BAYESIAN INFERENCE FOR SURVIVAL DATA WITH NONPARAMETRIC HAZARDS AND VAGUE PRIORS

by

Michael J. Symons

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1491

October 1985

BAYESIAN INFERENCE FOR SURVIVAL DATA WITH NONPARAMETRIC HAZARDS AND VAGUE PRIORS

Michael J. Symons

Department of Biostatistics
University of North Carolina
Chapel Hill, North Carolina 27514

ABSTRACT

Statistical inference is reviewed for survival data applications with hazard models having one parameter per distinct failure time and using Jeffreys' (1961) vague priors. Distinction between a discrete hazard and a piecewise exponential model is made. Bayes estimators of survival probabilities are derived. For a single sample and a discrete hazard, the Bayes estimator is shown to be larger than Nelson's (1972) which in turn is larger than Kaplan-Meier's (1958) estimator. With a piecewise exponential model, the Bayes estimator is also shown to be larger than that using maximum likelihood. Presuming a proportional hazards formulation to incorporate covariate information and a discrete underlying hazard model, the marginal posterior distribution of the regression parameters is proportional to Breslow's (1974) approximation to the marginal likelihood of Kalbfleisch and Prentice (1973). A refinement of Breslow's (1974) approximate likelihood is obtained when a piecewise exponential model is used for the underlying hazard. These results serve as illustrations of differences between estimators obtained from a frequentist's approach and a Bayes strategy with vague priors. Further, the Bayes results have practical advantages.

KEY WORDS: Discrete Hazard, Kaplan-Meier Estimator, Nelson's Method, Martingales, Piecewise Exponential Hazard, Nuisance Parameters, Maximum Likelihood.

1. Introduction and Summary

For a positive failure time T of an individual with p measured covariables $\underset{\sim}{z} = (z_1,\ldots,z_p)$, Cox (1972) proposed that the distribution of T could be modelled by specifying the hazard of T given $\underset{\sim}{z}$ as

$$\lambda(t|\underset{\sim}{z}) = \lim_{\Delta t \to 0} \Pr\{T < t + \Delta t \mid T > t, \underset{\sim}{z}\}/\Delta t = \lambda_0(t) \exp(\underset{\sim}{z}\underset{\sim}{\beta}), \qquad (1)$$

where $\underset{\sim}{\beta}$ is a column vector of p regression coefficients and $\lambda_0(t)$ is an arbitrary, unspecified underlying hazard function. Two nonparametric models are considered for the underlying hazard. Each model permits one parameter for every distinct failure time, i.e., models whose parameter spaces are not fixed but depend upon the data. A discrete hazard, having only a finite spike at each failure time, and a piecewise exponential hazard, specifically with a constant risk between failures, are considered in turn. For the parameters of each underlying hazard and the regression coefficients, vague priors following Jeffreys (1961) are combined with the likelihood of the data.

The data in survival applications are from n individuals placed on test. Let k ($\leq$ n) be the number of distinct failure times with possible multiplicity d_j ($\geq$ 1) at the jth ordered failure time $t_{(j)}$. The times of censoring may also be available, but the censoring mechanism is assumed independent of the failure process and so is not informative for the regression coefficients $\underset{\sim}{\beta}$. Consequently, the other primarily pertinent data are the sets, $\mathcal{R}_j$, of individuals alive just before $t_{(j)}$ and having size R_j. For the piecewise exponential model, the survival times to censoring are also incorporated. The likelihood for the jth failure time is given by

$$L_j = \prod_{\ell \in \mathcal{R}_{j-1}} \left[\exp\left\{-\int_{t_{(j-1)}}^{t_{(j)}} \lambda(u|\underset{\sim}{z})du\right\} \lambda\left(t_{(j)}|\underset{\sim}{z}\right)^{\delta_{j\ell}}\right], \qquad (2)$$

where $\delta_{j\ell} = 1$ for each individual failing at $t_{(j)}$ and zero otherwise. Notice that $t_{(0)} = 0$ and $\mathcal{R}_0$ is composed of all n individuals placed on test. Since $\mathcal{R}_{j-1}$ contains $\mathcal{R}_j$, the likelihood (2) will suffice for each of the two nonparametric hazards considered in detail later.

For a single sample without covariables, Bayes estimators are derived for each underlying hazard. With the discrete hazard, the posterior mean of the survival probability is at least as large as the Kaplan-Meier (1958) or Nelson (1972) estimators. With a piecewise exponential hazard, a Bayes estimator is at least as large as that obtained by maximum likelihood. Variance estimators and an example are provided.

For inference on the regression coefficients, the underlying hazard is composed of nuisance parameters. As a result, the marginal posterior distributions for $\underset{\sim}{\beta}$ are determined for each underlying hazard model and using only vague prior information. With the discrete hazard model and distinct failures, Cox's (1972) conditional likelihood is obtained. If there are multiplicities, Breslow's (1974) approximation to Kalbfleisch and Prentice's (1973) marginal likelihood is recovered. With a piecewise exponential underlying hazard, a refinement of Breslow's (1974) approximate likelihood is the basis for inference.

2. Single Sample

The single sample problem without concomitant information corresponds to the situation where the covariates are the same for each individual, or without loss of generality, $\underset{\sim}{z}$ is identically a vector of zeroes.

2.1 Inference on a Discrete Underlying Hazard

Consider a discrete hazards model, i.e., one where the cumulative hazard $\Lambda(t)$ has a positive, finite jump at each failure time. Specifically,

$$h(t) = \Lambda(t) - \Lambda(t\text{-}o) = \begin{cases} \lambda_j & \text{, at time } t_{(j)}; \\ 0 & \text{, at other times.} \end{cases} \tag{3}$$

Alternatively, this is an expression of unit mass at the jth ordered failure time $t_{(j)}$ for a weight function $G_j(u)$ so that

$$\Lambda\big(t_{(j)}\big) - \Lambda\big(t_{(j-1)}\big) = \int_{t_{(j-1)}}^{t_{(j)}} \lambda(u)dG_j(u) = h\big(t_{(j)}\big) = \lambda_j \ , \tag{4}$$

noting $t_{(0)} = 0$. This imposes a step function shape on the cumulative hazard and survival distribution and is meant to be an innocuous representation of either function. Further, the weight function $G_j(u)$ makes superfluous any partial survival experience of those items censored between $t_{(j-1)}$ and $t_{(j)}$. It then formalizes the convention of discarding survival time for censored data between failures, for example as used by Kalbfleisch and Prentice (1972) and Breslow (1972). At the other extreme is the convention used by Mantel (1966), where survival to time $t_{(j+1)}$ is credited to all those surviving the jth failure time.

The likelihood for the k increments: $\underset{\sim}{\lambda} = (\lambda_1,\ldots,\lambda_k)$ to the cumulative hazard is the following product over n individuals:

$$L(\underset{\sim}{\lambda}) = \prod_{i=1}^{n} \left[\exp\left\{-\int_{o}^{t_i} \lambda(u)du\right\} \lambda(t_i)^{\delta_i}\right]. \tag{5}$$

In words, each individual's contribution is the survival probability through their follow-up time to t_i and their failure probability at t_i if they failed: $\delta_i = 1$. Otherwise $\delta_i = 0$ indicates censoring at t_i and their likelihood contribution is their probability of survival to t_i. This likelihood can be organized as a product of factors, each as in (2) with $\underset{\sim}{z} = \underset{\sim}{0}$, at each failure time. With the discrete hazard described in (3) and (4), the likelihood becomes

$$L(\underset{\sim}{\lambda}) = \prod_{j=1}^{k} \left[\exp(-\lambda_j R_j)\ \lambda_j^{d_j}\right]. \tag{6}$$

Following Lindley (1965), the posterior distribution of $\underset{\sim}{\lambda}$ is obtained by normalizing the product of the likelihood $L(\underset{\sim}{\lambda})$ and a prior distribution of $\underset{\sim}{\lambda}$. In the spirit of Jeffreys (1961), an independent prior is presumed for each λ_j, reflecting vague information on its positive magnitude. The form

$$f'(\underset{\sim}{\lambda}) \propto \prod_{j=1}^{k} \lambda_j^{-1} \tag{7}$$

is uniformly vague on each $\ln \lambda_j$. The posterior then is of the form

$$f''(\underset{\sim}{\lambda}) = \prod_{j=1}^{k} f''(\lambda_j). \tag{8}$$

The λ_j are independent in the posterior and each is distributed as a gamma random variable. By inspection,

$$f''(\lambda_j) = \frac{R_j^{d_j}}{\Gamma(d_j)} \lambda_j^{d_j-1} \exp(-R_j\lambda_j), \tag{9}$$

which reduces to an exponential when all the failures are distinct, i.e., $d_j = 1$.

When a squared error loss function is most relevant, the preferred Bayesian estimator is the posterior mean of the probability of survival. With a discrete hazards form, the probability of surviving to time t is

$$S(t) = \prod_{j|t>t_{(j)}} \exp(-\lambda_j). \tag{10}$$

Averaging S(t) over the posterior distribution of $\underset{\sim}{\lambda}$ becomes the product of the following integrals:

$$S_{DH}(t) = \prod_{j|t>t_{(j)}} \left\{ \int_0^\infty \frac{R_j^{d_j}}{\Gamma(d_j)} \lambda_j^{d_j-1} \exp\left[-(R_j+1)\lambda_j\right] d\lambda_j \right\}, \tag{11}$$

reducing to

$$S_{DH}(t) = \prod_{j|t>t_{(j)}} \left\{ \frac{R_j}{R_j+1} \right\}^{d_j}. \tag{12}$$

Even with distinct failures, this estimator is different from Kaplan-Meier's (1958) or Nelson's (1972) estimators. From (6) the maximum likelihood (ML) estimator of λ_j is $\hat{\lambda}_j = d_j/R_j$, for $j=1,2,\ldots,k$. Then the ML estimator of the survival probability (10) is obtained by replacing each λ_j by $\hat{\lambda}_j$. This yields Nelson's (1972) estimator, i.e.,

$$S_N(t) = \prod_{j|t>t_{(j)}} \exp(-\hat{\lambda}_j) = \exp\{-\sum_{j|t>t_{(j)}} \hat{\lambda}_j\}. \tag{13}$$

More precisely, this is the survival function estimator based upon the cumulative hazard, $H(t) = \int_o^t h(u)du$, estimator given by Nelson (1972). Lawless (1982) and Elandt-Johnson and Johnson (1980) give excellent descriptions of Nelson's method. Kaplan and Meier's (1958) product limit estimator,

$$S_{KM}(t) = \prod_{j|t>t_{(j)}} (1 - \hat{\lambda}_j), \tag{14}$$

can be viewed as a linearized series expansion of each factor in Nelson's estimator. [Elandt-Johnson and Johnson (1980) note an approximate relationship between $S_N(t)$ and $S_{KM}(t)$ based upon the series linearization of $\ln(1-\hat{\lambda}_j)$.] Alternatively, the Kaplan-Meier estimator is the maximum likelihood estimator with the discrete hazards model using a binomial likelihood in which $(1-\lambda_j)^{R_j-d_j}$ replaces $\exp(-R_j\lambda_j)$ in the likelihood (4); see Kaplan and Meier (1958) or Kalbfleisch and Prentice (1980) for presentations.

The estimators $S_{DH}(t)$, $S_N(t)$, and $S_{KM}(t)$ have magnitudes in the same order as their presentation. By comparing the jth factor in (12) and (13), it follows that the Bayes estimator is larger than Nelson's. And Nelson's estimator is in turn larger than Kaplan-Meier's, since the first term omitted in the series expansion of $\exp\{-\hat{\lambda}_j\}$ in (13) is $1/2\ \hat{\lambda}_j^2$.

Variances for these survival probability estimators are given next. For the Bayes estimator, the variance of the probability of surviving to time t is with respect to the posterior distribution of $\underset{\sim}{\lambda}$. The posterior expectation of the square of S(t) in (10) is obtained following the same manipulations as in (11). Then the variance is given by $V(y) = E(y^2) - E^2(y)$, yielding

$$V''(S(t)) = \prod_{j|t>t_{(j)}} \left\{\frac{R_j}{R_j+2}\right\}^{d_j} - \left\{ \prod_{j|t>t_{(j)}} \left[\frac{R_j}{R_j+1}\right]^{d_j} \right\}^2 . \tag{15}$$

The posterior variance of S(t) reflects the uncertainty contained in the posterior distribution of $\underset{\sim}{\lambda}$. The multiplicity of failures, d_j, and numbers at risk, R_j, are observed constants from the sample.

Approximate variance formulae for $S_N(t)$ and $S_{KM}(t)$ condition on the numbers at risk, but treat the d_j as random variables. For Nelson's estimator, a variance estimator can be obtained from the information matrix for $\underset{\sim}{\lambda}$ in the standard way. From the second partials of the log-likelihood, the variance of $\hat{\lambda}_j$ based upon either the observed or expected information is estimated by

$$\hat{V}(\hat{\lambda}_j) = d_j/R_j^2 . \tag{16}$$

Since the jumps in the cumulative hazard are unique to each sample, averaging over samples not obtained is irrelevant, hence the equality of observed and expected information for this problem. Now $-\ln\{S_N(t)\}$ is the sum of the independent $\hat{\lambda}_j$ and using a linearized Taylor series approximation,

$$\hat{V}(S_N(t)) = [S_N(t)]^2 \sum_{j|t>t_{(j)}} \left\{\frac{d_j}{R_j^2}\right\} . \tag{17}$$

The approximate variance for the Kaplan-Meier estimator is given by Greenwood's formula:

$$\hat{V}[S_{KM}(t)] = [S_{KM}(t)]^2 \sum_{j|t>t_{(j)}} \{d_j/[R_j(R_j-d_j)]\} . \tag{18}$$

See Greenwood (1926), Kalbfleisch and Prentice (1980), Elandt-Johnson and Johnson (1980), or Lawless (1982).

Table 1 contains the calculations for the Bayes, Nelson, and Kaplan-Meier estimators of the survival function for the data in Kaplan-Meier (1958). There were four distinct failures occurring at times $t_{(1)}=0.8$, $t_{(2)}=3.1$, $t_{(3)}=5.4$, and $t_{(4)}=9.2$. The R_j are determined given the censored losses at times 1.0, 2.7, 7.0, and 12.1. The ordering of the three estimators is clearly illustrated. Notice that if at the last failure time, all those at risk failed, the Kaplan-Meier estimator becomes zero. Neither the Nelson, nor Bayes, estimator would be so dramatically affected. As the Kaplan-Meier estimator is known to have a negative bias, $S_N(t)$ and $S_{DH}(t)$ offer a potential advantage by the functional relationships established above. Further comments on this point are provided in the Discussion section.

The variance estimates for the survival probabilities estimated in Table 1 are presented in Table 2. Those for the Kaplan-Meier and Nelson methods condition on the R_j and presume the d_j are random. Therefore, the variances for these estimators are smaller than when not conditioning on R_j. Kuzma (1967) showed this in a comparison of the Kaplan-Meier variance with that derived by Chiang (1960) in a stochastic study of life table functions. However, with less than 40% censored losses in the sample, the underestimation was negligible. With 50% or more censoring, as in this example, the variances approximated for $S_N(t)$ and $S_{KM}(t)$ should be used cautiously.

From a different perspective, the Bayesian variance (15) reflects the dispersion in the posterior distribution (9) for the function of the λ_j in S(t) given by (10). The smaller variance for the Bayes method is descriptive of uncertainty in the observed sample, i.e., conditioning on the R_j and d_j, rather than that in the process of sampling under model (3).

2.2 Inference on a Piecewise Exponential Underlying Hazard

The piecewise exponential hazards model has been suggested as an alternative to the discrete hazards model by several authors, especially Breslow (1972). Between failure times this hazard is constant, specifically

$$h(t) = \lambda_j \tag{19}$$

for time t in the interval $I_j = (t_{(j-1)}, t_{(j)})$ and noting $t_{(0)} = 0$. Still let $\underset{\sim}{\lambda} = (\lambda_1,\ldots,\lambda_k)$ as with the discrete hazards model, but now λ_j is the constant hazard during I_j rather than the spike in the discrete hazard at $t_{(j)}$.

The likelihood (5) with the piecewise constant hazard (19) becomes

$$L(\underset{\sim}{\lambda}) = \prod_{j=1}^{k} \Big[\prod_{\ell \in R_{j-1}} \exp\{-\int_{I_{j\ell}} \lambda_j dt\} \prod_{\ell \in R_j} \{\lambda_j\}^{\delta_{j\ell}} \Big] , \tag{20}$$

where $I_{j\ell}$ is the interval of follow-up time for individual ℓ between $t_{(j-1)}$ and $t_{(j)}$, i.e., $I_{j\ell} = (t_{(j-1)}, \min(t_\ell, t_{(j)}))$. Rewriting (20),

$$L(\underset{\sim}{\lambda}) = \prod_{j=1}^{k} \big[\exp\{-\lambda_j V_j\}\ \lambda_j^{d_j}\big] , \tag{21}$$

where

$$V_j = \sum_{\ell \in R_{j-1}} w_{j\ell} ; \tag{22}$$

and

$$w_{j\ell} = \begin{cases} t_\ell - t_{(j-1)} , & \text{for } t_\ell < t_{(j)}; \\ t_{(j)} - t_{(j-1)} , & \text{for } t_\ell \geq t_{(j)}. \end{cases} \tag{23}$$

Or, $w_{j\ell}$ is the width of I_j survived by individual ℓ and V_j is the volume of "person-years" accumulated during the interval I_j by the R_{j-1} individuals at risk of failure just before $t_{(j-1)}$.

The object of estimation is the survival function, i.e., with the piecewise exponential hazard,

$$S(t) = \Pr[T>t] = \prod_{j\varepsilon J(i)} \exp\{-\lambda_j[t_j^* - t_{(j-1)}]\}, \tag{24}$$

where J(i) is the set of integers $j=1,2,\ldots,i$ with $t(i-1) < t \leq t_{(i)}$. Then

$$t_j^* = \begin{cases} t(j), & \text{for } j \leq i-1; \\ t\ , & \text{for } j = i. \end{cases}$$

As with the discrete hazard model, a preferred Bayes estimator of the survival probability (24) is its average over the normalized product of the likelihood (21) and vague prior (7). Following the derivation of $S_{DH}(t)$ in (12),

$$S_{PE}(t) = \prod_{j\varepsilon J(i)} \left[\int_o^\infty \frac{V_j^{d_j}}{\Gamma(d_j)} \lambda_j^{d_j-1} \exp\{-\lambda_j[V_j + (t_j^* - t_{(j-1)})]\}\, d\lambda_j\right], \tag{25}$$

which becomes

$$S_{PE}(t) = \prod_{j\varepsilon J(i)} \left\{\frac{V_j}{V_j + (t_j^* - t_{(j-1)})}\right\}^{d_j}. \tag{26}$$

For comparison, maximum likelihood estimation of the corresponding survival probability replaces each λ_j in (24) by its ML estimator, $\hat{\lambda}_j = d_j/V_j$ from (21). The ML estimator of the survival probability S(t) is then

$$\hat{S}_{PE}(t) = \prod_{j\varepsilon J(i)} \exp\{-d_j(t_j^* - t_{(j-1)})/V_j\} \tag{27}$$

The jth factor of the Bayes estimator (26) raised to the negative d_j power is $1 + (t_j^* - t_{(j-1)})/V_j$ while that for the ML estimator (27) is $\exp\{(t_j^* - t_{(j-1)})/V_j\}$. Since this factor for the Bayes estimator of the survival probability is the linear portion of the corresponding exponential factor in $\hat{S}_{PE}(t)$, it follows that the Bayes estimator is at least as large as the ML estimator for every follow-up time t. Further, the larger the fraction of the n individuals who are placed on test and are observed to fail, the more the Bayes estimator will exceed its maximum likelihood analogue. With more factors having small accumulations of person-years, the factors $(t_j^* - t_{(j-1)})/V_j$ will be larger.

Table 3 contains the Bayes and maximum likelihood estimates of the survival curve estimates at the four failure times in the Kaplan-Meier (1958) data. As algebraically shown above, the Bayes estimates are larger than those by maximum likelihood. Further, the survival curve estimates based upon the piecewise exponential model are generally larger than those in Table 1 from the discrete hazards model. This is due to the increments to survival for the observations censored between failure times. However, when the censoring for each of the R_{j-1} individuals is at $t_{(j)}$, a later failure time, or after the last failure time, then $S_{DH}(t)$ will equal $S_{PE}(t)$ at the times of failure. Between failure times $S_{DH}(t)$ is constant and $S_{PE}(t)$ has a decreasing exponential shape. With this Mantel (1966) distribution of censoring times, $S_N(t)$ and $\hat{S_{PE}}(t)$ are likewise related.

Variances for the Bayes estimator (26) and maximum likelihood estimator (27) can be obtained using the same technique as with the discrete hazards model in (15 and (17), respectively. The results are

$$V''(S(t)) = \prod_{j \epsilon J(i)} \left\{ \frac{V_j}{V_j + 2(t_j^* - t_{(j-1)})} \right\}^{d_j} - \left[\prod_{j \epsilon J(i)} \left\{ \frac{V_j}{V_j + (t_j^* - t_{(j-1)})} \right\}^{d_j} \right]^2 \quad (28)$$

and

$$\hat{V}[\hat{S_{PE}}(t)] = [\hat{S_{PE}}(t)]^2 \sum_{j \epsilon J(i)} \left\{ \frac{d_j (t_j^* - t_{(j-1)})^2}{V_j^2} \right\} . \quad (29)$$

The variance for the ML estimator is an approximation because the V_j have been considered non-random and there was a linearization of an exponential series.

Table 4 contains the variances computations for the survival function estimates in Table 3. The additional increments in survival result in variance estimates with the Bayes and maximum likelihood approaches that are no larger, and generally smaller, than their corresponding estimates in Table 2 for the discrete hazards model. The Bayes and maximum likelihood estimators of variance are descriptive of different types of uncertainty, specifically in the dispersion of the likelihood and prior product and in repeated sampling, respectively. So their relative sizes are not so important, but since the Bayes approach conditions on the d_j and V_j, it is natural that its variances are smaller than those from maximum likelihood.

3. Multiple Samples or Covariates

By the use of covariates, applications composed of several samples or heterogeneous individuals may be addressed. The incorporation of this additional information is made presuming Cox's (1972) log-linear, proportional formulation given by (1).

3.1 Inference on the Regression Parameters with a Discrete Underlying Hazard

The adaptation of the discrete hazards model to permit the inclusion of covariates is accomplished by multiplying λ_j in (3) and (4) by $\exp(\underset{\sim}{z}\underset{\sim}{\beta})$. The expansion of likelihood (6) to include the p regression coefficients $\underset{\sim}{\beta}$, as well as the k increments: $\underset{\sim}{\lambda} = (\lambda_1,\ldots,\lambda_k)$ to the cumulative hazard, gives

$$L(\underset{\sim}{\lambda}, \underset{\sim}{\beta}) = \prod_{j=1}^{k} \left[\exp\{-\lambda_j \sum_{\ell \in R_j} \exp(\underset{\sim}{z}_\ell \underset{\sim}{\beta})\} \lambda_j^{d_j} \exp\{\underset{\sim}{s}(j)\underset{\sim}{\beta}\}\right], \tag{30}$$

where the sum of the covariates for those d_j individuals failing at $t_{(j)}$ is

$$\underset{\sim}{s}(j) = \sum_{\ell \in R_j} \delta_{j\ell}\, \underset{\sim}{z}_\ell. \tag{31}$$

As inference on the vector of regression coefficients is of primary importance, the marginal posterior distribution of $\underset{\sim}{\beta}$ is obtained by integrating over the k nuisance parameters $\underset{\sim}{\lambda}$ in the joint posterior of $\underset{\sim}{\lambda}$ and $\underset{\sim}{\beta}$. A joint prior on $\underset{\sim}{\lambda}$ and $\underset{\sim}{\beta}$ is specified which is initially vague, again following Jeffreys (1961). Since the parameter range for each regression coefficient is the real line, a prior $f'(\underset{\sim}{\lambda}, \underset{\sim}{\beta})$ proportional to (7) is uniformly vague on each $\ln \lambda_j$ and on each regression coefficient. Then as per Lindley (1972), integrating the product of the likelihood (30) and prior $f'(\underset{\sim}{\lambda}, \underset{\sim}{\beta})$ over each λ_j yields an unnormalized marginal posterior for $\underset{\sim}{\beta}$ as follows:

$$f''(\underset{\sim}{\beta}) \propto \prod_{j=1}^{k} \left\{ \frac{\exp[\underset{\sim}{s}(j)\underset{\sim}{\beta}]}{[\sum_{\ell \in R_j} \exp(\underset{\sim}{z}_{\ell}\underset{\sim}{\beta})]^{d_j}} \right\} , \tag{32}$$

Several observations on this result are appropriate.

(a) With distinct failures, i.e., $d_j = 1$ for all k failure times, the marginal posterior distribution of $\underset{\sim}{\beta}$ yields equivalent inference as Cox's (1972) approach in the sense that the mode of (32) corresponds to the maximum likelihood estimator with Cox's (1972, 1975) conditional, or partial, likelihood and Kalbfleisch and Prentice's (1973) marginal likelihood. Further, variance estimation based upon an approximated quadratic shape of the logarithm of this marginal posterior, or the observed information of Cox's (1972) likelihood, is also equivalent.

(b) With multiple failures at any $t_{(j)}$, i.e., $d_j > 2$ for at least one j, Bayesian inference is equivalent in mode, and approximated quadratic shape of the logarithm, of this marginal posterior as that based upon the likelihood described by Breslow (1974). Therefore, this Bayesian approach suggests that Breslow's (1974) approximation to Kalbfleisch and Prentice's (1973) marginal likelihood has a statistical basis as well as its value for computational practicality. Statistical practice relies on (32), e.g., as in the SAS procedure PHGLM. The essence of Breslow's (1972, 1974) approximation was also noted by Peto (1972), but more from a view of numerically approximating a probability than from likelihood considerations.

(c) Under a quadratic loss structure, the posterior mean of β is preferred to the posterior mode as a point estimator. Even with a single covariate, numerical integration is needed for the required computations. However, symmetry of the posterior could be checked by an evaluation of the third partial of the natural logarithm of the posterior at the mode. Values close to zero would suggest symmetry in the posterior distribution and this is when the posterior mode should reasonably approximate the posterior mean.

(d) The above presentation is for baseline covariates, but results with time dependent covariates are immediately available when covariables z_ℓ are replaced by $z_{j\ell}$, i.e., by their value at the time of the jth failure.

3.2 Inference on the Regression Parameters with a Piecewise Exponential Hazard

The piecewise exponential hazard with concomitant information in log-linear, proportional form is $\lambda_j \exp(z\beta)$ rather than simply λ_j in (19). Likelihood (21) becomes

$$L(\lambda, \beta) = \prod_{j=1}^{k} \left[\exp\{-\lambda_j V_j(\beta)\} \lambda_j^{d_j} \exp\{s(j)\beta\}\right], \tag{33}$$

where

$$V_j(\beta) = \sum_{\ell \in R_{j-1}} w_{j\ell} \exp\{z_\ell \beta\} \tag{34}$$

and $w_{j\ell}$ is described with (23). The volume of "person-years", $V_j(\beta)$, is weighted by the individual's covariate values and reduces to (22) when the z_ℓ are all zeros.

The marginal posterior for β is the integrated product of the likelihood (33) and a parameter space delineating joint prior, $f'(\lambda, \beta)$, which is taken to be proportional to (7). The result is

$$f''(\beta) \propto \prod_{j=1}^{k} \left[\frac{\exp\{s_{(j)}\beta\}}{\left[\sum_{\ell \in R_{j-1}} w_{j\ell} \exp\{z_\ell \beta\}\right]^{d_j}} \right] , \tag{35}$$

where $s_{(j)}$ is described with (31) and the shorthand notation, $V_j(\beta)$, is suppressed to show the analogy with (32).

This result is proportional to a refinement of Breslow's (1974) likelihood by incorporating the follow-up time between failures survived by those individuals who do not fail on study. Its significance is not in the refinement, but in suggesting that Breslow's (1974) result can be considered as an alternative likelihood, rather than only a practical approximation to Kalbfleisch and Prentice's (1973) marginal likelihood. The basis for this consideration is conditional on the observed data and is obtained by averaging over the uncertainty in the nuisance parameters with only vague expression of prior knowledge. Further discussion is included in the next section.

4. Discussion

This last section is composed of three subsections: in the first: inference on the underlying hazard for a single sample without covariates is reviewed; in the second: inference for the regression coefficients is briefly revisited; and in the third: general inference issues in the paper are considered.

4.1 Nonparametric Hazard Estimators

The Bayesian estimator (12) of the discrete hazard is contrasted with the result of Sursarla and Van Ryzin (1976), who derived a nonparametric Bayesian estimator of survival curves from incomplete observations. Their approach was a decision theoretic one assuming a Dirichlet process prior. The limit of their Bayes estimator, as their exponential prior approached zero, reduced to the Kaplan-Meier estimator when the failures are distinct, i.e., when all $d_j=1$. The Bayes estimator (12) is the posterior mean of the desired function of parameters from a discrete hazards model. The principles of decision theory guided the selection of this Bayes approach in choosing the posterior mean as the point estimator, but it is not as directed as the exponential shape of the prior which was specified by Sursarla and Van Ryzin. In that sense, $S_{DH}(t)$ could be regarded as a more primitive nonparametric Bayesian estimator of the survival function. In further contrast with the Sursarla and Van Ryzin paper, the difference between $S_{DH}(t)$ and $S_{KM}(t)$ is stressed rather than their similarity. In view of the known negative bias of $S_{KM}(t)$, alternative estimators seemingly should be sought.

The Kaplan-Meier product limit estimator of the survival function has been the historic, nonparametric choice. Since Chen, et al (1982) documented that $S_{KM}(t)$ has a negative bias, either Nelson's estimator or the Bayes estimator may be an improvement in this regard. However, neither a maximum likelihood, nor a Bayes, approach generally delivers unbiased estimators. An assessment of bias in a repeated sampling framework for $S_{DH}(t)$ and $S_N(t)$ could be made following the Chen, et al (1982) paper. A mean squared error comparison may be even more revealing, given the variance estimates in Table 2. A repeated sampling based comparison study with practical sized samples for alternative nonparametric estimators of the survival function should be an informative effort. (A Master's student, Mr. Thomas Coleman, is performing such a study at the University of North Carolina).

From the perspective of the statistical theory of counting processes, Aalen (1978) identifies Nelson's estimator as an application of martingales and notes a close relationship with the Kaplan-Meier estimator. Yandell (1983) provides a review of "delta sequence" estimators, of which the kernel rate estimator with censored survival data is the Nelson (1972), or Aalen (1978), empirical cumulative rate. Informally, the weight functions $G_j(u)$ are descriptive of this approach. The asymptotic unbiasedness, strong consistency, and asymptotic normality of the Nelson (1972) estimator are sketched by Yandell (1983). However, for the purposes of model examination with small samples or stratification in the analysis of larger studies, comparisons of such estimators with small and moderate sized samples is needed, e.g., as per Chen, et al (1982).

4.2 Inference on the Regression Coefficients

A Bayesian approach to this problem was also taken by Kalbfleisch (1978). However, in his approach a gamma density for the prior distribution of the cumulative hazard was used. It required specifying an initial guess at the underlying hazard and another constant reflecting the weight to be given to that guess. With distinct failures and to a first order approximation, his marginal posterior distribution of $\underset{\sim}{\beta}$ was proportional to the marginal likelihood of Kalbfleisch and Prentice (1973) when the weighting constant tended to vague information on the underlying hazard. With tied failure times, Kalbfleisch's results were more complicated.

Using Jeffreys' (1961) vague priors, these difficulties were not encountered. Vague specifications of prior knowledge were to delineate simply the range of each parameter and to be relatively indifferent to various portions of each parameter's range. The parameters of the cumulative hazard were more directly reflected in the prior. Further, the prior was guided by the discrete nature of the likelihood of the data, rather than imposing features onto the likelihood, for example, an absolutely continuous shape to the cumulative hazard.

The Bayesian approach with vague priors resulted in a different solution than that of the frequents' for this problem when tied observations are present. In simplier problems these two strategies typically agree in solution and differ only in their interpretations of that solution. From the Bayesian perspective, there is no concern for the possible distinct order of the observed failures in the sample. Rather there is always a desire for more precise measurements, in this case of the failure times. Bayes inference is conditional on the available data, however precise, or imprecise, they may be.

4.3 Two Other Points of Inference

It should be clear that the parameter space for this approach increases with the sample size. To be nonparametric in the handling of the underlying hazard, this is necessarily so. Cox (1972), for example, pointed out that maximum likelihood theory is troubled by such situations in his reply to Breslow's (1972) proposed piecewise exponential model. With large samples, there will be sufficient information to discriminate reasonably between parametric models for the underlying hazard. The problem with small samples is that the flexibility of several models pose difficult discrimination problems. After the available biological requirements for the application have been satisfied, candidate models could be examined by an asymptotic likelihood method, or an equivalent approach. For example, a life table could be constructed on follow-up time and a cumulative hazard plot created; see Elandt-Johnson and Johnson (1980). Then the weighted least squares methodology of Grizzle, Starmer, and Koch (1969) could be used to test the adequacy of the model examined in the cumulative hazard plot. Their error chi-square provides a measure of goodness-of-fit with large samples.

The use of a flexible parametric model for the underlying hazard may be supported from other considerations as well. For example, see the application in Green and Symons (1983) where the regression coefficient estimates and their standard errors were almost identical from Cox's (1972) approach and maximum likelihood estimation with a Weibull hazard and the same log-linear incorporation of covariates.

There are interesting issues related to the estimation of the underlying hazard when concomitant information is available. A frequentists' approach is sketched by Kalbfleisch and Prentice (1980). Specifically, with Breslow's (1974) piecewise exponential model, the joint likelihood of $\underset{\sim}{\lambda}$ and $\underset{\sim}{\beta}$ is maximized, subject to the constrant that the regression coefficients take on the value of the maximum likelihood estimator using Breslow's (1974) likelihood. Bailey (1983) considers the joint estimation of $\underset{\sim}{\lambda}$ and $\underset{\sim}{\beta}$. With no tied failure times, his asymptotic argument supports the use of Kaplan and Meier's (1958) product limit estimator. Breslow (1972) reports an application where the covariate had a marked effect on survival, but that his estimator, a Kaplan-Meier form of estimator (actually Nelson's) obtained by setting the regression coefficient to zero with Cox's (1972) model, and Cox's (1972) more complicated estimator of the underlying hazard all agree remarkably well. Although the number of failures was not reported by Breslow (1972), it appears that a large sample was available. Yandell's (1983) asymptotic consideration confirms that Nelson's (1972) estimator would also agree quite well with the other two estimators.

With large samples and in the presence of mild prior information, Bayes estimators are known to be similar to maximum likelihood estimators. Asymptotically the two approaches are the same, as sketched by Lindley (1965). But, Bayesian considerations offer some insights on the frequentists' results. Unfortunately, a closed form for the marginal posterior distribution of $\underset{\sim}{\lambda}$ is not available by integrating over the parameter space of $\underset{\sim}{\beta}$, either the product of (28) and (7) for the discrete hazard, or the product of (31) and (7) for the piecewise exponential hazard. Nevertheless, it is the marginal posterior of $\underset{\sim}{\lambda}$, $f''(\underset{\sim}{\lambda})$, that is desired. For comparison with the frequentists' conditional approach, the conditional posterior

$$f''(\underset{\sim}{\lambda} | \underset{\sim}{\beta} = \hat{\underset{\sim}{\beta}}) = \frac{f(\lambda, \underset{\sim}{\beta} = \hat{\underset{\sim}{\beta}})}{f(\underset{\sim}{\beta} = \hat{\underset{\sim}{\beta}})}, \qquad (34)$$

could be compared with $f''(\underset{\sim}{\lambda})$. For example with a single covariate a series expansion of the joint posterior could be integrated term by term and used to provide bounds on $f''(\underset{\sim}{\lambda})$ to any required accuracy. But the empirical results of Breslow (1972) and the asymptotic ones of Yandell (1983) suggest that $f''(\underset{\sim}{\lambda})$ and $f''(\underset{\sim}{\lambda} | \underset{\sim}{\beta} = \hat{\underset{\sim}{\beta}})$ will be very similar. Or, asymptotically $\underset{\sim}{\lambda}$ and $\underset{\sim}{\beta}$ are independent. As noted in the example at the end of section 2.1, differences between the Bayes and frequentists' approaches may be noteworthy for this problem, at least with smaller sample sizes.

ACKNOWLEDGEMENTS

Initial interest in this topic grew out of very stimulating talks on applications of martingales to survival analysis given by Vidyadhar Mandrekar of Michigan State University and colleague Mohammed Habib. Discussions with P.K. Sen and the expert typing of Ernestine Bland are gratefully acknowledged. Supported by NHLBI Contract NO1-HR-1-2243-L.

References

Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. Annals of Statistics, 6: 701-726.

Bailey, K.R. (1983). The asymptotic joint distribution of regression and survival parameter estimates in the Cox regression model. The Annals of Statistics, 11: 39-48.

Breslow, N.E. (1972). Contribution to the discussion of paper by D.R. Cox. Journal of the Royal Statistical Society, Series B, 34: 216-7.

Breslow, N.E. (1974). Covariance analysis of censored survival data. Biometrics, 30: 89-99.

Chen, Y.Y., Hollander, M., and Landberg, N.A. (1982). Small-sample results for Kaplan-Meier estimator. Journal of the American Statistical Association, 17: 141-144.

Chiang, C.L. (1960). A stochastic study of the life table and its applications: I. Probability distributions of biometric functions. Biometrics, 16: 618-635.

Cox, D.R. (1972). Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B, 34: 187-220.

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

Elandt-Johnson, R.C. and Johnson, N.L. (1980). Survival Models and Data Analysis. Wiley, New York.

Green, M.S. and Symons, M.J. (1983). A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. Journal of Chronic Diseases, 36: 715-724.

Greenwood, M. (1926). The natural duration of cancer. Reports of Public Health and Medical Subjects, Volume 33. Her Majesty's Stationery Office, London.

Grizzle, J.E., Starmer, F.C., and Koch, G.G. (1969). Analysis of categorical data by linear models. Biometrics, 25: 489-504.

Harrell, F.E. Jr. (1983). PHGLM. SUGI Supplemental Library User's Guide 1983 Edition. SAS Institute, Box 8000, Cary, N.C., 27511-8000, 268-294.

Jeffreys, H. (1961). Theory of Probability. Clarendon, Oxford.

Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data. Journal of the Royal Statistical Society, Series B, 40: 214-221.

Kalbfleisch, J.D. and Prentice, R.L. (1972). Contribution to discussion of paper by D.R. Cox. Journal of the Royal Statistical Society, Series B, 34: 215-6.

Kalbfleisch, J.D. and Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model. Biometrika, 60: 267-278.

Kalbfleisch, J.D. and Prentice, R.L. (1980). The Statistical Analysis of Failure Time Data. Wiley, New York.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53: 457-481.

Kuzma, J.W. (1967). A comparison of two life table methods. Biometrics, 23: 51-64.

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

Lindley, D.V. (1965). Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference. Cambridge University Press.

Lindley, D.V. (1972). Contribution to discussion of paper by D.R. Cox. Journal of the Royal Statistical Society, Series B, 34: 208-9.

Mantel, N. (1966). Estimation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports, 50: 163-170.

Nelson, W.B. (1972). Theory and applications of hazard plotting for censored failure data. Technometrics, 14: 945-965.

Peto, R. (1972). Contribution to discussion of paper by D.R. Cox. Journal of the Royal Statistical Society, Series B, 34: 205-7.

PHGLM. (1980). SAS Supplemental Library User's Guide. SAS Institute, North Carolina.

Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. Journal of the American Statistical Association, 71: 897-902.

Yandell, B.S. (1983). Nonparametric inference for rates with censored survival data. The Annals of Statistics, 11: 1119-1135.

TABLE 1. Bayes, Nelson, and Kaplan-Meier Estimators of the Survival Function with the Discrete Hazards Model for the Data in Kaplan-Meier (1958)

				Bayes (12)		Nelson (13)		Kaplan-Meier (14)	
j	$t_{(j)}$	d_j	R_j	$\left(\frac{R_j}{R_j+1}\right)^{d_j}$	$S_{DH}\left(t_{(j)}\right)$	$\exp(-\hat{\lambda}_j)$	$S_N\left(t_{(j)}\right)$	$1-\hat{\lambda}_j$	$S_{KM}\left(t_{(j)}\right)$
1	0.8	1	8	0.888889	0.8889	0.882497	0.8825	0.875000	0.8750
2	3.1	1	5	0.833333	0.7407	0.818731	0.7225	0.800000	0.7000
3	5.4	1	4	0.800000	0.5926	0.778801	0.5627	0.750000	0.5250
4	9.2	1	2	0.666667	0.3951	0.606531	0.3413	0.500000	0.2625

TABLE 2. Variance Computations for Survival Function Estimates in Table 1

j	R_j	Bayes				Nelson			Kaplan-Meier		
		$[\frac{R_j}{R_j+2}]^{d_j}$	$\prod_{j\mid t>t(j)} [\frac{R_j}{R_j+2}]^{d_j}$	$[S_{DR}(t(j))]^2$	$V^*[(15)]$	$S_N(t(j))$	$\frac{d_j}{R_j^2}$	$\hat{V}[(17)]$	$S_{KM}(t(j))$	$\frac{d_j}{[R_j(R_j-d_j)]}$	$\hat{V}[(18)]$
1	8	0.800000	0.800000	0.790124	0.0099	0.882497	0.015625	0.0122	0.875000	0.017857	0.0137
2	5	0.714286	0.571429	0.548697	0.0227	0.722528	0.040000	0.0290	0.700000	0.050000	0.0332
3	4	0.666667	0.380952	0.351165	0.0298	0.562705	0.062500	0.0374	0.525000	0.083333	0.0417
4	2	0.500000	0.190476	0.156074	0.0344	0.341298	0.250000	0.0429	0.262500	0.500000	0.0449

TABLE 3. Bayes and Maximum Likelihood Estimates of the Survival Function with the Piecewise Exponential Hazard for the Data in Kaplan-Meier (1958)

					Bayes Estimator (26)		Maximum Likelihood Estimator (27)	
j	$t_{(j)}$	d_j	R_{j-1}	V_j^*	$\left[\frac{V_j}{V_j + (t_{(j)} - t_{(j-1)})}\right]^{d_j}$	$S_{PE}(t_{(j)})$	$\exp\{-d_j(t_{(j)}-t_{(j-1)})V_j\}$	$\hat{S}_{PE}(t_{(j)})$
1	0.8	1	8	6.4	0.888889	0.8889	0.882497	0.8825
2	3.1	1	8	13.6	0.855346	0.7603	0.844410	0.7452
3	5.4	1	5	9.2	0.800000	0.6082	0.778801	0.5804
4	9.2	1	4	9.2	0.707692	0.4305	0.661634	0.3840

*The widths of these intervals survived by those at risk at time $t_{(j-1)}$ are as follows: $w_{1\ell}$ = 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8; $w_{2\ell}$ = 0, 0.2, 1.9, 2.3, 2.3, 2.3, 2.3, 2.3; $w_{3\ell}$ = 0, 2.3, 2.3, 2.3, 2.3; $w_{4\ell}$ = 0, 1.6, 3.8, 3.8.

TABLE 4. Variance Computations for the Survival Function Estimates in Table 3

		Bayes				Maximum Likelihood		
j	v_j	$\frac{v_j}{v_j + 2(t_{(j)}-t_{(j-1)})}$	$\prod_{<j} [\frac{v_j}{v_j+2(t_{(j)}-t_{(j-1)})}]^{d_j}$	$[S_{PE}(t_{(j)})]^2$	$V^{-}[(28)]$	$\hat{S}_{PE}(t_{(j)})$	$\frac{d_j(t_{(j)}-t_{(j-1)})^2}{v_j^2}$	$\hat{V}[(29)]$
1	6.4	0.800000	0.800000	0.790124	0.0099	0.882497	0.015625	0.0122
2	13.6	0.747253	0.597802	0.578068	0.0197	0.745189	0.028601	0.0246
3	9.2	0.666667	0.398535	0.369963	0.0286	0.580354	0.062500	0.0359
4	9.2	0.547619	0.218245	0.185288	0.0330	0.383982	0.170605	0.0409