Chapter 4

Background: Survival analysis

4.1 Summary

We recall some elementary definitions concerning probability distributions, putting an emphasis toward one minus the usual cumulative distribution function, i.e., the survival function. This is also sometimes called the survivorship function. The closely related hazard function has, traditionally, been the most popular function around which to construct models. For multistate models it can be helpful to work with intensity functions, rather than hazard functions since these allow the possibility of moving in and out of states. This is facilitated by the very important function, $Y(t)$, the "at-risk" indicator. A number of special parametric cases of proportional hazards models are presented. The issue of censoring and the different kinds of censoring is discussed. The "at-risk" indicator $Y_i(w,t)$, taking the value one when the subject i is at risk of making a transition of a certain kind, indicated by w, makes it particularly simple to address more complex issues in survival such as repeated events, competing risks, and multistate modelling. We consider some tractable parametric models, the exponential model in particular.

4.2 Motivation

Survival time T will be a positive random variable, typically right skewed and with a non-negligible probability of sampling large values, far above the mean. The fact that an ordering, $T_1 > T_2$, corresponds to a solid physical interpretation has led some authors to consider that time is somehow different from other continuous random variables, reminiscent of discussion among early twentieth century physicists about the nature of time "flowing inexorably in and of itself." These characteristics are sometimes put forward as a reason for considering techniques other than the classic techniques of linear regression. From a purely statistical viewpoint, this reasoning is incorrect. Elementary transformations fix the skewness problems which, in consequence, reveal themselves as quite superficial. Nor is there any worthwhile, statistical, distinction between time and, say, height or weight. The reason for considering particular techniques, outside of the classical ones of linear regression, is the presence of censoring. In early work censoring came to be viewed as a nuisance feature of the data collection, hampering our efforts to study the main relationships of interest. A great breakthrough occurred when this feature of the data, the censoring, was modelled by the "at-risk" function. Almost immediately it became clear that all sorts of much more involved problems; competing risks, repeated events, correlated outcomes, could all be handled with almost no extra work. Careful use of the "at-risk" indicator was all that would be required. At the heart then of survival analysis is the idea of being at risk for some event of interest taking place in a short time frame (for theoretical study this short time will be made arbitrarily small). Transition rates are then very natural quantities to consider. In epidemiology these ideas have been well rooted for a half-century where age-dependent rates of disease incidence have been the main objects under investigation.

4.3 Basic tools

Time and risk

The insurance example in the introduction highlights an obvious, but important, issue. If driver A, on average, has a higher daily risk than driver B, then his mean time to be involved in an accident will be shorter. Conversely, if driver B has a longer mean time to accident, then he has, on average, a lower daily risk. For many examples we may tend to have in mind the variable time and how it is affected by other variables. But we can think equally well in terms of risk over short time periods, a viewpoint that we will see generalizes more readily to be able to deal with complicated situations. The connection between time and risk is outlined more formally below.

Hazard and related functions

The purpose here is to continue the introduction of preliminary notions and some basic concepts. Before discussing data and estimation we consider the problem in its most simplified form as that of the study of the pair of random variables (T, Z), T being the response variable "survival" of principal interest and Z an associated "explanatory" variable. There would be little difficulty in applying the host of techniques from linear regression to attacking this problem were it not for the presence of a "censoring" variable C. The particularity of C is that, when observed, i.e., $C = c$, we are no longer able to observe values of T for which $T > c$. Also, in most cases, when T is observed, we are no longer able to observe C. Nonetheless an observation on one tells us something about the other, in particular that it must assume some greater value.

Although the joint distribution of (T, Z) can be of interest, we are particularly interested in the conditional distribution of T given Z. First let us consider T alone. The probability density function of T is defined as

$$f(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \Pr(t < T < t + \Delta t), \tag{4.1}$$

where $\lim_{\Delta t \to 0^+}$ means that Δt goes to 0 only through positive values. We define as usual $F(t) = \int_0^t f(u)du$. The survivorship function is written as $S(t) = 1 - F(t)$. If we view the density as the unconditional failure rate, we can define a conditional failure rate as being the same quantity after having accounted for the fact that the individual has already survived until the time point t. We call this $\lambda(t)$ and we define

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \Pr(t < T < t + \Delta t | T > t). \tag{4.2}$$

It helps understanding to contrast equation (4.2) and (4.1) and we can see that $\lambda(t)$ and $f(t)$ are closely related quantities. In a sense the function $f(t)$ for all values of t is seen from the standpoint of an observer sitting at $T = 0$, whereas, for the function $\lambda(t)$, the observer moves along with time looking at the same quantity but viewed from the position $T = t$. Analogous to a density, conditioned by some event, we can define

$$\lambda(t|C > t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \Pr(t < T < t + \Delta t | T > t, C > t). \tag{4.3}$$

The conditioning event $C > t$ is of great interest since, in practical investigations, all our observations at time t have necessarily been conditioned by the event. All associated probabilities are also necessarily conditional. But note that, under an independent censoring mechanism, $\lambda(t|C > t) = \lambda(t)$. This result underlies the great importance of certain assumptions, in this case that of independence between C and T. The conditional failure rate, $\lambda(t)$, is also sometimes referred to as the hazard function, the force of mortality, the instantaneous failure rate or the age-specific failure rate. If we consider a small interval then $\lambda(t) \times \Delta t$ closely approximates the probability of failing in a small interval for those aged t, the approximation improving as Δt goes to zero. If units are one year then these are yearly death rates. The cumulative hazard function is also of interest and this is defined as $\Lambda(t) = \int_0^t \lambda(u)du$. For continuous $\lambda(t)$, using elementary calculus we can see that:

$$\lambda(t) = f(t)/S(t)\,, \;\; S(t) = \exp\{-\Lambda(t)\}\,, \;\; f(t) = \lambda(t)\exp\{-\Lambda(t)\}.$$

Although mathematically equivalent, we may prefer to focus attention on one function rather than another. The survival function, $S(t)$, is the function displaying most clearly the information the majority of applied workers are seeking. The hazard function, $\lambda(t)$, of central concern in much theoretical work, provides the most telling visual representation of time effects. An important function, of theoretical and practical interest, is the conditional survivorship function,

$$S(t, u) = \Pr(T > t | T > u) = \exp\{\Lambda(u) - \Lambda(t)\}\,, \quad (u < t).$$

From this it is clear that $S(t, u) = S(t)/S(t)$ and that $S(u, u) = 1$ so that it is as though the process had been restarted at time $t = u$. Other quantities that may be of interest in some particular contexts are the mean residual lifetime, $m(t)$, and the mean time lived in the interval $[0, t]$, $\mu(t)$, defined as

$$m(t) = E(T - t | T \geq t), \qquad \mu(t) = \int_0^t S(u)du. \tag{4.4}$$

Like the hazard itself, these functions provide a more direct reflection on the impact of having survived until time t. The mean residual lifetime provides a very interpretable measure of how much more time we can expect to survive, given that we have already reached the timepoint t. This can be useful in actuarial applications. The mean time lived in the interval $[0, t]$ is not so readily interpretable, requiring a little more thought (it is not the same as the expected lifetime given that $T < t$). It has one strong advantage in that it can be readily estimated from right censored data in which, without additional assumptions, we may not even be able to estimate the mean itself. The functions $m(t)$ and $\mu(t)$ are mathematically equivalent to one another as well as the three described above and, for example, a straightforward integration by parts shows that $m(t) = S^{-1}(t) \int_t^\infty S(u)du$ and that $\mu(\infty) = E(T)$. If needed, it follows that the survivorship function can be expressed in terms of the mean residual lifetime by

$$S(t) = m^{-1}(t)m(0)\exp\left(-\int_0^t m^{-1}(u)du\right).$$

We may wish to model directly in terms of $m(t)$, allowing this function to depend on some vector of parameters θ. If the expression for $m(t)$ is not too intractable then, using $f(t) = -S'(t)$ and the above relationship between $m(t)$ and $S(t)$, we can write down a likelihood for estimation purposes in the situation of independent censoring. An interesting and insightful relationship (see for instance the Kaplan-Meier estimator) between $S(t)$ and $S(t, u)$ follows from considering some discrete number of time points of interest. Thus, for any partition of the time axis, $0 = a_0 < a_1 <, \dots, a_n = \infty$, we see that

$$S(a_j) = S(a_{j-1})S(a_j, a_{j-1}) = \prod_{\ell \le j} S(a_\ell, a_{\ell-1}).$$

The implication of this is that the survival function $S(t)$ can always be viewed as the product of a sequence of conditional survival functions, $S(t, u)$. Although more cumbersome, a theory could equally well be constructed for the discrete case whereby $f(t_i) = \Pr(T = t_i)$ and $S(t_i) = \sum_{\ell \ge i} f(t_\ell)$. We do not explore this here.

Intensity functions and compartment models

Modern treatment of survival analysis tends to focus more on intensity than hazard functions. This leads to great flexibility, enabling, for

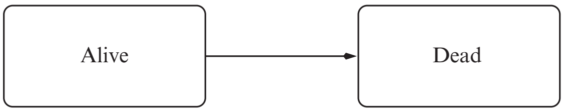

Figure 4.1: A simple alive/dead transition model.

example, the construction of simple models to address questions in complex situations such as repeated events (Andersen and Gill 1982). We believe that both concepts can be useful and we will move back and forth between them according to the application. Intensity functions find their setting in the framework of stochastic processes where the random nature of T is suppressed, t being taken simply as an index to some stochastic process. The *counting process* $N(t)$, takes the value 0 at $t = 0$, remaining at this same value until some time point, say $T = u$, at which the event under study occurs and then $N(t) = 1\,(t \geq u)$. We can then define, in an infinitesimal sense, i.e., the equality only holds precisely in the limit as dt goes to zero through positive values

$$\Pr\{(N(t) - N(t - dt) = 1|\mathcal{F}_{t-dt})\} = \alpha(t)dt \tag{4.5}$$

where $\mathcal{F}_{t-dt}$, written as $\mathcal{F}_{t-}$ when we allow $dt > 0$ to be arbitrarily close to zero, is the accumulated information, on all processes under consideration, observed up until time $t - dt$. The observed set $\mathcal{F}_{t-}$ is referred to as the history at time t. The set is necessarily non decreasing in size as t increases, translating the fact that more is being observed or becoming known about the process. The Kolmogorov axioms of probability, in particular sigma additivity, may not hold for certain noncountable infinite sets. For this reason probabilists take great care, and use considerable mathematical sophistication, to ensure, in broad terms, that the size of the set $\mathcal{F}_{t-}$ does not increase too quickly with t. The idea is to ensure that we remain within the Kolmogorov axiomatic framework, in particular that we do not violate sigma additivity. Much of these concerns have spilled over into the applied statistical literature where they do not have their place. No difficulties will arise in applications, with the possible exception of theoretical physics, and the practitioner, unfamiliar with measure theory, ought not be deterred from applying the techniques of stochastic processes simply because he or she lacks a firm grasp of concepts such as filtrations. It is hard to imagine an application in which a lack of understanding of the term “filtration” could have led to error. On the other hand, the more accessible notions of history, stochastic process, and conditioning sets are central and of great importance both to understanding and to deriving creative structures around which applied problems can be solved. Viewing t as an index to a stochastic process rather than simply the realization of a random variable T, and defining the intensity process $\alpha(t)$ as above, will enable great flexibility and the possibility to model events dynamically as they unfold.

At risk functions $Y(t)$, $Y(w,t)$ and multistate models

The simplest case we can consider occurs when following a randomly chosen subject through time. The information in $\mathcal{F}_{t-}$ tells us whether or not the event has yet occurred and if the subject is still at risk i.e., the set $\mathcal{F}_{t-}$ is providing the same information as an observation on the function $Y(t)$ where we take $Y(t)$ to be left continuous, assuming the value one until the occurrence of an event, or removal from observation, at which time it assumes the value zero. If the simple fact of not having been removed from the study, the event $(C > t)$ is independent of the event $(t < T < t + dt)$, then conditioning on $Y(t) = 1$ is the same as conditioning on $T > t$. Referring then to Equation (4.2) it is clear that if $Y(t) = 0$ then $\alpha(t) = 0$ and, if $Y(t) = 1$ then $\alpha(t) = \lambda(t)$. Putting these two results together we have

$$\alpha(t) = Y(t)\lambda(t). \tag{4.6}$$

This relation is important in that, under the above condition, referred to as the independent censoring condition, the link between the intensity function and the hazard function is clear. Note that the intensity function is random since Y is random when looking forward in time. Having reached some time point, t say, then $\alpha(t)$ is fixed and known since the function $Y(u)$, $0 < u < t$ is known and $Y(t)$ is left continuous.

We call $Y(\cdot)$ the "at risk" function (left continuous specifically so that at time t the intensity function $\alpha(t)$ is not random). The idea generalizes readily and in order to cover a wide range of situations we also allow Y to have an argument w where w takes integer values counting the possible changes of state. For the ith subject in any study we will typically define $Y_i(w,t)$ to take the value 1 if this subject, at time t, is at risk of making a transition of type w, and 0 otherwise. Figure 4.2 summarizes a situation in which there are four states of interest, an absorbing state, death, and three states from which an individual is able to make a transition into the death state. Transitions

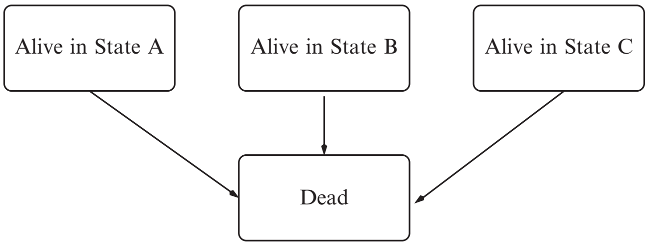

Figure 4.2: A simple compartment model with an absorbing state.

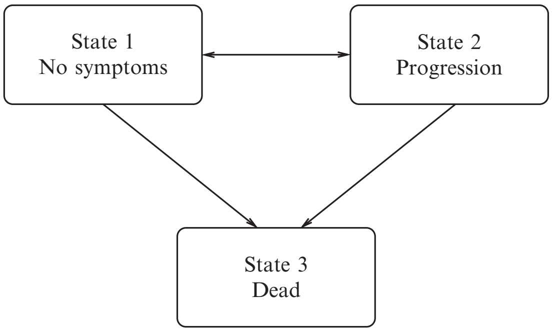

Figure 4.3: A simple compartment model with a single absorbing state.

among the three nondeath states themselves are not allowed. Later we will consider different ways of modeling such a situation, depending upon further assumptions we may wish or not wish to make.

In Figure 4.3 there is one absorbing state, the death state, and two non absorbing states between which an individual can make transitions. We can define $w = 1$ to indicate transitions from state 1 to state 2, $w = 2$ to indicate transitions from state 2 to state 1, $w = 3$ to indicate transitions from state 1 to state 3 and, finally, $w = 4$ to indicate transitions from state 2 to state 3. Note that such an enumeration only deals with whether or not a subject is at risk for making the transition, the transition probabilities (intensities) themselves could depend on the path taken to get to the current state. We can then appreciate why it can be helpful to frame certain questions in terms of compartment models, intensity functions and the risk function. Rather complex situations can be dealt with quite straightforwardly, the figures illustrating simple cases where we can use the argument w in $Y_i(w,t)$ to indicate, at any t, which kinds of transition any given subject i is available to make. In Figure 4.4 there are two absorbing states, one of

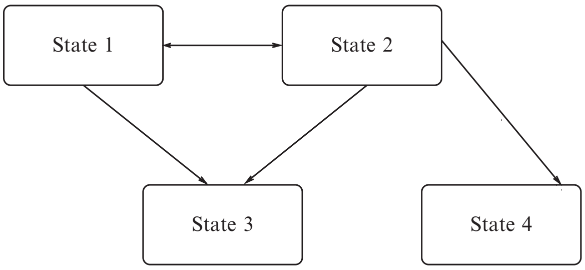

Figure 4.4: A complex compartment model with two absorbing states.

which can only be reached from state 2. The transition rate between state 2 and state 4 may or may not depend on the number of times a subject moves between states 1 and 2. Allowing for transitions between states greatly adds to the flexibility of any model so that, in Figure 4.2, although the explanatory variable (state) has three levels, the model is, in principle, much simpler than that described in Figure 4.3 where the explanatory variable can assume only two states.

At-risk indicator $Y(w,t)$ and repeated events

Some studies have the particularity that an occurrence of the event of interest does not remove the subject from further observation. Additional events, of the same or of different types, may happen. An example is benign breast disease, potentially followed by malignant disease. A patient may have several incidences of benign breast disease at different intervals of time. Following any one of these incidences, or even before such an incidence takes place the subject may become incident for malignant disease. If our interest is essentially focussed on the incidence of malignant disease then we would treat the time-dependent history of benign breast disease as a potential explanatory variable for incidence of malignant disease. However, we may also be interested in modelling directly the repeated incidence of benign breast disease in its own right. Clearly a patient can only be at risk of having a third incident of benign breast disease if she has already suffered two earlier incidents. We can model the rate of incidence for the j th occurrence of benign disease as,

$$\alpha_j(t) = Y(j,t)\lambda_j(t - t_{j-1}), \tag{4.7}$$

where $t_0 = 0$ and t_j is the observed occurrence of the j th event. Different options may be considered for modeling $\lambda_j(t)$. Usually there will be at least one covariate, Z, indicating two distinct prognostic groups, possibly established on the basis of different treatments. The model will involve coefficients multiplying Z and thereby quantifying treatment affect. Allowing these coefficients to also depend upon j provides the broadest generality and is equivalent to analyzing separate studies for each of the occurrences. Stronger modeling, imposing greater structure, might assume that the coefficients do not depend upon j, in which case the information provided by a subject having three incident cases is comparable to that of three independent subjects each providing information on a single incident. So-called marginal models have been proposed in this context. Here, it would be as though the subject, after an event, starts the clock from zero and, aside from covariate information, is deemed to be in the same position as another subject who has just entered the study without having yet suffered a single event. A lot of information would appear to be thereby gained but the set-up seems rather artificial and implausible. Starting the clock from zero, after each event, is sensible but it is more realistic to assume that the underlying hazard rates, i.e., those not adjusted by covariate information, would change with the number of prior incidents. In other words the most sensible model would condition on this information allowing the baseline hazard rate to change according to the number of events counted so far.

4.4 Some potential models

Simple exponential

The simple exponential model is fully specified by a single parameter λ. The hazard function, viewed as a function of time, does not in fact depend upon time so that $\lambda(t) = \lambda$. By simple calculation we find that $\Pr(T > t) = \exp(-\lambda t)$. Note that $E(T) = 1/\lambda$ and, indeed, the exponential model is often parameterized directly in terms of the mean $\theta = E(T) = 1/\lambda$. Also $\text{Var}(t) = 1/\lambda^2$. This model expresses the physical phenomenon of no aging or wearing out since, by elementary calculations, we obtain $S(t+u, u) = S(t)$; the probability of surviving a further t units of time, having already survived until time u, is the same as that associated with surviving the initial t units of time. The property is sometimes referred to as the lack of memory property of the exponential model.

For practical application the exponential model may suggest itself in view of its simplicity or sometimes when the constant hazard assumption appears realistic. A good example is that of a light bulb which may only fail following a sudden surge in voltage. The fact that no such surge has yet occurred may provide no information about the chances for such a surge to take place in the next given time period. If T has an exponential distribution with parameter λ then λT has the so-called standard exponential distribution, i.e., mean and variance are equal to one.

Recall that for a random variable Y having normal distribution $\mathcal{N}(\mu, \sigma^2)$ it is useful to think in terms of a simple linear model $Y = \mu + \sigma\epsilon$, where ϵ has the standard distribution $\mathcal{N}(0,1)$. As implied above, scale changes for the exponential model lead to a model still within the exponential class. However, this is no longer so for location changes so that, unlike the normal model in which linear transformations lead to other normal models, a linear formulation for the exponential model is necessarily less straightforward. It is nonetheless of interest to consider the closest analogous structure and we can write

$$Y = \log T = \alpha + bW, \tag{4.8}$$

where W has the standard extreme value density $f(w) = \exp\{w - \exp(w)\}$. When $\alpha = 0$ we recover an exponential model for T with parameter b, values other than zero for α pushing the variable T out of the restricted exponential class into the broader Weibull class discussed below.

Proportional hazards exponential

In anticipation of the central topic of this book (that of heterogeneity among the subjects under study) imagine that we have two groups, indicated by a binary variable $Z = 0$ or $Z = 1$. For $Z = 0$ the subjects follow an exponential law with parameter λ_0. For $Z = 1$ the subjects follow an exponential law with parameter λ_1. It is clear that for the hazard functions there exists real β $(= \log \lambda_1 - \log \lambda_0)$ such that

$$\lambda(t|Z) = \lambda(t|Z = 0)\exp(\beta Z) = \lambda_0 \exp(\beta Z). \tag{4.9}$$

The important point to note here is that the ratio of the hazards, $\lambda(t|Z = 1)/\lambda(t|Z = 0)$ does not involve t. It also follows that $S(t|Z = 1) = S(t|Z = 0)^\alpha$ where $\alpha = \exp(\beta)$. The survival curves are power transformations of one another. This is an appealing parameterization since, unlike a linear parameterization, whatever the true value of β, the constraints that we impose upon $S(t|Z=1)$ and $S(t|Z=0)$ in order to be well-defined probabilities, i.e., remaining between 0 and 1, are always respected. Such a model is called a proportional hazards model. For three groups we can employ two indicator variables, Z_1 and Z_2, such that, for group 1 in which the hazard rate is equal to λ_0, $Z_1 = 0$ and $Z_2 = 0$, for group 2, $Z_1 = 1$ and $Z_2 = 0$ whereas for group 3, $Z_1 = 0$ and $Z_2 = 1$. We can then write;

$$\lambda(t|Z) = \lambda_0 \exp(\beta_1 Z_1 + \beta_2 Z_2), \tag{4.10}$$

where $\lambda_0 = \lambda(t|Z_1 = Z_2 = 0)$. It is worthwhile bringing the reader's attention to just where the constraints of the model express themselves here. They concern the hazard rates for all groups, which are assumed to be constant. Given this constraint there are no further constraints concerning the relationship between the groups. Suppose, though, that we were to consider a further group, group 4, defined by $Z_1 = 1$ and $Z_2 = 1$. In order to add a fourth group without introducing a further binary coding variable Z_3, we introduce the constraint that the hazard for group 4 is simply expressed in terms of the hazards for groups 2 and 3. Such assumptions are commonly made in routine data analysis but, nonetheless, ought come under critical scrutiny. We return to this issue in later chapters. The extension to many groups follows in the same way. For this we take Z to be a p dimensional vector of indicator variables and β a vector of parameters having the same dimension as Z, the product βZ in Equation 4.9 now implying an inner product, i.e., $\beta Z = \sum_{i=1}^{p} \beta_i Z_i$. In this case the proportional hazards exponential model (4.9) implies that every group follows some simple exponential law, a consequence being that the survivorship function for any group can be expressed as a power transformation of any other group. Once again, it is important to keep in mind just which assumptions are being made, the potential impact of such assumptions on conclusions, and techniques for bringing under scrutiny these assumptions. The proportional hazards constraint then appears as a very natural one in which we ensure that the probabilities $S(t|z)$ and subsequent estimates always remain between 0 and 1. A linear shift added to $S(t|0)$ would not allow for this. We do nonetheless have a linear shift although on a different, and thereby more appropriate, scale and we can write

$$\log - \log S(t|Z) = \log - \log S(t|0) + \sum_{i=1}^{p} \beta_i Z_i .$$

This formulation is the same as the proportional hazards formulation. Noting that $-\log S(T|Z = z)$ is an exponential variate some authors prefer to write a model down as a linear expression in the transformed random variable itself with an exponential error term. This then provides a different link to the more standard linear models we are familiar with.

Piecewise exponential

The lack of flexibility of the exponential model will often rule it out as a potential candidate for application. Many other models, only one or two of which are mentioned here, are more tractable, a property stemming from the inclusion of at least one additional parameter. Even so, it is possible to maintain the advantages of the exponential model's simplicity while simultaneously gaining in flexibility. One way to achieve this is to construct a partition of the time axis $0 = a_0 < a_1 < \ldots < a_k = \infty$. Within the jth interval $(a_{j-1}, a_j)\,, (j = 1, \ldots, k)$ the hazard function is given by $\lambda(t) = \lambda_j$. We can imagine that this may provide quite a satisfactory approximation to a more involved smoothly changing hazard model in which the hazard function changes through time. We use $S(t) = \exp\{-\Lambda(t)\}$ to obtain the survival function where

$$\begin{aligned} \Lambda(t) &= \sum_{j=1}^{k} I(t \geq a_j) \lambda_j (a_j - a_{j-1}) \\ &\quad + \sum_{j=1}^{k} I(a_{j-1} \leq t < a_j) \lambda_j (t - a_{j-1}). \end{aligned} \tag{4.11}$$

Properties such as the lack of memory property of the simple exponential have analogues here by restricting ourselves to remaining within an interval. Another attractive property of the simple exponential is that the calculations are straightforward and can be done by hand and, again, there are ready analogues for the piecewise case. Although the ready availability of sophisticated computer packages tends to eliminate the need for hand calculation, it is still useful to be able to work by hand if for no other purposes than those of teaching. Students gain invaluable insight by doing these kind of calculations the long way.

Proportional hazards piecewise exponential

In the same way as for the simple exponential model, for two groups, indicated by a binary variable $Z = 0$ or $Z = 1$, each having constant piecewise rates on the same intervals, it is clear that there exists β_j such that, for $t \in [a_{j-1}, a_j)$,

$$\lambda(t|Z) = \lambda(t|Z = 0)\exp(\beta_j Z) = \lambda_0(t)\exp\{\beta(t)Z\}, \qquad (4.12)$$

where we now have a function $\beta(t) = \sum_{j=1}^{k} \beta_j I(a_{j-1} \leq t < a_j)$. This can be described as a nonproportional hazards model and, if, under a further restriction that $\beta(t)$ is a constant function of time, i.e., $\beta_1 = \beta_2 = \cdots = \beta_k = \beta$, then, as for the simple exponential model, we have $S(t|Z = 1) = S(t|Z = 0)^{\alpha}$ where $\alpha = \exp(\beta)$ and, once again, such a model is called a proportional hazards model. The model can once more be described in terms of a linear translation on $\log - \log S(t|z)$.

Weibull model

Another way to generalize the exponential model to a wider class is to consider a power transformation of the random variable T. For any positive γ, if the distribution of T^{γ} is exponential with parameter λ, then the distribution of T itself is said to follow a Weibull model whereby

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1}\exp\{-(\lambda t)^{\gamma}\}$$

and $S(t) = \exp -(\lambda t)^{\gamma}$. The hazard function follows immediately from this and we see, as expected, that when $\gamma = 1$ an exponential model with parameter λ is recovered. It is of interest to trace out the possible forms of the hazard function for any given λ. It is monotonic, increasing for values of γ greater than 1 and decreasing for values less than 1. This property, if believed to be reflected in some given physical situation, may suggest the appropriateness of the model for that same situation. An example might be the time taken to fall over for a novice roller blade enthusiast - the initial hazard may be high, initially decreasing somewhat rapidly as learning sets in and thereafter continuing to decrease to zero, albeit more slowly.

The Weibull model, containing the exponential model as a special case, is an obvious candidate structure for framing questions of the sort - is the hazard decreasing to zero or is it remaining at some constant level? A null hypothesis would express this as $H_0 : \gamma = 1$.

Straightforward integration shows that $E(T^r) = \lambda^{-r}\Gamma(1+r/\gamma)$ where $\Gamma(\cdot)$ is the gamma function,

$$\Gamma(p) = \int_0^\infty u^{p-1}e^{-u}du \quad p > 0.$$

For p integer $\Gamma(p) = (p-1)!$ The mean and the variance are $\lambda^{-1}\Gamma(1+1/\gamma)$ and $\lambda^{-2}\Gamma(1+2/\gamma) - E^2$, respectively. The Weibull model can be motivated from the theory of statistics of extremes. The distribution coincides with the limiting distribution of the smallest of a collection of random variables, under broad conditions on the random variables in question (Kalbfleisch and Prentice 1980, page 48).

Proportional hazards Weibull

Once again, for two groups indicated by a binary variable $Z = 0$ or $Z = 1$, sharing a common γ but different values of λ, then there exists a β such that $\lambda(t|Z)/\lambda(t|Z = 0) = \exp(\beta Z)$. Since, as above, the right-hand side of the equation does not depend on t, then we have a proportional hazards model. This situation and the other two described above are the only common parametric models that come under the heading proportional hazards models by simply expressing the logarithm of the location parameter linearly in terms of the covariates. The situation for more than two groups follows as before. Consider however a model such as

$$\lambda(t|Z) = \lambda\gamma(\lambda t)^{\gamma-1}\exp(\beta Z), \tag{4.13}$$

in which Z indicates three groups by assuming the values $Z = 1, 2, 3$.

Unlike the model just above in which three groups were represented by two distinct binary covariates, Z_1 and Z_2, we have only one covariate. In the context of estimation and a given set of data we will almost invariably achieve greater precision in our estimates when there are less parameters to estimate. We would then appear to gain by using such a model. As always though, any such gain comes at a price and the price here is that we have made much stronger assumptions. We are assuming that the signed "distance" between groups 1 and 2, as measured by the logarithm of the hazard, is the same as the signed distance between groups 2 and 3. If this is not the case in reality then we are estimating some sort of compromise, the exact nature of which is determined by our estimating equations. In an extreme case in which the distances are the same but the signs are opposite we might erroneously conclude that there is no effect at all. At the risk of being repetitive, it cannot be stressed too much just how important it is to identify the assumptions we are making and how they may influence our conclusions. Here the assumptions concern both the parametric form of the underlying risk as well as the nature of how the different groups are related. Allowing a shape parameter γ to be other than one provides a more flexible model for the underlying risk than that furnished by the simple exponential model. The choice of covariate coding, on the other hand, is more restrictive than the earlier choice. All of this needs to be studied in applications. An interesting point is that, for the three group case defined as above, the "underlying" hazard, $\lambda(t|Z = 0) = \lambda\gamma(\lambda t)^{\gamma-1}$ does not correspond to the hazard for any of the three groups under study. It is common in practice to consider a recoding of Z, a simple one being $Z-\bar{Z}$, so that the underlying hazard will correspond to some kind of average across the groups. For the case just outlined, another simple recoding is to rewrite Z as $Z - 2$, in which case the underlying hazard corresponds to the middle group, the other two groups having hazard rates lower and greater than this, respectively.

Log-minus-log transformation

As a first step to constructing a model for $S(t|Z)$ we may think of a linear shift, based upon the value of Z, the amount of the shift to be estimated from data. However, the function $S(t|Z)$ is constrained, becoming severely restricted for both $t = 0$ and for large t where it approaches one and zero respectively. Any model would need accommodate these natural constraints. It is usually easiest to do this by eliminating the constraints themselves during the initial steps of model construction. Thus, $\log S(t|Z) = -\Lambda(t)$ is a better starting point for modeling, weakening the hold the constraints have on us. However, $\log - \log S(t|Z) = \log \Lambda(t)$ is better still. This is because $\log \Lambda(t)$ can take any value between $-\infty$ and $+\infty$, whereas $\Lambda(t)$ itself is constrained to be positive. The transformation $\log - \log S(t|Z)$ is widely used and is called the log-minus-log transformation. The above cases of the exponential and Weibull proportional hazards models, as already seen, fall readily under this heading.

Other models

The exponential, piecewise exponential and Weibull models are of particular interest to us because they are especially simple and of the proportional hazards form. Nonetheless there are many other models which have found use in practical applications. Some are directly related to the above, such as the extreme value model in which

$$S(t) = \exp\left(-\exp\left(\frac{t-\mu}{\sigma}\right)\right),$$

since, if T is Weibull, then $\log T$ is extreme value with $\sigma = 1/\gamma$ and $\mu = \log\lambda$. These models may also be simple when viewed from some particular angle. For instance, if $M(s)$ is the moment-generating function for the extreme value density then we can readily see that $M(s) = \Gamma(1+s)$. A distribution, closely related to the extreme value distribution (see Johnson and Johnson 1980), and which has found wide application in actuarial work is the Gompertz where

$$S(t) = \exp\left(\beta\alpha^{-1}(1-e^{\alpha t})\right).$$

The hazard rates for these distributions increase with time, and, for actuarial work, in which time corresponds to age, such a constraint makes sense for studying disease occurrence or death. The normal distribution is not a natural candidate in view of the tendency for survival data to exhibit large skewness, not forgetting that times themselves are constrained to be positive. The log normal distribution has seen some use but is most often replaced by the log-logistic, similar in shape apart from the extreme tails, and much easier to work with. The form is particularly simple for this model and we have

$$S(t) = \left(1+(\alpha t)^{\gamma}\right)^{-1}.$$

For two groups, sharing a common γ but different values of α it is interesting to note that the hazard ratio declines monotonically with time t to its asymptotic value of one. Such a model may be appropriate when considering group effects which gradually wane as we move away from some initial time point.

Parametric proportional hazards models

In principle, for any parametric form, the above providing just a very few examples, we can make a straightforward extension to two or more groups via a proportional hazards representation. For example, if the survivorship functions of two groups are $S(t|Z=1)$ and $S(t|Z=0)$ then we can introduce the parameter α to model one group as a power transform of the other. Rewriting α to include Z via $\alpha = \exp(\beta Z)$ then we have an expression involving the regressors,

$$\log - \log S(t|Z) = \log - \log S(t|Z = 0) + \beta Z. \tag{4.14}$$

All parameters, including β, can be estimated using standard techniques, maximum likelihood in particular, the only restriction being that we require some conditions on the censoring variable C. In practice, standard techniques are rarely used, most likely as a consequence of the attractive proposal of Cox (1972) whereby we can estimate β without having to consider the form of $S(t|Z = 1)$ or $S(t|Z = 0)$. As attractive as the Cox approach is though, we should not overlook the fact that, in exchange for generality concerning the possible parametric forms of functions of interest, such as $S(t|Z)$, making inferences on these population quantities becomes that much more involved. Parametric proportional hazards models may be an area that merits renewed interest in applications.

4.5 Censoring

The most important particularity of survival data is the presence of censoring. Other aspects such as the positivity and skewness of the main random variable under study, time T, and other complex situations such as repeated measures or random effects, are not of themselves reasons for seeking methods other than linear regression. Using transformations and paying careful attention to the structure of the error, linear models are perfectly adequate for dealing with almost any situation in which censoring does not arise. It is the censoring that forces us to consider other techniques. Censoring can arise in different ways.

We typically view the censoring as a nuisance feature of the data, and not of direct interest in its own right, essentially something that hinders us from estimating what it is we would like to estimate. In order for our endeavors to succeed we have to make some assumptions about the nature of the censoring mechanism. The assumptions may often be motivated by convenience, in which case it is necessary to give consideration as to how well grounded the assumptions appear to be as well as to how robust are the procedures to departures from any such assumptions. In other cases the assumptions may appear natural given the physical context of interest, a common case being the uniform recruitment into a clinical trial over some predetermined time interval. When the study closes patients for whom the outcome of interest has not been observed are censored at study close and until that point occurs it may be reasonable to assume that patients are included in the study at a steady rate.

It is helpful to think of a randomly chosen subject being associated with a pair of random variables (T, C), an observation on one of the pair impeding observation on the other, while at the same time indicating that the unobserved member of the pair must be greater than the observed member. This idea is made more succinct by saying that only the random variable $X = \min(T, C)$ can be fully observed. Clearly $\Pr(X > x) = \Pr(T > x, C > x)$ and we describe censoring as being independent whenever

$$\Pr(X > x) = \Pr(T > x, C > x) = \Pr(T > x)\Pr(C > x). \quad (4.15)$$

Type I censoring

Such censoring most often occurs in industrial or animal experimentation. Items or animals are put on test and observed until failure. The study is stopped at some time T^*. If any subject does not fail it will have observed survival time at least equal to T^*. The censoring times for all those individuals being censored is then equal to T^*. Equation (4.15) is satisfied and so this is a special case of independent censoring, although not very interesting since all subjects, from any random sample, have the same censoring time.

Type II censoring

The proportion of censoring is determined in advance. So if we wish to study 100 individuals and observed half of them as failures we determine the number of failures to be 50. Again all censored observations have the same value T^* although, in this case, this value is not known in advance. This is another special case of independent censoring.

Type III censoring

In a clinical trial patients enter randomly. A model for entry is often assumed to be uniform over a fixed study period, anywhere from a few months to several years but determined in advance. Survival time is the time from entry until the event of interest. Subjects can be censored because (1) the end of the study period is reached, (2) they are lost to follow-up (3) the subject fails due to something unrelated to the event of interest. This is called random censoring. So, unlike for *Type I* or *Type II* censoring, for a random sample $C_1, \ldots, C_n$, the C_i could all be distinct.

For a random sample of pairs $(T_i, C_i),\ i = 1, \ldots, n$, we are only able to observe $X_i = \min(T_i, C_i)$. A fundamental result in this context was discovered by Tsiatis (1975). The result says that, for such data, we are unable to estimate the joint distribution of the pair (T, C). Only the marginal distributions can be estimated under the independent censoring assumption, the assumption itself not being testable from such data. It is common then to make the assumption of independent censoring, sometimes referred to as non informative censoring, by stipulating that

$$\Pr(X_i > x) = \Pr(T_i > x, C_i > x) = \Pr(T_i > x) \Pr(C_i > x). \quad (4.16)$$

The assumption is strong but not entirely arbitrary. For the example of the clinical trial with a fixed closing date for recruitment it seems reasonable to take the length of time from entry up until this date as not being associated with the mechanism generating the failures. For loss to follow-up due to an automobile accident or due to leaving the area, again the assumption may be reasonable, or, at least, a good first approximation to a much more complex, unknown, and almost certainly unknowable, reality.

Informative censoring

When censoring is informative, which we can take to be the negation of non-informative, then it is no longer possible to estimate the main quantities of interest without explicitly introducing some model for the censoring. The number of potential models relating C and T is infinite and, in the absence of special knowledge, it can be helpful to postulate some simple relationship between the two, the proportional hazards model itself having been used in this context (Koziol and Green 1976, Slud and Rubinstein 1983). Obvious examples might be surrogate endpoints in the study of the evolution of AIDS following treatment, where, for falling CD4 cell counts, below a certain point patients can be withdrawn from study. Censoring here is clearly informative. This will be the case whenever the fact of removing a subject, yet to experience the event of interest, from study implies a change in risk. Informative censoring is necessarily more involved than non informative censoring and we have to resort to more elaborate models for the censoring itself in order to make progress. If, as might be the case for a clinical trial where the only form of censoring would be the termination of the study, we know for each subject, in advance, their censoring time C, we might then postulate that

$$\log - \log S(t) = \log - \log(S(t|C < t) + \beta I(C > t).$$

This would be a proportional hazards model for a dependent censoring mechanism. More generally we would not know C in advance of making observations on T, but we could write down a similar model in terms of intensity functions, viewing the censoring indicator as a predictable stochastic process. For the purposes of estimation we may require empirical quantities indicating how the risk changes once censoring is observed, and for this we need to be able to compare rates between those censored at some point and those who are not. Mostly, once censoring has occurred, it is no longer possible to observe the main event under study so that, for data of this nature, we are not able to estimate parameters of interest without further assumptions. These assumptions are usually that the censoring is independent of the failure process or that it is conditionally independent given covariate values. The paper of Tsiatis (1975) demonstrates this intuitive observation formally.

Marginal and conditionally independent censoring

When considering many groups, defined by some covariate value Z, there are essentially two types of independence commonly needed. The stronger assumption is that of marginal independence in which the variables T, C, and Z are pairwise independent. The censoring distribution for C is the same for different values of Z. A weaker assumption that is often made, is that of conditional independence. Here, the pair (T, C) are independent given Z. In other words, for each possible value of Z, the pair (T, C) is independent, but the censoring distribution C can be different for different values of Z.

Finite censoring support

Many mathematical issues simplify immediately when the failure variable T is continuous, as we generally suppose, but that the censoring variable is restricted to having support on some finite subset. We can imagine that censoring times are only allowed to take place on the set $\{a_0, a_1, \ldots, a_k\}$. This is not a practical restriction since we can make the division (a_j, a_{j-1}) as fine as we wish. We will frequently need to consider the empirical distribution function and analogues (Kaplan-Meier estimate, Nelson-Aalen estimate) in the presence of censoring. If we adopt this particular censoring set-up of finite censoring support, then generalization from the empirical distribution function to an analogue incorporating censoring is very straightforward. We consider this in greater detail when we discuss the estimation of marginal survival.

4.6 Competing risks as a particular type of censoring

Recalling the "at-risk" indicator function, $Y_i(w, t)$, which takes the value one if, at time t, the i th subject is at risk of making a transition of type w, and is zero otherwise, we can imagine a simple situation in which w takes only one of two values. Calling these $w = 1$ and $w = 2$, consider a constraint whereby $Y_i(1, t) = Y_i(2, t)$. In words, if the ith subject is at risk of one kind of transition, then he or she is also at risk of the other kind. If the subject is no longer at risk then this means that they are not at risk for either kind of transition. Thus, if a subject suffers an event of type $w = 1$ then he is no longer considered at risk of suffering an event of type $w = 2$, and conversely.

This is the situation of so-called competing risks. As long as the subject is at risk, then either of the event types can occur. Once one type of event has occurred, then it is no longer possible to observe an occurrence of an event of the other type. Such a construction fits in immediately with the above models for survival involving censoring. If at time $t = t_1$ an event of type $w = 1$ takes place, then, as far as events of type $w = 2$ are concerned, the subject is simply censored at $t = t_1$. In Figure 4.5 a subject may be at risk of death from stroke or at risk from either stroke or cirrhosis of the liver. Once one of the types of death has occurred, then the other type of event can no longer be observed. We will assume that the subject is censored at this point, in as much as our attention focuses on the second type of event, and the above discussion on the different censoring models applies in the same way. We will need make some assumptions, most often that of

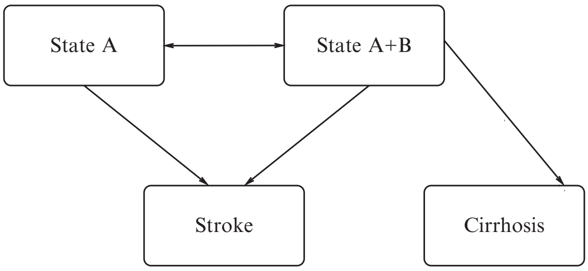

Figure 4.5: A situation of competing risks for subjects in states A+B.

independent censoring or that of independent censoring conditional on covariate information in order to make progress.

4.7 Exercises and class projects

1. Using the definition for $\lambda(t) = f(t)/S(t)$, show that $S(t) = \exp\{-\Lambda(t)$ and that $f(t) = \lambda(t)\exp\{-\Lambda(t)$.

2. For a Weibull variate with parameters λ and k, derive an expression for the conditional survivorship function $S(t+u, u)$. How does this function vary with t for fixed u? With u for fixed t?

3. Use numerical integration to calculate the mean residual lifetime $m(t)$ and the mean time lived in the interval $[0, t]$, $\mu(t)$ for the Weibull with parameters 2 and 1.5. With parameters 2 and 0.7. Plot these as functions of time t.

4. Consider two groups each of which follows a Weibull distribution, i.e., $f(t) = \lambda\gamma(\lambda t)^{\gamma-1}\exp\{-(\lambda t)^\gamma\}$. For the first group, $\lambda = \lambda_1$, $\gamma = \gamma_1$. For the second, $\lambda = \lambda_2$, $\gamma = \gamma_2$. Under which conditions will this situations be described by proportional hazards?

5. Undertake a numerical and graphical study of the conditional survivorship function, $S(t+u, u)$, for the Weibull model, the extreme value model, the Gompertz model and the log-logistic model. What conclusions can be drawn from this?

6. Repeat the previous class project, focusing this time on the mean residual lifetime. Again what conclusions can be drawn from the graphs.

7. Consider a disease with three states of gravity (state 1, state 2 and state 3), the severity corresponding to the size of the number. State 4 corresponds to death and is assumed to follow state 3. New treatments offer the hope of prolonged survival. The first treatment, if it is effective, is anticipated to slow down the rate of transition from state 2 to state 3. Write down a compartmental model and a survival model, involving a treatment indicator, for this situation. A second treatment, if effective, is anticipated to slow down all transition rates. Write down the model for this. Write down the relevant null and alternative hypotheses for the two situations.

8. Consider a nondegenerative disease with several states; $1, 2, \ldots,$ counting the occurrence of these together with a disease state indicating a progression to something more serious, e.g., benign and malignant tumors or episodes of mild asthma with the possibility of progression to a more serious respiratory ailment. Write down possible models for this and how you might formulate tests of hypotheses of interest under varying assumptions on the role of the less serious states.

9. Suppose we have data; $T_1, \ldots, T_n$, from a Weibull distribution in which the shape parameter γ is known to be equal to 1.3. Use the delta-method to find an estimate for the variance of the estimated median (transform to a standard form).

10. For a proportional hazards Weibull model describe the relationship between the respective medians.

11. Investigate the function $S(t, u)$ for different parametric models described in this chapter. Draw conclusions from the form of this two-dimensional function and suggest how we might make use of these properties in order to choose suitable parametric models when faced with actual data.

12. Consider two possible structures for a parametric proportional hazards model;

$$\begin{aligned} \log S(t|Z) &= \log\{S[t|E(Z)]\} \exp(\beta Z) \\ \log S(t|Z) &= \log\{ES[t|Z]\} \exp(\beta Z). \end{aligned}$$

How do the interpretations differ and what difficulties are likely to be encountered in fitting either of the models?

13. Consider a clinical trial comparing two treatments in which patients enter sequentially. Identify situations in which an assumption of an independent censoring mechanism may seem a little shaky.

14. On the basis of a single data set, fit the exponential, the Weibull, the Gompertz and the log-normal models. On the basis of each model estimate the mean survival. On the basis of each model estimate the 90th percentile. What conclusions would you draw from this.

15. Suppose our focus of interest is on the median. Can you write down a model directly in terms of the median. Would there be any advantage/drawback to modeling in this way rather than modeling the hazard and then obtaining the median via transformations of the hazard function?