B.A.J. **2,** III, 703–726 (1996)

AN ACTUARIAL SURVEY OF STATISTICAL MODELS FOR DECREMENT AND TRANSITION DATA

III: COUNTING PROCESS MODELS

BY A. S. MACDONALD, B.Sc., Ph.D., F.F.A.

ABSTRACT

Counting processes and their compensators are introduced at a heuristic level. The martingale property of stochastic integrals with respect to a compensated counting process leads to moment estimates and asymptotic normal distributions for statistics arising in multiple state, non-parametric and semi-parametric models. The place of survival models in actuarial education is discussed.

KEYWORDS

Censoring; Compensators; Counting Processes; Kaplan-Meier Estimate; Martingales; Multiple State Models; Nelson-Aalen Estimate; Partial Likelihood; Survival Analysis

INTRODUCTION

In Part I, Sections 1-4 and Part II, Sections 5-7 we described statistical models for small segments of lifetimes and for complete lifetimes, all in the presence of censoring. Many of these can be formulated as special cases of counting process models, in which setting results about the distributions of parameter estimates can be obtained much more simply than in the original models. The key is the adoption of a dynamic viewpoint; survival data arise from processes which evolve over time, so models based on random processes rather than random variables are better representations.

All of the material in this part is standard in the recent statistical literature, and it might be useful background for an actuary working with mortality or morbidity data. Acknowledgements and references were given in Part I.

Part I appears in *British Actuarial Journal*, Volume **2,** Part I and Part II in *B.A.J.* **2,** II.

8. COUNTING PROCESS MODELS

In this section we describe some recent developments which unify multiple state models, the Kaplan-Meier estimate and the Cox model. The key is the introduction of *counting processes*. Unfortunately, most of the available literature assumes knowledge of stochastic processes well beyond the level of the U.K. actuarial syllabus. The basic ideas, however, are often very simple. Our aim here is to guide the reader through the sequence of constructions leading to the applications, introducing the technical terms of the subject heuristically. The reader might regard the following as a glossary for accessing a proper treatment,

such as the monographs by Fleming & Harrington (1991) and by Andersen *et al.* (1993). The machinery we will need includes the following:

(a) conditional probabilities and expectations;
(b) counting processes and multivariate counting processes;
(c) martingales;
(d) predictable processes;
(e) compensators of counting processes;
(f) moment properties of counting processes; and
(g) stochastic integrals.

Some of these tools also appear in models from financial economics, so it is worth repeating that the ideas behind these constructs are extremely simple, even when rigorous proof is delicate; the actuary should not be discouraged by the technical appearance of much of the literature.

8.1 *Conditional Probabilities and Expectations*

In survival analysis, we observe events and acquire information as time passes; in other words we need some mathematical structure which can represent this accrual of information; such a structure is a *filtered probability space*. At each time t, we represent all the information which has been revealed up to and including time t by the symbol $\mathcal{F}_t$. Similarly, we denote all the information revealed up to, but not including, time t by $\mathcal{F}_{t-}$. In the case of a stochastic process, the revelation of information usually means the past history of the process. The idea is that, given a stochastic process $\mathbf{H}_t$, say, we can write:

(a) conditional probabilities $\mathrm{P}[\mathbf{H}_{t+s} \in \mathscr{A} | \mathcal{F}_t]$, interpreted as "the probability that $\mathbf{H}$ is in the subset $\mathscr{A}$ of Ω (the sample space) at time $t + s$, given all the information up to and including time t"; and
(b) conditional expectations $\mathrm{E}[\mathbf{H}_{t+s} | \mathcal{F}_t]$, interpreted as "the expected value of $\mathbf{H}$ at time $t + s$, given all the information up to and including time t".

We suppose that information, once acquired, is not forgotten, so anything known at time t is also known at time $t + s$ for $s > 0$.

If we are dealing with a finite number of random variables, or a discrete-time stochastic process, the $\mathcal{F}_t$ notation is no more than a compact way to write conditional probabilities and expectations which we could have written explicitly in terms of the random variables themselves, and, in general, the reader will not be led astray if he or she regards symbols like $\mathcal{F}_t$ in this way. In continuous time, however, the idea of filtrations is essential for technical reasons. It is particularly important to distinguish between $\mathcal{F}_{t-}$ and $\mathcal{F}_t$, as we shall see in the case of predictable processes.

8.2 *Counting Processes*

A counting process is an integer-valued stochastic process $\mathbf{N}_t$ indexed by time

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

t, where $\mathbf{N}_t$ counts the number of events of specified type which have taken place up to time t. The prototype of a counting process is the Poisson process.

In Part I, Section 3 we introduced the indicator $\mathbf{D}_i$ of the event that the ith life was observed to die between the ages of $x + a_i$ and $x + b_i$. Suppose we extend the definition of $\mathbf{D}_i$ to a family of random variables $\{\mathbf{N}_{i,t}\}$ where $(a_i \leq t \leq b_i)$, as follows:

$$\mathbf{N}_{i,t} = \begin{cases} 1 \text{ if the } i\text{th life is observed to die not later than age } x+t \\ 0 \text{ if the } i\text{th life is not observed to die by age } x+t. \end{cases}$$

(We use the notation $\mathbf{N}$ rather than $\mathbf{D}$ because this agrees with most of the literature on counting processes.) Then the family $\{\mathbf{N}_{i,t}\}$ is a counting process, taking on the values 0 or 1. The sample path of the process indicates the event of death, and, if death occurs, the time of death; we will refer to it as an *indicator of death*.

Some models encountered in survival analysis are based on *multivariate* counting processes. A multivariate counting process is a collection of K (say) counting processes, with the additional assumption that no two component processes jump simultaneously. For example, if we extend the random variables defined in connection with the illness-death model of Part I, Section 3 to stochastic processes, in the same manner as $\mathbf{N}_{i,t}$ above, then the collection $(\mathbf{N}_{i,t}, \mathbf{U}_{i,t}, \mathbf{S}_{i,t}, \mathbf{R}_{i,t})$ is a multivariate counting process with $K = 4$. It is evident from the model specification that no two components can jump simultaneously.

It is helpful to characterise the stochastic processes which we will use by their increments over infinitesimal intervals. All of the 'obvious' results in this heuristic approach can be proved formally. At time t, we condition on knowledge up to, *but not including*, time t; the information accruing at time t cannot affect anything we can decide *at* time t. In other words, our knowledge at time t is represented by $\mathscr{F}_{t-}$. Then the (multivariate) counting process is characterised by:

$$\mathrm{E}[d\mathbf{N}_t \mid \mathscr{F}_{t-}] = \mathrm{P}[d\mathbf{N}_t = 1 \mid \mathscr{F}_{t-}] \tag{1}$$

because $\mathbf{N}_t$ has jumps of size +1, and the component processes do not jump simultaneously. Slightly more technically, this characterisation depends on the process being right-continuous. If it were left-continuous, then its value at time t would be the limit of its values up to time t, and $d\mathbf{N}_t$ (given $\mathscr{F}_{t-}$) would be zero.

A rigorous discussion of the probabilistic aspects of counting processes can be found in Jacobsen (1982) or Karr (1991).

8.3 *Martingales*

Essentially, a martingale is a stochastic process without expected drift. Counting processes, being non-decreasing, are not martingales. Formally, a

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

martingale is a stochastic process $\mathbf{M}_t$, whose absolute value process $|\mathbf{M}_t|$ has finite expected values $\mathrm{E}[|\mathbf{M}_t|]$ for all $t \geq 0$, and for which:

$$\mathrm{E}[\mathbf{M}_{t+s} \mid \mathscr{F}_t] = \mathbf{M}_t \tag{2}$$

for all $s > 0$. Alternatively, a martingale has zero expected future increments over finite intervals:

$$\mathrm{E}[\mathbf{M}_{t+s} - \mathbf{M}_t \mid \mathscr{F}_{t-}] = 0 \tag{3}$$

for $s > 0$, or over infinitesimal intervals:

$$\mathrm{E}[d\mathbf{M}_t \mid \mathscr{F}_{t-}] = 0. \tag{4}$$

Martingales have two properties which play a central part in the sequel. As we observe a population, we can record the evolution of any statistics of interest. The history of each such statistic is a stochastic process. In some cases, this stochastic process can be written in terms of a stochastic integral with respect to a martingale integrator (described later). The two relevant properties, and their consequences, are as follows:

(a) *Stochastic integrals preserve martingales.* If we form a stochastic integral with a martingale $\mathbf{M}_t$ as the integrator (for a suitable integrand $\mathbf{H}_t$) then the resulting process:

$$\int_0^t \mathbf{H}_s d\mathbf{M}_s$$

is also a martingale. So, if we can express a statistic as a stochastic integral with respect to a martingale, the martingale property tells us that the statistic has conditional expectation zero. This allows us to find *asymptotically unbiased* estimators of parameters or distributions.

(b) *Martingale central limit theorems.* Under reasonably general conditions, it can be shown that a sequence of martingales, where each successive member of the sequence represents a statistic drawn from a larger population, converges in distribution to a Gaussian process (one whose finite-dimensional distributions are all multivariate Normal). Applied to survival data statistics expressed in terms of stochastic integral martingales, these results provide asymptotic Normal distributions.

For a survey of martingales and their actuarial uses, see Smith (1991). The plan here is to construct martingales from the quantities arising in our survival models, and then to express relevant statistics in terms of stochastic integrals. In order to define a stochastic integral we need predictable processes.

8.4 *Predictable Processes*

A predictable (sometimes called previsible) process is one whose value at time t is known, given the information available up to, but not including, time t. The prototype of a predictable process is one whose sample paths are left-continuous, since its value at time t is certain, given its values at all preceding times. Not all predictable processes are left-continuous, however.

The technical condition satisfied by a predictable process is that it is $\mathscr{F}_{t-}$ — measureable at time t. This allows us to characterise predictable processes by their properties over infinitesimal intervals, of which the most useful is the following: let $\mathbf{H}_t$ be a predictable process and let $\mathbf{G}_t$ be any other process. Then:

$$\mathrm{E}[\mathbf{H}_t\mathbf{G}_t \mid \mathscr{F}_{t-}] = \mathbf{H}_t\mathrm{E}[\mathbf{G}_t \mid \mathscr{F}_{t-}] \tag{5}$$

or, over an infinitesimal interval:

$$\mathrm{E}[\mathbf{H}_t d\mathbf{G}_t \mid \mathscr{F}_{t-}] = \mathbf{H}_t\mathrm{E}[d\mathbf{G}_t \mid \mathscr{F}_{t-}] \tag{6}$$

In terms of the increment $d\mathbf{H}_t$, the predictable property is:

$$\mathrm{E}[d\mathbf{H}_t \mid \mathscr{F}_{t-}] = d\mathbf{H}_t. \tag{7}$$

Counting processes are not predictable processes.

8.5 *Compensators of Counting Processes*

Let $\mathbf{S}_t$ be a stochastic process for which $\mathrm{E}[|\mathbf{S}_t|] < \infty$ $(t \geq 0)$ as for a martingale, but for which $\mathrm{E}[\mathbf{S}_{t+s} \mid \mathscr{F}_t] \geq \mathbf{S}_t$ (s, $t \geq 0$). In other words, the martingale property is weakened so that the conditional expectation of $\mathbf{S}_t$ is merely non-decreasing. Such a process is called a *submartingale*. A fundamental result — the Doob-Meyer decomposition (Fleming & Harrington, 1991, Chapter 1) — says that, under suitable conditions, a positive submartingale can be decomposed in a unique way as:

$$\mathbf{S}_t = \mathbf{A}_t + \mathbf{M}_t$$

where:

(a) $\mathbf{A}_t$ is a non-decreasing right-continuous predictable process (which obviously must be of bounded variation); and
(b) $\mathbf{M}_t$ is a martingale.

Now, note that a counting process $\mathbf{N}_t$ with jumps of +1 is a positive submartingale; apply this result and we have that $\mathbf{A}_t$ exists such that:

$$\mathbf{M}_t = \mathbf{N}_t - \mathbf{A}_t \tag{8}$$

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

is a martingale. The process $\mathbf{A}_t$ is called the *compensator* of the counting process $\mathbf{N}_t$. There follow some simple examples of compensators, which are more familiar than might be supposed from the definition:

(a) *A homogeneous Poisson process.* Given a Poisson process $\mathbf{P}_t$ with parameter λ, the compensator is the function λt. To see this, note that by the memoryless property of the Poisson process:

$$\mathrm{E}[\mathbf{P}_{t+s} \mid \mathbf{P}_t] = \mathbf{P}_t + \lambda s$$

so that:

$$\mathrm{E}[\mathbf{P}_{t+s} - \lambda(t+s) \mid \mathbf{P}_t - \lambda t] = \mathbf{P}_t - \lambda t$$

or, in other words, $\mathbf{P}_t - \lambda t$ is a martingale. However, λt is a non-decreasing predictable (in fact, deterministic) process; hence by the uniqueness of the Doob-Meyer decomposition λt must be the compensator of $\mathbf{P}_t$.

(b) *An inhomogeneous Poisson process.* Given a Poisson process $\mathbf{P}^*_t$ with time-dependent parameter $\lambda(t)$, we know that:

$$\mathrm{E}[\mathbf{P}^*_{t+s} \mid \mathbf{P}^*_t] = \mathbf{P}^*_t + \int_t^{t+s} \lambda(u)du.$$

Define $\Lambda_t = \int_0^t \lambda(u)du$. Then we have:

$$\mathrm{E}[\mathbf{P}^*_{t+s} - \Lambda_{t+s} \mid \mathbf{P}^*_t - \Lambda_t] = \mathbf{P}^*_t - \Lambda_t$$

or, in other words, $\mathbf{P}^*_t - \Lambda_t$ is a martingale. By uniqueness again, Λ_t is the compensator of $\mathbf{P}^*_t$. Note that this is also a deterministic function. It can be shown that any counting process with a continuous deterministic compensator must be a Poisson process (Karr, 1991).

(c) *An indicator of death.* Let $\mathbf{T}_x$ be the random lifetime of a life aged x, and let $\mathbf{N}_t$ be the counting process taking the value 1 if the life has died not later than age $x+t$ and 0 otherwise. Define:

$$\mathbf{Y}_t = \mathbf{I}_{\{\mathbf{T}_x \geq t\}} = \begin{cases} 1 \text{ if } \mathbf{T}_x \geq t \\ 0 \text{ if } \mathbf{T}_x < t. \end{cases}$$

The choice of inequalities in the definition of $\mathbf{Y}_t$ ensures that it is a predictable process. Suppose that the force of mortality μ_{x+t} is continuous. Then define:

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

$$\mathbf{A}_t = \int_0^t \mathbf{Y}_s \mu_{x+s} ds = \int_0^t \mathbf{Y}_s d\Lambda_s. \tag{9}$$

Intuitively, we have 'stopped' the integrated hazard process $\Lambda_t = \int_0^t \mu_{x+s} ds$ at the time of death. (This formulation of the hazard is a special case of Aalen's multiplicative intensity model which will be discussed later.) The continuity of μ_{x+t} makes it obvious that $\mathbf{A}_t$ is predictable. Then we have:

$$\mathrm{E}[\mathbf{N}_{t+s} - \mathbf{A}_{t+s} \mid \mathscr{F}_t] = \mathbf{N}_t - \mathbf{A}_t \tag{10}$$

where $\mathscr{F}_t$ is a filtration representing all the information about the processes $\mathbf{N}_t$ and $\mathbf{A}_t$ up to and including time t. Hence $\mathbf{A}_t$ is the compensator of $\mathbf{N}_t$; note that it is a stochastic process. The intuitive explanation of this is that $\mathbf{N}_t$ behaves like the inhomogeneous Poisson process with parameter μ_{x+t} until time $\mathbf{T}_x$, when the process jumps. Since no more jumps are possible, the process $\mathbf{N}_t$ is identically 1 after that time, and it is clear from equation (8) that the compensator must also be constant after that time.

(d) *Transitions in a multiple state model.* Consider, for example, the illness-death model of Part I, Section 3. We can describe the movements of a single life in terms of a multivariate counting process with 4 components, each component counting the number of transitions of a given type. Note that the components counting transitions between the able and ill states can jump more than once.

Each component is a counting process in its own right, and has a compensator. As an example, let $\mathbf{N}_t$ be the component counting the number of transitions able $\rightarrow$ dead. This process can only jump while the life is in the able state, so define:

$$\mathbf{Y}_t^a = \mathbf{I}_{\{\text{In able state at time } t^-\}}. \tag{11}$$

$\mathbf{Y}_t^a$ is a predictable process indicating presence in the able state. Now define:

$$\mathbf{A}_t = \int_0^t \mathbf{Y}_s^a \mu_{x+s} ds = \int_0^t \mathbf{Y}_s^a d\Lambda_s \tag{12}$$

and it is easy to see that $\mathbf{A}_t$ is the compensator of $\mathbf{N}_t$. The compensators of the other components are defined in a similar way.

Note that, in the last two examples, the indicator processes $\mathbf{Y}_t$ and $\mathbf{Y}_t^a$ are related to the waiting times defined in Part I, Section 3. For example, in the illness-death model the waiting time in the able state (for a single individual) is just:

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

$$\mathbf{V}_t = \int_0^t \mathbf{Y}_s^a ds. \tag{13}$$

In fact, this integral is the same as that which is estimated in census formulae for exposed to risk.

A general definition of the integrated hazard process, where the underlying lifetime distribution may be mixed, is:

$$\Lambda_t = \int_0^t \frac{dF(t)}{1 - F(t^-)}. \tag{14}$$

The 'stopped' process, defined as in the last term of equation (9), does not now have left-continuous sample paths, it has jumps at the discontinuities of Λ_t, at which points it is right-continuous. It is, however, predictable, and it is the compensator of the corresponding counting process; see Fleming & Harrington (1991, Chapter 1) for a proof of this slightly surprising result. (The proof depends on the fact that $\mathbf{Y}_t$ is predictable.)

8.6 *Moment Properties of Counting Processes*

Given a counting process $\mathbf{N}_t$ with compensator $\mathbf{A}_t$, we need to know how the moments of the process behave. The martingale central limit theorem will guarantee convergence to a Gaussian process, but this still leaves the problem of computing the asymptotic mean and second moments of the limit process. We will show that we can do so if $\mathbf{A}_t$ is known, which explains why it is useful to specify models in terms of compensators.

In this and the following sections, $\mathbf{M}_t$ will denote a compensated counting process; that is $\mathbf{M}_t = \mathbf{N}_t - \mathbf{A}_t$:

(a) *First moments.* Since $\mathbf{M}_t$ is a martingale, the first moment behaviour of $\mathbf{N}_t$ is very simple; it is given by equation (10) or by:

$$\mathrm{E}[d\mathbf{N}_t \mid \mathscr{F}_{t^-}] = \mathrm{E}[d\mathbf{A}_t \mid \mathscr{F}_{t^-}] \tag{15}$$

(b) *The predictable variation process.* Second moment results require an application of the Doob-Meyer decomposition to the non-negative submartingale $\mathbf{M}_t^2$. This shows that, under some mild conditions, there is a non-decreasing right-continuous predictable process, denoted $\langle \mathbf{M} \rangle_t$, such that $\mathbf{M}_t^2 - \langle \mathbf{M} \rangle_t$ is a martingale. $\langle \mathbf{M} \rangle_t$ is called the *predictable variation process* of $\mathbf{M}_t$. The origin of the term 'variation' can be seen by writing:

$$\mathrm{E}[d(\mathbf{M}_t^2) \mid \mathscr{F}_{t^-}] = \mathrm{E}[d\langle \mathbf{M} \rangle_t \mid \mathscr{F}_{t^-}] = d\langle \mathbf{M} \rangle_t$$

by the predictability of $\langle \mathbf{M} \rangle_t$; but:

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

$$\mathrm{E}[d(\mathbf{M}_t^2) \mid \mathscr{F}_{t^-}] = \mathrm{E}[d(\mathbf{M}_t)^2 \mid \mathscr{F}_{t^-}]$$

because $\mathbf{M}_t$ is a martingale (heuristic hint: expand $\mathrm{E}[d(\mathbf{M}_t)^2 \mid \mathscr{F}_{t^-}]$ and note that terms in $\mathbf{M}_{t^-}$ can be taken outside the expectation, being known at time t^-). Hence:

$$d\langle \mathbf{M} \rangle_t = \mathrm{Var}[d\mathbf{M}_t \mid \mathscr{F}_{t^-}].$$

In words, $\langle \mathbf{M} \rangle_t$ integrates the variances of the increments of $\mathbf{M}_t$, conditional on the information accruing as time passes. This result is valid for any suitable martingale, but if $\mathbf{M}_t$ is a compensated counting process we can show more. Suppose that $\mathbf{A}_t$ has continuous sample paths (we will not consider the more general case here). First, consider the case that $\mathbf{N}_t$ can jump at or after time t. Then:

$$\begin{aligned} \mathrm{Var}[d\mathbf{M}_t \mid \mathscr{F}_{t^-}] &= \mathrm{Var}[d\mathbf{N}_t - d\mathbf{A}_t \mid \mathscr{F}_{t^-}] \\ &= \mathrm{Var}[d\mathbf{N}_t \mid \mathscr{F}_{t^-}] \\ &= d\mathbf{A}_t(1 - d\mathbf{A}_t) \approx d\mathbf{A}_t \end{aligned} \tag{17}$$

because:

(a) the left continuity of the compensator implies that $\mathbf{A}_t$ is non-random, given $\mathscr{F}_{t^-}$; and

(b) $d\mathbf{N}_t$ is a $0-1$ random variable with expected value $d\mathbf{A}_t$, given $\mathscr{F}_{t^-}$.

The approximate equality in equation (17) is, in fact, an equality, as can be shown in a more rigorous treatment. Second, in the case that $\mathbf{N}_t$ cannot jump at or after time t the same equality is obvious.

(c) *The predictable covariation process.* Given two martingales $\mathbf{M}_t$ and $\mathbf{M}_t^*$ satisfying some mild conditions, there is a right-continuous predictable process $\langle \mathbf{M}, \mathbf{M}^* \rangle_t$ such that:

$$\mathbf{M}_t\mathbf{M}_t^* - \langle \mathbf{M}, \mathbf{M}^* \rangle_t$$

is a martingale. $\langle \mathbf{M}, \mathbf{M}^* \rangle_t$ is called the *predictable covariation process* of $\mathbf{M}_t$ and $\mathbf{M}_t^*$; the name is explained heuristically by the property:

$$d\langle \mathbf{M}, \mathbf{M}^* \rangle_t = \mathrm{Cov}[d\mathbf{M}_t, d\mathbf{M}_t^*]. \tag{18}$$

If $\langle \mathbf{M}, \mathbf{M}^* \rangle_t = 0$ then $\mathbf{M}_t$ and $\mathbf{M}_t^*$ are said to be *orthogonal.*

These results are useful when the compensator is known or can be estimated from observable quantities. Furthermore, we gain insight into the Poisson central limit theorem, referred to in Part I, Section 4 (Hoem, 1987). One of the conditions required by the martingale central limit theorem is that the predictable variation process (of a suitable sequence of martingales) converges to a

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

deterministic process. Under certain assumptions, this can be proved for our compensated counting processes, and in most cases of interest the resulting process is continuous; but for counting processes the predictable variation process is the compensator, and, as we remarked before, any process with a continuous deterministic compensator is a Poisson process.

8.7 *Application to Multiple State Models*

The key to the application of these results to parameter estimation in multiple state models is that $\mathbf{N}_t$ is observable, so if $\mathbf{A}_t$ is observable apart from an unknown parameter, we can solve $\mathbf{M}_T = 0$ for the parameter (where T is the end-point of the investigation), and then the moment results of equations (15) and (17) can be used in the martingale central limit theorem. As an illustration, we will generalise Sverdrup's equations, (Part I, equation (21)), for first moments and (Part I, equations (22) and (35)) for second moments in the illness-death model of Part I, Section 3 (Sverdrup, 1965).

Take the transition able → dead as an example, assuming:

(a) a constant force of mortality μ; and

(b) that the ith life is able at age x.

We change the notation slightly from that in Part I, Section 3, and here let $\mathbf{N}_{i,t}$ denote the observed number of deaths. Define an indicator process $\mathbf{Y}^a_{i,t}$ as in equation (11) indicating that the life is in the able state. The compensator of $\mathbf{N}_{i,t}$ is given by equation (12) with constant μ. Putting $t=0$ in equation (10) then gives:

$$\begin{aligned} \mathrm{E}[\mathbf{N}_{i,t} \mid \mathscr{F}_0] &= \mathrm{E}[\mathbf{A}_{i,t} \mid \mathscr{F}_0] \\ &= \mu \mathrm{E}\left[\int_0^t \mathbf{Y}^a_{i,s} ds \mid \mathscr{F}_0\right] \\ &= \mu \mathrm{E}[\mathbf{V}_{i,t} \mid \mathscr{F}_0] \end{aligned} \tag{19}$$

where $\mathbf{V}_{i,t}$ is the waiting time in the able state up to time t. If we end the investigation at $t = 1$, then equation (19) is Sverdrup's (Part I, equation (21)). For the variance, apply equation (17) to the martingale $\mathbf{M}_{i,t} = \mathbf{N}_{i,t} - \mu \mathbf{V}_{i,t}$.

$$\begin{aligned} \mathrm{E}[(\mathbf{M}_{i,t})^2 \mid \mathscr{F}_0] &= \int_0^t \mathrm{E}[d(\mathbf{M}^2_{i,s}) \mid \mathscr{F}_0] \\ &= \int_0^t \mathrm{E}[\mathrm{E}[d(\mathbf{M}^2_{i,s}) \mid \mathscr{F}_{s^-}] \mid \mathscr{F}_0] \\ &= \int_0^t \mathrm{E}[d\langle \mathbf{M}_i \rangle_s \mid \mathscr{F}_0] \end{aligned} \tag{20}$$

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

$$= \int_0^t \mathrm{E}[d(\mu \mathbf{V}_{i,s}) \mid \mathscr{F}_0]$$

$$= \int_0^t \mathrm{E}[d\mathbf{N}_{i,s} \mid \mathscr{F}_0] = \mathrm{E}[\mathbf{N}_{i,t} \mid \mathscr{F}_0] \tag{21}$$

which is Sverdrup's (Part I, equation (22)). Further, if we extend the random variables defined in Part I, Section 3 to stochastic processes in the same manner as above, we see that (for example):

$$\mathrm{E}[d\langle(\mathbf{N}_i - \mu \mathbf{V}_i), (\mathbf{U}_i - \nu \mathbf{W}_i)\rangle_t \mid \mathscr{F}_t] = 0 \tag{22}$$

because the life can only be in one state at a time, and upon integration (and conditioning on $\mathscr{F}_0$) we obtain Sverdrup's (Part I, equation (35)). In fact:

$$\langle(\mathbf{N}_i - \mu \mathbf{V}_i), (\mathbf{U}_i - \nu \mathbf{W}_i)\rangle_t = 0$$

that is, $(\mathbf{N}_{i,t} - \mu \mathbf{V}_{i,t})$ and $(\mathbf{U}_{i,t} - \nu \mathbf{W}_{i,t})$ are orthogonal martingales.

By shifting our viewpoint from random variables in a model with a fixed observational period — the multiple state model of Part I, Section 3 — to stochastic processes describing events evolving over time, we find the natural setting for problems in which the future is conditioned upon the past.

8.8 *Stochastic Integrals*

Given two deterministic functions $F(t)$ and $G(t)$, we can form the Stieltjes integral of F with respect to G over an interval $[a, b]$:

$$\int_a^b F(t)\, dG(t).$$

The Appendix gives a brief explanation of the Stieltjes integral, and, in particular, notes that, if $G(t)$ is the distribution function of a random variable $\mathbf{G}$ taking values in $[a, b]$, whether continuous, discrete or mixed, then:

$$\int_a^b F(t) dG(t) = \mathrm{E}[F(\mathbf{G})].$$

Given two stochastic processes $\mathbf{F}_t$ and $\mathbf{G}_t$, defined on the same sample space, a single realisation yields a sample path from each process, namely a function

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

F(t) and a function *G(t)*. A natural and useful question is to ask if it is possible to form a Stieltjes integral of *F(t)* with respect to *G(t)* over the intervals [0, *t*) ($t \geq 0$). If this could be done, then the integral:

$$\mathbf{S}_t = \int_0^t \mathbf{F}_s d\mathbf{G}_t(t)$$

would itself be a stochastic process; a function of time through the upper limit of integration, and a function of the sample space through the sample paths. Under what conditions is it possible to form such a *stochastic integral*?

It is enough that the process $\mathbf{F}_t$ should be predictable and bounded, and the process $\mathbf{G}_t$ should be of bounded variation. More precisely, we require the sample paths of the processes to have these properties (with probability 1), since we are forming integrals path-by-path. All of the processes which we use in survival analysis have these properties; in particular a counting process $\mathbf{N}_t$ and its compensator $\mathbf{A}_t$ are of bounded variation, therefore so is the martingale $\mathbf{M}_t = \mathbf{N}_t - \mathbf{A}_t$.

Integration with respect to a martingale $\mathbf{M}_t$ has an important consequence. Let $\mathbf{H}_t$ be a bounded predictable process. Then:

$$\begin{aligned} \mathrm{E}\left[d\int_o^t \mathbf{H}_s d\mathbf{M}_s \mid \mathscr{F}_{s^-} \right] &= \mathrm{E}[\mathbf{H}_t d\mathbf{M}_t \mid \mathscr{F}_{t^-}] \\ &= \mathbf{H}_t \mathrm{E}[d\mathbf{M}_t \mid \mathscr{F}_{t^-}] \qquad (23) \\ &= 0 \qquad (24) \end{aligned}$$

where equation (23) follows from the predictability of $\mathbf{H}_t$, and equation (24) because $\mathbf{M}_t$ is a martingale. In other words, the integral itself is a martingale, as was mentioned previously.

The stochastic integrals used here are very simple. Because the sample paths of counting processes and compensators are well-behaved, we can form the integrals path-by-path using ordinary, deterministic calculus. The difficulties arise when an integral is formed with respect to a process whose sample paths are not of bounded variation, of which Brownian motion is the prototype. Then ordinary calculus fails, and the stochastic integral (most often the Itô integral) must be defined by a limiting argument. None of these difficulties apply here; our integrals are very straightforward and we have no need of any truly stochastic calculus.

In particular, a stochastic integral with respect to a counting process is no more than the sum of the integrand evaluated at the jump points of the counting process, weighted by the sizes of the jumps. Assuming jumps of +1, we have:

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

$$\int_0^t \mathbf{H}_s d\mathbf{N}_s = \sum_{s \in \mathcal{J}} \mathbf{H}_s \tag{25}$$

where $\mathcal{J}$ is the (path-dependent) set of jump points of $\mathbf{N}_t$.

If we wish to study the distributions of statistics by writing them as stochastic integrals, we have to find the predictable variation process of the integral, since this embodies the second moment behaviour. There is a simple result where the integrator is a martingale:

$$\left\langle \int \mathbf{H}_s d\mathbf{M}_s \right\rangle_t = \int_0^t \mathbf{H}_s^2 d\langle \mathbf{M} \rangle_s \tag{26}$$

and there is a similar result for the predictable covariation process of two stochastic integrals, which we will not show here.

8.9 *Aalen's Multiplicative Intensity Model*

We have noted above the forms of the compensators for certain simple counting processes, including that which jumps just once, at the time of death. We did not allow for censoring, however. Suppose now that observation of a random lifetime $\mathbf{T}_i$ is subject to censoring at a random time $\mathbf{U}_i$, which we suppose to be independent of $\mathbf{T}_i$. (Censoring at the end of a fixed period of observation is a special case.) We observe $\mathbf{T}_i^{\min} = \min(\mathbf{T}_i, \mathbf{U}_i)$, and we define an indicator $\mathbf{D}_i$ which takes the value 1 if death is observed and 0 otherwise. The counting process representing an observed death is now defined as:

$$\mathbf{N}_{i,t} = \begin{cases} 1 \text{ if } \mathbf{T}_i^{\min} \le t \text{ and } \mathbf{D}_i = 1 \\ 0 \text{ otherwise.} \end{cases}$$

The compensator of $\mathbf{N}_{i,t}$ is obtained by 'stopping' the integrated hazard when it becomes impossible for the process to jump in future, which happens:

(a) when death occurs; or

(b) when censoring occurs.

A convenient device is the following. Define:

$$\mathbf{Y}_{i,t} = \mathbf{I}_{\{\mathbf{T}_i^{\min} \ge t\}} = \begin{cases} 1 \text{ if } \mathbf{T}_i^{\min} \ge t \\ 0 \text{ if } \mathbf{T}_i^{\min} < t \end{cases}$$

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

or, in words, $\mathbf{Y}_{i,t}$ indicates that the ith life is under observation. Note carefully the inequalities; we have defined $\mathbf{Y}_{i,t}$ to be 1 if $\mathbf{T}_i^{\min} \geq t$, that is, to jump to 0 just after time $\mathbf{T}_i^{\min}$, so that $\mathbf{Y}_{i,t}$ will be a predictable process. Now define the hazard by:

$$\lambda_{i,t}^* = \mathbf{Y}_{i,t}\lambda_{i,t} \tag{27}$$

where $\lambda_{i,t}$ is hazard rate associated with the random lifetime $\mathbf{T}_i$ in the presence of censoring. For simplicity, we are assuming that $\lambda_{i,t}$ is continuous. Then the compensator of $\mathbf{N}_{i,t}$ is:

$$\mathbf{A}_{i,t} = \int_0^t \lambda_{i,s}^* ds. \tag{28}$$

The hazard $\lambda_{i,t}^*$ is the product of:
(a) a deterministic hazard rate; and
(b) a stochastic indicator.

This is the basis of *Aalen's multiplicative model* (Aalen, 1978). The inclusion of the indicator $\mathbf{Y}_{i,t}$ allows us to specify the counting process $\mathbf{N}_{i,t}$ by specifying the form of its compensator, and so to apply the machinery outlined above to censored observations. It is a very simple idea, but a very powerful one, easily applied to more general models. In any multiple state model, for example, in which $\mathbf{N}_{i,t}$ counts the number of jumps between two given states, we can define the process $\mathbf{Y}_{i,t}$ to be 1 whenever the life is in the originating state at time t^-, and 0 otherwise; it is simple to allow for further random censoring (Andersen & Borgan, 1985).

$\lambda_{i,t}$ corresponds to:
(a) the transition intensity in a Markov model for two decrements; or
(b) the crude hazard rate of the competing risks model.

Let $\lambda_{i,t}^n$ denote the net hazard rate, that is, the hazard associated with the random lifetime $\mathbf{T}_i$. We are assuming the independence of $\mathbf{T}_i$ and $\mathbf{U}_i$ here, so that $\lambda_{i,t} = \lambda_{i,t}^n$, but, more generally, it can be shown that $\lambda_{i,t} = \lambda_{i,t}^n$ if and only if:

$$\mathbf{N}_{i,t} - \int_0^t \mathbf{Y}_{i,s}\lambda_{i,s}^n ds$$

is a martingale. So the counting process approach provides a necessary and sufficient condition (weaker than the independence of $\mathbf{T}_i$ and $\mathbf{U}_i$) for the equality of the crude and net hazards.

Next, we assume that the $\mathbf{N}_{i,t}$ ($i = 1, 2, \ldots, N$) comprise a collection of independent counting processes. If the underlying hazards $\lambda_{i,t}$ are all equal, as in

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

the observation of a homogeneous group of lives of the same age, then we can define the aggregated processes:

$$\mathbf{N}_t = \sum_{i=1}^{i=N} \mathbf{N}_{i,t} \tag{29}$$

$$\mathbf{Y}_t = \sum_{i=1}^{i=N} \mathbf{Y}_{i,t} \tag{30}$$

$$\mathbf{A}_t = \sum_{i=1}^{i=N} \mathbf{A}_{i,t} = \int_0^t \mathbf{Y}_s \lambda_s ds \tag{31}$$

$$\mathbf{M}_t = \mathbf{N}_t - \mathbf{A}_t \tag{32}$$

where λ_t is the common underlying hazard. We now show how this construction allows the properties of non-parametric estimates and the Cox model to be explored. In each case we show how certain model quantities arise naturally as stochastic integrals, and state without proof some consequences.

8.10 *Application to Non-Parametric Estimation*

We can interpret the aggregated counting process, above, as representing observations on N independent lives with a common (unknown) underlying hazard rate. Then:

$$\int_0^t \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s} d\mathbf{M}_s \tag{33}$$

is a martingale, where $\mathbf{I}$ is an indicator process, taking the value 1 when the stated condition is true and 0 otherwise. Its purpose here is to avoid division by zero; by convention we take the integrand to be 0 when $\mathbf{Y}_s = 0$. Now notice that:

$$\begin{aligned}\int_0^t \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s} d\mathbf{M}_s &= \int_0^t \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s} d\mathbf{N}_s - \int_0^t \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s} d\mathbf{A}_s \\ &= \int_0^t \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s} d\mathbf{N}_s - \int_0^t \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s} \mathbf{Y}_s \lambda_s ds \\ &= \int_0^t \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s} d\mathbf{N}_s - \int_0^t \mathbf{I}_{\{\mathbf{Y}_s>0\}} \lambda_s ds.\end{aligned} \tag{34}$$

Each term on the right hand side of equation (34) has a simple interpretation. The

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

second term is the integrated hazard, Λ_t, over the range of ages for which lives are under observation. The first term is:

$$\int_0^t \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s} d\mathbf{N}_s = \sum_{\mathcal{J}} \frac{1}{\mathbf{Y}_s} \tag{35}$$

where $\mathcal{J}$ is the set of times at which deaths occur; that is, the jump points of $\mathbf{N}_t$. Compare this with (Part II, equation (17)); we see that it is the Nelson-Aalen estimate. The martingale property (equation (33)) then tells us that the Nelson-Aalen estimate is an unbiased estimate of the second term of equation (34), which differs from the integrated hazard by:

$$\int_0^t \lambda_s ds - \int_0^t \mathbf{I}_{\{\mathbf{Y}_s>0\}} \lambda_s ds = \int_0^t \mathbf{I}_{\{\mathbf{Y}_s=0\}} \lambda_s ds \tag{36}$$

so that the Nelson-Aalen estimate is approximately an unbiased estimate of the integrated hazard, the bias arising from the probability that all the lives should have left observation. Clearly, in a large enough study this probability can be made small, and, in fact, it decreases exponentially as N increases (for a proof of this and subsequent results in this section, see Fleming & Harrington (1991, Chapter 3)). The following notation is convenient in the sequel:

$$\Lambda_t = \int_0^t \lambda_s ds$$

$$\Lambda_t^* = \int_0^t \mathbf{I}_{\{\mathbf{Y}_s>0\}} \lambda_s ds$$

$$\hat{\Lambda}_t = \int_0^t \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s} d\mathbf{N}_s .$$

Λ_t is the object of the estimation, and $\hat{\Lambda}_t$ is the observable statistic. It is not an unbiased estimate of Λ_t, but it is an unbiased estimate of Λ_t^*. The variance of $\hat{\Lambda}_t - \Lambda_t^*$ is given by the predictable variation formula:

$$\left\langle \int \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s} d\mathbf{M}_s \right\rangle_t = \int_0^t \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s^2} d\langle \mathbf{M} \rangle_s$$

$$= \int_0^t \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s^2} d\mathbf{A}_s . \tag{37}$$

Since $\Lambda_t^* - \Lambda_t$ is small, a good estimate of the mean squared error $\mathrm{E}[(\hat{\Lambda}_t - \mathrm{E}[\hat{\Lambda}_t])^2]$ is:

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

$$\int_0^t \frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s^2} d\mathbf{N}_s. \tag{38}$$

It is the hazard λ_t rather than the integrated hazard Λ_t which is usually the object of interest, especially for computational work. The Nelson-Aalen estimate lends itself naturally to estimation of λ_t by means of *kernel smoothing* (similar to moving average smoothing); see Ramlau-Hansen (1983). If $K(t)$ is a suitable kernel function, zero outside $[-1, 1]$ and integrating to unity, then:

$$\hat{\lambda}_t = \frac{1}{b}\int K\left(\frac{t-s}{b}\right) d\hat{\Lambda}_s \tag{39}$$

is an estimate of λ_t. It can be shown to be asymptotically consistent as the sample size increases and the bandwidth b tends to zero (the bandwidth corresponds to the number of terms in a moving average formula); the proofs rely on the fact that:

$$\frac{1}{b}\int K\left(\frac{t-s}{b}\right)(d\hat{\Lambda}_s - d\Lambda_s^*) = \frac{1}{b}\int K\left(\frac{t-s}{b}\right)\frac{\mathbf{I}_{\{\mathbf{Y}_s>0\}}}{\mathbf{Y}_s} d\mathbf{M}_s \tag{40}$$

is a martingale. Note that these integrals do not provide estimates of λ_t for t in the lower and upper extremities of length b.

The Nelson-Aalen estimate is easier to handle than the Kaplan-Meier estimate because of the very simple martingale representation (equation (33)). We can derive the Kaplan-Meier estimate by product-integration, but this does not lead straightaway to its statistical properties. For completeness, we list below some of the properties of the Kaplan-Meier estimate which are most easily proved by the methods of this section. Let $\hat{F}_t$ be the Kaplan-Meier estimate, and $\hat{S}_t = 1 - \hat{F}_t$ the corresponding estimate of the survival distribution.

(a) It can be shown that:

$$\frac{\hat{S}_t}{S_t} - 1 = \int_0^t \frac{\hat{S}_{s^-}}{S_s}\left(\frac{d\mathbf{N}_s}{\mathbf{Y}_s} - d\Lambda_s\right) \tag{41}$$

from which, by writing:

$$d\Lambda_s = \mathbf{I}_{\{\mathbf{Y}_s>0\}} d\Lambda_s + \mathbf{I}_{\{\mathbf{Y}_s=0\}} d\Lambda_s$$

it follows after some algebra that the bias of $\hat{S}_t$ is:

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

$$\mathrm{E}\left[\mathbf{I}_{\{T<t\}}\frac{\hat{S}_T(S(T)-S(t))}{S(T)}\right]\geq 0$$

where $T = \inf\{s\colon \mathbf{Y}_s = 0\}$ is the time at which the last life leaves observation. The estimate $\hat{F}_t$, therefore, has negative bias, arising from the probability that all lives exit during the period of observation, so again this bias reduces exponentially as $N \to \infty$.

(b) If $\mathbf{Y}_t \to \infty$ as $N \to \infty$ for all $t \in [0, b]$, then $\int \mathbf{Y}_s^{-1}\, d\mathbf{N}_s$ estimates Λ_t consistently and $\hat{F}_t$ estimates F_t consistently for $t \in [0, b]$. A consistent estimator is one whose distance from the 'true' value tends to zero in probability; here we take the 'distance' between two functions $f(t)$ and $g(t)$ to be the supremum norm $\sup_{t\in[0,b]} |f(t) - g(t)|$.

(c) The martingale central limit theorem shows that the process $\int \mathbf{Y}_s^{-1}\, d\mathbf{N}_s$ converges in distribution to a Gaussian process with mean Λ_t and variance given by equation (38).

Some of these results were previously proved by methods other than those described here, in particular see Breslow & Crowley (1974), but, in general, the proofs depended on individual calculations and the connections were not obvious.

8.11 *Application to Logrank Statistics*

A wide variety of logrank statistics can be written as stochastic integrals with respect to martingales which represent sums of compensated counting processes. Consider the two-sample logrank statistic $\sum(d_{1j} - e_{1j})$ of Part II, Section 6. The ith sample can be represented by a collection of counting processes:

$$(\mathbf{N}_{i1}(t),\ \mathbf{N}_{i2}(t),\ldots,\ \mathbf{N}_{iN_i}(t)) \quad (i = 1,\ 2)$$

with associated indicators $\mathbf{Y}_{ij}(t)$ and compensators $\mathbf{A}_{ij}(t)$. Processes with a single subscript (i = 1, 2) represent the aggregation of the component processes of the corresponding sample, and processes with no subscript represent the aggregation of the component processes of both samples. Then:

$$\begin{aligned}\sum(d_{1j} - e_{1j}) &= \int_0^t d\mathbf{N}_1(s) - \int_0^t \frac{\mathbf{Y}_1(s)}{\mathbf{Y}_1(s)+\mathbf{Y}_2(s)}\, d(\mathbf{N}_1(s)+\mathbf{N}_2(s)) \\ &= \int_0^t \frac{\mathbf{Y}_2(s)}{\mathbf{Y}_1(s)+\mathbf{Y}_2(s)}\, d\mathbf{N}_1(s) - \int_0^t \frac{\mathbf{Y}_1(s)}{\mathbf{Y}_1(s)+\mathbf{Y}_2(s)}\, d\mathbf{N}_2(s)\end{aligned}$$

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

$$= \int_0^t \frac{\mathbf{Y}_1(s)\mathbf{Y}_2(s)}{\mathbf{Y}_1(s)+\mathbf{Y}_2(s)} \left(\frac{d\mathbf{N}_1(s)}{\mathbf{Y}_1(s)} - \frac{d\mathbf{N}_2(s)}{\mathbf{Y}_2(s)} \right)$$

$$= \int_0^t \mathbf{K}(s) \left(\frac{d\mathbf{N}_1(s)}{\mathbf{Y}_1(s)} - \frac{d\mathbf{N}_2(s)}{\mathbf{Y}_2(s)} \right) \tag{42}$$

where:

$$\mathbf{K}(s) = \frac{\mathbf{Y}_1(s)\mathbf{Y}_2(s)}{\mathbf{Y}_1(s)+\mathbf{Y}_2(s)}.$$

If the two samples have underlying hazard rates $\lambda_1(s)$ and $\lambda_2(s)$, equation (42) is:

$$\int_0^t \frac{\mathbf{K}(s)}{\mathbf{Y}_1(s)} d\mathbf{M}_1(s) - \int_o^t \frac{\mathbf{K}(s)}{\mathbf{Y}_2(s)} d\mathbf{M}_2(s) + \int_0^t \mathbf{K}(s)(\lambda_1(s) - \lambda_2(s))ds. \tag{43}$$

Under the hypothesis of equal hazards, equation (43) states that the logrank statistic is a martingale, and as the sample sizes increase the martingale central limit theorem establishes the asymptotic Normal distribution of the logrank statistic.

The same method can be applied to weighted logrank statistics, in which a weight function *W(s)* is applied to the integrand in equation (42) (Gill, 1980). The shape of the weight function determines whether the statistic emphasises earlier or later observed lifetimes; this generalises the various logrank statistics mentioned in Part II, Section 6.

8.12 *Application to the Cox Model*

Regression problems can also be put into a counting process framework, by specifying the hazard as a function of the covariates. Here, let the vector $\mathbf{Z}_{i,t}$ be the covariate *process* associated with the ith life. The covariates need not be constants, and need not even be deterministic functions of time, although these are clearly special cases. $\mathbf{Z}_{i,t}$ must be predictable, however, because it appears in the following definition of the compensator $\mathbf{A}_{i,t}$:

$$\mathbf{A}_{i,t} = \int_0^t \mathbf{Y}_{i,s} \exp(\beta \mathbf{Z}_{i,s}^T)\lambda_0(s)ds \tag{44}$$

where β is the vector of regression coefficients. As in Cox's original formulation, the object is to estimate β and possibly the baseline hazard $\lambda_0(s)$. Now we can rewrite the partial likelihood:

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

$$L(\beta) = \prod_{j=1}^{j=k} \frac{\exp(\beta z_j^T)}{\sum_{i \in R(t_j)} \exp(\beta z_i^T)} \tag{45}$$

as:

$$L(\beta) = \prod_{i=1}^{i=N} \left[\frac{\mathbf{Y}_{i,s} \exp(\beta \mathbf{Z}_{i,s}^T)}{\sum_{j=1}^{j=N} \mathbf{Y}_{j,s} \exp(\beta \mathbf{Z}_{j,s}^T)} \right]^{d\mathbf{N}_{i,t}} \tag{46}$$

Then the key step is to write the score function:

$$u(\beta) = \left(\frac{\partial \log L(\beta)}{\partial \beta_1}, \ldots, \frac{\partial \log L(\beta)}{\partial \beta_p} \right) \tag{47}$$

(suitably formulated as a stochastic process) as a sum of stochastic integrals. Take, for example, the first component $\partial \log L(\beta)/\partial \beta_1$ of the score function. This can be written:

$$\frac{\partial \log L(\beta)}{\partial \beta_1} = \sum_{i=1}^{i=N} \int_0^t \left[\mathbf{Z}_{i,s,1} - \frac{\sum_{j=1}^{j=N} \mathbf{Y}_{j,s} \exp(\beta \mathbf{Z}_{j,s}^T) \mathbf{Z}_{j,s,1}}{\sum_{j=1}^{j=N} \mathbf{Y}_{j,s} \exp(\beta \mathbf{Z}_{j,s}^T)} \right] d\mathbf{N}_{i,s}. \tag{48}$$

It is easy to show that, under the hypothesis $\beta = \beta^0$, equation (48) is equivalent to:

$$\sum_{i=1}^{i=N} \int_0^t \left[\mathbf{Z}_{i,s,1} - \frac{\sum_{j=1}^{j=N} \mathbf{Y}_{j,s} \exp(\beta^0 \mathbf{Z}_{j,s}^T) \mathbf{Z}_{j,s,1}}{\sum_{j=1}^{j=N} \mathbf{Y}_{j,s} \exp(\beta^0 \mathbf{Z}_{j,s}^T)} \right] d\mathbf{M}_{i,s}. \tag{49}$$

Then the martingale central limit theorem can be applied to show that this score process, evaluated at β^0, has an asymptotic Gaussian process limit. Subject to some regularity conditions, this property is the key to the asymptotic normal properties of likelihood estimators (Cox & Hinkley, 1974, Chapter 9; Fleming & Harrington, 1991, Chapter 8). Therefore, the estimator obtained by setting equation (47) equal to zero and solving for $\hat{\beta}$ has the same properties. The formulation *via* the censoring processes $\mathbf{Y}_{i,t}$ and the covariate processes $\mathbf{Z}_{i,t}$ means that this result holds under a wide variety of censoring schemes and dependencies. Gill (1984) gave a simple introduction to the counting process approach to the Cox model.

Stochastic process approaches are of value when the data have dependencies over time, making analysis of simpler models difficult. It is interesting that the explanatory power of a martingale approach is seen in equally striking fashion in

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

the proof of Hattendorf's Theorem in life contingencies (Papatriandafylou & Waters, 1984).

9. Actuarial Education

The approach we have adopted might seem to differ from the traditional actuarial education; we have not even mentioned rate intervals! The reader might ask how much the student should be expected to know about these statistical models (we exclude the topics surveyed in Section 8 from consideration).

Were we to accept the utilitarian view of Haycocks & Perks quoted in the introduction to Part I, the syllabus might cover only those computational tools which the average actuary would use. On the other hand, we might worry that statisticians and others are fast disappearing over the horizon and taking our heritage with them. We make three general points below, and then consider the actuary's education in the subject matter of this review.

(a) Models beget computational tools and not vice versa. Pitfalls await those who would extend tools to new problems and new data, without considering the models upon which the tools are based.

(b) Actuarial science gives rise to a distinctive jargon where it deals with distinctive problems, but otherwise it would be helpful to use the same language as everyone else, and to beware of assuming actuarial ownership of widespread techniques.

(c) If actuarial work should evolve in future, in the face of competition from other numerate professionals, education has to emphasise principles, even at the expense of methods. The view that deterministic methods, l_x and related functions, and the calculation of exposed to risk form an adequate approach to actuarial science is no longer tenable. Likewise, the temptation to strip down the mathematical basis of the subject to the bare minimum needed to access current applications should be strongly resisted.

In our view, the relevant part of the syllabus should be based on the Markov models of Part I, Section 3, emphasising estimation and computation as aspects of a coherent modelling process, not as separate subjects. In such a context, the place of exposed to risk should be clear.

It is desirable to present alternative models, not least to make the point that there is no one correct model for all problems. The Binomial and Poisson models are obvious candidates; brief mention of the actuarial estimate would be more than enough comment on initial exposed to risk. An introduction to non-parametric estimation and the Cox model is also required; it is difficult to call any treatment adequate that does not include these tools.

Competing risks models offer no advantages over multiple state models, though they are common enough in the literature that the student should, perhaps, be aware of their existence. The approach to multiple decrements should make clear the distinction between observable and unobservable quantities, and calculations

involving the latter should be clearly identified as such and not confused with estimation.

REFERENCES

References were given in part I.

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

APPENDIX

STIELTJES INTEGRALS

Recall the definition of the Riemann integral of a non-negative function $F(t)$ over the interval $[a, b)$. We define a partition of the interval to be a set of points:

$$a = t_0 < t_1 \cdots < t_{k-1} < t_k = b$$

and we define the mesh of the partition to be:

$$\sup_{j=0}^{j=k-1} (t_{j+1} - t_j).$$

The lower and upper Riemann sums are respectively:

$$\sum_{j=0}^{j=k-1} (t_{j+1} - t_j) \inf_{t \in [t_j, t_{j+1})} F(t) \quad \text{and} \quad \sum_{j=0}^{j=k-1} (t_{j+1} - t_j) \sup_{t \in [t_j, t_{j+1})} F(t)$$

and the Riemann integral is defined as the limit of the lower and upper Riemann sums as the mesh of the partition tends to zero, provided these exist and are finite and equal. It is visualised as the area under the graph of $F(t)$.

The Stieltjes integral is a generalisation of the familiar Riemann integral, allowing the function values $F(t)$ to be weighted by the rate of change of an integrating function $G(t)$. The lower and upper Riemann sums are modified as follows:

$$\sum_{j=0}^{j=k-1} (G(t_{j+1}) - G(t_j)) \inf_{t \in [t_j, t_{j+1})} F(t) \quad \text{and} \quad \sum_{j=0}^{j=k-1} (G(t_{j+1}) - G(t_j)) \sup_{t \in [t_j, t_{j+1})} F(t)$$

and the Stieltjes integral is defined as the limit of these sums as for the Riemann integral. It is often called the Riemann-Stieltjes integral, and is written:

$$\int_{[a,b)} F(t) dG(t) \quad \text{or just} \quad \int_{[a,b)} F(t) dG.$$

The definition above treats an interval of the form $[a, b)$; the integral can be extended to closed intervals and unbounded intervals.

If $G(t)$ is a probability distribution, the Stieltjes integral is very easily visualised. First suppose that $G(t)$ is the distribution of a continuous random variable **G**, with density function $g(t)$, taking values in the interval $[a, b)$.

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press

Multiply and divide the jth term in each of the lower and upper sums by $(t_{j+1}-t_j)$, and notice that, as the mesh of the partitions tends to zero:

$$\lim_{(t_{j+1}-t_j)\to 0} \frac{G(t_{j+1})-G(t_j)}{t_{j+1}-t_j}$$

tends everywhere to the derivative $g(t)$. Hence, in this case, the Stieltjes integral is:

$$\int_{[a,b)} F(t)dG(t) = \int_{[a,b)} F(t)g(t)dt = \mathrm{E}[F(\mathbf{G})].$$

Second, suppose that $G(t)$ is the distribution of a discrete random variable $\mathbf{G}$ distributed on the points $t_1 < t_2 < \ldots < t_k$, with probability p_j at the point t_j $(1 \le j \le k)$. Then $G(t)$ has a jump of size p_j at the point t_j $(1 \le j \le k)$, and is constant everywhere else, so the Stieltjes integral is:

$$\int_{[a,b)} F(t)dG(t) = \sum_{j=1}^{j=k} F(t_j)p_j = \mathrm{E}[F(\mathbf{G})].$$

In other words, the Stieltjes integral is a natural way in which expected values of functions of random variables can be written down.

Of more direct relevance to counting process data, consider the case that $G(t)$ is a single sample path of some counting process $\mathbf{N}_t$. Then the Stieltjes integral is:

$$\int_{[a,b)} F(t)dG(t) = \sum_{t\in\mathcal{J}} F(t)$$

where $\mathcal{J}$ is the set of jump points of $\mathbf{N}_t$ (which depends on the particular sample path). Thus, statistics which consist of sums of quantities observed when specified events occur (such as transitions between states) can be simply represented as Stieltjes integrals; an example is the Nelson-Aalen estimate. There are other, more general, theories of integration which we will not discuss.

https://doi.org/10.1017/S1357321700003524 Published online by Cambridge University Press