), (2.11) can be derived directly without recourse to the discretization argument, as at the beginning of section 3, but in complicated situations (2.5) is useful for clarifying the exact meaning of the likelihood expressions.

Because the Kaplan-Meier estimate is very lumpy it is not believable in most real situations. (Usually the statistician will graph (2.7) and implicibly or explicitly draw a smooth cdf through the step function.) It is more realistic, though also more difficult, to parameratize a smooth family of hazards $h(t, \underset{\sim}{\gamma})$ in some way by means of an unknown C dimensional parameter vector $\underset{\sim}{\gamma}$, and estimate $\underset{\sim}{\gamma}$ from the data. Differentiating (2.11) gives the maximum likelihood equations

$$\int_0^\infty \{\underset{\sim}{J}(t)-n(t)h(t,\underset{\sim}{\gamma})\} \frac{\partial \log h(t,\underset{\sim}{\gamma})}{\partial\gamma_c} dt = 0 \quad c=1,2,\ldots,C . \qquad (2.12)$$

The factor $\{\underset{\sim}{J}(t)-n(t)h(t,\underset{\sim}{\gamma})\}dt$ may be thought of as "observed minus expected number of failures in the interval $[t,t+dt]$". Going back to the discrete model makes this explicit. Equations (2.12) are the same as those for estimating the probability vector $\underset{\sim}{\pi}(\underset{\sim}{\gamma})$ of a parameterized family of multinomials,

$$\sum_t (O_t-E_t) \frac{\partial \log \underset{\sim}{\pi}_t(\underset{\sim}{\gamma})}{\partial\gamma_c} = 0 \quad c=1,2,\ldots,C , \qquad (2.13)$$

where t now indexes the multinomial categories, and O_t-E_t indicates observed minus expected in category t.

3. Cox's Likelihood Function.

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

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

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 than (1.1); for the moment it will not be indicated in the notation.

The likelihood function of the observed data is now

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

where as before t_j is the $j^{\underline{th}}$ 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.

Equation (3.2) is derived from standard Poisson process arguments by noting that the probability of no event between t_{j-1} and t_j is $\exp\{-\int_{t_{j-1}}^{t_j} (\sum_{\mathcal{R}(t)} \theta_i(t))h(t)dt\}$, while the probability of the single event "i_j fails at t_j" is proportional to $\theta_{i_j}(t_j)h(t_j)$. This assumes that h_{i_j} is continuous at t_j, and that the risk sets are continuous from the left and change only finitely often in any finite interval. Of course (3.2) can be derived more carefully by discretization as in section 2. Aalen (1975) gives a rigorous derivation.

We will rewrite (3.2) to emphasize its relation to the Cox likelihood and the likelihood (2.9) for the Kaplan-Meier situation. Define

$$\tilde{h}(t) = \left(\frac{\sum_{i=1}^{n} \theta_i(t)}{n} \right) h(t) \ , \qquad (3.3)$$

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

$$\tilde{n}(t) \equiv n \frac{\sum_{i\in\mathcal{R}(t)} \theta_i(t)}{\sum_{i=1}^{n} \theta_i(t)} \ . \qquad (3.4)$$

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

The likelihood function (3.2) can now be written as

$$f_{\theta,h}(\text{data}) = \left\{ \prod_{j=1}^{J} \frac{\theta_{i_j}(t_j)}{\sum_{\mathcal{R}(t)} \theta_i(t_j)} \right\} \left\{ e^{-\int_0^\infty \tilde{n}(t)\tilde{h}(t)dt} \prod_{j=1}^{J} \tilde{n}(t_j)\tilde{h}(t_j) \right\} \tag{3.5}$$

The first factor is the Cox likelihood while the second factor is analogous to (2.9).

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

$$\frac{\theta_i(t)}{\theta_{i'}(t)} = \frac{e^{\underset{\sim}{\beta}\underset{\sim}{z}_i(t)}}{e^{\underset{\sim}{\beta}\underset{\sim}{z}_{i'}(t)}} , \tag{3.6}$$

where $\underset{\sim}{\beta}$ is a $1 \times B$ unknown parameter vector and $\underset{\sim}{z}_i(t)$ a $B \times 1$ possibly time-varying vector of observed covariates. This makes the Cox likelihood

$$\prod_{j=1}^{J} \frac{e^{\underset{\sim}{\beta}\underset{\sim}{z}_{i_j}(t)}}{\sum_{i\in\mathcal{R}(t_j)} e^{\underset{\sim}{\beta}\underset{\sim}{z}_i(t_j)}} \tag{3.7}$$

as in the 1972 paper; (3.4) becomes

$$\tilde{n}(t) \equiv \tilde{n}(t,\underset{\sim}{\beta}) = n\,\frac{\sum\limits_{\mathcal{R}(t)} e^{\underset{\sim}{\beta}\,\underset{\sim}{z}_i(t)}}{\sum\limits_{1}^{n} e^{\underset{\sim}{\beta}\,\underset{\sim}{z}_i(t)}}\ . \tag{3.8}$$

Notice that (3.6) is weaker than the assumption $\theta_i(t) = \exp\{\underset{\sim}{\beta}\,\underset{\sim}{z}_i(t)\}$ 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 $\tilde{h}(t)$ is assumed to be of the form

$$\tilde{h}(t,\underset{\sim}{\gamma}) = e^{\underset{\sim}{\gamma}\,\underset{\sim}{w}(t)} \tag{3.9}$$

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

$$f_{\underset{\sim}{\beta},\underset{\sim}{\gamma}}(\text{data}) = \left\{\prod_{j=1}^{J} \frac{e^{\underset{\sim}{\beta}\,\underset{\sim}{z}_{i_j}(t_j)}}{\sum\limits_{\mathcal{R}(t_j)} e^{\underset{\sim}{\beta}\,\underset{\sim}{z}_i(t_j)}}\right\} \tag{3.10}$$
$$\left\{e^{-\int_0^\infty \tilde{n}(t,\underset{\sim}{\beta})\tilde{h}(t,\underset{\sim}{\gamma})dt} \prod_{j=1}^{J} \tilde{n}(t_j,\underset{\sim}{\beta})\tilde{h}(t_j,\underset{\sim}{\gamma})\right\}$$

See Remark I, section 5.

Cox's 1975 paper shows that the first factor can be treated an an ordinary likelihood function for the purpose of large sample inference. In particular the "maximum likelihood estimator" of $\underset{\sim}{\beta}$ obtained by maximizing (3.7) will asymptotically have mean $\underset{\sim}{\beta}$ and covariance matrix 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 $\underset{\sim}{\beta}$. The quotation marks serve as a reminder that (3.7) is not really a likelihood function. (For example it is <u>not</u> in general the likelihood of the reduced data set $(\mathcal{R}(t_1),i_1)$, $(\mathcal{R}(t_2),i_2),\dots,(\mathcal{R}(t_J),i_J)$.)

In what follows we will calculate the actual Fisher information for $\underset{\sim}{\beta}$ from (3.10), and give a heuristic demonstration that asymptotically this gives the same results as the calculation based just on (3.7), assuming that the class of hazards $\tilde{h}(t,\underset{\sim}{\gamma})$ is moderately large. This shows that the maximum likelihood estimate of $\underset{\sim}{\beta}$ 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.

For conveneince 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 5. Define

$$E_\beta^{\mathcal{R}(t)} z \equiv \frac{\sum\limits_{i\in\mathcal{R}(t)} z_i(t)e^{\beta z_i(t)}}{\sum\limits_{i\in\mathcal{R}(t)} e^{\beta z_i(t)}}, \quad E_\beta z \equiv \frac{\sum\limits_{i=1}^{n} z_i(t)e^{\beta z_i(t)}}{\sum\limits_{i=1}^{n} e^{\beta z_i(t)}} \qquad (3.11)$$

and

$$\mathrm{Var}_\beta^{\mathcal{R}(t)} z = \sum_{i\in\mathcal{R}(t)} [z_i(t)-E_\beta^{\mathcal{R}(t)} z]^2 e^{\beta z_i(t)} \Big/ \sum_{i\in\mathcal{R}(t)} e^{\beta z_i(t)} \qquad (3.12)$$

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

<u>Lemma</u>. The Fisher information for estimating β in (3.10) is

$$\inf_{\underset{\sim}{g}} \int_0^\infty \mathcal{E}\left(\left\{ \mathrm{Var}_\beta^{\mathfrak{R}(t)}z + [(E_\beta^{\mathfrak{R}(t)}z - E_\beta z) - \underset{\sim}{g}\,\underset{\sim}{w}(t)]^2 \right\} \tilde{n}(t,\beta)\tilde{h}(t,\underset{\sim}{\chi})\right) dt \ , \tag{3.13}$$

where the infinum is over all choices of the C dimensional vector $\underset{\sim}{g}$, and "$\mathcal{E}$" indicates expectation over the randomness in the risk sets $\mathfrak{R}(t)$. The same expression without the term in square brackets is the Fisher information for β based just on the Cox likelihood (3.7). (Proof given in section 6.)

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

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

where (3.14)

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

Let $\tilde{N}(t,\beta)$ be the expectation, over the randomness in $\mathfrak{R}(t)$, of $\tilde{n}(t,\beta)$,

$$\tilde{N}(t,\beta) \equiv \mathcal{E}\ \tilde{n}(t,\beta) = \frac{\mathcal{E} \sum_{\mathfrak{R}(t)} e^{\beta z_i(t)}}{\sum_{i=1}^{n} e^{\beta z_i(t)}} \ , \tag{3.15}$$

and define $B \equiv \tilde{n}(t,\beta)/\tilde{N}(t,\beta)$, $A \equiv E_\beta^{\mathcal{R}(t)} z - E_\beta z$, and $a \equiv \underset{\sim}{g}\,\underset{\sim}{w}(t)$. Using (3.14), the integrand of (3.13) can be expressed as

$$\left\{ \mathcal{E}\, \frac{\tilde{n}(t,\beta)}{\tilde{N}(t,\beta)}\, \mathrm{Var}_\beta^{\mathcal{R}(t)} z + \mathbf{Var}_{\tilde{n}/\tilde{N}}\, E_\beta^{\mathcal{R}(t)} z + [v_\beta(t) - \underset{\sim}{g}\underset{\sim}{w}(t)]^2 \right\} \tilde{N}(t,\beta)\tilde{n}(t,\underset{\sim}{\chi}) \tag{3.16}$$

where

$$v_\beta(t) \equiv \mathcal{E}_{\tilde{n}/\tilde{N}} (E_\beta^{\mathcal{R}(t)} z - E_\beta z) \tag{3.17}$$

and "$\mathbf{Var}_{\tilde{n}/\tilde{N}}\, E_\beta^{\mathcal{R}(t)} z$" indicates a weighted variance, as in (3.14), the random quantity being $\mathcal{R}(t)$.

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

$$v_\beta(t) = \frac{\sum_1^n P_i(t) z_i(t) e^{\beta z_i(t)}}{\sum_1^n P_i(t) e^{\beta z_i(t)}} - \frac{\sum_1^n z_i e^{\beta z_i}}{\sum_1^n e^{\beta z_i}} . \tag{3.18}$$

Notice that $P_i(t)$ is also a function of β and $\underset{\sim}{\chi}$, and possibly other extraneous random factors.

The principle implied by the lemma and (3.16), admittedly in a rough manner, is this: if, as the number items tested goes to infinity, the function $v_\beta(t)$ 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.

Suppose for a moment that $v_\beta(t) = \underset{\sim}{g}\underset{\sim}{w}(t)$ for all t for some choice of g. This eliminates the term in square brackets from (3.16). The additional information for estimating β <u>not</u> in the Cox likelihood corresponds to the term $\mathrm{Var}_{\tilde{n}/\tilde{N}}\, E_\beta^{\mathcal{R}(t)} z$. Intuitively this comes from local variations in $\tilde{n}(t,\beta)$ 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 $\tilde{h}(t,\underset{\sim}{\chi})$ since this is necessarily a fixed (non-random) function of time. However the magnitude of this term tends to be $O(1/\tilde{N})$ compared to the term $(\tilde{n}/\tilde{N})\mathrm{Var}_\beta^{\mathcal{R}(t)} z$ from the Cox likelihood, essentially because $E_\beta^{\mathcal{R}(t)} z$ is the average of about $\tilde{N}$ random quantities. (See Remark K, section 5.)

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

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

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

In order for the Cox likelihood to estimate β with reasonable efficiency in finite samples it is necessary for $v_\beta(t)$ to be in or at least near $\mathcal{L}(\underset{\sim}{w})$. 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 $v_\beta(t)$ is monotonic. For $\beta > 0$, $v_\beta(t)$ will decrease monotonically in time as those items with large values of z_i are selectively removed by earlier failure. Censoring can distort $v_\beta(t)$, but not seriously unless a large proportion of the items have the same fixed censoring time. (See Remark K, section 5.) In the absence of firm prior knowledge it may be reasonable to assume that $\tilde{h}(t,\underset{\sim}{\chi}) = \exp\{\underset{\sim}{\chi}\ \underset{\sim}{w}(t)\}$ can be any smooth monotonic function, which in this case guarantees the asymptotic efficiency of the Cox likelihood.

Of course there are situations in which the Cox likelihood by itself produces seriously inefficient inferences. For example $\mathcal{L}(\underset{\sim}{w})$ might be known to be the class of linear functions w_1+w_2t while $v_\beta(t)$ is some considerably more complicated function. In theory at least the statistician can always calculate the actual MLE of β from (3.10) in such cases. Kalbfleisch (1974) gives an efficiency calculation in one such case, which reinforces faith in using Cox's likelihood by itself.

4. Estimating the Hazard Rates. Suppose we are willing to rely on the first factor in (3.10), the Cox likelihood, for the estimation of $\underset{\sim}{\beta}$. We can treat the estimate obtained in this way, say $\underset{\sim}{\beta}^*$, as if it were the true value of $\underset{\sim}{\beta}$, and then use the second factor in (3.10) to estimate $\underset{\sim}{\chi}$. The maximum likelihood equations (2.12) are

$$\int_0^\infty \{\underset{\sim}{J}(t)-\tilde{n}(t,\underset{\sim}{\beta}^*)\tilde{h}(t,\underset{\sim}{\chi})\} \frac{\partial \log \tilde{h}(t,\underset{\sim}{\chi})}{\partial \gamma_c}\, dt = 0 \qquad c=1,2,\ldots,C \, . \qquad (4.1)$$

As a limiting case where $\tilde{h}$ is taken to be completely arbitrary we get as an estimate of the cdf $\tilde{F}(t) \equiv \exp\{-\int_0^t \tilde{h}(s)ds\}$ the Kaplan-Meier estimator (2.8),

$$\tilde{F}^*(t) = \prod_{t_j \leq t} (1 - \frac{1}{\tilde{n}(t_j, \underset{\sim}{\beta}^*)}) . \qquad (4.2)$$

Let $\underset{\sim}{\chi}^*$ be the "maximum likelihood" estimator of $\underset{\sim}{\chi}$ obtained from (4.1), the quotes indicating that $\underset{\sim}{\chi}^*$ is really only the conditional maximizer given the value $\underset{\sim}{\beta}^*$ obtained from the Cox likelihood. From (3.1), (3.3) and (3.6) we get

$$h_i(t) = n \frac{\theta_i(t)}{\sum_{i'=1}^{n} \theta_{i'}(t)} \tilde{h}(t) = n \frac{e^{\underset{\sim}{\beta}\underset{\sim}{z}_i(t)}}{\sum_{i'=1}^{n} e^{\underset{\sim}{\beta}\underset{\sim}{z}_{i'}(t)}} \tilde{h}(t, \underset{\sim}{\chi}) \qquad (4.3)$$

so the corresponding estimate of the hazard rate for item i is

$$h_i^*(t) = n \frac{e^{\underset{\sim}{\beta}^*\underset{\sim}{z}_i(t)}}{\sum_{i'=1}^{n} e^{\underset{\sim}{\beta}^*\underset{\sim}{z}_{i'}(t)}} \tilde{h}(t, \underset{\sim}{\chi}^*) . \qquad (4.4)$$

The Kaplan-Meier case (4.2) represents a limiting form where $\tilde{h}(t, \underset{\sim}{\chi}^*)$ approaches a sum of delta functions at $t_1, t_2, \ldots, t_J$ satisfying

$$e^{-\int_{t_j-}^{t_j+} \tilde{h}^*(s)ds} = 1 - \frac{1}{\tilde{n}(t_j, \underset{\sim}{\beta}^*)} . \qquad (4.5)$$

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

$$e^{-\int_{t_j-}^{t_j+} h_i^*(s)ds} = [1 - \frac{1}{\tilde{n}(t_j, \underset{\sim}{\beta}^*)}]^{\phi_i^*(t_j)} \qquad (4.6)$$

where $\phi_i^*(t_j) \equiv ne^{\underset{\sim}{\beta}^* \underset{\sim}{z}_i(t_j)} \Big/ \sum_1^n e^{\underset{\sim}{\beta}^* z_{i'}(t_j)}$. The estimate of the i$\underline{\text{th}}$ cdf is

$$F_i^*(t) = \prod_{t_j \leq t} [1 - \frac{1}{\tilde{n}(t_j, \beta^*)}]^{\phi_i^*(t_j)}$$
$$\qquad (4.7)$$
$$\approx \exp - \left[\sum_{t_j \leq t} \left(e^{\underset{\sim}{\beta}^* \underset{\sim}{z}_i(t_j)} \Big/ \sum_{i'=1}^n e^{\underset{\sim}{\beta}^* \underset{\sim}{z}_{i'}(t_j)} \right) \right] .$$

This last form being essentially the same as that derived in Breslow (1974). (See remark C, section 5.)

5. Some Remarks.

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

$$\inf_{\underset{\sim}{G}} \int_0^\infty \mathcal{E} \left(\left\{ \mathrm{Cov}_{\underset{\sim}{\beta}}^{\mathcal{R}(t)} \underset{\sim}{z} + [(E_\beta^{\mathcal{R}(t)} \underset{\sim}{z} - E_\beta \underset{\sim}{z}) - \underset{\approx}{G} \underset{\approx}{w}(t)][(E_\beta^{\mathcal{R}(t)} \underset{\sim}{z} - E_\beta \underset{\sim}{z}) - \underset{\approx}{G} \underset{\approx}{w}(t)]' \right\} \right.$$
$$\left. \cdot \ \tilde{n}(t, \underset{\sim}{\beta}) \tilde{h}(t, \underset{\sim}{\gamma}) \right) dt \ , \qquad (5.1)$$

the infinum being taken over all B × C matrices $\underset{\sim}{G}$.

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

$$\mathrm{Prob}\{T_i < 1\} = \frac{e^{\underset{\sim}{\beta}\underset{\sim}{z}_i}}{1+e^{\underset{\sim}{\beta}\underset{\sim}{z}_i}} . \tag{5.2}$$

This implies

$$\theta_i(\underset{\sim}{\beta}) \propto \log(1+e^{\underset{\sim}{\beta}\underset{\sim}{z}_i}) \tag{5.3}$$

rather than $\theta_i(\underset{\sim}{\beta}) \propto \exp\{\underset{\sim}{\beta}\underset{\sim}{z}_i\}$.

C. If m is a large positive number then

$$\log(1 - \frac{1}{m}) = - \frac{1}{m-c(m)} \tag{5.4}$$

where $c(m) = 1/2 - 1/12m + \cdots$. Expression (2.8) for the Kaplan-Meier estimator can be written as

$$F^*(t) = e^{-\sum_{t_j \le t} 1/[n(t_j)-c(n(t_j))]} . \tag{5.5}$$

Ignoring the correction term $c(n(t_j))$ leads to the last expression in (4.7)

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

E. Suppose in (2.11) that $\underset{\sim}{J}(t)$ is approximated by any continuous function. A simple calculation shows that the maximizing choice of $h(t)$ is

$$h(t) = \underset{\sim}{J}(t)/n(t) \tag{5.6}$$

giving (2.11) a maximum value of

$$\int_0^\infty \underset{\sim}{J}(t)\{\log \underset{\sim}{J}(t) - 1\}dt \tag{5.7}$$

which does not depend on $n(t)$.

This suggests that in the case where $\tilde{h}(t,\underset{\sim}{\gamma})$ is completely unspecified the second factor in (3.10) contributes nothing to the maximum likelihood estimation of $\underset{\sim}{\beta}$, so that the maximizer $\underset{\sim}{\beta}^*$ for the first factor should be the overall maximizer. The flaw in this argument is that (5.6) does not yield the actual maximizer, namely the Kaplan-Meier form $1/[n(t_j)-c(n(t_j))]$, as $\underset{\sim}{J}(t)$ approaches the limit (2.10), but rather the approximate maximizer $1/n(t_j)$ mentioned in remark D. (This difficulty relates to the fact that (2.11) holds only for suitably smooth functions h.) It is true, as the discussion in section 3 shows, that the unrestricted MLE for $\underset{\sim}{\beta}$ from (3.10) will be asymptotically identical to that obtained from the Cox likelihood alone.

F. The likelihood expressions (3.2), (3.5), (3.10) assume that the risk sets $\mathcal{R}(t)$ are themselves uninformative for $\underset{\sim}{\beta}$ and $\underset{\sim}{\gamma}$. 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 $\underset{\sim}{\beta}$ or $\underset{\sim}{\gamma}$. Subject to these restrictions a malevelent 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; that 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 $\theta_1,\theta_2,\theta_3$ 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. As easy calculation gives the probability of observing the partial ordering "T_1 less than $\min\{T_2,T_3\}$" to be

$$\text{Prob}\{(1,2,3)\cup(1,3,2)\} = e^{-(\theta_1+\theta_2+\theta_3)\int_0^{1.5}h(t)dt}\ \frac{\theta_1}{\theta_1+\theta_2+\theta_3} \qquad (5.8)$$

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

G. 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).

H. Real censored data problems are often discrete; items are reported to fail during intervals, not by exact times. (In the senior citizen study for example, 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 $j^{\underline{th}}$ 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 $j^{\underline{th}}$ factor in the Cox likelihood is

$$\frac{\theta_{i_{j1}}(t_j)\theta_{i_{j2}}(t_j)\cdots\theta_{i_{jm}}(t_j)}{\prod_{\ell=0}^{m-1}\left[\sum_{i\in\mathcal{R}(t_j)}\theta_i(t_j)-\frac{\ell}{m}\sum_{h=1}^{m}\theta_{i_{jh}}(t_j)\right]} \quad . \tag{5.9}$$

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

I. The parameterization (3.6), (3.9) which lead 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 $\underset{\sim}{\beta}^{(o)}, \underset{\sim}{\gamma}^{(o)}$ require that the possible $\underset{\sim}{\gamma}$ vectors corresponding to $\underset{\sim}{\beta} = \underset{\sim}{\beta}^{(o)}$ include an open set around $\underset{\sim}{\gamma}^{(o)}$.

Other parameterizations are possible. For example we can parameterize θ_i and h directly as at (1.1), leading to the likelihood expression

$$\left\{\prod_{j=1}^{J}\frac{\theta_{i_j}(t_j,\underset{\sim}{\beta})}{\sum_{\mathcal{R}(t_j)}\theta_i(t_j,\underset{\sim}{\beta})}\right\} \tag{5.10}$$

$$\left\{e^{-\int_o^\infty\left(\sum_{\mathcal{R}(t)}\theta_i(t,\underset{\sim}{\beta})\right)h(t,\underset{\sim}{\gamma})dt}\prod_{j=1}^{J}\left(\sum_{\mathcal{R}(t_j)}\theta_i(t_j,\underset{\sim}{\beta})\right)h(t_j,\underset{\sim}{\gamma})\right\} .$$

The forms $\theta_i(t,\underset{\sim}{\beta}) = \exp\{\underset{\sim}{\beta}\underset{\sim}{z}_i(t)\}$, $h(t,\underset{\sim}{\gamma}) = \exp\{\underset{\sim}{\gamma}\underset{\sim}{w}(t)\}$ can be shown to give virtually the same results as those derived in sections 3 and 4.

A parameterization which seems appealing is to let $\bar{h}(t,\underset{\sim}{\gamma}) \equiv (\sum_{\mathcal{R}(t)} \theta_i(t,\underset{\sim}{\beta})/n(t))h(t)$ 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 $\theta_i(t,\underset{\sim}{\beta}) = \exp\{\underset{\sim}{\beta}\underset{\sim}{z}_i(t)\}$, $h(t,\underset{\sim}{\gamma}) = \exp\{\underset{\sim}{\gamma}\,\underset{\sim}{w}(t\ )\}$. This makes the second factor in (5.10) equal to

$$e^{-\int_0^\infty n(t)\bar{h}(t,\underset{\sim}{\gamma})dt} \prod_{j=1}^{J} n(t_j)\bar{h}(t_j,\underset{\sim}{\gamma}) \ , \tag{5.11}$$

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

J. Consider the case where the $\underset{\sim}{z}_i$ are not functions of time. In large samples the ratio of successive spacings $(t_{j+1}-t_j)/(t_j-t_{j-1})$ will tend to be distributed as

$$\frac{\sum_{i\in\mathcal{R}(t_{j-1})} e^{\underset{\sim}{\beta}\underset{\sim}{z}_i}}{\sum_{i\in\mathcal{R}(t_j)} e^{\underset{\sim}{\beta}\underset{\sim}{z}_i}} F_{2,2} \ , \tag{5.12}$$

where $F_{2,2}$ indicates an F variate with both degrees of freedom equal to two. This depends on $\underset{\sim}{\beta}$ in a weak way. It is an example of extra information for $\underset{\sim}{\beta}$ not available from the Cox likelihood, and not dependent on $\tilde{h}$ being restricted to a limited parametric family.

K. The function $\tilde{N}(t,\beta)$ defined at (3.15) equals

$$n \sum_{i=1}^{n} P_i(t)e^{\beta z_i(t)} \Big/ \sum_{i=1}^{n} e^{\beta z_i(t)} \qquad (5.13)$$

when $P_i(t) = \text{Prob}\{i\epsilon\mathcal{R}(t)\}$. If item i has a fixed censoring time c_i, after which it is taken off test if it hasn't already failed, then $P_i(t)$ goes discontinuously to zero at $t = c_i$. From (3.18) we see that this produces a discontinuity of order $O(1/\tilde{N})$ in $v_\beta(t)$ as $\tilde{N}$ goes to infinity, assuming that the $z_i(t)$ are bounded.

Suppose that the various items act independently of each other in terms of failures and censorings. Then standard expansion methods show that the quantity $\mathcal{V}ar_{\tilde{n}/\tilde{N}} E_\beta^{\mathcal{R}(t)} z$ which figures in (3.16) approximately equals

$$\frac{1}{\tilde{N}} \frac{\sum_{i=1}^{n} P_i Q_i \phi_i^2 (z_i - R)^2}{\tilde{N}}, \qquad (5.14)$$

where t, β, and $\underset{\sim}{\chi}$ have been dropped from the notation, $Q_i \equiv 1-P_i$, and

$$\phi_i \equiv n\, e^{\beta z_i} \Big/ \sum_{i'=1}^{n} e^{\beta z_{i'}}, \quad R \equiv \sum_{i=1}^{n} P_i z_i e^{\beta z_i} \Big/ \sum_{i=1}^{n} P_i e^{\beta z_i}. \qquad (5.15)$$

(5.14) is $O(1/\tilde{N})$ as $\tilde{N}$ goes to infinity, again assuming the $z_i(t)$ bounded.

6. 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),

$$S_\beta \equiv \frac{\partial \log f_{\beta,\underset{\sim}{\chi}}}{\partial \beta} = \sum_{j=1}^{J} [z_{i_j}(t_j) - E_\beta^{\mathcal{R}(t_j)} z] + \int_0^\infty [E_\beta^{\mathcal{R}(t)} z - E_\beta z][\underset{\sim}{J}(t) - \tilde{n}(t,\beta)\tilde{h}(t,\underset{\sim}{\chi})]dt \qquad (6.1)$$

and

$$S_{\gamma_c} \equiv \frac{\partial \log f_{\beta,\underset{\sim}{\chi}}}{\partial \gamma_c} = \int_0^\infty w_c(t)[\underset{\sim}{J}(t) - \tilde{n}(t,\beta)\tilde{h}(t,\underset{\sim}{\chi})]dt \ ,$$

where $\underset{\sim}{J}(t) = \sum_{j=1}^{J} \delta(t-t_j)$ as before.

For any arbitrary choice of $\underset{\sim}{g} = (g_1, g_2, \ldots, g_C)$ we can write

$$S_\beta - \sum_{c=1}^{C} g_c S_{\gamma_c} = \int_0^\infty [U(t) + (v(t) - \underset{\approx}{gw}(t))]dK(t) \qquad (6.2)$$

where

$$U(t) = \begin{cases} z_{i_j}(t) - E_\beta^{\mathcal{R}(t)} z & \text{if} \quad t \in \{t_1, t_2, \ldots, t_J\} \\ 0 & \text{otherwise ,} \end{cases}$$

$$v_\beta(t) = E_\beta^{\mathcal{R}(t)} z - E_\beta z \qquad (6.3)$$

and

$$dK(t) = [\underset{\sim}{J}(t) - \tilde{n}(t,\beta)\tilde{h}(t,\underset{\sim}{\chi})]dt \ .$$

Given the observed value of $\mathcal{R}(t)$, $v_\beta(t)$ is a fixed number while $U(t)$ is a random variable with mean 0 and variance

$$\text{variance}\{U(t)|\mathcal{R}(t)\} = \begin{array}{ll} \text{Var}_\beta^{\mathcal{R}(t)} z & \text{if } t \in \{t_1, t_2, \ldots, t_J\} \\ 0 & \text{otherwise ,} \end{array} \tag{6.4}$$

$\text{Var}_\beta^{\mathcal{R}(t)} z$ as given in (3.12). Also, still assuming $\mathcal{R}(t)$ given,

$$dK(t) = \begin{cases} 1-\tilde{n}(t,\beta)\tilde{h}(t,\underset{\sim}{\chi})dt \text{ with } & \text{prob. } \tilde{n}(t,\beta)\tilde{h}(t,\underset{\sim}{\chi})dt \\ -\tilde{n}(t,\beta)\tilde{h}(t,\underset{\sim}{\chi})dt \text{ with } & \text{prob. } 1-\tilde{n}(t,\beta)\tilde{h}(t,\underset{\sim}{\chi})dt \text{ .} \end{cases} \tag{6.5}$$

Notice that the two cases for $dK(t)$ correspond to the two cases for $U(t)$ given in (6.3).

Putting (6.3)-(6.5) together gives

$$\begin{aligned} &E\{([U(t)+(v_\beta(t)-\underset{\approx}{gw}(t))]dK(t)^2|\mathcal{R}(t)\} \\ &\quad = \{\text{Var}_\beta^{\mathcal{R}(t)} z+[(E_\beta^{\mathcal{R}(t)} z-E_\beta z)-\underset{\approx}{gw}(t)]^2\}\tilde{n}(t,\beta)\tilde{h}(t,\underset{\sim}{\chi})dt \text{ ,} \end{aligned} \tag{6.6}$$

and, for $t' < t$,

$$E\{([U(t')+(v_\beta(t')-\underset{\approx}{gw}(t'))])([U(t)+(v_\beta(t)-\underset{\approx}{gw}(t))])|\mathcal{R}(s),\ 0 \leq s \leq t\} = 0 \tag{6.7}$$

Therefore, writing the integral (6.2) 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 one over the Cramér-Rao lower bound for β, by definition the Fisher information for β, is the infinum of the expected value of $(S_\beta - \sum_{c=1}^{C} g_c S_{\gamma_c})^2$ over all choices of $\underset{\sim}{g}$. This proves the first part of the lemma. The second part follows by a similar argument, made easier by the fact that (3.7) does not involve $\underset{\sim}{\gamma}$.

References

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, 437-453.

Cox, D.R. (1972). Regression Models and Life-Tables (with Discussion). Jour. Royal Statistical Soc, Series B, 34, 187-220.

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

Kalbfleisch, J. and Prentice, R. (1973). Marginal likelihoods based on Cox's regression and life model. Biometrika, 60, 267-279.

Kalbfleisch, J. (1974). Some efficiency calculations for survival distributions. Biometrika, 61, 31-38.

Mantel, N., and Haenzel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of desease. Jour. Nat. Cancer Inst. 22, 719-748.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant procedures. Jour. Royal Stat. Soc, Series A, 135, 185-206.