B.A.J. **2**, II, 429–448 (1996)

AN ACTUARIAL SURVEY OF STATISTICAL MODELS FOR DECREMENT AND TRANSITION DATA

II: COMPETING RISKS, NON-PARAMETRIC AND REGRESSION MODELS

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

ABSTRACT

This paper surveys some statistical models of survival data. Competing risks models are described; the unidentifiability of net decrements suggests a sceptical approach to the use of underlying single decrement tables. Approaches based on observations of complete lifetimes (with censoring) are surveyed including the Kaplan-Meier and Nelson-Aalen estimates. Regression models for lifetimes depending on covariates are discussed, in particular the Cox model and partial likelihood estimation.

KEYWORDS

Competing Risks; Cox Model; Kaplan-Meier Estimate; Nelson-Aalen Estimate; Partial Likelihood; Proportional Hazards; Survival Analysis

INTRODUCTION

Part I, Sections 1–4 of this paper described some models used to analyse survival data over short segments of lifetimes. In this part, Section 5 describes a statistical model for multiple decrement data, the competing risks model, and compares it with the multiple state models of Part I. Sections 6 and 7 describe approaches to the analysis of complete lifetime data. Section 6 discusses non-parametric methods, including the widely-used Kaplan-Meier or product-limit estimate, while Section 7 describes semi-parametric methods, with emphasis on the Cox proportional hazards model.

All of the material in this part is standard in the statistical literature, and it might usefully be included in the actuary's toolkit. Acknowledgements and references were given in Part I.

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

5. COMPETING RISKS MODELS

5.1 *Specification of the Model*

In this section we consider the extension of models for one decrement to two or more decrements. It should be clear from Part I, Section 3 that multiple decrements can be handled simply within the multiple state framework. There is, however, an extensive literature on a different approach known as *competing risks* models; see Gail (1975) or David & Moeschberger (1978). These models

lead to the so-called dependent and independent rates of decrement, long familiar to actuaries.

One formulation of a model for several decrements is a pair of random variables, say $\mathbf{T}$ and $\mathbf{J}$. $\mathbf{T}$ records the time spent under observation, and $\mathbf{J}$ is an integer-valued random variable indicating the reason for the cessation of observation. For example $\mathbf{J}=0$ might indicate censoring at the end of the investigation, $\mathbf{J}=1$ death and $\mathbf{J}=2$ withdrawal. See Bowers *et al.* (1986) for an example of this approach, and note also that the data for the two state model in Part I, Section 3 were of this type. More generally, let there be M decrements, and let $\mathbf{J}=1, 2,\ldots, M$ indicate the observed decrement or $\mathbf{J}=0$ censoring at a fixed time.

A competing risks model associates a failure time random variable with each decrement. Let $\mathbf{T}_j$ (for $j=1, 2,\ldots, M$) be the time to failure (from the start of observation) under the jth decrement. Let $\mathbf{T}_0$ be a fixed time at which censoring will occur ($\mathbf{T}_0$ could be ∞). Then we observe:

$$\mathbf{T} = \min(\mathbf{T}_0, \mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_M)$$
$$\mathbf{J} = \{j : \mathbf{T}_j = \mathbf{T}\}.$$

In some cases, the failure times $\mathbf{T}_1$, $\mathbf{T}_2,\ldots,$ $\mathbf{T}_M$ have an obvious physical representation. The time to failure of a machine with several components, or the time to the first death of several lives, might naturally be modelled as the minimum of several failure times, but if the $\mathbf{T}_1$, $\mathbf{T}_2,\ldots,$ $\mathbf{T}_M$ are 'lifetimes' in respect of one person, more imagination is needed. For example, if $\mathbf{J}=1$ represents death and $\mathbf{J}=2$ represents withdrawal from insurance, it is difficult to give any meaning to the event $\mathbf{T}_2 > \mathbf{T}_1$.

5.2 *Crude and Net Hazards*

Associated with each decrement ($j=1, 2,\ldots, M$) are two forces. In actuarial notation, the dependent force of decrement (known as the *crude hazard rate* to statisticians) is:

$$(a\mu)_t^j = \lim_{dt\to 0^+} \frac{\mathrm{P}[t < \mathbf{T} \le t+dt, \mathbf{J}=j | \mathbf{T} > t}{dt} \tag{1}$$

and the independent force of decrement (known as the *net hazard rate* to statisticians) is:

$$\mu_t^j = \lim_{dt\to 0^+} \frac{\mathrm{P}[t < \mathbf{T}_j \le t+dt | \mathbf{T}_j > t}{dt}. \tag{2}$$

The net hazards correspond to each decrement acting alone, while the crude

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

hazards correspond to each decrement acting in the presence of the others. In many problems, the mathematics is simplified if the crude and net hazards are equal. It can be shown that independence of the failure times $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_M$ is sufficient (though not necessary) to give equality of the crude and net hazards. For sufficiency, note that:

$$\mathrm{P}[t < \mathbf{T} \leq t + dt,\ \mathbf{J} = j | \mathbf{T} > t] \leq \mathrm{P}\left[t < \mathbf{T}_j \leq t + dt,\ \bigcap_{i \neq j} \mathbf{T}_i > t | \mathbf{T} > t\right]$$

$$= \frac{\mathrm{P}[t < \mathbf{T}_j \leq t + dt] \times \prod_{i \neq j} \mathrm{P}[\mathbf{T}_i > t]}{\prod_{i=1}^{i=M} \mathrm{P}[\mathbf{T}_i > t]}$$

$$= \frac{\mathrm{P}[t < \mathbf{T}_j \leq t + dt]}{\mathrm{P}[\mathbf{T}_j > t]}$$

$$= \mathrm{P}[t < \mathbf{T}_j \leq t + dt | \mathbf{T}_j > t]$$

and, on dividing by dt and taking right limits, we have $(a\mu)_t^j \leq \mu_t^j$. Also:

$$\mathrm{P}[t < \mathbf{T} \leq t + dt,\ \mathbf{J} = j | \mathbf{T} > t] \geq \mathrm{P}\left[t < \mathbf{T}_j \leq t + dt,\ \bigcap_{i \neq j} \mathbf{T}_i > t + dt | \mathbf{T} > t\right]$$

$$= \frac{\mathrm{P}[t < \mathbf{T}_j \leq t + dt] \times \prod_{i \neq j} \mathrm{P}[\mathbf{T}_i > t + dt]}{\prod_{i=1}^{i=M} \mathrm{P}[\mathbf{T}_i > t]}.$$

By the right-continuity of distribution functions:

$$\lim_{dt \to 0^+} \mathrm{P}[\mathbf{T}_i > t + dt] = \mathrm{P}[\mathbf{T}_i > t]$$

so dividing by dt and taking right limits gives $(a\mu)_i^j \geq \mu_t^j$. This point is often glossed over in textbooks; for example Neill (1977) said:

> " ... the force of decrement in the multiple-decrement table is not based on a time interval and is not affected by the other decrements, giving $(a\mu)_x^k = \mu_x^k$."

Unfortunately this would imply that crude and net hazards are always equal, which is false. The equality of $(a\mu)_t^j$ and μ_t^j is not a fact; it is either an assumption (Makeham, 1874) or the consequence of some other assumption. Bailey & Haycocks (1946, Section 11; 1947) gave a clear treatment of this point, albeit not a statistical one; their emphasis on the forces of decrement as the fundamental quantities of the model and probabilities as derived quantities anticipated some modern developments. The assumption of independent failure

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

times is often very strong and even unrealistic, an obvious example being a selective decrement such as withdrawal from assurance.

5.3 *Identifiability*

Turning to inference, the fact that we only observe $\mathbf{T}$ and $\mathbf{J}$ leads to a major problem; the joint distribution of the failure times $\mathbf{T}_1$, $\mathbf{T}_2$,..., $\mathbf{T}_M$ cannot be identified from these data. This is perhaps not surprising, since we have specified a model in terms of random variables which we cannot observe directly. In terms of the hazard rates, only the crude hazards are observable; the net hazards are intrinsically unobservable.

Let $S_{T_1,\ldots,T_M}(t_1,\ldots t_M)$ be the joint survivor function:

$$S_{T_1,\ldots,T_M}(t_1,\ldots,t_M) = \mathrm{P}\left[\bigcap_{j=1}^{j=M} \mathbf{T}_j > t_j\right]. \tag{3}$$

Associated with this are the marginal survivor functions for each decrement:

$$S_{T_j}(t_j) = S_{T_1,\ldots,T_M}(0,\ldots,0,\ t_j,\ 0,\ldots,0) \tag{4}$$

and the overall survivor function:

$$S_T(t) = \mathrm{P}[\mathbf{T} > t] = \mathrm{P}\left[\bigcap_{j=1}^{j=M} \mathbf{T}_j > t\right] = S_{T_1,\ldots,T_M}(t,t,\ldots,t). \tag{5}$$

The overall survivor function is observable, but, in general, the marginal survivor functions are not. It is easily shown that, in the absence of simultaneous failures, $(a\mu)_t^1 + \cdots + (a\mu)_t^M$ is the overall hazard associated with $S_T(t)$, and on integrating we have:

$$S_T(t) = \prod_{j=1}^{j=M} \exp\left(-\int_0^t (a\mu)_s^j \, ds\right). \tag{6}$$

We have made no assumption about the independence of the failure times; they may be dependent; but now define a set of independent failure times $\mathbf{U}_j$, for $j = 1,\ldots, M$, by specifying their survivor functions as:

$$S_{U_j}(t) = \exp\left(-\int_0^t (a\mu)_s^j \, ds\right) \tag{7}$$

and define $\mathbf{U}_0 = \mathbf{T}_0$, then these new failure times define a competing risks model under which the minimum failure time $\mathbf{U} = \min(\mathbf{U}_0, \mathbf{U}_1,\ldots, \mathbf{U}_M)$ and the type of failure $\mathbf{K} = \{k{:}\mathbf{U}_k = \mathbf{U}\}$ have the same joint distribution as $\mathbf{T}$ and $\mathbf{J}$. Therefore, we cannot tell from any amount of data whether the process being observed is represented by independent or dependent competing risks. This is the

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

identifiability problem; see Tsiatis (1975), Elandt-Johnson & Johnson (1980, Chapter 9), Kalbfleisch & Prentice (1980, Chapter 7) or Cox & Oakes (1981, Chapter 9). Robinson gave a simple example in the discussion of Broffit (1984). Crowder (1991) showed that, even if it is possible to observe each cause of exit in the absence of the others as well as in their presence, knowledge of all these survival functions still does not allow the model to be identified. Carrière (1994) gave an example of the wide range of possibilities resulting from the (hypothetical) elimination of deaths from heart/cerebrovascular disease, all of which were consistent with the data.

It is sometimes possible to obtain lifetime distributions which can be identified from the observed minimum by imposing greater structure upon the model, such as a given parametric form. For example, Arnold & Brockett (1983) showed that the bivariate Makeham model with survival function given by:

$$S(x,y) = \exp(-c_1 x - c_2 y - c_3 \max(x,y) - d_1 \exp(s_1 x) - d_2 \exp(s_2 y)) \qquad (8)$$

is identifiable; that is, the observable $S(t,t)$ and crude (dependent) survival distributions allow all the parameters c_1, c_2, c_3, d_1, d_2, s_1, s_2 to be estimated. (The model can be interpreted as the joint lifetimes of two lives with simultaneous deaths modelled by a Poisson process with parameter c_3; without the last two terms it is the bivariate exponential model of Marshall & Olkin (1967).)

Arnold & Brockett (1983) also showed that if the underlying lifetime distributions depend on a structure variable $\mathbf{W}$, but conditional on $\mathbf{W} = \lambda$, are independent with proportional hazard rates $(a\mu)_t^j = \lambda \delta_j h(t)$, then knowledge of $h(t)$ or knowledge of the distribution of $\mathbf{W}$ results in an identifiable model. Heckman & Honore (1989) showed identifiability in a family of proportional hazards models with covariates. Proportional hazards are important in survival analysis, and will be discussed in Section 7.

5.4 *Multiple State or Competing Risks?*

Competing risks models have attracted considerable criticism in the statistical literature:

(a) They are founded on unobservable quantities —- the underlying lifetimes or the net hazards —- with the consequent problem of identifiability. Sometimes the lifetimes have a physical interpretation, especially in reliability studies of systems of components, but in survival analysis this is less usual.

(b) The assumption of independent failure times is often made in order to simplify the mathematics, even if it is manifestly unreasonable. The actuarial assumption $(a\mu)_t^j = \mu_t^j$ is a case in point.

Aalen (1987) described the approach as leading to "distortion of the statistical analysis, and to artificial problems, like the question of identifiability". He also pointed out the infeasibility of a competing risks approach to more general transitions. Prentice *et al.* (1978) said:

> "It therefore seems important to concentrate on the [crude hazard] functions for statistical modelling as they lead to procedures that have a clear interpretation regardless of the interrelation between causes of failure and yet are identical with the more traditional results, based upon independent latent failure times, in circumstances in which an independence assumption is justifiable. ... It is perhaps surprising that this approach has received so much attention in the literature."

Net hazards are motivated by the wish to compute the effect of each decrement acting alone, or the effect of the removal of one or more decrements. They are, therefore, included in the model specification, despite the fact that they are unobservable and irrelevant for inference. For inference, it is sensible to restrict attention to what is observable, and estimate the crude hazard rates $\{(a\mu)_t^j\}$. Then the multiple state model seems to be the most natural approach. If we must compute unobservable quantities, it seems better to treat the resulting calculations as hypothetical. The example by Carrière (1994) made this point very clear.

5.5 *Multiple Decrements and the Actuarial Estimate*

Our treatment of the actuarial estimate in Part I, Section 4 was based on a single decrement. In practice, there are usually at least two decrements and the effect of using the actuarial estimate is to treat deaths as exposed until the end of the year of death and all other exits, including 'enders', as exposed until the time of exit. The unequal treatment of one of the decrements arises, essentially, from an attempt to estimate the so-called single decrement table associated with the chosen decrement (see Benjamin & Pollard (1980, Chapter 6) or Bowers *et al.* (1986, Chapter 9)). In terms of the competing risks model, this is equivalent to estimating one of the marginal survival functions given by equation (4), but the unidentifiability of these functions makes this impossible. It is hardly surprising that the calculation of initial exposed to risk has been a source of boundless confusion to students over the years. Redington was reported to say, in the discussion of Bailey & Haycocks (1947), that:

> "In the ordinary census formula, $\frac{1}{2}P$ at the beginning of the year and $\frac{1}{2}P'$ at the end of the year were obvious and easily remembered; but when adjustments had to be made to give the deaths a full year's exposure the fog descended."

The fact that such problems simply do not arise in the multiple state approach supports our view that it should be the actuary's model of choice.

6. NON-PARAMETRIC ESTIMATION

6.1 *The Kaplan-Meier (Product-Limit) Estimator*

Non-parametric estimation for uncensored data was described in Part I, Section 2, namely to estimate the survivor function by the proportion still alive at each future time. In this section we develop this important idea to allow for censoring.

We will consider lifetimes as a function of time t without mention of a starting age x. The following could be applied equally to newborn lives, to lives aged x

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

at outset, or to lives with some property in common at time $t = 0$, for example diagnosis of a medical condition. Medical studies are often based on time since diagnosis or time since the start of treatment, and if the patient's age enters the analysis it is usually as an explanatory variable in a regression model.

Suppose we observe a population of N lives in the presence of non-informative censoring, and suppose we observe m deaths. Let:

$$t_1 < t_2 < \ldots < t_k$$

be the ordered times at which deaths were observed. We do not assume that $k = m$, so more than one death might be observed at a single failure time. Suppose that d_j deaths are observed at time t_j, $(1 \le j \le k)$, so that $d_1 + d_2 + \ldots + d_k = m$. Observation of the remaining $N - m$ lives is censored; suppose that c_j lives are censored between times t_j and t_{j+1}, $(0 \le j \le k)$ where we define $t_0 = 0$ and $t_{k+1} = \infty$ to allow for censored observations after the last observed failure time; then $c_0 + c_1 + \ldots + c_k = N - m$. Strictly, we regard all the c_j censored observations as falling in the open interval (t_j, t_{j+1}). Suppose that the times at which observations are censored within this interval are $t_{j1}, t_{j2}, \ldots, t_{jc_j}$ (which need not be distinct). It will also be convenient to define n_j to be the number of lives who are alive and at risk at time t_j^-, that is, just before the jth observed lifetime $(1 \le j \le k)$. To obtain the likelihood of these observations, without making any prior assumptions about the form of $F(t)$, proceed as follows:

(a) *Deaths.* The probability that a death occurs at time t_j is $F(t_j) - F(t_j^-)$.

(b) *Censored observations.* The probability that a life should survive to be censored at time t_{jl} is $1 - F(t_{jl})$, under non-informative censoring.

Therefore the total likelihood is:

$$\prod_{j=1}^{j=k} (F(t_j) - F(t_j^-))^{d_j} \prod_{j=0}^{j=k} \prod_{l=1}^{l=c_j} (1 - F(t_{jl})). \tag{9}$$

We ask what *function* $F(t)$ will maximise this likelihood, constrained only by the requirement that it should be a distribution function. Since any distribution function is non-decreasing, each factor $(1 - F(t_{jl}))$ will be maximised if $F(t_{jl}) = F(t_j)$, while we must have $F(t_j) > F(t_j^-)$ at each observed lifetime or the likelihood will be zero. Therefore any maximum likelihood estimate of $F(t)$ will be a step function, with jumps at each observed lifetime.

It is convenient to extend to discrete distributions the definition of a hazard function given in Part I, Section 2 for continuous distributions. Suppose $F(t)$ has probability masses at the points $t_1, t_2, \ldots t_k$. Then define:

$$\lambda_j = \mathrm{P}\left[\mathbf{T} = t_j | \mathbf{T} \ge t_j\right] \ (1 \le j \le k). \tag{10}$$

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

This definition is valid for discrete or mixed distributions. If we assume that **T** has a discrete distribution then:

$$1-F(t)=\prod_{t_j\le t}(1-\lambda_j)$$

so that, with the conventions that $F(0) = 0$ and $d_0 = 0$, the likelihood (9) can be written:

$$\begin{aligned}&\prod_{j=1}^{j=k}\left(\frac{(F(t_j)-F(t_j^-))}{1-F(t_j^-)}\right)^{d_j}\prod_{j=0}^{j=k}\left((1-F(t_j^-))^{d_j}\prod_{l=1}^{l=c_j}(1-F(t_{jl}))\right)\\&=\prod_{j=1}^{j=k}\lambda_j^{d_j}\prod_{j=0}^{j=k}(1-F(t_j^-))^{d_j}(1-F(t_j))^{c_j}\\&=\prod_{j=1}^{j=k}\lambda_j^{d_j}(1-\lambda_j)^{n_j-d_j}.\end{aligned}\tag{11}$$

This is proportional to a product of independent binomial likelihoods, so that the maximum is attained by setting:

$$\hat{\lambda}_j=\frac{d_j}{n_j}\quad(1\le j\le k)\tag{12}$$

$$\hat{F}(t)=1-\prod_{t_j\le t}(1-\hat{\lambda}_j).\tag{13}$$

This is the *Kaplan-Meier* or *product-limit* estimate (Kaplan & Meier, 1958). It can be viewed in several ways:

(a) In studying the probability of death over small age intervals, we can choose to divide up the time axis in any way we like. A convenient choice is to have a very short time interval containing each t_j (short enough to exclude any of the censored times t_{jl}) and longer time intervals containing only censored observations. The only information gained from the latter is that there were no deaths, so there is no reason to assume anything except that $F(t)$ is constant within these intervals, while the former contribute Binomial estimates of the hazard at the observed lifetimes.

(b) Alternatively, we might choose finer and finer partitions of the time axis, and estimate $(1 - F(t))$ as the product of the probabilities of surviving each sub-interval. Then, with the above definition of the discrete hazard, we obtain the Kaplan-Meier estimate as the mesh of the partition tends to zero. This is the origin of the name 'product-limit' estimate, by which the Kaplan-Meier estimate is sometimes known.

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

Only those at risk at the observed lifetimes $\{t_j\}$ contribute to the estimate. It follows that it is unnecessary to start observation on all lives at the same time or age; the estimate is valid for data truncated from the left, provided the truncation is non-informative in the sense that entry to the study at a particular age or time is independent of the remaining lifetime. (Note that left-truncation is not the same as left-censoring.)

6.2 *Comparing Lifetime Distributions*

Since Kaplan-Meier estimates are often used to compare the lifetime distributions of two or more populations — for example, in comparing medical treatments — their statistical properties are important. Approximate formulae for the variance of $\tilde{F}(t)$ are available. Greenwood's formula (Greenwood, 1926):

$$\mathrm{Var}[\tilde{F}(t)] \approx (1-\hat{F}(t))^2 \sum_{t_j \le t} \frac{d_j}{n_j(n_j-d_j)} \tag{14}$$

is reasonable over most t, but might tend to understate the variance in the tails of the distribution. An alternative given by Peto *et al.* (1977) was:

$$\mathrm{Var}[\tilde{F}(t)] \approx \frac{(1-\hat{F}(t))^2 \hat{F}(t)}{n(t)} \tag{15}$$

where $n(t)$ is the number of lives at risk at time t. This approximation overstates the variance. These formulae can be used to estimate confidence intervals for $\tilde{F}(t)$ or of transformations of $\tilde{F}(t)$, using a Normal approximation. For example, it is common to calculate confidence intervals of $\log(-\log(1-\tilde{F}(t)))$, since this has an unrestricted range, and then to transform back to obtain confidence intervals of $\tilde{F}(t)$. The results depend on the transformation of $\tilde{F}(t)$ which is chosen. For examples and further discussion, see Collett (1994) and Cox & Oakes (1984).

For testing differences between two estimated lifetime distributions, *logrank* tests are commonly used. These are based on the differences between the actual deaths in one population at the observed lifetime t_j, say d_{1j}, and those expected on the basis of the combined observations, say e_{1j}. The simplest logrank statistic is $\sum(d_{1j}-e_{1j})$, where the summation is over all the observed lifetimes in the combined sample. Other logrank statistics can be calculated by weighting the terms in this sum to give more emphasis to early deaths or later deaths; for example a *generalised Wilcoxon statistic* (Gehan, 1965) is $\sum n_{1j}(d_{1j}-e_{1j})$. Both statistics, when standardised, have for large samples an approximate unit Normal distribution. See Collett (1994) for further details.

A difficulty only recently resolved is that the variances of the logrank and Wilcoxon statistics are calculated by summing the variances of the $(d_{1j}-e_{1j})$ over

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

all the observed lifetimes, assuming independence; for example the variance of the logrank statistic with two samples is taken to be:

$$\sum_{\text{All lifetimes}} \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$$

where now $d_j = d_{1j} + d_{2j}$ and so on. This cannot be strictly correct, since the expected deaths at each observed lifetime depend on the number then at risk, which is not independent of the preceding numbers of deaths. It is, therefore, not clear that approximations based on the central limit theorem, which lead to the asymptotic Normal distribution of the standardised statistic, are applicable; but a problem caused by conditioning on previous events suggests a stochastic process approach, and in Part III, Section 8 it is shown that stochastic processes provide the most natural setting for most of the topics discussed in this paper. In passing, we should note that these and other difficulties have never prevented statisticians from applying useful results in a pragmatic way while waiting for the probabilists to clean up the theory.

Breslow (1993) gave a useful summary of developments since Kaplan & Meier (1958) introduced their estimate.

6.3 *The Actuarial Estimate Revisited*

Sometimes the data are provided in grouped form, so that the $\{t_j\}$ do not represent observed lifetimes, but simply partition the observation period. We change the notation slightly; let $0 = t_0 < t_1 < \ldots < t_k < t_{k+1} = \infty$ be such a partition; let d_j be the number of deaths and c_j the number of right-censored observations in the interval $(t_{j-1}, t_j]$ $(1 \le j \le k+1)$; finally let n_j be the number of lives at risk just after time t_{j-1}. Then an 'actuarial' non-parametric estimate of $F(t)$, along the lines of equations (39) and (40) of Part I, Section 4 is:

$$\hat{F}(t) = 1 - \prod_{j \ge 1, t_j \le t} \left(1 - \frac{d_j}{n_j - \frac{1}{2}c_j}\right). \tag{16}$$

Greenwood's formula, and other tests for comparing survival curves, are adapted (approximately) by substituting $n_j' = n_j - c_j/2$ for n_j where appropriate. See Benjamin & Pollard (1980) or Collett (1994) for details.

6.4 *The Nelson-Aalen Estimate and the Product Integral*

An alternative non-parametric approach is to estimate the integrated hazard:

$$\Lambda_t = \int_0^t \mu_s ds + \sum_{t_j \le t} \lambda_j$$

where the integral deals with the continuous part of the distribution and the sum with the discrete part. Then an estimate of $F(t)$ can be based on the relationship:

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

$$F(t) = 1 - \exp\left(-\int_0^t \mu_s ds\right) \times \prod_{t_j \le t} (1 - \lambda_j).$$

(Some authors define the integrated hazard instead as:

$$\Lambda_t = \int_0^t \mu_s ds - \sum_{t_j \le t} \log(1 - \lambda_j)$$

so that $F(t)$ can be written as $1 - e^{-\Lambda t}$; see for example, Cox & Oakes (1984).) The *Nelson-Aalen estimate* of the integrated hazard is:

$$\hat{\Lambda}_t = \sum_{t_j \le t} \frac{d_j}{n_j}. \tag{17}$$

The Kaplan-Meier estimate can be approximated in terms of $\hat{\Lambda}_t$:

$$\hat{F}_t = 1 - \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right) \tag{18}$$

$$\approx 1 - \exp\left(-\sum_{t_j \le t} \frac{d_j}{n_j}\right). \tag{19}$$

In a broad sense, estimation of Λ_t instead of $F(t)$ corresponds to the use of a multiple state model instead of a Binomial-type model; attention is focused on the hazard function instead of on the probabilities.

The link between the Nelson-Aalen and Kaplan-Meier estimates is expressed neatly in terms of *product integrals*. Regarded as interval functions (i.e. functions from a suitably defined space of intervals of the real line to the real numbers) Λ_t is additive and $S(t) = 1 - F(t)$ is multiplicative. Additive and multiplicative interval functions are related by suitable integrals. Let $\alpha(d)$ be an additive interval function on the positive real line (for simplicity), and let $\mathscr{D}$ be any partition of an interval $\mathscr{I}$, the longest member of which has length $\delta(\mathscr{D})$. Then it can be shown that the product integral:

$$\beta(\mathscr{I}) = \mathscr{P}_{\mathscr{I}}[1 - d\alpha(t)] = \lim_{\delta(\mathscr{D}) \to 0} \prod_{d \in \mathscr{D}} (1 - \alpha(d))$$

exists and is a multiplicative interval function. If $\alpha([0,t])$ is a continuous function of t then the product integral has the form:

$$\beta(\mathscr{I}) = \exp(-\alpha(\mathscr{I})).$$

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

The familiar equation ${}_tp_x = \exp(-\int_0^t \mu_{x+s}ds)$ is, therefore, an example of a product integral; or, if $\alpha([0, t])$ is a step function of t, with jumps $\Delta\alpha_i$ at some set of jump points $\{t_i\}$, then:

$$\beta(\mathscr{I}) = \prod_{t_i \in \mathscr{I}} (1 - \Delta\alpha_i)$$

is a multiplicative interval function. Applying these results to the Nelson-Aalen and Kaplan-Meier estimates, and noting that the former is a step function, we see that equation (18) is the exact representation of $\hat{F}(t)$ and equation (19) is an approximation based on continuity. The Kaplan-Meier estimate is simply the product integral of the Nelson-Aalen estimate. For a survey, see Gill & Johansen (1990).

7. Regression Models

7.1 *Covariates*

Non-parametric approaches are limited in their ability to deal with some important questions in survival analysis, such as the effect of *covariates* on survival. A covariate is any quantity recorded in respect of each life, such as age, sex, type of treatment, level of medication, severity of symptoms and so on. If the covariates partition the population into a small number of homogeneous groups, it is possible to compare Kaplan-Meier or other non-parametric estimates in respect of each population, but a more direct and transparent method is to construct a model in which the effects of the covariates on survival are modelled directly; a regression model. In this section, we will assume that the values of the covariates in respect of the ith life are represented by a $1 \times p$ vector, z_i.

The most widely used regression model in recent years has been the *proportional hazards* model, also known as the Cox model (Cox, 1972), and most of this section is devoted to that model.

7.2 *Fully Parametric Models*

In a fully parametric model, the strong assumption is made that the lifetime distribution belongs to a given family of parametric distributions, and the regression problem is reduced to estimating the parameters from the data. Distributions commonly used are the exponential (constant hazard), Weibull (monotonic hazard), Gompertz-Makeham (exponential hazard) and log-logistic ('humped' hazard). The same distributions are often used as loss distributions with insurance claims data (Hogg & Klugman, 1984), but censored observations complicate the likelihoods considerably and numerical methods are usually required. For the distributions above the likelihoods can be written down (though not always solved) explicitly, which is not the case for some other well-known loss distributions such as the log-normal.

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

Parametric models can be used with a homogeneous population (the one-sample case) instead of the approaches of Section 6, or can be fitted to a moderate number of homogeneous groups, in which case confidence intervals for the fitted parameters give a test of differences between the groups which should be better than non-parametric procedures. However, fully parametric models are difficult to apply without foreknowledge of the form of the hazard function, which might be the very object of the study; for that reason a semi-parametric approach is more popular.

7.3 *The Cox Model*

The Cox model (Cox, 1972) proposes the following form of hazard function for the ith life (where, in keeping with statistical habit, we denote hazards by λ rather than μ):

$$\lambda(t;z_i) = \lambda_0(t)\exp(\beta z_i^T). \tag{20}$$

β is a $1 \times p$ vector of *regression parameters*, so that through the scalar product βz_i^T the influence of each factor in z_i enters the hazard multiplicatively. $\lambda_0(t)$ is the *baseline hazard*. In this simple formulation only $\lambda_0(t)$ depends on time, but the model can also be formulated with time-dependent covariates.

Under the Cox model, the hazards of different lives with covariate vectors z_1 and z_2 are in the same proportion at all times:

$$\frac{\lambda(t;z_1)}{\lambda(t;z_2)} = \frac{\exp(\beta z_1^T)}{\exp(\beta z_2^T)}$$

giving rise to the name *proportional hazards* model. Cox's formulation is not the only model with proportional hazards, we could formulate a model $\lambda(t;z_i) = \lambda_0(t)g(z_i)$ where $g(z)$ is any function of z, but not t. However, Cox's model ensures that the hazard is always positive, and gives a linear model for the log-hazard which is very convenient in both theory and practice.

The utility of this model arises from the fact that the general 'shape' of the hazard function for all individuals is determined by the baseline hazard, while the exponential term accounts for differences between individuals. So, if we are not primarily concerned with the precise form of the hazard, but with the effects of the covariates, we can ignore $\lambda_0(t)$ and estimate β from the data irrespective of the shape of the baseline hazard; this is termed a *semi-parametric* approach. So useful and flexible has this proved, that the Cox model now dominates the literature on survival analysis, and it is probably the tool to which a statistician would turn first for the analysis of survival data.

To estimate β it is usual to maximise the following *partial likelihood*. Let $R(t_j)$ denote the set of lives which are at risk just before the jth observed lifetime, and

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

for the moment assume that there is only one death at each observed lifetime, that is $d_j = 1$ $(1 \le j \le k)$. The partial likelihood (Cox, 1972, 1975) is:

$$L(\beta) = \prod_{j=1}^{j=k} \frac{\exp(\beta z_j^T)}{\sum_{i \in R(t_j)} \exp(\beta z_i^T)}. \tag{21}$$

Intuitively, each observed lifetime contributes the probability that the life observed to die should have been the one out of the $R(t_j)$ lives at risk to die, conditional on one death being observed at time t_j. Note that the baseline hazard cancels out and the partial likelihood depends only on the order in which deaths are observed. (The name 'partial' likelihood arises because those parts of the full likelihood involving the times at which deaths were observed and what was observed between the observed deaths are thrown away.) Maximisation of this expression has to proceed numerically, and most statistics packages have procedures for fitting a Cox model; see Collett (1994) for a recent review.

In practice there might be ties in the data, that is:

(a) some $d_j > 1$; or

(b) some observations are censored at an observed lifetime.

It is usual to deal with (b) by including the lives on whom observation was censored at time t_j in the risk set $R(t_j)$, effectively assuming that censoring occurs just after the deaths were observed. Accurate calculation of the partial likelihood in case (a) is messy, since all possible combinations of d_j deaths out of the $R(t_j)$ at risk at time t_j ought to contribute, and an approximation due to Breslow (1974) is often used, namely:

$$L(\beta) = \prod_{j=1}^{j=k} \frac{\exp(\beta s_j^T)}{\left(\sum_{i \in R(t_j)} \exp(\beta z_i^T)\right)^{d_j}} \tag{22}$$

where s_j is the sum of the covariate vectors z of the d_j lives observed to die at time t_j. For some other approximations see Collett (1994). Kalbfleisch & Prentice (1973, 1980) also discussed estimation of β based on marginal likelihood.

Remarkably, the partial likelihood behaves much like a full likelihood (Cox, 1975); it yields an estimator for β which is asymptotically (multivariate) Normal and unbiased, and whose asymptotic variance matrix can be estimated by the inverse of the observed information matrix. The *efficient score* function, namely the vector function:

$$u(\beta) = \left(\frac{\partial \log L(\beta)}{\partial \beta_1}, \ldots, \frac{\partial \log L(\beta)}{\partial \beta_p} \right) \tag{23}$$

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

plays an important part; in particular solving $u(\hat{\beta}) = 0$ furnishes the maximum likelihood estimate $\hat{\beta}$. The observed information matrix $\mathcal{I}(\hat{\beta})$ is then the $p \times p$ matrix of second partial derivatives:

$$\mathcal{I}(\beta)_{ij} = \frac{\partial^2 \log L(\beta)}{\partial \beta_i \, \partial \beta_j} \qquad (1 \leq i, j \leq p) \tag{24}$$

evaluated at $\hat{\beta}$. This 'good behaviour' on the part of the partial likelihood function is a consequence of the fact that it arises naturally in a counting process framework (see Part III, Section 8) and an even more remarkable fact is that the same counting process framework also brings together multiple state models and non-parametric estimation.

A useful feature of most computer packages for fitting a Cox model is that the information matrix evaluated at $\hat{\beta}$ is usually produced as a by-product of the fitting process (it is used in the Newton-Raphson algorithm) so standard errors of the components of $\hat{\beta}$ are available. These are helpful in evaluating the fit of a particular model.

7.4 *Model Fitting*

In a practical problem, several possible explanatory variables might present themselves, and part of the modelling process is the selection of those which have significant effects. Therefore criteria are needed for assessing the effects of covariates, alone or in combination.

A common criterion is the *likelihood ratio statistic*. Suppose we need to assess the effect of adding further covariates to the model. In general, suppose we fit a model with p covariates, and another model with $p+q$ covariates which include the p covariates of the first model. Each is fitted by maximising a likelihood; let L_p and L_{p+q} be the maximised log-likelihoods of the first and second models respectively. The likelihood ratio statistic is then $-2(L_p - L_{p+q})$, and it has an asymptotic χ^2 distribution on q degrees of freedom, under the hypothesis that the extra q covariates have no effect in the presence of the original p covariates. Strictly this statistic is based upon full likelihoods, but when fitting a Cox model it is used with partial likelihoods.

For example, suppose we have considered a model for the effect of hypertension on survival, in which z has two components, with the level of z_1 representing sex and the level of z_2 representing blood pressure. Suppose we want to test the hypothesis that cigarette smoking has no effect, allowing for sex and blood pressure. Then we could define an augmented covariate vector $z' = (z_1, z_2, z_3)$ in which z_3 is a factor (say, 0 for non-smoker and 1 for smoker) and refit the model. The likelihood ratio statistic $-2(L_2 - L_3)$ then has an asymptotic χ^2 distribution on 1 degree of freedom, under the null hypothesis (which is that the new parameter $\beta_3 = 0$).

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

The likelihood ratio statistic is the basis of various model-building strategies, in which:

(a) we start with the *null model* (one with no covariates) and add possible covariates one at a time; or

(b) we start with a *full model* which includes all possible covariates, and then try to eliminate those of no significant effect.

In addition, it is necessary to test for *interactions* between covariates, in case their effects should depend on the presence or absence of each other. Some examples of model building strategies, and the interpretation of likelihood ratio statistics, are given by Collett (1994).

The likelihood ratio statistic is a standard tool in model selection; for example Forfar, McCutcheon & Wilkie (1988) used it to choose members of a Gompertz-Makeham family of functions for parametric graduations.

7.5 *Further Aspects of the Cox Model*

The Cox proportional hazards model has been intensively studied since it was proposed, and the methodology has been extended and applied to a wide range of problems. Here we can only list a few interesting points about its use; many of these were suggested by Cox (1972).

(a) The Cox model is flexible because it can be used without estimating the baseline hazard $\lambda_0(t)$. If we want an estimate of the hazards $\lambda(t;x)$ however, $\lambda_0(t)$ must be estimated. One approach is to assume a parametric form for $\lambda_0(t)$. The Weibull distribution is often used, since, for the addition of only two parameters, it encompasses a range of decreasing and increasing hazards, and it includes the exponential distribution as a special case. The Weibull distribution with scale parameter α and shape parameter γ has the hazard rate:

$$\lambda(t;\alpha,\gamma) = \alpha\gamma t^{\gamma-1}. \tag{25}$$

Different values of α result in different Weibull distributions, but with the same shape. If we model α multiplicatively in the covariates z:

$$\alpha(\beta) = \alpha_0 e^{\beta z^T}$$

then we can fit α_0, γ and the vector β simultaneously to obtain the entire hazard function $\lambda(t; z)$. Put another way, if the baseline hazard is assumed to be Weibull, then any hazard proportional to it is the hazard of a Weibull distribution with the same parameter γ. This is the proportional hazards property of the Weibull distribution.

An alternative approach, possibly more suitable if there is no reason to assume a Weibull baseline hazard, is to estimate the vector β, as before, and then apply the same reasoning as in the derivation of the Kaplan-Meier

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

estimate to find a non-parametric estimate of $\lambda_0(t)$. See Kalbfleisch & Prentice (1980) for details.

(b) The Cox model can be used with covariates which change over time. This might be necessary if the outcome were thought to depend more on the current value of a covariate than on its value at the start of observation. In this case, the hazards of different lives are no longer proportional, but the baseline hazard cancels out in the partial likelihood for β, as before, so the model retains its usefulness. Although not more difficult to handle in theory, time-varying covariates introduce some practical difficulties. If, in equation (21), the covariate vectors z_j are made functions of time, it is clear that their values must be known at each observed lifetime t_j, not just for the life who dies at that time, but for all the $R(t_j)$ lives then at risk. This imposes fairly stringent requirements on the observational plan, or requires approximate solutions. See Collett (1994) for suggested approaches.

(c) The proportional hazards assumption can be tested by finding Kaplan-Meier estimates of the lifetime distributions of groups of lives with different values of the covariates, and then for each group plotting $-\log(1-\hat{F}(t))$ against $\log t$. Under proportional hazards, these plots should be parallel. Alternatively, fit a time-varying covariate z_{p+1}; if the hazards are proportional the fitted parameter β_{p+1} should be zero. Non-proportional hazards might be the result of the aggregation of several groups, within each of which the assumption is valid, but each of which has a different baseline hazard. (Actuaries will recognise this as a species of spurious selection.) If the groups can be identified, a *stratified* analysis might be possible, in which the parameter vector β is fitted to all groups simultaneously, but each group contributes a separate factor to the partial likelihood. The procedure was described by Kalbfleisch & Prentice (1980).

(d) Several residuals are available to assist in model checking. The *Cox-Snell* residual in respect of the ith life in an uncensored sample is simply:

$$\exp(\hat{\beta} z_i^T)\hat{\Lambda}_{t_i} = -\log \hat{S}_i(t_i) \tag{26}$$

where β and the integrated baseline hazard Λ_t are represented by their estimates, and $S_i(t)$ is the survival function in respect if the ith life. It is easily shown that $S_i(\mathbf{T}_i)$ has an exponential distribution with mean 1, so the Cox-Snell residuals should be close to a sample from that distribution, if the model is adequate. If observation of the ith life is censored at time t_i^c, the Cox-Snell residual can be modified to $\exp(\hat{\beta} z_i^T)\hat{\Lambda} t_i^c + 1$, which represents the expected value of the residual at the (unobserved) actual lifetime, since the exponential distribution is 'memoryless'.

Alternatives to the Cox-Snell residuals are martingale residuals, deviance residuals and score residuals. In each case (as with ordinary linear regression) the residuals are statistics whose distribution is known in a correct model and

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

which should display no dependence on the covariates or pattern over time; evidence of any such feature suggests that the model is inadequate.

7.6 *Regression using GLIM*

Renshaw (1988) applied the Cox model to the Prudential impaired lives data set. Since the data were in respect of insured lives, a natural baseline hazard was one based on the Continuous Mortality Investigation Bureau (CMIB) analyses of assured lives without impairment. Renshaw formulated a proportional hazards model as follows:

$$\lambda(t;z) = \lambda_0^*(t)\exp(\beta z^T) \tag{27}$$

where $\lambda_0^*(t)$ was based upon extrapolation of the A1967-70 mortality table. This method lies between the Weibull proportional hazards model, in which the baseline hazard has a parametric form which is estimated along with β, and Cox's original formulation, in which the baseline hazard is unknown and β alone is estimated. $\lambda_0^*(t)$ is a known baseline hazard which, of course, need not be estimated; its inclusion in the regression might be expected to improve the estimation of β.

Data for each major impairment were analysed separately. Values of possible covariates were recorded for each life in each analysis, allowing the data for each impairment to be split into homogeneous cohorts according to the possible factors. The contribution to the likelihood for β from the ith life in the jth cohort, assuming that observation extends from age x_{ij} to age $x_{ij} + \mathbf{T}_{ij}$ is then:

$$L_{ij}(\beta) = (\lambda_0^*(x_{ij} + t_{ij})\exp(\beta z_j^T))^{d_{ij}} \exp\left(- \int_{x_{ij}}^{x_{ij}+t_{ij}} \lambda_0^*(s)\exp(\beta z_j^T)ds \right) \tag{28}$$

where $\mathbf{D}_{ij}$ is the usual indicator of death or censoring, and the covariate vector z_j is indexed by j alone because, by definition, it is constant within the jth cohort. We can write this conveniently as:

$$L_{ij}(\beta) = (\lambda_{ij}(t_{ij})\exp(\beta z_j^T))^{d_{ij}} \exp\left(-\int_0^{t_{ij}} \lambda_{ij}(s)\exp(\beta z_j^T)ds \right) \tag{29}$$

where $\lambda_{ij}(t) = \lambda_0^*(x_{ij} + t)\ \exp(\beta z_j^T)$ is the hazard in the jth cohort measured from age x_{ij}. As is usual in medical studies, age enters the hazard as a factor in z. The unusual feature of this model is that the baseline hazard is also a function of age, since acceptance of the risk, not just time since acceptance of the risk. From equation (29), the total log-likelihood over all cohorts is:

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

$$\log L(\beta) = \text{constant} + \sum_j (d_j(\log m_j + \beta z_j^T) - \exp(\log m_j + \beta z_j^T)) \quad (30)$$

where $d_j = \sum_i d_{ij}$ is the total number of deaths in the jth cohort and

$$m_j = \sum_i \int \lambda_{ij}(s)ds$$

is the aggregated integrated hazard in the jth cohort. Therefore, if we define $\phi_j = m_j \exp(\beta z_j^T)$, we can write:

$$\log L(\beta) = \text{constant} + \sum_j (d_j \log \phi_j - \phi_j). \quad (31)$$

Renshaw (1988) derived equation (31) and pointed out that it was identical in form to a log-likelihood of independent Poisson random variables $\{\mathbf{D}_j\}$ with means $\{\phi_j\}$. Whitehead (1980) obtained the same result in the more general case of an unknown baseline hazard. In the generalised linear model (GLM) approach, the $\{\mathbf{D}_j\}$ are modelled as Poisson response variables with means $\{\phi_j\}$ and a log-link function $\log \phi_j = \log m_j + \beta z_j^T$. This is one of the simplest GLMs, and the regression process can be carried out by standard computer packages for GLMs such as GLIM or GENSTAT. For an example, including a discussion of model selection, see Renshaw (1988), and for a comprehensive analysis of a large part of the Prudential data set see England (1993).

The significance of this work lies in the application of sound statistical methodologies to actuarial data. Renshaw (1988) was able to extend considerably the previous analysis of Clarke (1978), using tools which are very much in the mainstream of practical statistics. England (1993) compared multiplicative models with the additive models generally used for underwriting impaired lives, and found them to be superior. Statistical methods such as the Cox model and GLMs have developed as computing power has made them useable; thus they are new and unfamiliar to actuaries; but much data which used to defy analysis can now be modelled within a sound statistical framework, and as time passes these methods should be added to the actuary's toolbox.

7.7 *Other Regression Models*

Alternative models have been proposed in case a fully parametric model is not justified, or the proportional hazards property appears not to hold. We describe briefly two of the more important examples; the *accelerated lifetime* model and the *discrete logistic* model.

The accelerated lifetime model supposes that the covariates act multiplicatively on the lifetime itself, which can be expressed as:

$$S_i(t) = S_0(g(z_i)t) \quad \text{or equivalently} \quad \lambda(t;z_i) = g(z_i)\lambda_0(g(z_i)t). \quad (32)$$

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press

As with the Cox model, the choice $g(z_i) = \exp(\beta z_i^T)$ is often convenient. In reliability testing, where items are tested to destruction under different operating conditions, the model has a direct physical interpretation. In medical studies also, the idea that a disease proceeds at different speeds in different individuals is easily understood. Application of the model does not introduce any new principles, and we omit further details except to observe that the Weibull distribution is the most general parametric family which is closed under both the proportional hazards and the accelerated lifetime properties.

The discrete logistic model (Cox, 1972) is equivalent to the Cox model when the lifetimes are distributed on a discrete set. It is specified by:

$$\frac{\lambda(t;z_i)}{1-\lambda(t;z_i)} = \exp(\beta z_i^T)\frac{\lambda_0(t)}{1-\lambda_0(t)} \tag{33}$$

where the hazards here are the discrete hazards of equation (10). (The denominators are needed because $\exp(\beta z_i^T)$ is unbounded.) The partial likelihood is the same as that suggested by Cox (1972) for tied data in the continuous model, and in the limit as the distances between the points of support of the discrete lifetimes tend to zero, the Cox model is obtained.

REFERENCES

References were given in Part I.

https://doi.org/10.1017/S1357321700003469 Published online by Cambridge University Press