Austral. J. Statist., 31(3), 1989, 365–384

A DISCRETE TIME PARAMETRIC MODEL FOR THE ANALYSIS OF FAILURE TIME DATA

GEOFF ADAMS[1] AND RAY WATSON[2]

[1]*Department of Mathematics, University of Western Australia*
[2]*Statistics Department, University of Melbourne*

Summary

This paper introduces a parametric discrete failure time model which allows a variety of smooth hazard function shapes, including shapes which are not readily available with continuous failure time models. The model is easy to fit, and statistical inference is simple. Further, it is readily extended to allow for differences between subjects while retaining the ease of fit and simplicity of statistical inference. The performance of the discrete time analysis is demonstrated by application to several data sets.

Key words: Censoring; discrete failure time data analysis; linear log odds; logistic model; parametric hazard estimation; smooth hazards.

1. Introduction

Discrete failure time models have largely been overlooked or ignored in the literature. They are most often treated briefly almost as an addendum to the continuous case. In fact, they have most often found application to grouped continuous data — the underlying modelling assumption being that of a continuous failure time distribution. The application of discrete models has been traditionally restricted to engineering applications where the failure process involved discrete trials.

Discrete time models are worth greater consideration. A number of problems which occur with continuous time models are overcome by using a discrete time model. For example, ties are expected to arise and are handled quite straightforwardly — no special procedures are necessary. Further, it can be argued that a discrete time model is a more natural model in many instances, as most biological populations live according to a natural discrete time unit: the day.

In Section 2 we introduce a simple parametric model which can approximate a wide variety of hazard function shapes, and which provides smooth estimates of the hazard and survivor functions. In Section 3, estimation and inference procedures for the model are discussed. A generalization of the model which is introduced in Section 4 allows the failure time distribution to differ between subjects, and in Section 5 the inclusion of time dependent covariates and an alternative generalization are discussed.

2. The Basic Model

Denote by T a random variable representing failure time. For convenience, the time units will be referred to as days, though other time units may of course be used. It is assumed that T is a non-negative integer-valued random variable with probability function $p(t)$ and survivor function $S(t)$.

The hazard mass function $h(t)$ is defined as the probability of failure at time t, conditional upon survival until time t, i.e.,

$$h(t) = \Pr\{T = t \mid T \geq t\} = p(t)/S(t) \quad (t = 0, 1, \ldots) .$$

One important consideration when modelling the hazard mass function $h(t)$ is that, since $h(t)$ is a probability, the range of values that it can take is restricted to the interval $[0, 1]$. This has obvious implications on the type of functions that can be used to model the hazard mass function.

The discrete parametric model to be considered here is derived as follows. Let $G(x)$ and $g(x)$ denote the distribution function and the probability density function of an arbitrary continuous distribution which is symmetric about zero; i.e.,

$$g(x) = g(-x) , \quad G(x) = 1 - G(-x) \quad (-\infty < x < \infty) . \tag{1}$$

The hazard mass function is then defined by

$$h(t) = G(\xi(t)) \quad (t = 0, 1, \ldots) , \tag{2}$$

where $\xi(t)$ is an unknown function of t. Note that, as a result of (1) and (2),

$$1 - h(t) = 1 - G(\xi(t)) = G(-\xi(t)) .$$

A wide variety of hazard mass functions can be obtained by assuming particular functions for $\xi(t)$, some of which may involve unknown parameters

that will need to be estimated from the data. In particular, the choice of $\xi(t)$ as a low order polynomial,

$$\xi(t) = \alpha_0 + \alpha_1 t + \cdots + \alpha_m t^m = \boldsymbol{\alpha}' \mathbf{e}_t \,, \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_m)'$ and $\mathbf{e}_t = (1, t, \ldots, t^m)'$, leads to a simple and useful parametric family of hazard mass functions.

While the estimation procedures described in Section 3 do not depend on the particular form for the function $G(x)$, when applying the model (2) to failure time data, a particular form for $G(x)$ must be chosen. The most suitable form for the function $G(x)$ has been found to be the logistic function $G(x) = e^x/(1 + e^x) = 1/(1 + e^{-x})$. This choice combines tractability with computational simplicity. It also has the non-trivial advantage of providing convenient statistics for the parameters $\boldsymbol{\alpha}$.

The model (2) has a number of advantages. Firstly, the hazard mass function is a smooth function if the function $\xi(t)$ is smooth. Smooth hazards are preferable when examining and interpreting results. Secondly, by an appropriate choice of the function $\xi(t)$, these models are capable of approximating a wide variety of hazard curves. In particular, the choice of $\xi(t)$ as a low order polynomial in t gives considerable freedom to the behaviour of the hazard mass function — even with $\xi(t)$ a quadratic — while retaining the smoothness of the function. Such a choice allows not only constant, monotone increasing and monotone decreasing curves to be fitted, but also U-shaped and bell-shaped curves. Thirdly, since both $h(t)$ and $1 - h(t)$ occur in the likelihood function and simple expressions exist for both functions, estimation is relatively straightforward. In addition, these models involve only a finite number of parameters, and so standard maximum likelihood theory can be applied.

Polynomial models for the continuous time hazard density function $\lambda(t)$ have been examined by Krane (1963), Kodlin (1967), Gehan & Siddiqui (1973), Bain (1974), Gross & Clark (1975, Section 4.8), Taulbee (1979) and Clayton (1983). These models have the general form

$$\psi(\lambda(t)) = \lambda_0 + \lambda_1 t + \cdots + \lambda_m t^m \,, \tag{4}$$

where $\psi(x)$ is a monotonic "link" function. The main models of this type are the Rayleigh distribution, with $\lambda(t) = \lambda_0 + \lambda_1(t)$, and the Gompertz distribution, with $\lambda(t) = \exp(\lambda_0 + \lambda_1 t)$. While the model (4) is also capable of giving a wide variety of hazard shapes, it suffers from the disadvantage that $\lambda(t)$ is not automatically constrained to be non-negative with either

the linear link function or the inverse link function of the Pareto distribution.

One problem with the model (2) is that there is no convenient expression for the survivor function

$$S(t) = \prod_{s=0}^{t-1} \big(1 - h(t)\big) = \prod_{s=0}^{t-1} G\big(-\xi(s)\big) \,. \tag{5}$$

However, it must be pointed out that this problem is shared by most other discrete parametric models; and by all models allowing such flexibility of hazard function shape.

3. Estimation and Inference in Discrete Time

3.1. Maximum Likelihood Estimation

Suppose that there are initially n subjects under observation, and let n_t be the number of subjects at risk (alive and uncensored) at the beginning of the t th day, so that $n_0 = n$. Also, let f_t denote the number of subjects which fail on the t th day. If the censoring is non-informative, then the likelihood function is proportional to

$$L = \prod_{t=0}^{\infty} \big\{\big(1 - h(t)\big)^{n_t - f_t} h(t)^{f_t}\big\} \,. \tag{6}$$

Even if the censoring is informative, (6) has the interpretation of a partial likelihood (Cox, 1975). The maximum likelihood estimates $\{\hat{h}(t)\}$ are easily obtained by maximizing the log-likelihood function. Clearly,

$$\hat{h}(t) = f_t/n_t \quad (t = 0, 1, \ldots) \,, \tag{7}$$

the proportion of those subjects still alive at the beginning of the t th day that fail during the t th day. This approach is essentially non-parametric, and leads to a survivor function estimate

$$\hat{S}(0) = 1 \,, \quad \hat{S}(t) = \prod_{s=0}^{t-1} \big(1 - \hat{h}(s)\big) \quad (t = 1, 2, \ldots) \,, \tag{8}$$

which is equivalent to the product limit estimate of Kaplan & Meier (1958).

However, it is often the case that, either because the sample is small or because the data are widely spread, that the behaviour of the estimated

hazard mass function is fairly erratic. (If $n_t > 0$ and there are no failures on the tth day then $\hat{h}(t) = 0$.) The parametric model (2), introduced in Section 2, forces the hazard mass function estimate to be smooth while still allowing a wide range of shapes. Thus the proposed model can be regarded as producing a form of smoothing over time of the estimated hazard mass function.

By substituting $G(\xi(t))$ for $h(t)$ in (6), the likelihood function for the model (2) can be written

$$L(\alpha) = \prod_{t=0}^{\infty} \{G(\xi(t))^{f_t} G(-\xi(t))^{s_t}\} = \prod_{t=0}^{\infty} \{G(\boldsymbol{\alpha}'\mathbf{e}_t)^{f_t} G(-\boldsymbol{\alpha}'\mathbf{e}_t)^{s_t}\}, \quad (9)$$

where $\xi(t)$ is given by (3). Since there is only a finite number of parameters $\boldsymbol{\alpha}$, standard maximum likelihood methods can be applied to (9). The log-likelihood function is given by

$$\log L(\boldsymbol{\alpha}) = \sum_{t=0}^{\infty} \{f_t \log G(\xi(t)) + s_t \log G(-\xi(t))\}. \quad (10)$$

Note that when $G(x) = 1/(1 + e^{-x})$, the log-likelihood function (10) can be rewritten as

$$\log L(\boldsymbol{\alpha}) = \sum_{k=0}^{m} \alpha_k \left(\sum_{t=0}^{\infty} f_t t^k \right) - \sum_{t=0}^{\infty} n_t \log \{1 + \exp(\boldsymbol{\alpha}'\mathbf{e}_t)\},$$

indicating that $(r_0, \ldots, r_m)$, where $r_k = \sum_{t=0}^{\infty} f_t t^k$, is sufficient for $(\alpha_0, \ldots, \alpha_m)$.

The maximum likelihood estimate $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ is obtained as the solution to the system of equations

$$\frac{\partial \log L(\alpha)}{\partial \alpha_k} = \sum_{t=0}^{\infty} t^k \{f_t u(\xi(t)) - s_t u(-\xi(t))\} = 0 \quad (k = 0, \ldots, m),$$

where

$$u(x) = \frac{\partial \log G(x)}{\partial x} = \frac{g(x)}{G(x)}. \quad (11)$$

Note that the hazard mass function estimate at time t is then given by $G(\hat{\boldsymbol{\alpha}}'\mathbf{e}_t)$.

The solution $\hat{\boldsymbol{\alpha}}$ can be obtained using the Newton–Raphson iterative procedure. This involves the observed information matrix $I(\alpha)$, whose (j,k) element is given by

$$-\frac{\partial^2 \log L(\alpha)}{\partial \alpha_j \partial \alpha_k} = \sum_{t=0}^{\infty} t^{j+k} \{ f_t v(\xi(t)) + s_t v(-\xi(t)) \} , \tag{12}$$

where

$$v(x) = \frac{\partial^2 \log G(x)}{\partial x^2} = \left[\frac{g(x)}{G(x)} \right]^2 - \frac{g'(x)}{G(x)} . \tag{13}$$

Note that if $G(x) = 1/(1+e^{-x})$, $u(x) = G(-x)$ and $v(x) = G(x)G(-x)$.

Exact distribution theory for the estimator $\hat{\boldsymbol{\alpha}}$ is not feasible. However, provided that certain regularity conditions on $\boldsymbol{\alpha}$ and the likelihood function are satisfied, standard likelihood theory provides asymptotic results, which can be used for inference on $\boldsymbol{\alpha}$. The asymptotic distribution of the estimator $\hat{\boldsymbol{\alpha}}$ is multivariate normal with mean $\boldsymbol{\alpha}$ and covariance matrix $\mathcal{I}(\boldsymbol{\alpha})^{-1}$, where $\mathcal{I}(\boldsymbol{\alpha}) = \mathrm{E}\left(I(\boldsymbol{\alpha})\right)$ is the Fisher information.

3.2. Estimating the Hazard Mass and Survivor Functions

In most situations the primary interest will be in estimating the hazard mass function (2) and the survivor function (5), and obtaining variance estimates and confidence intervals for these functions. Both functions involve the polynomial $\xi(t)$, which is estimated by $\hat{\xi}(t) = \hat{\boldsymbol{\alpha}}' \mathbf{e}_t$. The asymptotic variance of $\hat{\xi}(t)$ is $\mathbf{e}_t' \mathcal{I}(\boldsymbol{\alpha})^{-1} \mathbf{e}_t$. This is estimated by $\mathbf{e}_t' I(\hat{\boldsymbol{\alpha}})^{-1} \mathbf{e}_t$.

The estimated hazard mass function is then

$$\hat{h}(t) = G\big(\hat{\xi}(t)\big) = G(\hat{\boldsymbol{\alpha}}' \mathbf{e}_t) ,$$

with asymptotic variance

$$\mathrm{var}\left[\hat{h}(t)\right] = g\big(\xi(t)\big)^2 \mathbf{e}_t' \mathcal{I}(\boldsymbol{\alpha})^{-1} \mathbf{e}_t .$$

The variance of $\hat{h}(t)$ can be estimated by

$$\text{est. var}\left[\hat{h}(t)\right] = g\big(\hat{\xi}(t)\big)^2 \mathbf{e}_t' I(\hat{\boldsymbol{\alpha}})^{-1} \mathbf{e}_t .$$

An approximate 95% confidence interval for the hazard mass function at some specified time t is

$$\hat{h}(t) \pm 1.96 \big\{ \text{est. var}\left[\hat{h}(t)\right] \big\}^{1/2} . \tag{14}$$

For values of $\hat{h}(t)$ close to zero or one, such an approximate confidence interval may include impossible values outside the range [0,1]. This problem can be avoided by first obtaining a confidence interval for $\hat{\xi}(t)$ and then applying the function G. This leads to the approximate 95% confidence interval

$$G\left(\hat{\boldsymbol{\alpha}}'\mathbf{e}_t \pm 1.96\left[\mathbf{e}_t' I(\hat{\boldsymbol{\alpha}})^{-1}\mathbf{e}_t\right]^{1/2}\right) \tag{15}$$

for $h(t)$.

Similarly, the survivor function $\hat{S}(t)$ is also straightforward to estimate:

$$\hat{S}(t) = \prod_{s=0}^{t-1}\left[1 - G(\hat{\xi}(s))\right] = \prod_{s=0}^{t-1} G(-\hat{\xi}(s)) .$$

However, estimating the variance of $\hat{S}(t)$ is not so straightforward. Using

$$\operatorname{var}\left[\log \hat{S}(t)\right] = \sum_{r=0}^{t-1}\sum_{s=0}^{t-1} \operatorname{cov}\left[\log G(-\hat{\xi}(r)), \log G(-\hat{\xi}(s))\right]$$

and the usual approximations, we obtain

$$\operatorname{var}\left[\hat{S}(t)\right] \simeq S(t)^2 \sum_{r=0}^{t-1}\sum_{s=0}^{t-1} u(-\xi(r))u(-\xi(s))\mathbf{e}_r'\mathcal{I}(\boldsymbol{\alpha})^{-1}\mathbf{e}_s .$$

Thus an estimate of the variance of $\hat{S}(t)$ is

$$\text{est. var}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{r=0}^{t-1}\sum_{s=0}^{t-1} u(-\hat{\xi}(r))u(-\hat{\xi}(s))\mathbf{e}_r' I(\hat{\boldsymbol{\alpha}})^{-1}\mathbf{e}_s = \hat{S}(t)^2 R(t) . \tag{16}$$

For large values of t calculation of (16) will be (computer) time consuming since it involves a double summation from 0 to $t-1$ for each value of t. However, (16) can be calculated iteratively since

$$\begin{aligned} R(t+1) = R(t) &+ \left[u(-\hat{\xi}(t))\right]^2 \mathbf{e}_t' I(\hat{\boldsymbol{\alpha}})^{-1}\mathbf{e}_t \\ &+ 2u(-\hat{\xi}(t)) \sum_{s=0}^{t-1} u(-\hat{\xi}(s))\mathbf{e}_t' I(\hat{\boldsymbol{\alpha}})^{-1}\mathbf{e}_s , \end{aligned}$$

which involves only a single summation from 0 to $t-1$ for each value of t and so results in a considerable saving for large values of t.

Confidence intervals for the survivor function can also be calculated by two methods. The first assumes that the asymptotic distribution of $\hat{S}(t)$ is normal, so that an approximate 95% confidence interval for $S(t)$ may be calculated as

$$\hat{S}(t) \pm 1.96\{\text{est. var}\,[\hat{S}(t)]\}^{1/2}\,. \tag{17}$$

As in the hazard function case, this approximate confidence interval may contain impossible values for extreme values of $\hat{S}(t)$. The second method involves using the transformation $\varphi(t) = \log(-\log S(t))$ for which the range is unrestricted. The asymptotic variance of $\hat{\varphi}(t)$ is estimated by

$$\text{est. var}\,[\hat{\varphi}(t)] = \frac{\text{est. var}\,[\hat{S}(t)]}{[\hat{S}(t)\log\hat{S}(t)]^2} = \frac{R(t-1)}{[\log\hat{S}(t)]^2}\,,$$

and a 95% confidence interval for $\varphi(t)$ is $\hat{\varphi}(t) \pm 1.96\text{est. var}\,[\hat{\varphi}(t)]\}^{1/2}$. The corresponding 95% confidence interval for $S(t)$ is thus

$$[\hat{S}(t)]^{\exp\{\pm 1.96\{\text{est. var}\,[\hat{\varphi}(t)]\}^{1/2}\}}\,, \tag{18}$$

which takes values in [0,1].

3.3. Choosing the Degree m of the Polynomial

One problem with using a polynomial function for $\xi(t)$ is the selection of the degree m of the polynomial. Unless there are *a priori* grounds for the specification of m, the smallest m which gives a satisfactory fit to the data should be used. In our experience a satisfactory fit is usually achieved by a linear or quadratic polynomial. In fact, it seems rarely worth considering models with $m > 3$.

Problems of a similar kind arise in connection with polynomial regression models. Procedures utilized for this purpose have been described in many of the texts on linear models. Among these, the so-called step-up (or forward) and step-down (or backward) procedures are most common. The F-tests used in such procedures are in fact the likelihood ratio tests. For failure time data, procedures for selecting a subset of covariates based on likelihood ratio tests have been discussed by Greenberg, Bayard & Byar (1974) and Elandt-Johnson & Johnson (1980, Section 13.8).

The degree m can be selected using a likelihood ratio step-up procedure. The significance of the α_k-term can be tested using the likelihood

TABLE 1

Days until the development of vaginal cancer in rats

142,	156,	163,	198,	204*,	205,	232,	233,	233,	233,
233,	239,	240,	261,	280,	296,	296,	323,	344*,	

* Censored.

ratio statistic, which under the hypothesis $\alpha_k = 0$, is asymptotically distributed as χ^2_1. The step-up procedure is continued until two successive terms are non-significant. The degree of the polynomial is then chosen to be the last value of k for which the inclusion of the α_k-term resulted in a significant increase in the maximized log-likelihood function. This is a standard termination procedure for polynomial regression models (Fruend & Minten, 1979, p.168), and was also advocated by Taulbee (1979) for his polynomial hazard model.

3.4. An Example

The proposed model (2) is demonstrated by application to a data set from Pike (1966) involving a laboratory investigation into vaginal cancer in female rats.

Table 1 gives the number of days T from insult with the carcinogen DMBA until the appearance of a carcinoma. At the time the data were collected only 19 out of the 21 rats had developed cancer, so that two of the times in Table 1 (marked *) are censoring times. Two major problems are encountered when analysing these data. Firstly, there are a large number of tied values and secondly, there appears to be a threshold period in which the carcinoma cannot appear. To handle this second problem, Pike and others (e.g., Kalbfleisch & Prentice, 1980; Lawless, 1982) have used $T-100$ in their analyses. Our model (2) is sufficiently flexible to handle this initial failure-free period and, of course, ties present no problem.

The model (2) with polynomials $\xi(t)$ up to degree four were fitted and the step-up procedure described in Section 3.3 used to determine the degree of the polynomial. Table 2 gives the resulting test values $\Lambda(\alpha_k)$, to be referred to a χ^2_1 distribution. Examination of the table shows that the inclusion of cubic and quartic degree terms into the model does not lead to a significant increase in the maximized log-likelihood function and so a quadratic polynomial is sufficient for $\xi(t)$. The estimated hazard model is

$$\hat{h}(t) = \frac{\exp(-14.8 + 0.0736t - 0.000122t^2)}{1 + \exp(-14.8 + 0.0736t - 0.000122t^2)} . \qquad (19)$$

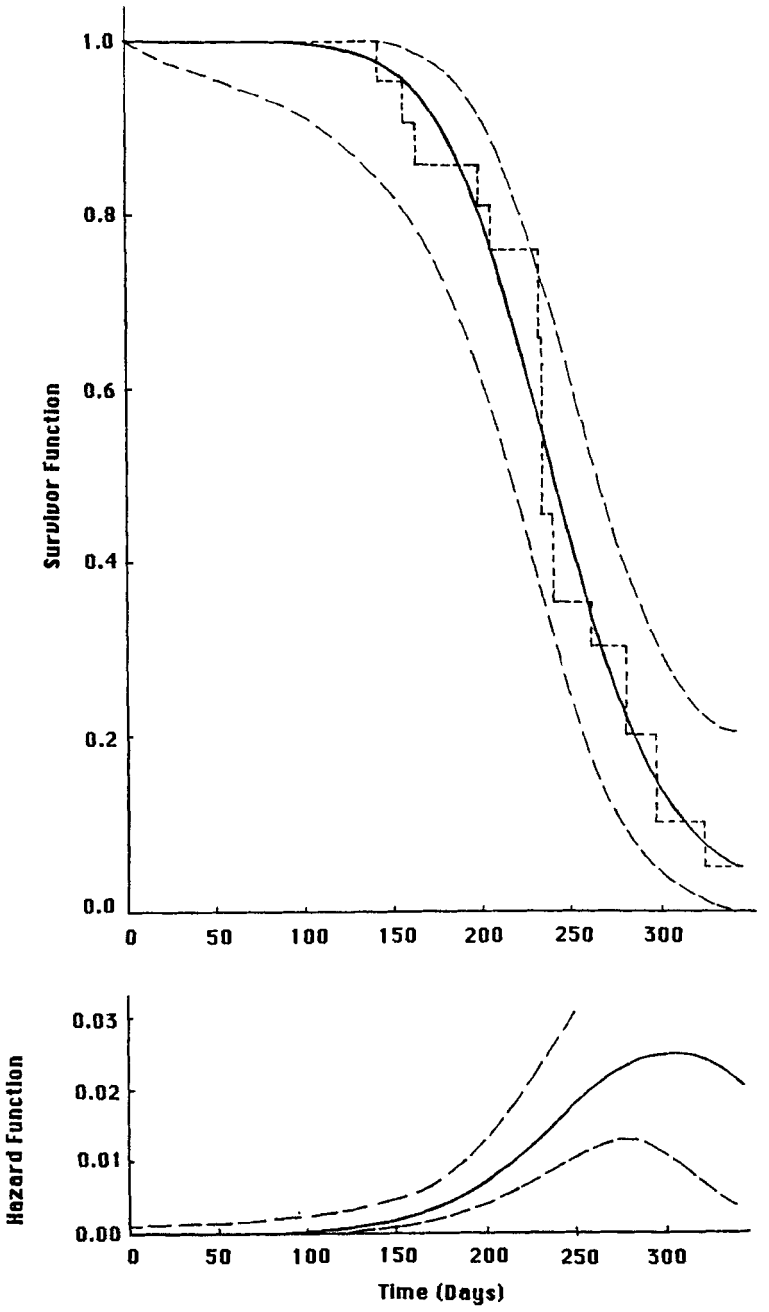

Fig. 1.—Estimates and confidence intervals for the hazard mass function and the survivor function for the carcinogenesis data.

TABLE 2

Results of likelihood ratio tests for determining the degree m of the polynomial $\xi(t)$ for the carcinogenesis data

Model	Log-likelihood	$\Lambda(\alpha_k)$
Constant	−125.013	
Linear	−106.214	37.60
Quadratic	−104.234	3.96
Cubic	−104.229	0.01
Quartic	−104.054	0.35

The survivor function is estimated by

$$\hat{S}(t) = \prod_{s=0}^{t-1} \left[1 + \exp(-14.8 + 0.0736s - 0.000122s^2)\right]^{-1} . \qquad (20)$$

Figure 1 shows the estimated hazard mass function (19) and also the survivor function (20) with the product limit estimate for comparison. It is seen that the quadratic model provides a good fit to the empirical survival probabilities. The figure also shows the 95% confidence bands for the hazard mass function obtained from (15) and for the survivor function obtained from (18). It is observed that the upper confidence limit for the hazard function increases sharply for $t > 300$, reflecting the uncertainty involved in hazard function estimation based on small numbers of subjects.

4. Allowing for Differences Between Subjects

In Section 2 the distribution of T was assumed to be the same for all subjects. Here it is allowed to differ between subjects. Each subject k is assumed to have associated with it a vector $\mathbf{z}_k = (z_{k_1}, \ldots, z_{k_p})'$ of p observed concomitant variables (or covariates). The vector $\mathbf{z}_k$ may include both quantitative variables, such as age, and qualitative variables, such as sex and treatment group. The hazard mass function for subject k is then denoted by $h(t; \mathbf{z}_k)$ and the survivor function by $S(t; \mathbf{z}_k)$.

Discrete failure time models that allow for differences between subjects by the inclusion of covariate information are basically of two types. The first type consists of models obtained by grouping of the continuous time proportional hazards model (Cox, 1972), and can be characterized by the linear model

$$\log\left[-\log h(t; \mathbf{z})\right] = \gamma_t + \boldsymbol{\beta}'\mathbf{z} \quad (t = 0, 1, \ldots) , \qquad (21)$$

where $\gamma_t = \log\left[-\log h(t;0)\right]$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. Although the model (21) retains the same relative risk parameter, $\exp(\boldsymbol{\beta}'\mathbf{z})$, as the continuous time proportional hazards model, the resultant hazard mass functions are not proportional. Models of this type have been proposed by Kalbfleisch & Prentice (1973), who used a marginal likelihood approach to estimate $\boldsymbol{\beta}$, and Prentice & Gloeckler (1978), who estimated $\boldsymbol{\beta}$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots)'$ from the joint likelihood.

The second type is characterized by the linear model

$$\log\left\{\frac{h(t;\mathbf{z})}{1-h(t;\mathbf{z})}\right\} = \mu_t + \boldsymbol{\beta}'\mathbf{z} \quad (t = 0, 1, \ldots)\,, \tag{22}$$

where $\mu_t = \log\left\{h(t,0)/[1-h(t;0)]\right\}$, which results in a linear logistic model with an arbitrary logistic location parameter μ_t for the hazard mass function. The term $\exp(\boldsymbol{\beta}'\mathbf{z})$ now represents the odds ratio of failure at time t rather than the relative risk of failure at time t. This model was proposed by Cox (1972), who used a partial likelihood approach to estimate $\boldsymbol{\beta}$. However, calculation of the partial likelihood requires excessive compuation when there are large numbers of ties at the failure times. Thompson (1977) proposed obtaining estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots)'$ by maximization of the joint likelihood. He also allowed the covariates to be time dependent. Arjas (1985) and Arjas & Haara (1987) have considered a dynamic form of the model (22) that incorporates time dependent covariates and which does not align observations to some common starting point. No baseline hazard function is specified — all dependencies on time-related quantities being accommodated into the covariates.

Maximum likelihood estimation from the joint likelihood is relatively straightforward for either of the models (21) and (22) (see Lawless, 1982, Section 7.3). The approach is again distribution-free in the sense that no parametric form is assumed for the baseline hazard mass function $h(t;0)$. Both models can be fitted using standard packages such as GLIM or BMDP, and for discrete data (or grouped data with fine intervals) should yield similar estimates of the $\boldsymbol{\beta}$ coefficients. However, for discrete data the dimensions of $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ will be large if the study runs for a long time, and so the models (21) and (22) can easily become overparameterized. Also any estimation method that requires calculation of the matrix of second partial derivatives to obtain estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ or $\boldsymbol{\mu}$, will result in an excessive amount of computation.

A third type of model for the hazard mass function is proposed which is a generalization of the model (2) of Section 2 to incorporate covariates.

Let the baseline hazard mass function be such that

$$h(t;\mathbf{0}) = G\big(\xi(t)\big) \quad (t = 0, 1, \ldots) ,$$

where G satisfies (1). Then the hazard mass function for subject k with covariate vector $\mathbf{z}_k$ is defined as

$$h(t;\mathbf{z}_k) = G\big(\xi(t) + \boldsymbol{\beta}'\mathbf{z}_k\big) \quad (t = 0, 1, \ldots) . \tag{23}$$

The function $\xi(t)$ can be approximated by the polynomial (3), so that

$$h(t;\mathbf{z}_k) = G(\boldsymbol{\alpha}'\mathbf{e}_t + \boldsymbol{\beta}'\mathbf{z}_k) = G(\boldsymbol{\theta}'\mathbf{x}_{tk}) ,$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})'$ and $\mathbf{x}_{tk} = (\mathbf{e}_t, \mathbf{z}_k)'$.

Again, the most suitable form for the function $G(x)$ has been found to be the logistic function, $G(x) = 1/(1 + e^{-x})$. This choice for $G(x)$ combines tractability with computational simplicity, and also seems to be a reasonable analogue of the exponential function used for the continuous time proportional hazards model. Also, as was found in Section 3, there are convenient sufficient statistics for $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})'$ in the logistic case. Furthermore, this choice of $G(x)$ means that the model (23) corresponds to the linear log-odds model (22) proposed by Cox (1972); the advantages being a smaller number of parameters and the natural inclusion of a parametric baseline hazard of the type introduced in Section 2. Note, that in the logistic case, the model (23) is similar to the model used by Dinse & Lagakos (1983) for analysing tumor prevalence data.

Let $\mathcal{F}_t$ denote the set of subjects which fail on the t th day. Then the log-likelihood corresponding to the model (23) is given by

$$\log L(\boldsymbol{\theta}) = \sum_{t=0}^{\infty} \left\{ \sum_{k \in \mathcal{F}_t} \log G(\boldsymbol{\theta}'\mathbf{x}_{tk}) + \sum_{k \in \overline{\mathcal{F}}_t} \log G(-\boldsymbol{\theta}'\mathbf{x}_{tk}) \right\} , \tag{24}$$

while the first and second derivatives of the log-likelihood are now

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_i} = \sum_{t=0}^{\infty} x_{ikt} \left\{ \sum_{k \in \mathcal{F}_t} u(\boldsymbol{\theta}'\mathbf{x}_{kt}) - \sum_{k \in \overline{\mathcal{F}}_t} u(-\boldsymbol{\theta}'\mathbf{x}_{kt}) \right\}$$

$$\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = \sum_{t=0}^{\infty} x_{ikt} x_{jkt} \left\{ \sum_{k \in \mathcal{F}_t} v(\boldsymbol{\theta}'\mathbf{x}_{kt}) - \sum_{k \in \overline{\mathcal{F}}_t} v(-\boldsymbol{\theta}'\mathbf{x}_{kt}) \right\} ,$$

where the functions $u(x)$ and $v(x)$ are given by (11) and (13), respectively.

As in the homogeneous case, the parameter $\boldsymbol{\theta}$ can be estimated by the maximum likelihood method. The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ must be obtained by iteration. A suitable choice for the initial values is $\hat{\boldsymbol{\theta}}_0 = (\hat{\boldsymbol{\alpha}}(0), 0)'$, where $\hat{\boldsymbol{\alpha}}(0)$ is the maximum likelihood estimate obtained assuming that $\boldsymbol{\beta} = 0$ (i.e., the maximum likelihood estimates of $\boldsymbol{\alpha}$ for the model (2) of Section 2).

The estimation procedure with covariates is considerably more (computer) time consuming than for the model (2), since for each time t, the contribution to the log-likelihood involves a sum over all subjects k still in the risk set $\mathcal{R}_t$. However, as Arjas (1985) pointed out, a valuable computational aspect of this type of model is that the contribution of subject k on day t to the likelihood can be treated separately from the contribution on day $t+1$, and also from the contribution of subject j on day t, each contribution only adding a term to the log-likelihood (24). Thus there is no need to form, as in Cox's proportional hazard model, risks sets consisting of subjects with matching baseline hazards, while also keeping track on their time-dependent covariates. Furthermore, as was remarked earlier, no tie breaking procedures are necessary. Thus the calculations required to evaluate the log-likelihood (24) are straightforward when compared with the calculations required to evaluate the corresponding likelihood for the continuous time proportional hazards model in the presence of ties. (See Kalbfleisch & Prentice (1980, Chapter 4) for details of calculations for the continuous model.)

Inference on $\boldsymbol{\theta}$ is based on standard likelihood asymptotic results. For example, to test the hypothesis $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$, where $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)'$, either the Wald statistic

$$\Delta(\boldsymbol{\theta}_2) = (\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{20})' I^{22}(\hat{\boldsymbol{\theta}})^{-1} (\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{20}) \tag{25}$$

where $I^{22}(\boldsymbol{\theta})$ denotes the sub-matrix of the inverse of the information matrix corresponding to $\boldsymbol{\theta}_2$, or the likelihood ratio statistic

$$\Lambda(\boldsymbol{\theta}_2) = -2 \log \left\{ \frac{L[\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_{20}), \boldsymbol{\theta}_{20}]}{L(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)} \right\} \tag{26}$$

can be used. In particular, the effects of the covariates can be assessed by setting $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)' = (\boldsymbol{\alpha}, \boldsymbol{\beta})'$ and then using either (25) or (26). Similarly selection of a subset of covariates can be done using either a step-down or a step-up procedure based on the likelihood ratio statistic (26). The k th

TABLE 3

Remission times (weeks) of leukæmia patients

Drug 6-MP	6,	6,	6,	6*	7,	9*	10,	10*	11*	13,	16,
	17*	19*	20*	22,	23	25*	32*	32*	34*	35*	
Placebo	1,	1,	2,	2,	3,	4,	4,	5,	5,	8,	8,
	8,	8,	11,	11,	12,	12,	15,	17,	22,	23	

* Censored.

component of $\boldsymbol{\beta}$ is tested by setting $(\theta_1, \theta_2)' = (\alpha, \beta_{-k})'$, where $\beta_{-k} = (\beta_1, \ldots, \beta_{k-1}, \beta_{k+1}, \ldots, \beta_p)'$, for the step-down procedure and by setting $(\theta_1, \theta_2)' = (\alpha, \beta_k)'$ for the step-up procedure.

Expressions for estimates and confidence intervals for the hazard mass and survivor functions are also readily obtained from the corresponding results in Section 3.2 by substituting $\boldsymbol{\theta}$ for $\boldsymbol{\alpha}$ and $\mathbf{x}_{kt}$ for $\mathbf{e}_t$. Note that estimates of the hazard mass and survivor functions can be obtained for any covariate pattern, and that if the covariate values in $\mathbf{z}_k$ are centred, the estimated of the hazard mass and survivor functions obtained for the model (2) are estimates of the baseline hazard mass and survivor functions.

The use of the model (23) is illustrated with two numerical examples.

Example 1. Freireich Leukæmia Data

The clinical trial data of Freireich *et al.* used by Gehan (1965), Cox (1972) and many others are analysed here. The data are remission times, measured in weeks, in two groups of 21 acute leukæmia patients. The trial compared the drug 6-metcaptopurine (6-MP) with a placebo with respect to the ability to maintain remission. Table 3 shows the ordered times for the two samples; censored values are indicated by asterisks. There are 12 censored observations, all falling in the 6-MP group. There are also numerous tied observations providing difficulties for other methods, although not for the present approach.

Preliminary computations suggested that a constant baseline hazard mass function is sufficient. Similar conclusions have also been reached by a number of other authors (e.g., Taulbee, 1979; Lawless, 1982; Cox & Oakes, 1984) who found that the data could be adequately fitted by an exponential failure time model. The model (23) with a constant baseline hazard and a single covariate indicating treatment group was fitted. The treatment group indicator variable z_k took values -1 for the placebo group

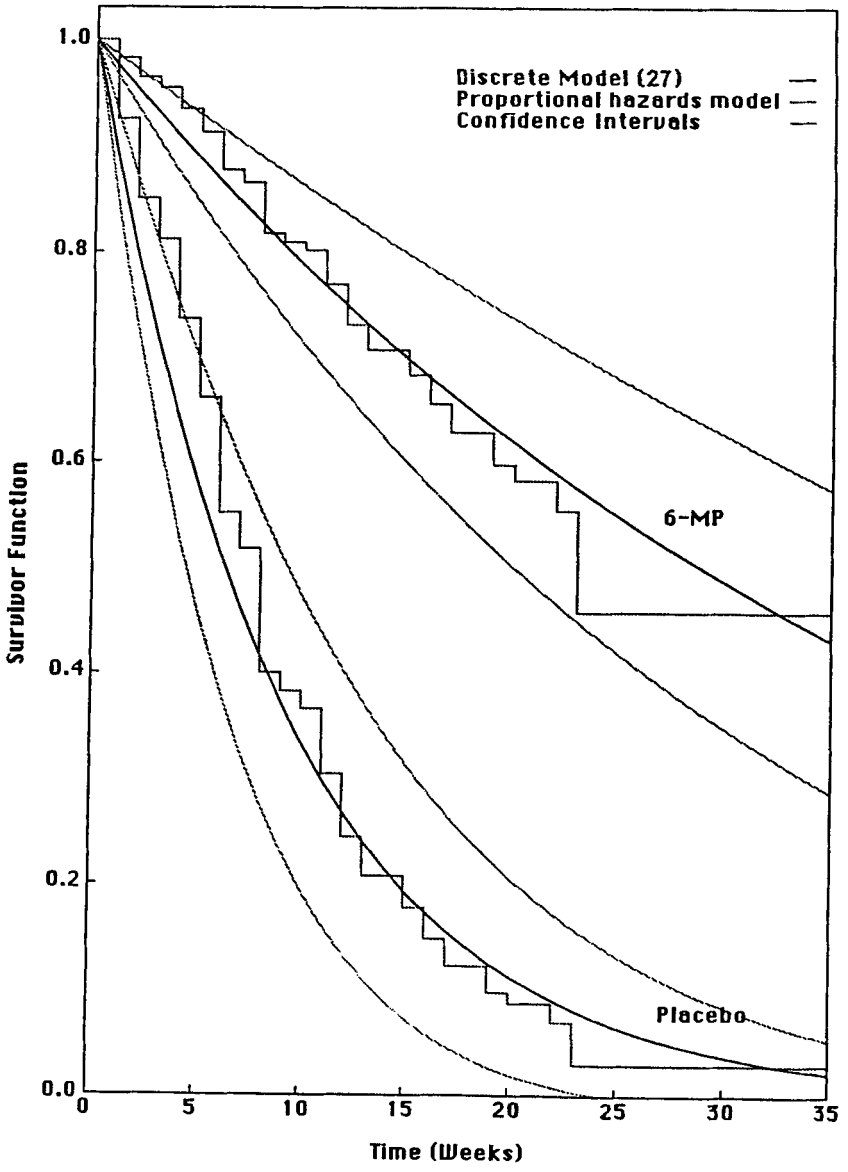

Fig. 2.—Survivor function estimates for the discrete model (27) and for the proportional hazards model for the Freireich Leukæmia data.

and +1 for the 6-MP group. The fitted hazard model is

$$h(t; \mathbf{z}_k) = \frac{\exp(-2.94 - 0.780 z_k)}{1 + \exp(-2.94 - 0.780 z_k)}, \tag{27}$$

and the estimated variance of maximum likelihood estimate $\hat{\beta}$ is 0.0416. The test of the hypothesis $\beta = 0$ based on the Wald statistic (25) gave $\Delta(\beta) = 14.62$, while the test based on the likelihood ratio statistic (26) gave $\Lambda(\beta) = 16.24$, both values being referred to a χ^2_1 distribution. Obviously β is significantly different from zero, and so there is a significant difference in the hazard rates of the two groups. A similar conclusion was reached by Cox (1972) who obtained the likelihood ratio of 14.85 for the test of the hypothesis $\beta = 0$. The odds ratio of failure at time t of the 6-MP group versus the placebo group is

$$\left\{\frac{h(t;+1)}{1-h(t;+1)}\right\}\left\{\frac{h(t;-1)}{1-h(t;-1)}\right\} = \exp(2\beta) = 4.76 ,$$

which is comparable with the odds ratio of 5.21 found by Cox (1972). Figure 2 compares the survivor function estimates $\hat{S}(t;+1)$ and $\hat{S}(t;-1)$ for the discrete model (27) with the corresponding estimates from the (continuous time) proportional hazards model. Also shown are the approximate 95% confidence intervals for the survivor functions $S(t;+1)$ and $S(t;-1)$.

Example 2. VA Lung Cancer Data

In the final example the U.S. Veterans Administration lung cancer trial data presented by Prentice (1973) are examined. In this trial, 137 males with advanced inoperable lung cancer were randomized to either a standard or a test chemotherapy. The data represent the time to death (in days). Only 9 of the 137 survival times were censored. The model (23) was fitted to the data with eight covariates and a linear baseline hazard. Single indicator variables distinguished treatment and prior therapy groups, and three indicator variables distinguished the four histological types of tumour (squamous, small cell, adeno and large cell). The other three covariates were performance status, number of months from diagnosis to entry into the study and the patient's age in years.

The results are summarized in Table 4. The table shows the Wald statistics (25) and the likelihood ratio statistics (26) for tests of various submodels, obtained by omitting each covariate in turn, against the full model with all eight covariates present. Also shown in Table 4 are the corresponding Wald and likelihood ratio statistics for the proportional hazards

TABLE 4

Asymptotic likelihood inference on lung cancer data

Covariate Omitted	Discrete Model (26) Wald Statistic	Discrete Model (26) Likelihood Ratio Statistic	Proportional Hazards Model Wald Statistic	Proportional Hazards Model Likelihood Ratio Statistic
Performance status	34.99	33.73	33.76	33.21
Disease duration	0.01	0.01	0.00	0.00
Age	0.34	0.34	0.80	0.79
Prior therapy	0.01	0.01	0.12	0.12
Cell type				
Squamous *v.* large				
Small cell *v.* large	17.86	18.61	18.02	18.71
Adeno *v.* large				
Treatment	1.15	1.15	1.86	1.87

model fitted to the data using the statistical package BMDP. There is reasonable agreement between the test statistic values for the discrete model and the proportional hazards model.

5. Concluding Remarks

The advantages of the discrete failure time models (2) and (23) proposed here are that they are capable of approximating a wide variety of hazard curves while remaining easy to fit and simple to interpret. Interpretation is aided by the smooth estimates for both the hazard mass function and the survivor functions that are a feature of these models.

It is when the covariate vector contains time dependent covariates that discrete models of the form (23) are superior. If $\mathbf{z}_{kt}$ denotes the covariate vector for subject k on the tth day, then the hazard mass function for subject k can still be written

$$h(t;\mathbf{z}_k) = G(\boldsymbol{\theta}'\mathbf{x}_{kt}) ,$$

where now $\mathbf{x}_{kt} = (\mathbf{e}_t, \mathbf{z}_{kt})'$. The estimation procedures and inference results given in Section 4 remain unchanged. Furthermore, the inclusion of time dependent covariates results in only a slight increase in the amount of computation required to calculate the log-likelihood and its derivatives;

this increase being that involved in evaluating the covariates at each time point. The inclusion of time dependent covariates allows the hazard mass functions for subjects with different covariate vectors to cross.

In the model (23) the effect of the covariates was to act additively on the function $\xi(t)$. An alternative model, in which the covariates act multiplicatively on the function $\xi(t)$, specifies that

$$h(t;\mathbf{z}_k) = G\big(\xi(t)\varphi(\mathbf{z}_k,\beta)\big) \tag{28}$$

where $\varphi(\mathbf{z}_k,\beta)$ is a known function of $\mathbf{z}_k$ and β, and G satisfies (1). Some possible forms for the function $\varphi(\mathbf{z}_k,\beta)$ are: $\exp(\boldsymbol{\beta}'\mathbf{z}_k)$, $(1+\boldsymbol{\beta}'\mathbf{z}_k)$, and $(1+\boldsymbol{\beta}'\mathbf{z}_k)^{-1}$ (e.g., Feigl & Zelen, 1965; Taulbee, 1979). On approximating the function $\xi(t)$ by the polynomial (3), the model (28) becomes

$$h(t;\mathbf{z}_k) = G\big(\alpha_0\varphi(\mathbf{z}_k,\boldsymbol{\beta}_0) + \alpha_1\varphi(\mathbf{z}_k,\boldsymbol{\beta}_1)t + \cdots + \alpha_m\varphi(\mathbf{z}_k,\boldsymbol{\beta}_m)t^m\big)\ .$$

For this model the hazard mass functions are allowed to cross either by allowing the parameter vector $\boldsymbol{\beta}$ in the function $\varphi(\mathbf{z}_k,\boldsymbol{\beta})$ to differ between powers of t, or by using time dependent covariates. The model is similar to the polynomial model proposed by Taulbee (1979) for the continuous hazard density function.

Acknowledgements

The major part of this paper was taken from the first author's M.Sc. thesis written at the University of Melbourne, under the supervision of the second author. Financial support for the first author was provided by a Commonwealth Postgraduate Research Award from the Australian Commonwealth Department of Education, and a University of Western Australia Postgraduate Scholarship.

References

ARJAS, E. (1985). Stanford heart transplantation data revisited: a real time approach. In *Modern Statistical Methods in Chronic Disease Epidemiology*. New York: Wiley.

ARJAS, E. & HAARA, P. (1987). A logistic regression model for hazard: asymptotic results. *Scand. J. Statist.* **14**, 1–18.

BAIN, L.J. (1974). Analysis for the linear failure rate distribution. *Technometrics* **16**, 551–559.

CLAYTON, D.G. (1983). Fitting a general family of failure time distributions using GLIM. *Appl. Statist.* **32**, 102–109.

COX, D.R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187–202.

COX, D.R. & OAKES, D. (1984). *Analysis of Survival Data*. London: Chapman & Hall.

DINSE, G.E. & LAGAKOS, S.W. (1983). Regression analysis of tumor prevalence data. *Appl. Statist.* **32**, 236–248.

ELANDT-JOHNSON, R.C. & JOHNSON, N.L. (1980). *Survival Models and Data Analysis*. New York: Wiley.

FEIGL, P. & ZELEN, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics* **21**, 826–838.

FEUND, R.J. & MINTEN, P.D. (1979). *Regression Methods*. New York: Marcel Dekker.

GEHAN, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 203–223.

GEHAN, E.A. & SIDDIQUI, M.M. (1973). Simple regression methods for survival time studies. *J. Amer. Statist. Assoc.* **68**, 848–856.

GREENBERG, R., BAYARD, S. & BYAR, D. (1974). Selecting concomitant variables using a likelihood ratio step-down procedure and a method of testing goodness of fit of an exponential survival model. *Biometrics* **30**, 601–608.

GROSS, A.J. & CLARK, V.A. (1975). *Survival Distributions: Reliability Application in the Biomedical Sciences*. New York: Wiley.

KALBFLEISCH, J.D. & PRENTICE, R.L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* **60**, 267–279.

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

KAPLAN, E.L. & MEIER, P. (1958). Nonparametric estimations from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.

KODLIN, D. (1967). A new response time distribution. *Biometrics* **23**, 227–239.

KRANE, S.A. (1963). Analysis of survival data by regression techniques. *Technometrics* **5**, 161–174.

LAWLESS, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: Wiley.

PIKE, M. (1966). A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics* **22**, 142–161.

PRENTICE, R.L. (1973). Exponential survival with censoring and explanatory variables. *Biometrika* **60**, 279–288.

PRENTICE, R.L. & GLOECKLER, L.A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrika* **34**, 57–67.

TAULBEE, J.D. (1979). A general model for the hazard rate with covariables. *Biometrics* **35**, 439–450.

THOMPSON, W.J., Jr. (1977). On the treatment of grouped observations in life studies. *Biometrics* **33**, 463–470.

Received July 1988; revised December 1988