Introduction to Event History Analysis

U. Pötter, G. Rohwer

March 1999

Contents

1 Notation **3**

2 Basic descriptions of durations **6**
2.1 Distribution function, density, and hazard rate 6
2.2 Censoring mechanisms 10
2.3 Observing events through time 11
2.4 Nelson–Aalen and Kaplan–Meier estimators 12
2.5 Functionals of distributions 16

3 Simple regression models **17**

4 Durations: Parameterization **22**
4.1 Covariate effects 23
4.1.1 Scale models 23
4.1.2 Proportional hazards models 27
4.1.3 Other transformation models 30
4.1.4 Comparing regression coefficients across models . . 32
4.1.5 Semi–parametric models of covariate effects 34
4.2 Classes of distributions 36
4.2.1 Exponential distribution 37
4.2.2 Weibull distribution 38

4.2.3 Log–logistic distribution 42
4.2.4 Gamma distribution 44
4.2.5 Mixtures 44
4.2.6 Combining models for covariate effects and distributions 50
4.3 Time dependent covariates 51
4.4 Censoring processes 54

5 Estimation **57**
5.1 Maximum likelihood 57
5.2 EM and the missing information principle 60
5.3 Partial likelihood 61

6 References **65**
6.1 Text Books 65
6.2 Articles 65

1 Notation

In this introduction we will approach the analysis of events in time through a description of the durations between events. This approach does not directly attack the problems of dynamic descriptions in the social sciences. However, while the dynamics in any subject area will in general require a specialized study, the study of durations is less demanding. Moreover, it can often be used to build up more complicated models involving several simultaneous durations, many types of events, or different time scales.

The building blocks will therefore be variables T designating durations. We will assume that these variables take values in the positive real numbers $\mathbb{R}_+$. We make unrestricted use of the properties of the real numbers, their additive and multiplicative structure, their order and completeness. While this allows for a convenient mathematical description, one should bear in mind that durations in the social realm, let alone observations pertaining to them, rarely have all the required properties. As long as the difference is born in mind and as long as the use of real numbers leads to convenient approximations, this will do no harm.

Since we are mainly interested in the statistical description of durations, the variables T are treated as random variables. For the following, this basically means that all possible information on the random variables are given by their *distribution function*

$$F(t) = \Pr(T \leq t). \tag{1}$$

The complement of this function,

$$G(t) = \Pr(T > t) = 1 - \Pr(T \leq t) = 1 - F(t) \tag{2}$$

is often called *survivor function.*[1] If need arises to use several distribution functions, these are denoted by capital letters F, H, M etc. We will write F^* for the distribution function of functions of T to emphasize the relation to the original variable. The functions themselves, however, may have arbitrary names.

The density function of T is defined as

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t)}{\Delta t},$$

[1] The name is most unfortunate in many applications. But we follow established custom.

provided the limit exists everywhere. If it does, the distribution will be called *continuous*. The density can be used more flexibly than the distribution function to express arbitrary probabilities. For any set A for which a probability is defined, we can write

$$\Pr(T \in A) = \int_A f(u)\, du.$$

For the distribution and survivor function, this means that

$$F(t) = \Pr(T \leq t) = \int_0^t f(u)\, du$$

and

$$G(t) = \Pr(T > t) = \int_t^\infty f(u)\, du,$$

as well as the reverse relation, giving the density in terms of the distribution or survivor function as

$$f(t) = \frac{\partial}{\partial u} F(u)|_{u=t} = -\frac{\partial}{\partial u} G(u)|_{u=t}.$$

Densities are denoted by f, h, and m, corresponding to the capital letters used for distributions.

To denote the distribution of a random variable T, we use the symbol

$$T \simeq_d F.$$

In some cases it is necessary to deal with discrete random variables. Suppose that $\tau_0 = 0 < \tau_1 < \tau_2 \ldots$ is a sequence of durations with $\Pr(T \in \{\tau_0, \tau_1, \tau_2, \ldots\}) = 1$. The sequence $\tau_0, \tau_1, \ldots$ therefore contains all values the random variable T can take. This is called a *discrete* distribution. To emphasize the similarity with the continuous case, we denote the distribution function by

$$F(t) = \Pr(T \leq t) = \sum_{\tau_i \leq t} \Pr(T = \tau_i).$$

This is a right continuous step function. For the probabilities of single durations we write

$$f(\tau_i) = \Pr(T = \tau_i),$$

so that

$$f(\tau_i) = F(\tau_i) - F(\tau_{i-}),$$

where $F(\tau_{i-})$ is the limit from the left of F, $\lim_{s\uparrow\tau_i} F(s)$. $F(\tau_i)-F(\tau_{i-})$ is the height of the i th jump of the step function F, stressing the similarity with the definition of a density. We can re-express F in terms of the f by

$$F(t) = \sum_{\tau_i<t} f(\tau_i).$$

A sample of durations is denoted by $t_1, \ldots, t_n$, in contrast to a sequence of numbers known beforehand, like the τ_i above. The empirical distribution function of a sample is

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I[t_i \le t],$$

where $I[A]$ is the indicator function taking the value 1 if A is the case, and 0 otherwise. This is a step function with jumps at the observations t_i, of height $1/n$. It is a discrete distribution function, giving probability $1/n$ to each of the observed durations $t_1, \ldots, t_n$ (which in this case need not be ordered).

We try to use a unified notation for integrals with respect to continuous distributions, discrete distributions, the mixed case, and empirical distributions. Specifically, the expectation with respect to the empirical distribution function is written as

$$\mathbb{E}_{\hat{F}_n}(h(T)) = \int h(t)\, d\hat{F}_n(t) = \frac{1}{n}\sum_i h(t_i).$$

Generally, if $M_n(t)$ is a step function with jumps of height $m_1, \ldots m_n$ at the points $t_1, \ldots, t_n$, we write

$$\int h(t)\, dM_n(t) = \sum_{i=1}^{n} m_i h(t_i),$$

so that the integral simply denotes a weighted sum of the values $h(t_i)$. If a distribution has both an absolutely continuous part and discrete atoms, an integral with respect to that distribution is the sum of the

integral with respect to its continuous part and the integral with respect to a step function. Thus,

$$\int h(t)\, dM_n(t) = \int h(t) m^c(t)\, dt + \sum_{i=1}^{n} m_i h(t_i),$$

where m^c is the density of the continuous part and m_i is the weight of the discrete atoms at the points t_i.

2 Basic descriptions of durations

We assume that durations are represented by random variables taking values in the nonnegative real numbers. This implies that two descriptions of social situations are treated as equal if the descriptions result in the same distribution function. Within such an approach, aspects of a situation requiring a more detailed description than what a summary function can provide are excluded. This allows for a unified presentation of some recurrent themes in event history analysis.

We start with the discussion of a central concept within the theory of positive random variables, the hazard rate. Another central feature pertaining to the observation of event histories is that these need not have come to an end by the time data are gathered. Such uncompleted sequences of events will be referred to as censored. How censored informations can be used in descriptions of summary functions like the distribution function is the main topic of the latter part of this section.

2.1 Distribution function, density, and hazard rate

Durations are most often conceived of as the time between specific events. Taking a certain primary event as the starting point, the problem is to give a description of the time to the next event of interest. If the clock is set to zero at the time of the primary event, this is equivalent to asking for the (positive) amount of waiting time for the next event's occurrence. The standard descriptions of this situation in terms of distribution functions etc. do not take into account the time position of an observer. The *hazard rate* function turns out to be useful in this context. In discrete

time, it is defined as

$$r(\tau_i) = \Pr(T = \tau_i | T \geq \tau_i) = \frac{f(\tau_i)}{G(\tau_{i-1})}, \tag{3}$$

so that the hazard rate is the conditional probability of an event at time τ_i, given that there was no event before τ_i. The conditioning event $\{T \geq \tau_i\}$ may be interpreted as the information of an observer just before time τ_i. If the event did not take place before time τ_i, the probabilistic description should be updated to the conditional probability, given this information. The hazard rate does this for the event $\{T = \tau_i\}$.

In continuous models one takes the appropriate limit and defines

$$r(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr(T \in [t, t + \Delta t) \mid T \geq t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{G(t)}, \tag{4}$$

which is the limit of the probability of the occurrence of an event in $[t, t + \Delta t)$, given that there was no event before time t. Note that we should have written $G(t_-)$ to represent the conditioning event $\{T \geq t\}$. But for continuous distributions, $G(t_-) = G(t)$. The hazard rate is therefore a measure of the current intensity of an event to occur. It is not a probability, however, since it can take values larger than 1. As can be seen from the definition, the hazard rate exists if and only if a density exists.

The distribution function, the survivor function, the density, and the rate function are equivalent descriptions for the probability distribution of a positive random variable. That is, given one of the functions, the others can be derived analytically. It is therefore possible to choose that summary function that best suits ones purpose.

First, in the discrete case, we can use the properties of conditional probabilities directly to express the survivor function in terms of the hazard rate. From

$$G(\tau_i | T \geq \tau_i) = \Pr(T > \tau_i | T \geq \tau_i) = 1 - r(\tau_i)$$

one gets, going backward in time,

$$\begin{aligned} G(\tau_i) &= \Pr(T > \tau_i) = \Pr(T > \tau_i | T \geq \tau_i) \Pr(T \geq \tau_i) \\ &= \Pr(T > \tau_i | T \geq \tau_i) \Pr(T > \tau_{i-1}) \\ &= \Pr(T > \tau_i | T \geq \tau_i) \Pr(T > \tau_{i-1} | T \geq \tau_{i-1}) \Pr(T > \tau_{i-1}) \\ &= \dots \\ &= \prod_{j=1}^{i} (1 - r(\tau_j)). \end{aligned} \tag{5}$$

The relation for the density therefore is

$$f(\tau_i) = G(\tau_{i-1}) - G(\tau_i) = r(\tau_i) \prod_{j=1}^{i-1} (1 - r(\tau_j)). \tag{6}$$

In the continuous case, the definition of the hazard rate leads to

$$r(t) = \frac{f(t)}{G(t)} = -\frac{\partial}{\partial t} \ln G(t). \tag{7}$$

On the other hand, solving the implied differential equation in $G(t)$ above gives an expression for the survivor function in terms of the hazard rate

$$G(t) = 1 - F(t) = \exp\left(- \int_0^t r(u)\, du \right). \tag{8}$$

Differentiating this relation gives the density function in terms of the hazard rate:

$$f(t) = r(t) G(t) = r(t) \exp\left(- \int_0^t r(u)\, du \right). \tag{9}$$

The differences between the formulae in the discrete and continuous case are in fact more apparent than real. It is possible to express the survivor function in the continuous case in an analogous way as in the discrete case (5), although doing so would require the introduction of some concepts that aren't needed in the following. However, a useful function with a definition that covers both the continuous and the discrete case is the *integrated hazard rate*. Using the integral representation introduced above, this can also be expressed as

$$H(t) = \int_0^t \frac{1}{1 - F(u_-)}\, dF(u), \tag{10}$$

where $F(u_-)$ denotes again the limit from the left, $\lim_{s\uparrow u} F(s)$. Care with the limits is needed here since we do not want to presuppose the existence of a density in which case F may contain jumps. The denominator in the integrand above just involves a careful formulation of the probability $\Pr(T \geq u)$ which need not be equal to $G(u) = 1 - F(u) = \Pr(T > u)$.

From the definition we have

$$H(t) = \int_0^t r(u)\, du \text{ and } H(t) = -\ln G(t)$$

in the continuous case and

$$H(t) = \sum_{\tau_j < t} r(\tau_j)$$

in the discrete case.

In the general case, moreover, if there is a jump of the distribution function at time t, so that $F(t) - F(t_-) > 0$, the corresponding jump in the integrated hazard is

$$H(t) - H(t_-) = \Pr(T = t | T \geq t).$$

The integrated hazard function represents a positive measure in its own right. The only difference from a probability measure is that it is generally not finite since $H(t) \to \infty$ for $t \to \infty$. It figures below in the context of estimation, since it equals the expected number of events in the time interval $[0, t)$ if durations between events are independent and follow the distribution F.

A last observation used variously below is that the expectation of the random variable T can be expressed in terms of the survivor function as

$$\mathbb{E}(T) = \int_0^\infty u f(u)\, du = \int_0^\infty G(u)\, du, \tag{11}$$

which follows from integration by parts if either side is finite. Recall the formulae for integration by parts. If $\int_0^t f(u)\, du = F(t)$ and $\int_0^t g(u)\, du = G(t)$, then

$$\begin{aligned} \int_0^t f(u)G(u)\, du &= \int_0^t G(u)\, dF(u) \\ &= [F(t)G(t) - F(0)G(0)] - \int_0^t F(u)\, dG(u) \\ &= [F(t)G(t) - F(0)G(0)] - \int_0^t F(u)g(u)\, du \end{aligned} \tag{12}$$

This is a slightly rewritten version of the differentiation rule for products of functions. If F or G contain jumps but are right continuous, the result can be rewritten as

$$\int_0^t G(u)\, dF(u) = [F(t)G(t) - F(0)G(0)] - \int_0^t F(u_-)\, dG(u) \tag{13}$$

and

$$\begin{aligned} \int_0^t G(u_-)\, dF(u) = [F(t)G(t) - F(0)G(0)] - \int_0^t F(u_-)\, dG(u) \\ - \sum_{u \leq t} (F(u) - F(u_-))(G(u) - G(u_-)). \end{aligned} \tag{14}$$

2.2 Censoring mechanisms

Suppose that the duration of interest T starts at time $t = 0$. Suppose further that the process is observed during the period $[0, c]$. This means that the event in question is observed to occur only if $T \leq c$, before the observation on the process ceases. If $T > c$, information on the timing of the event is not available. The observation is said to be censored. The data in this form of observations can be represented in two parts. $T^* = \min(T, c)$ records either the time of the event if it occurred before c, or the *censoring time* c if the event did not occur before c. Additionally, an indicator $D = I[T \leq c]$, being 0 if the observation is censored, 1 otherwise, is given.

Observations similar to the above situation arise regularly from event histories, where observation time necessarily ceases at some point in time. In order to achieve a description of T at least on the interval $[0, c]$, it is necessary to assume that the censoring time does not involve any information on the future course of the process. This will generally be true if the censoring time is fixed in advance.

A slightly more general censoring model allows the censoring time to be a random variable C. In this case a valid description of T is achievable if the random variables C and T are stochastically independent. The data are once again represented by the pair $(T^* = \min(T, C), D)$. This is called the *independent random censoring* model.

These two representations of the lack of information arising in event histories are rarely very accurate descriptions. While the observation period

might be considered fixed in advance, the starting times of the processes of interest are most often not fixed in calendar time. Observations can cease for other reasons than the planned end of a study, especially because subjects drop out of a study.

The requirement for valid descriptions of T will in such cases still be that censoring at a particular time will not give information on the future time course of the process of interest. Over the past decades, probabilistic models for censoring have been considerably extended and do cover some of the situations indicated above. However, these models cannot be empirically verified. They all rely on speculations of what might happen or might have happened in the past. However general they are, they still need considerable knowledge of the subject matter to judge their merits. In effect, all that the probabilistic censoring models provide is a framework in which the statistical models and estimators described below are known to work. They scarcely affect the estimators form. To simplify the following discussion of statistical models and estimators, we will therefore assume that either censoring takes places at a fixed predetermined time (or a fixed sequence of times) or that censoring times can be represented by random variables independent of T.

2.3 Observing events through time

One of the main objectives of statistical theory is to provide estimates of the distribution function or of other summary functions describing durations. To do so, one needs a representation of observations that connects data with the probabilistic descriptions in terms of random variables. We will assume that the observations consist of a sequence of pairs $(t_i, d_i), i = 1, \ldots, n$ that are realizations of n independent identically distributed random variables T with distribution function F, transformed by an independent censoring mechanism. That is, each (t_i, d_i) is interpreted as a realization from the pair of random variables (T^*, D), and any two observations (t_i, d_i) and (t_j, d_j) arise from independent but identical copies of (T^*, D). The observations, or functions of the observations, can then be considered as random variables derived from the random variables of interest. Therefore, the relations between functions of the data and distributional descriptions of durations can be treated by probabilistic methods. Moreover, since we deal with repetitions of the same variables, it is possible to think of the observations as being part of an indefinite sequence, allowing thus the application of limit theorems

from probability theory.

The above conceptions have no special relation with processes developing in time. However, we do need some notation expressing the evolution of observations of the variables through time, mimicking the way information is revealed to an observer. This will basically mean to count events and censorings up to some time point t. We will set

$$N_i(t) = I[t_i \leq t, d_i = 1] \quad (15)$$

$$R_i(t) = I[t_i > t] \quad (16)$$

$$E_i(t) = I[T_i = t, d_i = 1]. \quad (17)$$

The corresponding sums over the n observations are denoted by the same symbol without the subscript i. So, $N(t) = \sum_{i=1}^{n} N_i(t)$ etc. Then, $N(t)$ is the number of uncensored events before time t, and $E(t)$ is the number of events exactly at time t, excluding censored observations. Also, $R(t)$ is the number of observations that had neither an event or a censoring recorded before t. This is often referred to as the *number at risk* at time t, since $R_i(t) = 1$ implies that the event time is later than t. In the sequel, the same symbols N, R, E will occasionally be used to refer to the respective quantities when data are replaced by corresponding random variables. E.g., $N(t)$ is also used to refer to $\sum_{i=1}^{n} I[T_i \leq t, D_i = 1]$. The meaning should be clear from the context.

2.4 Nelson–Aalen and Kaplan–Meier estimators

In a discrete time setting, the estimation of a hazard rate is straight forward. By analogy with the definition of the hazard rate, one might put

$$\hat{r}(\tau_j) = \frac{E(\tau_j)}{R(\tau_j)}, \quad (18)$$

the number of events at τ_j divided by the number still at risk, or under observation just before τ_j, together with the convention $\hat{r}(\tau_j) = 0$ if $R(\tau_j) = 0$. Note that this estimator does not depend on censoring or event times before τ_j.

From this estimator it is easy to derive respective estimators for the survivor function, distribution function, density, and integrated hazard

rate, simply by plugging the estimator $\hat{r}$ into the respective expressions in terms of the hazard function. For example, one might use

$$\hat{G}(\tau_j) = \prod_{k \le j} (1 - \hat{r}(\tau_k)). \tag{19}$$

A simple idea to generalize such estimators from discrete to continuous time models is to group the observations of the continuous model in fixed time intervals, and then to proceed as in the discrete case. In a second step it might then be checked whether the procedure is still sensible when the length of the intervals shrinks towards 0 and whether it approaches the correct quantity. Suppose there is a partition of $\mathbb{R}_+$ into intervals $[\tau_{j-1}, \tau_j)$. If the length of the intervals is small, approximately

$$H(\tau_j) - H(\tau_{j-1}) \approx r(\tau_{j-1})(\tau_j - \tau_{j-1}).$$

Summing over the intervals to time t and using $\hat{r}$ from above as an estimator of the jumps in the previous formula, one arrives at an estimator for the integrated hazard,

$$\hat{H}(t) = \sum_{\tau_j \le t} \hat{r}(\tau_j).$$

If the length of the intervals approaches zero, most of the intervals will contain no or at most one observations. Therefore, one is lead to consider

$$\hat{H}(t) = \sum_{t_j \le t} \frac{d_j}{R(t_j)}, \tag{20}$$

where (t_j, d_j) now refers to the observations from the continuous model. This is the *Nelson–Aalen estimator* of the integrated hazard function. If the estimator is used in (8), the resulting estimator for the survivor function is

$$\tilde{G}(t) = \prod_{t_j \le t} e^{-d_j/R(t_j)}. \tag{21}$$

Another possibility to extend the estimator from the discrete case is to use once again the discrete time hazard estimator, but this time in conjunction with the discrete time formula (5). If the length of the grouping intervals $[t_{j-1}, \tau_j)$ shrinks to zero, and the number of events in each

interval tends to at most one, the resultant estimator for the survivor function is

$$\hat{G}(t) = \prod_{t_j \le t} (1 - d_j/R(t_j)). \tag{22}$$

This is the *Kaplan–Meier estimator* of the survivor function. Its relation to the Nelson–Aalen estimator is somewhat illuminated by observing that

$$e^{-x} \approx 1 - x$$

for small x, so that

$$e^{-d_j/R(t_j)} \approx (1 - d_j/R(t_j)),$$

and the two estimators should give similar results.

The Kaplan–Meier estimator has another derivation that connects it with general methods of estimation in censored data models. The starting point is not the hazard rate but the empirical distribution that would generally be used to estimate the survivor function in the absence of censoring. This is

$$\hat{G}_n(t) = \frac{1}{n} \sum_i I[t_i > t].$$

In the presence of censoring, the approach does not seem to be appealing. For censored observations and t past the censoring time it is not known whether in fact the duration was longer than t or not, so that $I[t_i > t]$ is not known. However, one can try to replace the unknown quantities by an estimate, say by its conditional expectation given the censoring time. This is reasonable, since

$$G(t) = \mathbb{E}(I[T > t]) = \mathbb{E}(\mathbb{E}(I[T > t]|T^*, D)). \tag{23}$$

Then

$$\hat{G}_n(t) = \frac{1}{n} \sum_i \mathbb{E}_{\hat{G}}(I[T > t]|T^* = t_i, D_i = d_i) \tag{24}$$

is an empirical analogue of (23), since the outer expectation can be replaced by the empirical distribution of the observations. Note that the

inner expectation on the right hand side depends on the distribution function. An estimator that solves the above equation is called *self consistent*. Since it is defined as a fixed point, a self consistent estimator can be computed iteratively by computing $\hat{G}^{k+1}$ from the right hand side based on $\hat{G}^k$.

Computations need an explicit formula for the right hand side expectation. It is given by

$$\begin{aligned} & \mathbb{E}_{\hat{G}}\left(I[T > t] | T^* = t_i, D_i = d_i\right) \\ = \quad & \hat{\Pr}(T > t | T^* = t_i, D_i = d_i) \\ = \quad & \begin{cases} 0 & t > t_i,\ d_i = 1 \\ 1 & t_i > t \\ \dfrac{\hat{\Pr}(T > t)}{\hat{\Pr}(T > t_i)} = \dfrac{\hat{G}(t)}{\hat{G}(t_i)} & \text{else} \end{cases} . \end{aligned} \tag{25}$$

The algorithm based on the self consistency equation will converge to the Kaplan–Meier estimator if it is initialized by a discrete distribution with equal mass on all observations, whether censored or not. If, on the other hand, some of the uncensored observations are initialized with zero mass, the algorithm will never assign positive mass to them. Therefore, the set of self consistent estimators contains more members than just the Kaplan–Meier estimator.

To end this section, three more remarks are in order. First, we note a pointwise variance formula for the Kaplan–Meier estimator. It is normally derived from likelihood considerations that are discussed later. The result is *Greenwood's formula*

$$\hat{\mathbb{V}}(\hat{G}(t)) = \hat{G}(t)^2 \left[\sum_{t_j \le t} \frac{d_j}{R(t_j)(R(t_j) - d_j)} \right] \tag{26}$$

Second, the above formulae assumed that there is at most one observation in any small interval used in the approximation of the continuous case. The assumption can be deduced from the assumption of a continuous model. As a consequence, the numbers $E(\tau_j)$ in (18) could be replaced by the event indicator d_j, and approximations based on $n \to \infty$ justified from this assumption. But in most data sets there are *ties*, that is, more than one event at at least some time points. If the number of such ties is small in comparison to the number at risk, replacing d_j by $E(t_j)$ will not alter the estimators of this section considerably.

Third, the Kaplan–Meier estimator of the survivor function does not approach 0 on $t > t_n$ when the largest observation t_n is censored. In technical terms, this will lead to a bias in the estimator. From a practical point of view, the values of the survivor function beyond the largest observation can never be ascertained. However, in some cases, e.g. when the evaluation of expectations is required, some further assumptions are needed.

2.5 Functionals of distributions

In some applications estimators of summary functions are more than is needed. Instead of the step functions produced by the Kaplan–Meier or Nelson–Aalen estimators one would like to have a summary in terms of quantiles, means, variances etc. All these quantities can be treated as functionals of the underlying survivor or distribution function. For example, the expectation is given by

$$\mathbb{E}(T) = \int_0^\infty u \, dF(u)$$

and the median by

$$\text{median}(T) = F^{-1}(1/2).$$

In these cases it seems natural to use the estimator of the survivor function and plug it into the formula for the respective functional. The case of the median is instructive. Since the estimator of the survivor function is a step function, there need not exist a value t with $\hat{G}(t) = 1/2$, or it need not be unique. Moreover, in contrast to the case of uncensored observations, the jump heights of the estimator are not constant. Therefore, in practice neighboring values $\hat{G}(t_j) > 1/2 > \hat{G}(t_{j+1})$ are linearly interpolated.

A much more difficult problem is the estimation of moments. Plugging an estimator of the distribution function into

$$\mathbb{E}(T) = \int_0^\infty G(u) \, du$$

will lead to finite values only if the largest observation is uncensored. Estimation of moments does not seem to be feasible without rather strong assumptions.

3 Simple regression models

Many instances of social research involving durations require more than a summary measure for their argument. Very often, the problem may be cast in terms of regression models, a formulation familiar from cross–sectional analyses. The basic idea is to summarize the differences between groups of subjects parsimoniously by indicating the impact of group membership on a measure of central tendency only, e.g. the mean. A common way to express this idea mathematically is to consider the conditional distribution of duration given group membership. If all the conditional distributions look alike except for a different central tendency, the differences in central tendency might be expressed by a single number, depending only on a linear combination of group membership indicators.

More formally, let Y denote a random quantity of interest. Suppose that conditional on some covariates x, indicating group membership, Y follows the linear regression

$$Y = x\beta + \epsilon\,, \tag{27}$$

where x is a $1 \times p$ vector of covariates including a constant, β is a $p \times 1$ vector of unknown regression coefficients, and ϵ is a random variable having mean zero and finite variance. In the following, we will discuss an extension of this familiar linear model and its estimation to the case of possibly censored duration data.[2]

Durations are inherently positive quantities. Inserting durations directly as dependent variables Y in the above equation may therefore create conceptual difficulties. Changing the "central tendency" of a positive quantity by adding or subtracting some quantity may lead to negative values, which are impossible.

In analogy to similar arguments used in connexion with discrete dependent variable, one might choose a transformation of the original duration

[2] The formulation (27), given in terms of random variables, is meant here and in the following to refer to the equality of conditional distributions only. All that is implied is that the conditional distribution of Y, given x, is of the form

$$F^*(y|x) = F_0^*(y - x\beta).$$

The random variable ϵ is only used to indicate a certain distribution. The ϵ in the above equation need not be defined on the same probability space as Y. Nor is an interpretation of ϵ as "unobserved cause" warranted.

to fit the positivity constraint in all cases. An easy transform of durations that will always lead to interpretable results is the logarithm of the durations. That is, we set $Y = \ln T$. When using this transform, the effect of the covariates on the original time scale corresponds to a scale change: values of $x\beta < 0$ correspond to shortened durations, values of $x\beta > 0$ to prolonged ones.

It can be shown that the logarithmic transformation is the only one that can express all combinations of effects or reverses of effects additively. Still, using the logarithms of durations is no panacea. After all, if the effects of covariates $x\beta$ as well as the durations can be ascertained only to within a certain interval, many other transforms are consistent with a realistic description and should be used if needed.

To distinguish between random variables referring to durations and those referring to some transforms, in the following the former will be denoted by T, the latter by Y. The same convention will be obeyed when dealing with realizations of the random variables. Distribution, survivor, density and rate functions of transformations Y of the durations of interest will, however, uniformly be denoted by a superscript * on F, G, f, r etc. If the transform $Y = g(T)$ is monotone, as the logarithmic transform is, we can also consider the censored versions of Y, which are given by $Z = \min(Y, g(C))$ with z denoting the realized value.

In the absence of censoring, one can estimate β by minimizing the least squares criterion

$$\sum_{i=1}^{n}(y_i - x_i\beta)^2 = n\int e^2\, d\hat{F}_n^*(e) = \sum_{i=1}^{n}\int (y - x_i\beta)^2\, d\tilde{F}_{ni}^*(y)\,, \tag{28}$$

where $\hat{F}_n^*(e)$ is the empirical distribution function of the residuals $e_i = y_i - x_i\beta$, and $\tilde{F}_{ni}^*(y) = I[y_i < y]$ is the empirical distribution of an observation y_i.

Both the second and third representation in the above formula can be used to generalize the least squares criterion by replacing the empirical distributions involved by versions appropriate for censored data. It turned out, however, to be advantageous to start with the least squares estimating equations

$$\sum_{i=1}^{n} x_i'(y_i - x_i\hat{\beta}) = 0 \quad \text{or} \quad \sum_{i=1}^{n} x_i' y_i = \left(\sum_{i=1}^{n} x_i' x_i\right)\hat{\beta} \tag{29}$$

instead of the least squares criterion (28). In 1979, Buckley and James proposed to replace the censored observations Z by the conditional expectation of Y given the observed (censored) data and the covariates:

$$Y^* = \mathbb{E}_\beta(Y \mid z, d, x) = dz + (1-d)\mathbb{E}_\beta(Y \mid Y \geq z,\, x)\,. \tag{30}$$

This is an example of a general strategy dealing with incomplete data. It consists of replacing the unknown values of observables by their expectations, using all information available from the data (here, $Y \geq z$) as well as the information provided by the model structure. In this case, the dependence on model structure is reflected by the dependence of the conditional expectation on the unknown parameter β.

Replacing Y in expression (29) by its conditional expectation gives

$$\frac{1}{n}\sum_i x_i' \mathbb{E}_{\hat\beta}(Y \mid z_i, d_i, x_i) = \frac{1}{n}\left(\sum_{i=1}^n x_i' x_i\right)\hat\beta\,. \tag{31}$$

The *Buckley–James estimator* $\hat\beta$ is defined as the solution of the normal score function for β when the expectation on the left hand side is computed using $\hat\beta$.

Using the model formula (27) and a fixed β, an empirical version of the conditional expectation can be evaluated:

$$\begin{aligned}
\hat{\mathbb{E}}_\beta(Y \mid z_i, d_i, x_i) &= \hat y_i(\beta) \qquad (32)\\
&= d_i z_i + (1-d_i)\hat{\mathbb{E}}_\beta(Y \mid Y_i \geq z_i, x_i)\\
&= d_i z_i + (1-d_i)\left(x_i\beta + \frac{\int_{e_i}^\infty e\, d\hat F^*_\beta(e)}{\hat G^*_\beta(e_i)}\right)\\
&= d_i z_i + (1-d_i)\left(\sum_{k=i}^n v_{ik}(\beta)(z_k - x_k\beta) + x_i\beta\right)
\end{aligned}$$

where $\hat F^*_\beta$ is the empirical distribution function (e.g. the Kaplan–Meier estimator) of the residuals, $\hat G^*_\beta$ is the empirical survivor function $1-\hat F^*_\beta$, and we have put

$$v_{ik}(\beta) = \begin{cases} \dfrac{w_k(\beta)}{\hat G^*_\beta(e_i)} & \text{if } e_i < e_k \\ 0 & \text{otherwise} \end{cases}$$

and

$$w_k(\beta) = \hat P_\beta(\epsilon = e_k)\,,$$

so that $w_i(\beta)$ is the height of the jump of the empirical distribution at the i th residual.[3] A solution $\hat\beta$ of the estimating equation (29) therefore satisfies:

$$\hat\beta = \left(\sum_{i=1}^n x_i' x_i\right)^{-1}\left(\sum_{i=1}^n d_i x_i' z_i + \sum_{i=1}^n (1-d_i) x_i' \hat y_i(\hat\beta)\right)\,. \tag{33}$$

This leads to a straightforward iterative procedure for the computation of $\hat\beta$:

1. Assign starting values $\hat\beta^0$.
2. Compute $\hat y_i(\hat\beta^j)$ according to (32) using the Kaplan–Meier procedure as estimator for the distribution of the residuals.
3. Compute $\hat\beta^{j+1}$ using the right hand side from (33). This is a simple least squares regression of the pseudo data $\hat y_i(\hat\beta^j)$ on the regressors x.
4. Go back to step 2 unless some convergence criterion is met.

To be numerically effective, this simple iterative strategy needs elaboration. Following the steps of the algorithm, the basic choices are:

1. Starting values may be obtained using the least squares estimator treating all observations as uncensored. Other choices (e.g. using only uncensored observations) are of course possible but do not seem to have a decisive influence on the procedure.
2. The Kaplan–Meier estimator is not uniquely defined on the whole real line if the largest residual is censored. Buckley and James suggest to always treat the largest residual as uncensored. This will lead to an underestimation of the regression constant, but should scarcely affect the other regression estimators. Further choices are discussed by e.g. Efron (1988).

[3] For ease of notation it is assumed here that the observations are ordered according to the magnitude of the corresponding residuals.

4. The iteration may not converge to a unique value. This is due to the fact that the right hand side of (33) is a piecewise linear function in β: Changing β does not change the weights $v_{ik}(\beta)$ unless the ranks of the residuals change. Therefore, the estimator may oscillate between several values $\hat{\beta}$. The discontinuity of (33) hampers the analytic treatment of the estimator. Moreover, the number of limiting values in finite samples is not predictable, but may potentially be rather large. Fortunately, the phenomenon seems to be of practical interest only in rather small samples, in situations where the effect of covariates is small, or when the convergence criterion is very strict. In situations where a unique estimator is required (e.g. simulations, using the procedure as building block for more complicated models, etc.) one may use the arithmetic mean of all limit values of the algorithm as an estimator. Otherwise, the different values of the limiting cycle of estimators are very close and it may suffice to report just one of them.

A very simple estimator of the variance of $\hat{\beta}$ may be obtained by restricting attention to the uncensored observations:

$$\begin{aligned} \widehat{\mathrm{var}}(\hat{\beta}) &= (x'\mathrm{diag}(d_i)x)^{-1}\,\hat{\sigma}^2_{BJ} \\ \hat{\sigma}^2_{BJ} &= \frac{1}{n_u - p}\sum_i \left(d_i e_i - \frac{1}{n_u}\sum d_i e_i\right)^2, \end{aligned} \tag{34}$$

where n_u is the number of uncensored observations. This is the same as the classical variance estimator in the linear model with uncensored data. Since the estimator is computed from the uncensored observations only, it will not be very efficient. Moreover, it implicitly assumes that the variances of the non censored residuals are homoscedastic. But this is true only if the censoring variable follows the same regression as the uncensored dependent variable Y. With respect to the last point, a better estimator of the residual variance is

$$\hat{\sigma}^{2*}_{BJ} = \frac{n_u}{n(n_u - p)}\sum_i \left(d_i e_i + (1 - d_i)\sum_k v_{ik}(\hat{\beta})e_k^2\right). \tag{35}$$

In this formulation, the censored squared residuals are replaced by their conditional expectations. Combining $\hat{\sigma}^{2*}_{BJ}$ with the first equation in (34) provides an estimator of the variance of $\hat{\beta}$ that is (asymptotically) equivalent to a bootstrap estimator when the resampling is done holding the censoring information fixed. Experience with the two variance estimators suggests that the second version is more stable and has often smaller mean squared error than the first version. Both, however, are generally conservative.

4 Durations: Parameterization

Broadly speaking, a parameterization expresses a possibly large set of distributions, regressions, or interdependencies through a few (real) numbers. Parameterizations may serve several purposes: They summarize aspects of the data, they focus attention on interesting specific features, they allow for easy formal manipulations, they simplify comparisons between situations, and they can be used for simulations. In the following, we will treat several choices of parameterizations for the two main building blocks of duration models: how covariates affect duration, and how the class of durations and their properties can be described.

Together, these two building blocks, if fully specified, uniquely determine the conditional distribution of the durations under consideration. From a probabilistic point of view, this is all one needs to know. Introducing a family of conditional distributions by using a parameterization sets the frame for inference procedures, discussions, and the critique of proposed models. But with event history data, even when dealing only with durations, there are two more aspects that need attention. First, taking the temporal reference of duration models seriously allows for the introduction of covariates that change over time. Whether the marginal distribution of conventional covariates are specified as part of the model building process or not seems to be largely a matter of convenience. With covariates changing over time, more care is needed. Without specifying their path through time, one cannot even derive such simple characterizations as the conditional moments of durations. Second, most observations of durations suffer from a deficiency of sample information due to censoring. Without a formal representation of this lack of information one cannot hope to successfully confront models and observations. Both aspects, time dependent covariates and models for the censoring process, will be discussed at the end of this section.

4.1 Covariate effects

Covariates reflect the many aspects judged important in the comparative description of durations. They may relate to properties of individuals or groups, to group membership or changing environments and situations. While a comprehensive classification of their possible roles does not seem to be warranted, an understanding of the way covariate effects can be introduced parametrically is necessary to make efficient use of the gain they provide over the direct inspection of subgroups. The introduction of covariates reduces the burden of comparing many different subgroups to an examination of a vector β of regression parameters. But the interpretation of this numerical summary depends heavily on how covariates are supposed to affect a proposed model.

4.1.1 Scale models

In section 3 we introduced a regression model for durations derived from the classical linear model techniques. The interpretation of covariate effects in this model can be based on a distribution function F_0 affected by a linear combination of covariates, $x\beta$:

$$\Pr(\ln T \leq \ln t \mid x;\beta) = F^*_{\ln T|x}(\ln t \mid x;\beta) = F^*_0(\ln t - x\beta) \tag{36}$$

The conditional distribution of the logarithm of duration given the covariates is a shift by an amount of $x\beta$ of some basic distribution F^*_0. The basic distribution F^*_0 corresponds to a situation with covariate values $x = 0$.[4] The conditional densities, if they exist, satisfy a similar relation

$$f^*_{\ln T|x}(\ln t \mid x;\beta) = f^*_0(\ln t - x\beta), \tag{37}$$

exemplified in figure (1):

It is clear that one can use either one of the graphs in figure (1) as a starting point and define the other as an appropriate shift. Therefore, interpreting the action of covariates as a shift of densities does not depend on the choice of $x = 0$ for the baseline distribution or density. Any other value x_0 can be chosen as reference point. Then the effect of covariates x on the density is a shift of the location of the density corresponding to covariates x_0 by an amount of $(x - x_0)\beta$.

[4] Possibly up to a further shift given by the intercept term β_0

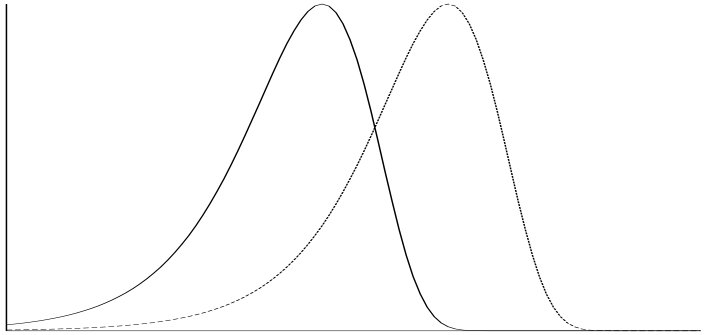

Figure 1: Densities of $\ln T$

The relation in terms of distributions (36) or densities (37) can also be reexpressed in terms of random variables as

$$\ln T = x\beta + \epsilon \quad \text{with } \epsilon \simeq_d F^*_0(.). \tag{38}$$

Written in this way, some of the operation on distributions or densities can be reduced to arithmetic operations on random variables. Often, this leads to more transparent formulations.[5]

While this interpretation is familiar from the classical linear model it only works for the logarithms of durations. It is certainly easier to have an interpretation in terms of durations, not just their logarithms. On the scale of durations, the distribution functions for different values of the covariates are related by

$$\begin{aligned} \Pr(T \leq t \mid x;\beta) &= \Pr(\ln T - x\beta \leq \ln t - x\beta) \\ &= \Pr(Te^{-x\beta} \leq te^{-x\beta}) = F_0(e^{-x\beta}t) \end{aligned} \tag{39}$$

Here, the basic distribution function F_0 once again corresponds to a situation with covariates $x = 0$. The graphs of the distribution functions

[5] The equality in (38) can in most cases only be interpreted as equality in distribution. This is what is needed for the interpretation above. But equality of random variables, if it can be ascertained, is a much stronger property. To take a simple example, if T_1 is uniformly distributed on $(0, 1)$, then $T_2 := 1 - T_1$ is also uniformly distributed on $(0, 1)$. But $\Pr(T_1 = T_2) = 0$. We will not discuss possible uses of equality of random variables in (40) because it can be rarely justified in social science applications.

for given x are squeezed or stretched along the t–axis, depending on whether $x\beta$ is negative or positive, but the lower end of their support, namely $t = 0$, is preserved. In terms of densities we get the relation

$$f(t \mid x;\beta) = f_0(te^{-x\beta})e^{-x\beta}.$$

The densities are not only scaled along the t–axis, but also their height changes. The figure (1) above, comparing densities for log durations, changes accordingly (see figure (2)). A direct comparison of the graphs

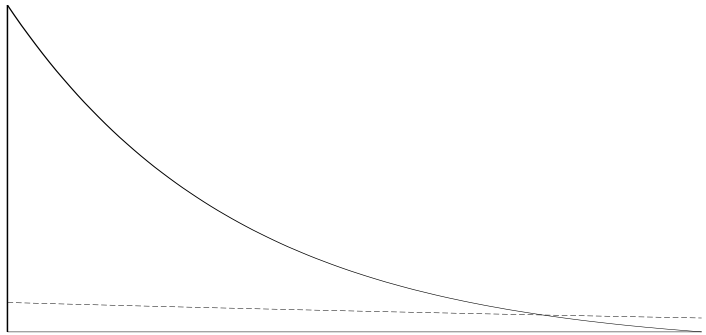

Figure 2: Densities of T

of the densities is not as easy as in the case of the logarithms of durations. But the relation (39) of the distributions suggests an interpretation in terms of random variables. Suppose there is a random variable T_0 with distribution function F_0, corresponding to durations with covariates $x = 0$. Then, durations T with covariate x are represented by

$$T = e^{x\beta}T_0 \text{ with } T_0 \simeq_d F_0(.). \tag{40}$$

This may be interpreted as a scaling of the underlying time structure: Positive values of $x\beta$ expand the time relative to the one on which T_0 is defined. Events develop slower on this time scale, so that durations are generally longer. On the other hand, negative values of $x\beta$ contract time relative to T_0 processes. Developments are faster and durations generally shorter.

Sometimes, especially in technical applications, this model of covariate effects can be linked directly to physical features of the environment:

machines working under higher load, at higher voltage, higher temperature etc. deteriorate faster. And these features of environment can be captured by respective covariate values. With such examples in mind, model (40), or its equivalent expressions (39) in terms of distribution functions, is often called *accelerated failure time model*. In view of the scale change expressed in (39) the term *scale model* is also used.

The relations between distributions, survivor functions, and rates in scale models can be summarized as follows:

$$\begin{aligned}
\Pr(T \le t \mid x;\beta) &= F(t \mid x;\beta) = \Pr(T_0 e^{x\beta} \le t) = \Pr(T_0 \le te^{-x\beta}) \\
&= F_0(te^{-x\beta}) \\
G(t \mid x;\beta) &= 1 - F(t \mid x;\beta) = G_0(te^{-x\beta}) \\
f(t \mid x;\beta) &= f_0(te^{-x\beta})e^{-x\beta} \\
r(t \mid x;\beta) &= \frac{f(t \mid x;\beta)}{G(t \mid x;\beta)} = r_0(te^{-x\beta})e^{-x\beta}
\end{aligned} \tag{41}$$

A further summary function is the quantile function defined as the (generalized) inverse of the distribution function:

$$Q(p) = F^{-1}(p) := \inf\{t \mid F(t) \ge p\}$$

From (41) we get

$$Q(p \mid x;\beta) = Q_0(p)e^{x\beta},$$

where $Q(p \mid x;\beta)$ is the quantile function corresponding to covariate value x and $Q_0(p)$ is the one corresponding to $x = 0$. The logarithms of the quantile functions are therefore related by

$$\ln Q(p \mid x;\beta) = \ln Q_0(p) + x\beta. \tag{42}$$

As a simple check of the appropriateness of the scale model one can plot the logarithms of empirical versions of the quantile function for different subgroups defined by x. The resulting graphs should be separated by a constant value.

Further consequences of (40) are simple relations for the moments of T, namely

$$\begin{aligned}
\mathbb{E}(T|x;\beta) &= e^{x\beta}\mathbb{E}(T_0) \\
\mathbb{E}(T^2|x;\beta) &= e^{2x\beta}\mathbb{E}(T_0) \quad \text{etc.},
\end{aligned} \tag{43}$$

so that the relation for the variances are

$$\mathbb{V}(T|x;\beta) = e^{2x\beta}\mathbb{V}(T_0)$$

Since the logarithms of durations form a location–shift family, the conditional variances $\mathbb{V}(\ln T|x;\beta)$ are constant. This homoscedasticity may also be used for model checking.

4.1.2 Proportional hazards models

Instead of looking at transforms of random variables as in (40) one can consider how covariates transform some baseline distribution or other summary function. The hazard rate is the most useful summary function from a dynamic point of view. Therefore it seems natural to examine transforms of a hazard rate $r(t|x;\beta)$. Since a hazard rate is nonnegative, its transforms by any covariate values should also be nonnegative. Moreover, a hazard rate corresponding to a proper distribution function should transform to one corresponding to a proper distribution function. In other words, if some baseline integrated hazard diverges to infinity, the same should be true for its transformed counterpart. The simplest way to achieve this is to multiply a baseline hazard rate by a positive function of the covariates. An obvious choice for the positive function is the exponential. With this choice we are let to the following model for covariate effects:

$$r(t \mid x;\beta) = e^{x\beta} r_0(t). \tag{44}$$

The model posits that positive values of $x\beta$ correspond to larger intensities in comparison to situations with $x\beta = 0$. With larger intensities for all t, events will tend to happen earlier and durations will be shorter. On the other hand, negative values of $x\beta$ give rise to smaller intensities, so that events tend to happen later, and durations will be longer.[6] An example is plotted in figure (3). For obvious reasons, models in which a positive function of covariates multiplies a baseline hazard rate are called *proportional hazards models*. The implied relations for the survivor functions, distributions, and densities are:

[6] Note that the sign of $x\beta$ has opposite consequences in a scale model.

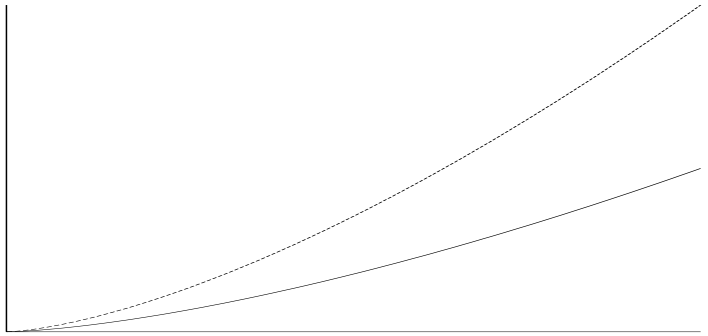

Figure 3: Proportional Hazard Rates

$$\begin{aligned}
\Pr(T > t \mid x;\beta) &= G(t \mid x,\beta) = e^{-\int_0^t r(u \mid x,\beta)\, du} \\
&= e^{-e^{x\beta} H_0(t)} = G_0(t)^{e^{x\beta}} \\
F(t|x;\beta) &= 1 - G(t|x;\beta) = 1 - G_0(t)^{e^{x\beta}} \\
f(t|x;\beta) &= e^{x\beta}\left(G_0(t)^{e^{x\beta}} - 1\right) f_0(t).
\end{aligned} \tag{45}$$

Of the above formulae, the first one, expressing the survivor functions given a covariate value as an exponentiation of a baseline survivor function, is the most useful. Even though the expectation of a positive random variable can be expressed through the integral of its survivor function as given in (11), there is no explicit formula relating the moments for different values of the covariates.[7]

Moreover, the quantile function, while easily computable, cannot be used directly for model checking purposes. On the other hand, the proportional hazards model can be written in regression form:

$$\ln H_0(T) = -x\beta + \epsilon \text{ with } \epsilon \simeq_d 1 - e^{-e^u} \quad , \tag{46}$$

[7] In this connexion, the formalism of Laplace transforms briefly treated in section 4.2.5 proves helpful.

where H_0 is the integrated hazard corresponding to the baseline distribution F_0 and ϵ follows an extreme value distribution with distribution function $1 - \exp(-\exp(u))$. For strictly increasing integrated hazards this follows from

$$\begin{aligned} G(t \mid x;\beta) &= \Pr(T > t \mid x;\beta) = \Pr(\ln H_0(T) > \ln H_0(t) \mid x;\beta) \\ &= \Pr(-x\beta + \epsilon > \ln H_0(t)) = \Pr(\epsilon > \ln H_0(t) + x\beta) \\ &= e^{-e^{\ln H_0(t) + x\beta}} = e^{-e^{x\beta} H_0(t)} \\ &= G_0(t)^{e^{x\beta}}. \end{aligned}$$

The regression representation (46) is useful for model comparison purposes. An increasing transformation of durations can be expressed as a homoscedastic linear regression with known extreme value distribution. On this transformed scale, the covariates shift the location of the standard extreme value distribution. Specifically, one can compare the proportional hazards and scale models using the regression representation. First, the scale model (47) can be expressed in regression form as

$$\ln T = x\beta + \sigma\epsilon. \tag{47}$$

Here, $\sigma\epsilon$ specifies a random variable with distribution equal to the distribution of the logarithm of the baseline random variable $\ln T_0$ in (40) and σ is used as an arbitrary but fixed scale parameter. If the extreme value distribution is chosen as the distribution of ϵ, it follows that

$$\frac{1}{\sigma} \ln T = \ln \left(T^{1/\sigma} \right) = x\beta/\sigma + \epsilon. \tag{48}$$

This is precisely of the regression form (46) for proportional hazards models with the special integrated hazard $H_0(t) \equiv t^{1/\sigma}$. The corresponding survivor function is

$$\begin{aligned} G(t \mid x;\beta) &= \Pr(T > t \mid x;\beta) = \Pr\left(\frac{1}{\sigma} \ln T > \frac{1}{\sigma} \ln t \mid x;\beta \right) \\ &= \Pr\left(\frac{x\beta}{\sigma} + \epsilon > \frac{1}{\sigma} \ln t \right) = e^{-\exp\left(\frac{1}{\sigma} \ln t - \frac{x\beta}{\sigma} \right)} \\ &= e^{-\exp\left(-\frac{x\beta}{\sigma} \right) t^{1/\sigma}}. \end{aligned} \tag{49}$$

and the relation of the hazards is

$$r(t|x;\beta) = e^{-\frac{x\beta}{\sigma}} r_0(t) \text{ with } r_0(t) = \frac{1}{\sigma} t^{1/\sigma - 1} \tag{50}$$

Starting with a scale model and assuming the extreme value distribution as baseline for the logarithm of durations one is lead to a proportional hazards model with integrated baseline hazard $H_0(t) \equiv t^{1/\sigma}$. In the proportional hazards parameterization the covariate effect is the negative of the one in the scale parameterization, corresponding to the fact that positive covariate effects in the scale model express longer durations while negative covariate effects in the proportional hazards model express lower hazards and thus longer durations. Moreover, the covariate effect in the proportional hazards expression is scaled by the scalar σ. It may be asked whether all scale models can be reexpressed as proportional hazards models or vice versa. This is not the case and the example above is the only one that is expressible both as scale and as proportional hazards model.

4.1.3 Other transformation models

The effect of covariates in proportional hazards models is to multiply some baseline hazard rate. Instead of a multiplicative transform of hazard rates one might be interested in other easily interpretable transforms, possibly based on other summary functions than hazards. In parallel to the well understood logit models for binary data one might e.g. look at the odds of an event before time t versus an event after time t. Using the logarithms of the odds as an appropriate scale for covariate effects, one is lead to the following relation between log odds for an event before vs. after time t:

$$\ln \frac{1 - G(t \mid x;\beta)}{G(t \mid x;\beta)} = x\beta + \ln \frac{1 - G_0(t)}{G_0(t)} \tag{51}$$

for some baseline survivor function G_0. In terms of odds,

$$\frac{1 - G(t \mid x;\beta)}{G(t \mid x;\beta)} = e^{x\beta} \frac{1 - G_0(t)}{G_0(t)}.$$

For positive $x\beta$, the odds for earlier events are larger than for the baseline survivor function. Since this is supposed to hold for all t, event

probabilities are larger and therefore durations are shorter. This model is generally refered to as *log–odds model*. The implied relations between survivor functions and hazard rates, respectively, are:

$$\begin{aligned} G(t \mid x;\beta) &= \frac{1}{1+e^{x\beta}\dfrac{1-G_0(t)}{G_0(t)}} \\ r(t|x;\beta) &= \frac{r_0(t)\cdot e^{x\beta}}{G_0(t)+(1-G_0(t))\cdot e^{x\beta}}. \end{aligned}$$

As can be seen from the relation of the rate functions, the hazard rates are not proportional. In fact, the relative rates $r(t|x_1;\beta)/r(t|x_2;\beta)$ for any two covariate values x_1 and x_2 converge to 1 as $t \longrightarrow \infty$. Therefore, the class of proportional hazards models and the class of log–odds models do not contain common members.

As in the case of both proportional hazards and scale models the log–odds models can be represented in regression form as

$$\ln\frac{1-G_0(T)}{G_0(T)} = -x\beta+\epsilon \text{ with } \epsilon \simeq_d \frac{1}{1+\exp(u)}, \tag{52}$$

where the error distribution is given by the survivor function of the logistic. Comparing this with the regression form (47) of a scale model it is seen that the only common member of the class of scale and log–odds models is the log–logistic distribution.

A slight generalization of the log–odds model, the γ–odds model, is given by

$$\begin{aligned} \frac{1-G^\gamma(t \mid x;\beta)}{\gamma G^\gamma(t \mid x;\beta)} &= e^{x\beta}\frac{1-G_0^\gamma(t)}{\gamma G_0^\gamma(t)} \quad \text{for } \gamma>0 \text{ and} \\ \ln\{G(t \mid x,\beta)\} &= e^{x\beta}\ln\{G_0(t)\} \quad \text{for } \gamma=0 \end{aligned} \tag{53}$$

The resulting survivor function is

$$G(t \mid x;\beta) = \frac{1}{\left(1+e^{x\beta}\dfrac{1-G_0^\gamma(t)}{G_0^\gamma(t)}\right)^{1/\gamma}}.$$

For $\gamma \longrightarrow 0$, this approaches a proportional hazards model, while for $\gamma = 1$ it reduces to the log–odds model. Since the γ–odds model interpolates between the proportional hazards and the log–odds models it is well suited for model assessment purposes. On the other hand, since the interpretation of covariate effects depends on the value of γ, and since this value is sometimes estimated from the data, it is less suited to express a well defined covariate effect.

4.1.4 Comparing regression coefficients across models

In the previous sections, several covariates are assumed to affect a model through a linear combination $x\beta = \beta_0+\beta_1x_1+\ldots+\beta_kx_k$ only. Linear combinations of covariates are the most popular choice for the description of joint effects. First, using a linear combination of covariates to represent joint effects is only rarely a real limitation of functional form. This is especially evident when the number of different covariate values is small, when interaction terms are introduced, or when fixed transforms of covariates (e.g. polynomials) are taken into account. Second, linear combinations are easily treated, both mathematically and algorithmically. Last but not least, the interpretation of β_i as the effect of a unit increase in the covariate value x_i on (a certain aspect of) the given model is very simple. Moreover, if covariates, say x_i and x_j, are defined on similar scales, a linear specification allows for a direct comparison of their effects via β_i and β_j.

On the other hand, as the discussion of covariate effects in the previous sections should have made clear, a direct comparison of regression coefficients across models is possible only in very special circumstances. For the family of distributions (49), which is both a scale and a proportional hazards model, the relation of coefficients turned out to be

$$\beta_{PH} = -\frac{\beta_{SC}}{\sigma},$$

where β_{PH} and β_{SC} are the vectors of regression coefficients in the proportional hazards and the scale model, respectively. A similar relation can be shown to hold for the family of distributions that are both a log–odds and a scale model. In both cases, the respective regression vectors are the same up to a scalar multiple. They are proportional.

This suggests to look at the *equivalent effects* $\gamma_{ij} := \beta_i/\beta_j$ for $\beta_j \neq 0$ instead of the regression coefficients themselves. The equivalent effects γ_{ij} express the change in the covariate value x_j required to achieve an equivalent effect on the model as a unit change in x_i. In the above example, the equivalent effects γ_{ij} do not change when the parameterization

is changed from a scale model to a proportional hazards model. In a simple linear regression, the γ_{ij} do not change when the scale of the dependent variable is changed. Also, when comparing several simple regressions with the same set of covariates but with the dependent variable measured differently, the γ_{ij} can be compared across models, while the interpretation of the β vectors changes with the scale of the dependent variable and the marginal distribution of the covariates in the different samples.

This constancy of γ_{ij} cannot be expected to hold across all contemplated models. Astonishingly, however, it holds approximately in a variety of circumstances. More specifically, it holds for small effects $|\beta| \sim 0$ when using different classes of covariate effects like proportional hazards or scale models. This approximation improves as the marginal distribution of the covariates becomes more symmetric. In the case of jointly normal covariates, the γ_{ij} are exactly constant across models, at least asymptotically. Moreover, the approximation results also cover the case of incomplete data, e.g. when grouped or discrete duration data are represented by continuous models.

Further insight into the role of the equivalent effects may be gained from considering a nonlinear scale model. Suppose the conditional expectation of $\ln T$ is given by a nonlinear function ϕ of a linear combination of covariates, that is

$$\mathbb{E}(\ln T \mid x; \beta, \phi) = \phi(x\beta) \tag{54}$$

Then, since

$$\frac{\partial}{\partial x_j}\phi(x\beta) = \frac{\partial \phi(x\beta)}{\partial x\beta}\beta_j,$$

$$\mathbb{E}\left(\frac{\partial}{\partial x_j}\phi(x\beta)\right) = \mathbb{E}\left(\frac{\partial \phi(x\beta)}{\partial x\beta}\right)\beta_j = c\beta_j,$$

where the expectations in the last equation are taken with respect to the marginal distribution of the covariates and c is a scalar constant depending on ϕ, β, and the distribution of the covariates. In other words, the coefficient vector β is proportional to the mean derivative of the regression function ϕ. Therefore, the equivalent effects γ_{ij} are also invariant with respect to different regression functions or marginal distributions of the covariates in this nonlinear scale model.

4.1.5 Semi–parametric models of covariate effects

While using linear combinations of covariates is sufficient in many situations, there are cases where more general specifications are warranted. One typical situation is when some covariates (e.g. age or income) take on many possible values and interest centers on the comparison of effects for all values of that covariates. It seems natural to replace the linear combination of covariates by some nonlinear function ϕ. This leads to models of the form

$$\phi(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1}x_{k-1} + g_k(x_k),$$

where g_k is some nonlinear function. These models are called *partly additive models*. One can of course add further nonlinear terms, and when all terms, linear or not, are denoted by g_i, model (4.1.5) can be written as

$$\phi(x_1, \ldots, x_k) = \beta_0 + g_1(x_1) + \ldots + g_k(x_k). \tag{55}$$

Since the constant term β_0 is only identified when the nonlinear functions g_i are constrained, one often uses the normalization $\mathbb{E}(g_i(X_i)) = 0$ or its empirical counterpart. Also, some assumptions on the smoothness of the g_i are generally added. For estimation purposes, one wants to consider observations close to a given covariate value x as giving information on the value of the function $\phi(x)$. And this is only possible if the function ϕ and therefore the g_i do not change too abruptly.

Note that the additive combination of covariate effects still allows for an interpretation of one effect when all others are kept constant. The effect of that covariate can usefully be expressed (plotted, etc.) without regard to the values of all other covariates.

Also, the additive structure can be used in a stepwise fitting procedure where each term g_i is treated separately. Namely, one may consider the effect of the covariates $g_j(x_j), j \neq i$ in any step of the fitting procedure as fixed. Since ϕ is additive, one can then fit the residual of the model given $g_j(x_j), j \neq i$ against the covariate x_i conditioned in the same way. This leads to a sequence of one dimensional estimating problems where each covariate is considered in turn. Such one dimensional problems are typically solved much easier than the general multidimensional regression problem where all covariates have to be considered simultaneously.

On the other hand, the partly additive model does not approximate all functional forms. Nor does it cover the important case of interactions.

To deal with nonlinear effects and interactions simultaneously, *regression trees* are often employed. Instead of using sums of smooth functions the idea is to express the regression function by step functions given by

$$\phi(x_1, \ldots, x_k) = \sum_{l=1}^{L} c_l I[(x_1, \ldots, x_k) \in R_l]. \tag{56}$$

Here, R_l is an element of a partition of the covariate space so that the regression function takes on the value c_l on the region R_l. The computational burden in constructing such a regression function is much reduced if the partition is made up from rectangles with sides parallel to the coordinate axes in the covariate space. Moreover, an easy interpretation of the partition becomes available in that case. The region to which an observation belongs can be determined by a sequence of simple binary decisions, each concerning only one variable. The regions are build by splitting the covariate space along one dimension according to whether the value of the jth covariate, say, is larger or smaller than a certain value. These splits can equivalently be represented by a tree: Suppose, e.g., $(x_1, x_2) \in \mathbb{R}^2$ and consider the partition of $\mathbb{R}^2$ into the rectangles $R_1 = \{x_1 \leq 0, x_2 \leq 0\}, R_2 = \{x_1 \leq 0, x_2 > 0\}, R_3 = \{0 < x_1 \leq 1, x_2 \leq 1\}, R_4 = \{0 < x_1 \leq 1, x_2 > 1\}$ and $R_5 = \{x_1 > 1, x_2 \in \mathbb{R}\}$. The regions are indicated in figure (4).

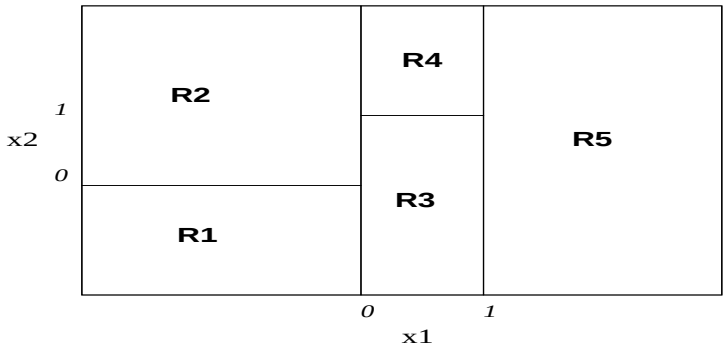

Figure 4: Partitions in a regression tree

The same information is given in the binary tree in figure (5), where the terminal nodes (or leaves) represent the respective regions. Note that

the choice of a root, the highest level in the tree representation, may not be unique. Moreover, the interpretation of a split on a lower level of the tree will depend on all those splits on higher levels that lead to that split (its "ancestors"). But the tree representation can be enhanced by adding statistical information on the subsets established at a node. This may be the degree of subgroup homogeneity with respect to duration, or the relative accuracy of prediction of a split etc. With such information added, regression trees are an effective regression summary.

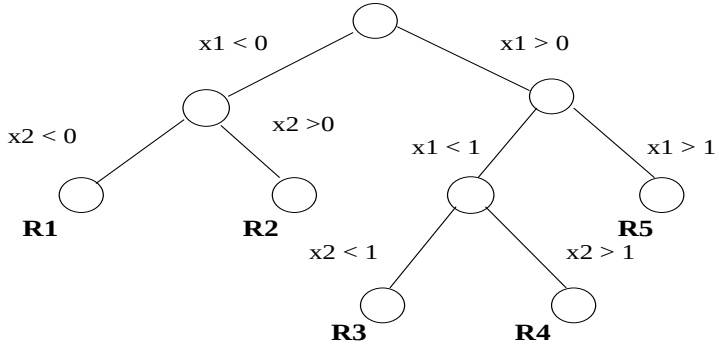

Figure 5: Regression tree

4.2 Classes of distributions

The conditional distribution of durations is fully specified if in addition to a parameterization of covariate effects the baseline distribution is defined. Traditionally, in the context of regression models, there are only rarely discussions on the choice of a baseline distribution. In that situation, most aspects of the statistical behavior of estimators depend on the first few moments of the distribution only. A peculiar feature of event history analysis is a much stronger interest in families of distributions and their properties. One of the reasons is that with censored observations, estimates of simple characteristics (e.g. expectations) will depend strongly on the choice of baseline distributions. Another reason is that models for durations are often only a first step in the analysis of more complex systems of events. In this case, the properties of the constituent distributions will constrain the properties of the whole system.

4.2.1 Exponential distribution

This last point is well demonstrated by the exponential distribution that frequently serves as a reference or as a starting point for the construction of more complicated models. The exponential distribution has density

$$f_a(t) = ae^{-at}\, a > 0. \tag{57}$$

Its distribution, survivor, and rate functions are

$$\begin{aligned} F_a(t) &= 1 - e^{-at} \\ G_a(t) &= e^{-at} \\ r_a(t) &= a, \end{aligned} \tag{58}$$

respectively. The constancy of the rate function is sometimes referred to as signifying no time dependence. Since the hazard function, if defined, uniquely determines the distribution, the exponential is the only class of distributions with this property.

The constancy of the hazard function is related to the basic characteristic of the exponential distribution, its *lack of memory* property. It states that at any given time t, the residual duration from t onward has the same distribution as the distribution itself. In other words, the information that an event did not occur before time t does not change the probability of its occurrence within $(t, t+s]$ from the initial probability $\Pr(T \in (0, s])$. Aging has no effect, and this is expressed by a constant intensity for the occurrence of an event, r_a. In the context of stochastic process models, this means that information on the past of a process does not add any information on its future beyond what is known about the state of the process at time t. This allows for the construction of process models with an easily understood dependence on the past. More formally, the lack of memory property of the exponential distribution follows from

$$\begin{aligned} \Pr(T > t+s | T > t) &= \frac{\Pr(T > t+s)}{\Pr(T > t)} \\ &= \frac{G(t+s)}{G(t)} = \frac{e^{-a(t+s)}}{e^{-at}} = G(s). \end{aligned} \tag{59}$$

Moreover, the exponential distribution is the only distribution with this property.[8]

A further simple but useful property is that for any positive random variable T with integrated hazard $H()$, its *hazard transform* $H(T)$ is exponentially distributed with parameter $a = 1$. Suppose, for simplicity, that the integrated hazard $H()$ is continuous and strictly increasing. Then,

$$\Pr(H(T) > t) = \Pr(T > H^{-1}(t)) = e^{H(H^{-1}(t))} = e^{-t}. \tag{60}$$

This transformation was already used when deducing the regression form of the proportional hazards model (46). In the present context, the hazard transform is often used as a device for model checking and for the comparison of distributions since it allows for the reduction of any distribution to the exponential. This may then serve as standard against which departures can be judged.

The expectation and variance of the exponential distribution are

$$\begin{aligned} \mathbb{E}(T) &= \frac{1}{a} \\ \mathbb{V}(T) &= \frac{1}{a^2}. \end{aligned}$$

It follows that the coefficient of variation, the ratio of the standard deviation to the mean, is unity. For this reason, the exponential may also serve as a baseline for judging relative dispersion.

4.2.2 Weibull distribution

Because of the lack of memory property, the exponential distribution is often not an appropriate representation of durations in the social sciences. Moreover, since it depends on one parameter only, it is not very flexible when fitted to data. A two parameter extension of the exponential distribution arises from the introduction of a second parameter transforming the time scale. A simple choice is the class of distributions

[8] Excluding the degenerate case $\Pr(T > t) \equiv 0$, this follows from Cauchy's equation. Writing $V(t) := \ln \Pr(T > t)$ and multiplying (59) by $\Pr(T > t)$ leads to $V(t+s) = V(t) + V(s)$. The only continuous solutions to this equation are the linear functions, $V(t) = V(1)t = -at$, say. The result follows upon exponentiation.

having survivor, distribution, density, and rate function

$$\begin{aligned} G_{a,b}(t) &= e^{-(at)^b} \\ F_{a,b}(t) &= 1 - e^{-(at)^b} \\ f_{a,b}(t) &= ba^b t^{b-1} e^{-(at)^b} \\ r_{a,b}(t) &= ba^b t^{b-1}, \end{aligned} \tag{61}$$

where $a, b > 0$. This class of distributions is referred to as the *Weibull* class of distributions. The parameter b is often called the Weibull parameter.

The rate function is monotone increasing or decreasing depending on whether $b > 1$ or $b < 1$. For $b = 1$, it reduces to the exponential distribution. The Weibull family therefore often serves as a representation for deviations from a constant hazard rate in the direction of monotone time dependence (compare figure 6). The Weibull class was already

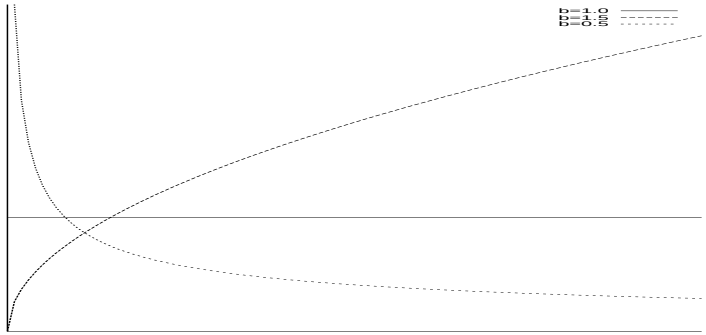

Figure 6: Weibull hazard rates

encountered, in thin disguise, when discussing the intersection of the proportional hazards class and the scale model for covariate effects. To recapture the representation used in (49) from the one given above, one only has to put

$$\begin{aligned} b &= \frac{1}{\sigma} \\ a &= e^{-x\beta}. \end{aligned}$$

Consequently, a representation using a proportionality factor for the hazard function instead of a scale factor is also possible. This can be achieved by setting

$$\begin{aligned} b^* &= b \\ a^* &= a^b \end{aligned}$$

resulting in the hazard function $r_0(t) = a^* t^{b^*-1}$. These different parameterizations of the class of Weibull distributions are equivalent in that they give rise to exactly the same class of distributions. Moreover, from an analytic point of view, transforming one parameterization into another in the foregoing example is a smooth operation. Still, the different types of parameterizations should be kept in mind. One reason is that existing software packages tend to use different versions of parameterizations implying different interpretations and the necessity to translate the interpretation for one parameterization into another. A second reason is that the statistical properties of some inference procedures, notably Wald's test for regression parameters, do change with the parameterization employed (see section 5 for some further comments).

A property of the Weibull class that makes it quite popular in several areas of application is its appearance as the asymptotic distribution of the minima of independent random variables. To start with, a simple consideration shows that the minima of Weibull distributions are themselves distributed according to the Weibull law. Suppose that there are n independent random variables $T_i, i = 1, \ldots, n$, each distributed according to the same Weibull law with parameters a, b as in (61). Then

$$\begin{aligned} \Pr(\min_i(T_1, \ldots, T_n) > t) &= \Pr(T_1 > t \cap \ldots \cap T_n > t) \\ &= \prod_{i=1}^{n} \Pr(T_i > t) \\ &= (\Pr(T_1 > t))^n, \end{aligned}$$

where the second equality follows from independence and the third from the assumed identical distribution. Inserting the survivor function of the Weibull distribution gives

$$\Pr(\min_i(T_1, \ldots, T_n) > t) = e^{-n(at)^b} = e^{-(n^{1/b}at)^b}.$$

That is, the minimum of n identically distributed, independent Weibull random variables follows again the Weibull distribution with the same Weibull parameter b and a scale parameter equal to $n^{1/b}a$. Therefore, the Weibull family is said to be closed under the forming of minima.

Of greater importance in social science applications is the more general fact that a similar result holds asymptotically without specifying an underlying class of distributions. Namely, for a large class of distributions it can be shown that their appropriately scaled minima tend to the Weibull distribution. More precisely, given a sequence of such random variables, $T_i, i = 1, \ldots$, there are sequences of numbers c_n and d_n such that the distribution of

$$d_n(\min_i(T_1, \ldots, T_n) - c_n)$$

tends to a Weibull distribution. This fact is sometimes exploited in modeling situations where one is interested in the time to the first arrival of a job offer, say, presupposing that there were many simultaneous applications for a job and the applicant chooses the offer that arrives first. In the social sciences, variants of the argument are invoked to justify the choice of the Weibull distribution in applications ranging from the theory of choice and the theory of search unemployment to theories of information processing in the human brain. In a more formal context, it is used to generate models for competing risks. Multivariate generalizations of the argument are employed in models involving a discrete response with only a few categories. It should be noted, however, that in contrast to the situation described by the central limit theorem, the norming constants c_n, d_n, and the rate of convergence depend heavily on the underlying distribution.[9]

The expectation and variance of the Weibull distribution can be derived from a change of variables by setting $u = (at)^b$. The Jacobian of the transformation $u \to t = u^{1/b}/a$ is given by $J(u) = u^{1/b-1}/ab$, so that

$$\mathbb{E}(T) = \int_0^\infty G_{a,b}(t)\,dt = \frac{1}{a}\int_0^\infty u^{1/b}e^{-u}\,du = \frac{\Gamma(1/b+1)}{a},$$

where $\Gamma()$ is the gamma function satisfying the functional equation $\Gamma(x+1) = x\Gamma(x)$. Specifically, $\Gamma(n) = (n-1)!$ for all integers $n > 0$. Repeating the same argument leads to

$$\mathbb{E}(T^2) = \int_0^\infty t^2 f_{a,b}\,dt = \frac{\Gamma(2/b+1)}{a^2}.$$

Thus the variance of the Weibull distríbution is given by

$$\mathbb{V}(T) = \frac{1}{a^2}\left(\Gamma(\frac{2}{b}+1) - \Gamma(\frac{1}{b}+1)^2\right)$$

An argument to the same effect, but perhaps closer in spirit to the probabilistic arguments used thus far, would be to refer to the moments of the exponential distribution via the hazard transform. Since $(aT)^b$ is exponentially distributed with unit parameter, the n–th moments of T is simply the n/b–th moment of the exponential distribution divided by a^n. This connexion will also be exploited in section 4.2.4.

4.2.3 Log–logistic distribution

A further two parameter class of distributions with some convenient properties is given by the following survivor, distribution, density, and hazard functions:

$$\begin{aligned} G_{a,b}(t) &= \frac{1}{1+(at)^b} \\ F_{a,b}(t) &= \frac{(at)^b}{1+(at)^b} \\ f_{a,b}(t) &= \frac{ba^bt^{b-1}}{[1+(at)^b]^2} \\ r_{a,b}(t) &= \frac{ba^bt^{b-1}}{1+(at)^b}, \end{aligned} \tag{62}$$

where $a, b > 0$. This is called the log–logistic class of distributions. If $b > 1$, the hazard function has a single maximum at $(b-1)^{1/b}/a$. If

[9] E.g., in the case of the minima of Weibull distributions, it is seen from the above result for n random variables that the normalizing sequence d_n needs to be of the form $n^{1/b}$. The norming thus changes for any change in the underlying common Weibull parameter. This situation should be compared with a simple version of the central limit theorem, where the asymptotic normal distribution for sums of independent identically distributed variables follows from a condition on the existence of moments, irrespective of other features of the underlying distributions. Moreover, the standard norming $1\sqrt{n}$ always applies. An argument based on extreme value theory, if only based on a rough asymptotic approximation, cannot sustain the same force of argument as similar ones based on the central limit theorem. A thorough but accessible discussion of the probabilistic aspects of the theory can be found in: J. Galambos: The Asymptotic Theory of Extreme Order Statistics; Wiley 1978.

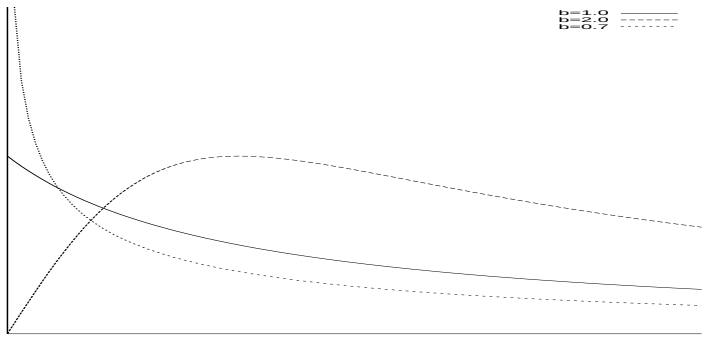

Figure 7: Log–logistic hazard rates

$b < 1$, the hazard function is decreasing. This is illustrated in figure 7. From the relation

$$f_{a,b}(t) = \frac{b}{t} G_{a,b}(t)[1 - G_{a,b}(t)]$$

one can reduce the problem of finding the moments of the log–logistic distribution to that of the moments of a beta distribution with density proportional to $x^{\alpha-1}(1-x)^{\beta-1}$ on the interval [0,1]. This is achieved by the substitution $u = G_{a,b}(t)$. It follows that

$$\mathbb{E}(T^n) = \frac{1}{a^n}\Gamma(1 + \frac{n}{b})\Gamma(1 - \frac{n}{b}) \tag{63}$$

Note that the n–th moment of the log–logistic distribution only exists if $b > n$. The log–logistic distribution therefore has heavier tails than the other distributions treated in this section.

The log–odds transform of the log–logistic distribution is linear in $\ln t$

$$\ln \frac{1 - G_{a,b}(t)}{G_{a,b}(t)} = b(\ln a + \ln t),$$

and this may be used for model checking and in characterizations involving the log–odds model as in (52). The log–logistic distribution and its parameterization will be further discussed in section 4.2.6.

4.2.4 Gamma distribution

Another two parameter family of distributions that is often applied is the gamma–distribution. It is given by:

$$\begin{aligned} G_{a,b}(t) &= 1 - \frac{1}{\Gamma(b)} \int_0^t a^b u^{b-1} e^{-au}\, du \\ F_{a,b}(t) &= \frac{1}{\Gamma(b)} \int_0^t a^b u^{b-1} e^{-au}\, du \\ f_{a,b}(t) &= \frac{1}{\Gamma(b)} a(at)^{b-1} e^{-at} \\ r_{a,b}(t) &= \frac{f_{a,b}(t)}{G_{a,b}(t)}, \end{aligned} \tag{64}$$

with $a, b > 0$. It reduces to the exponential distribution for $b = 1$. Its moments are

$$\begin{aligned} \mathbb{E}(T) &= \frac{b}{a} \\ \mathbb{V}(T) &= \frac{b}{a^2}. \end{aligned} \tag{65}$$

The presence of the incomplete gamma function in the survivor function makes it a rather cumbersome model to work with. But its usefulness in theoretical arguments derives from the fact that the family is closed under summation. If T_1, T_2 are independent gamma variates with the same scale parameter a but possibly different gamma parameters b_1 and b_2, then their sum $T_1 + T_2$ is once again a gamma variate with the same scale parameter a and gamma parameter $b_1 + b_2$. Since the exponential distribution corresponds to $b = 1$, an immediate consequence is that the sum of n independent exponential distributions with the same scale a is gamma distributed with scale a and gamma parameter $b = n$. This makes the gamma family an attractive candidate if an event is assumed to happen after the cumulative effects of several intermediate events. It is also used in the context of renewal processes and as a computationally convenient component in mixture models.

4.2.5 Mixtures

The weighted mean of two survivor functions G_1 and G_2,

$$G(t) = pG_1(t) + (1-p)G_2(t) \quad , p \in [0,1] \tag{66}$$

is again a survivor function. This is a useful and fundamental device to produce new distributions from given ones. It is often interpreted in terms of *heterogeneity*: Suppose that there is an indicator $V \in \{1, 2\}$ identifying two groups with different survivor functions G_1 and G_2. If $\Pr(V = 1) = p$, the marginal survivor function of the duration T is given by (66). The mixture therefore describes the survivor function if either the information on group membership V cannot be obtained or if one is interested in describing the situation without reference to group membership.

A special case of this model, the *mover–stayer model*, has a long tradition in sociological research. It posits that there is a subgroup that never experiences the type of event under consideration. In mobility research or demography, there are persons never changing their position or never marrying. Since these subgroups cannot be identified beforehand, the marginal survivor function is a mixture of the form

$$G(t) = pG_1(t) + (1 - p), \tag{67}$$

where the survivor function of the group not experiencing an event is unity. The above survivor function is sometimes called *defective*, because its limit for $t \to \infty$ is $1 - p > 0$. Equivalently, the corresponding distribution function converges to $p < 1$. As a result, the expectation in this model is infinity. Considering the above mentioned applications, the model is mildly unrealistic, if only because no one can live up to its expectation. Still, it might produce a useful approximation in some applications.

The idea of heterogeneity can be generalized by allowing not only discrete but general random variables. The realizations of these random variables are then often interpreted as characterizing a certain property of individuals. In the model building process, the heterogeneity variables are therefore treated on the same footing as other covariates. Thus the introduction of covariates in the general discussion of mixtures will make it possible to examine the effects of heterogeneity with respect to the different forms of covariate effects discussed earlier.

Suppose that in addition to the covariate vector x there is a random variable V, having the same distribution for all values of x, and influencing the conditional distribution of duration. If V is not included in the set of regressors, the resultant survivor function of T conditional on x is the expectation of the conditional distribution of T given x and V

with respect to the distribution of V, $M()$, say:

$$\begin{aligned} \Pr(T > t \mid x; \beta) &=: G(t \mid x; \beta) = \mathbb{E}_M\left(G_0(t \mid x, V; \beta)\right) \\ &= \int G_0(t \mid x, v; \beta)\, dM(v) \end{aligned} \tag{68}$$

The survivor function G is said to be a *mixture* derived from the *mixing* distribution M and the *mixed* distribution G_0. The mixed distribution G_0 is also called a *kernel*.

The effect of the mixing operation in the the present context is twofold. First, it generally changes the mixed distribution. If the mixed distribution belongs to some parameterized family, one gets a new family of distributions. If, in addition, the mixing distribution is allowed to come from a parameterized family, the mixing operation leads to a new parameterized family described by the parameters of both, the mixed and the mixing distribution. Second, except in special circumstances, the operation changes the way the included covariates act on the underlying family of distributions. Mixture models are therefore a useful tool to enlarge both the families of distributions and the classes of covariate effects considered.

The case that the random variable V acts as in a proportional hazards model on the kernel is of special importance. Suppose, therefore, that the hazard conditional on x and V is of the form

$$r(t|x, V; \beta) = V r_0(t|x; \beta) \quad \text{with } \Pr(V \geq 0) = 1. \tag{69}$$

Mixtures of this special form are called *proportional mixtures*. In technical or medical applications—where durations describe times to a failure or death and where the variable V refers to environmental effects—the model is often termed *frailty model*.

Computations and the derivation of characteristics of the mixed distribution can be eased considerably by noting the close connexion of this model with the *Laplace transform*. The Laplace transform of a positive random variable V is defined to be

$$\mathcal{L}_M(t) = \mathbb{E}_M\left(e^{-tV}\right) = \int e^{-tv}\, dM(v). \tag{70}$$

The function $\mathcal{L}_M(t)$ is thus seen to be the survival function of a proportional mixture with a unit exponential distribution as kernel. But there are extensive tables of Laplace transforms and many characterizations

of their properties.[10]This relationship can also be exploited for general proportional hazards models, since

$$\begin{aligned} G(t \mid x;\beta) &= \mathbb{E}_M\left(e^{-V\exp(x\beta)H_0(t)}\right) \\ &= \int e^{-ve^{x\beta}H_0(t)}\,dM(v) = \mathcal{L}_M\left(e^{x\beta}H_0(t)\right). \end{aligned} \tag{71}$$

That is, a proportional mixture of a proportional hazards model is the Laplace transform of the mixing distribution, evaluated at the integrated hazard, $e^{x\beta}H_0(t)$.

As an example, consider the gamma distribution as a mixing distribution with scale $a = \kappa$ and gamma parameter $b = \kappa$. Its density is

$$m_\kappa(v) = \frac{1}{\Gamma(\kappa)}\kappa^\kappa v^{\kappa-1}e^{-\kappa v}. \tag{72}$$

Setting $a = b = \kappa$ as above leads to the standardization $\mathbb{E}(V) = 1$ and $\mathbb{V}(V) = 1/\kappa$, compare (65). Its Laplace transform is of an especially simple form:

$$\mathcal{L}_M(t) = \frac{1}{\left(1+\frac{1}{\kappa}t\right)^\kappa}$$

Using the exponential distribution and proportional covariate effects as the kernel, one gets

$$G(t \mid x;\beta,\kappa) = \mathcal{L}_M\left(e^{x\beta}t\right) = \frac{1}{\left(1+\frac{1}{\kappa}e^{x\beta}t\right)^\kappa}.$$

The density and hazard function of this distribution are

$$\begin{aligned} f(t|x;\beta,\kappa) &= \frac{e^{x\beta}}{(1+\frac{1}{\kappa}e^{x\beta}t)^{\kappa+1}} \\ r(t|x;\beta,\kappa) &= \frac{e^{x\beta}}{1+\frac{1}{\kappa}e^{x\beta}t}. \end{aligned} \tag{73}$$

This family of distributions is called after *Pareto*. Note that the last formula above implies that the covariates do not act proportionally in the mixed distribution. Specifically, the ratio of any two hazard functions corresponding to two different values of the covariates converges to unity for $t \to \infty$.

[10]see: F. Oberhettinger/L. Badii: Tables of Laplace Transforms; Springer 1973 and W. Feller: An Introduction to Probability Theory and its Applications, vol. II; Wiley 1971, chap. XIII, for a general discussion.

Using the above gamma distribution as the mixing distribution in conjunction with a general proportional hazards model as the kernel leads to survivor functions of the form

$$G(t \mid x;\beta,\kappa) = \frac{1}{\left(1+\frac{1}{\kappa}e^{x\beta}H_0(t)\right)^\kappa}. \tag{74}$$

Now, the integrated hazard function $H_0(t)$ is monotone increasing. The same is true for the odds of survival $(1-G(t))/G(t)$, as well as for the transform $(1-G(t)^\gamma)/G(t)^\gamma$ that was used in the definition of the γ–odds model (53). In the absence of further restrictions on the rate or survivor function, the proportional mixture with gamma mixing distribution represents the same family of distributions and the same covariate effect as the γ–odds model. In other words, a proportional hazards model with gamma heterogeneity is formally and observationally equivalent to the γ–odds model. While the former posits a proportional effect of the covariates on the hazard function plus heterogeneity, the latter posits non–proportional effects but no heterogeneity. It follows that a good fit of the proportional mixture model cannot be regarded as empirical evidence for some form of heterogeneity. It may equally well be an indication of non–proportional covariate effects.

To end the discussion of proportional gamma mixture, we note its potential usefulness in the context of dependent durations. If, say, two durations are independent given covariate information and heterogeneity V, and if the heterogeneity term acts proportional on the hazard rate,

$$\begin{aligned} &\Pr(T_1 > t_1, T_2 > t_2 \mid x, V = v;\beta) \\ &= \Pr(T_1 > t_1 \mid x;\beta_1)^v \Pr(T_2 > t_2 \mid x;\beta_2)^v. \end{aligned} \tag{75}$$

If V follows the gamma distribution (72), the joint survivor function of T_1, T_2 is given by

$$\begin{aligned} &\Pr(T_1 > t_1, T_2 > t_2 \mid x;\beta) \\ &= \int \Pr(T_1 > t_1 \mid x;\beta_1)^v \Pr(T_2 > t_2 \mid x;\beta_2)^v\,dM(v) \\ &= \frac{1}{\left(1+\frac{1}{\kappa}H_1(t_1 \mid x;\beta_1)+\frac{1}{\kappa}H_2(t_2 \mid x;\beta_2)\right)^\kappa}, \end{aligned}$$

where H_1 and H_2 are the integrated hazard functions of the distributions in (75). The value of V may characterize a common property of individuals or a common environment. Since T_1 and T_2 share the same value of $V = v$, marginalizing the distribution of the two distributions with respect to V leads to dependent duration variables.

In the special case of a gamma heterogeneity term, κ can be seen as a measure of dependence, with $\kappa \to \infty$ corresponding to independence. At the same time, $\kappa \to \infty$ implies a vanishing variance for the mixing distribution, so that the values of V become concentrated around the value 1. In other words, the influence of common factors or the environment tends to a common value for all random variables considered. For an interpretation, however, it should be noted that κ features also in the marginal distributions of T_1 and T_2, respectively, via (74). Since the marginal distributions alone contain information on κ, the parameter cannot be said to pertain solely to dependence.

As the above examples demonstrate, proportional mixtures of proportional hazards models will in general lead to non–proportional covariate effects. It may be asked whether this is true for all mixing distributions M. The answer is in the negative. Using the relation given by the Laplace transform of a mixing distribution and the mixture (71), one needs only to consider Laplace transforms

$$\mathcal{L}_M(t) = e^{-t^\sigma} \text{ with } \sigma < 1.$$

It can be shown that such Laplace transforms do correspond to the distributions of positive random variables. But (71) then results in

$$G(t|x;\beta) = \mathcal{L}_M(e^{x\beta} H_0(t)) = e^{-\left(e^{x\beta} H_0(t)\right)^\sigma},$$

which is again a proportional hazards model. A sufficient condition to insure that proportional mixtures of proportional hazards models are not also in the class of proportional hazards models is to postulate a finite expectation for the heterogeneity term. This condition is sometimes stipulated when an empirical distinction between heterogeneity and proportional kernel is required. While this might be a reasonable assumption in special cases, there is obviously no way to decide problem empirically.

It remains to examine scale models—the second broad class of covariate effects—in conjunction with mixtures. Suppose, therefore, that the covariates as well as the heterogeneity term act as in a scale model,

multiplying an underlying duration variable T_0. In terms of logarithmic durations, the model can be written as

$$\begin{aligned} \ln T &= x\beta + \epsilon + U, \quad U \simeq_d M(u) \quad \text{or} \\ \ln T &= x\beta + \epsilon^*, \quad \epsilon^* = \epsilon + U \end{aligned} \tag{76}$$

On the transformed scale, mixing changes only the residual distribution. On this scale, introducing a mixing distribution does not lead to interesting consequences with regard to covariate effects. It simply increases the variability of the error term. Specifically, a *scale mixture* of a scale model of covariate effects is a scale model. However, the baseline distribution designated by T_0 is changed. This once again illustrates the interplay between the specification of covariate effects and mixtures.

4.2.6 Combining models for covariate effects and distributions

Given a class of distributions together with a parameterization, a rather direct way to introduce covariate effects is to make the parameters functions of the covariates. This will in some cases reduce to one of the classes of covariate effects discussed before. For example, in the parameterizations used here, the parameter a in the exponential, Weibull, log–logistic, and gamma distributions are scale parameters. Putting $a = \exp(-x\beta)$ leads to a scale model of covariate effects.

However, the second parameter in all the above two parameter classes does not have such an easy interpretation. Still, under certain circumstances it might be desirable to let these parameters be functions of some of the covariates, and flexible software packages allow for this possibility. Since the parameterization of a class of distributions is highly arbitrary and mostly follows custom, the interpretation of such models will require close scrutiny of the underlying parameterization.

Another possibility is to use one of the classes of covariate effects in conjunction with a class of distributions. E.g., none of the two parameter classes has a parameter representing proportional effects on the hazard rate. Introducing a proportional hazards model for the log–logistic distribution results in the hazard rate

$$r_{a,b}(t|x;\beta) = e^{x\beta} \frac{b a^b t^{b-1}}{1 + (at)^b} \tag{77}$$

with survivor function

$$G_{a,b}(t|x;\beta) = \frac{1}{(1+(at)^b)^{e^{x\beta}/a}}. \tag{78}$$

Comparison with the general proportional gamma mixture (74) reveals that this is the same as a gamma mixture of a Weibull model, where the variance of the mixing distribution is given by $\mathbb{V}(V) = 1/\kappa = a\exp(-x\beta)$ and the Weibull scale parameter in this interpretation is $\exp(x\beta/b)a^{(b-1)/b}$. Thus, in this model the covariate effect might be seen as either arising from a proportional hazards model or from the simultaneous determination of the variance of a mixing distribution and the scale.

It has sometimes been proposed to use both a proportional hazards and a scale model for covariate effects. While some covariates might multiply the hazard rate, others might multiply the scale of a model. Whether such a distinction is possible will depend on the class of distributions chosen. Both effects are basically the same within the Weibull class, while the log–logistic might be extended to allow both for a scale and a proportional hazards effect. However, extreme care is needed when the two covariate sets contain common members. First, the ability to distinguish the two effects hinges strongly on the family of distributions considered. Second, as can be seen from the case of the extended log–logistic distribution above, changes in proportional effects will also be reflected in the scale of the model. Third, both, higher rates and accelerated scales, while theoretically distinct concepts, lead to shorter durations. Since observations of durations are the only empirical basis for claims about covariate effects, estimators of the effects for the same covariates will tend to be highly correlated.

4.3 Time dependent covariates

One of the distinguishing aspects in the analysis of durations is the possibility to consider the impact of time varying covariates. Whether covariates represent the state of the environment, the stages of a decision process, or the contingencies of an actor, these changing circumstances can be incorporated in most duration models. The interpretation of their effects will depend not only on the form of covariate effects considered, but also on assumptions on the time path of these covariates.

Suppose first that the development of covariate values through time can be assumed to be fixed, or known in advance, or, at least, not to depend on the action of subjects figuring in the description of the durations of interest. This kind of covariates is called a *defined* time dependent covariate, if its time path can be ascertained without recourse to the actual event history. Otherwise, it is called *ancillary*.

Suppose next a proportional hazards model for the effect of such covariates. The effect of time dependent covariates can then be reflected in an immediate effect on the hazard at time t induced by the value of the covariate at the same time. Formally, this is written as

$$r(t|x(t)) = r_0(t)e^{x(t)\beta}. \tag{79}$$

An important special case arises when the process $x(t)$, considered as a function of time, is a step function. Let the process $x(t)$ be piecewise constant in the time intervals $0 < \tau_1 < \ldots < \tau_m < \infty$. Then the resultant conditional survivor function has a rather simple form since the integrated hazards can be evaluated piecewise also. For $\tau_m < t < \infty$, e.g.,

$$\begin{aligned} G(t|x(u)_{u\in[0,t)}) &= e^{-\int_0^t e^{x(u)\beta} r_0(u)\,du} \\ &= e^{-\left(e^{x_1\beta}\int_0^{\tau_1} r_0(u)\,du + \ldots + e^{x_m\beta}\int_{\tau_m}^{t} r_0(u)\,du\right)} \\ &= e^{-\alpha_1(H_0(\tau_1)-H_0(0)) - \ldots - \alpha_m(H_0(t)-H_0(\tau_m))}. \end{aligned} \tag{80}$$

An important application of this idea is used in a generalization of the class of exponential distributions. Fixing the values of the time intervals $\tau_0 = 0 < \tau_1 < \ldots < \tau_m < \infty$ and setting $\exp(x(u)\beta) = \alpha_k$ for u in the interval $[\tau_{k-1}, \tau_k)$ as above, while choosing the constant rate $r_0(u) \equiv 1$ gives rise to the *piecewise exponential distribution*. Its hazard rate is given by the function that is constant on the intervals $\tau_0 = 0 < \tau_1 < \ldots < \tau_m < \infty$, taking the value α_k on the k th interval. The hazard rate is therefore a step function. It follows that the survivor function is given by

$$G(t|x(u)_{u\in[0,t)}) = e^{-(\alpha_0\tau_1 + \alpha_1(\tau_2-\tau_1) + \ldots + \alpha_m(t-\tau_m))} \tag{81}$$

Choosing appropriate intervals and steps, it might be used to approximate other hazard rate functions. From (80), if $\tau_0 = 0 < \tau_1 < \ldots < \tau_m < \infty$ are the jump times of the covariate process $x(t)$, we get

$$\begin{aligned} & G(t|x(u)_{u\in[0,t)}) \\ = \;& e^{-e^{x(\tau_1)\beta}(H_0(\tau_0)-H_0(0)) - \ldots - e^{x(\tau_m)\beta}(H_0(t)-H_0(\tau_m))} \\ = \;& \prod_{k=1}^{m} \left(\frac{G_0(\tau_k)}{G_0(\tau_{k-1})}\right)^{e^{x(\tau_{k-1})\beta}} \left(\frac{G_0(t)}{G_0(\tau_m)}\right)^{e^{x(\tau_m)\beta}} \\ = \;& \prod_{k=1}^{m} (\Pr(T_0 > \tau_k | T_0 > \tau_{k-1}))^{e^{x(\tau_{k-1})\beta}} \times \\ & (\Pr(T_0 > t | T_0 > \tau_m))^{e^{x(\tau_m)\beta}} \end{aligned} \tag{82}$$

More generally, time dependent proportional covariates that are step functions with respect to time can be treated as in (80), if the integrated hazards have closed form expressions.

If the covariates act proportional on the hazard but are not step functions, one needs to be able to compute the integral with respect to time of $\exp(x(u)\beta)r_0(u)$ to get an expression for the survivor function and other summary functions. Models of this form with defined covariates are sometimes used to express deviations from the assumed proportional effect of covariates. A case in point is the use of the covariate $x(u) = x/(1+u)$ for some fixed covariate x. The covariate effect in a proportional hazards model is then

$$\phi(x(t);\beta) = e^{x\beta/(1+t)}. \tag{83}$$

The ratio of the hazard rates for two values of x, say x_1 and x_2, will then tend to one, in contrast to the proportional hazards model that was used as a starting point. Obviously, other forms of covariate effects or of classes of distributions can be obtained from deliberately choosing time dependent functions as covariates. As an example, consider $x(u) = \ln u$ in an exponential model. The rate then is $r(t|x;\beta) = \exp(\beta_0 + \beta_1 x(t))t = \exp(\beta_0)t^{\beta_1+1}$. In other words, the covariate transforms the exponential model into a Weibull model.

One may also start with a scale model of covariate effects. If time dependent variables are supposed to act immediately at each point in time, the physical interpretation of scale models leads the interpretation of covariate effects as changing the velocity of the underlying process as compared

to a uniform motion represented by the duration T_0. But change of velocity is acceleration. Therefore, putting

$$\Psi(t) = \int_0^t e^{-x(u)\beta}\, du,$$

one arrives at the expression

$$T = \Psi^{-1}(T_0),$$

which may be compared to the relation $T = \exp(x\beta)T_0$ for the scale model with fixed covariates. However, there is no special case similar to the piecewise constant case considered above. The integrals have to be worked out on a case by case bases. Moreover, if there are two or more time dependent covariates, the order of applying the respective transformations will matter. For both reasons, this type of transformation model is only rarely considered.

Defined or ancillary time dependent covariates can be used to extend the form of covariate effects and/or the class of distributions, and may have a direct interpretation as immediate effects of changing values of covariates. These simple interpretations are no longer available for *evolutionary covariates*. These covariates depend on the history of the whole process, and might not even be defined independently of the process under consideration. Simple examples are provided by measures that are outcomes of the process itself, like the amount of unemployment benefits received, when the interest centers on the duration of unemployment. Because of respective regulations, the amount of unemployment benefits will often simple be a re-expression of the duration of unemployment. In these cases, measures of effects can only be interpreted when the joint process is taken into account.

4.4 Censoring processes

The process that leads to censored observations is in general not of interest in itself. If censoring is judged to be non-informative, it neither enters into the construction of estimators nor in the interpretation of results.

On the other hand, censoring will certainly play a decisive role for the evaluation of estimators and for their precisions in any given sample. If in a sample of a hundred observations, two are censored, this is will

certainly signify a different information than that based on a hundred observations of which 90 are censored. Information on the censoring process is therefore needed for both the theoretical and the practical comparison of estimation procedures and results.

Of even greater practical importance is the fact that the probabilistic description of situations with censored observations is incomplete. If only the conditional distribution of the durations and the covariate effects are given, it is only possible to simulate complete observations, but never the impact of censored observations. This hampers the analysis of implications of assumed models as well as their criticism in a context where analytical results are especially difficult to obtain.

Specification of any censoring time independent of the duration time is sufficient to guarantee non-informative censoring. But a special case of that situation is very useful, both conceptually and empirically. Suppose, therefore, that censoring times and durations are independent. Moreover, assume that the rates of the durations and the censoring times do exist, and that they are proportional. Disregarding covariates for the moment, the assumption implies the existence of a constant a with

$$r_C(t) = a r_0(t), \tag{84}$$

where r_C is the rate function of the censoring time and r_0 is the rate function of the duration of interest. This special relationship between independent censoring times and durations is called the *Koziol–Green model*. The model has some simple but extremely useful consequences for simulations. First, the survivor function of the censored time $T^* = \min(C, T)$ is

$$\begin{aligned} \Pr(T^* > t) &= \Pr(C > t, T > t) \\ &= e^{-aH_0(t)} e^{-H_0(t)} = e^{-(1+a)H_0(t)}. \end{aligned}$$

In other words, all the distributions of T, C, and T^* have proportional hazards.

Second, the probability of censoring, $\Pr(D = 0)$, is equal to the ratio $a/(1+a)$:

$$\begin{aligned} \Pr(D = 0) &= \Pr(C < T) \\ &= \int \Pr(T > u) h_C(u)\, du \\ &= \int e^{-H_0(u)} a r_0(u) e^{-aH_0(u)}\, du \\ &= \int e^{-(1+a)H_0(u)} a r_0(u)\, du \\ &= \frac{a}{1+a} \int e^{-(1+a)H_0(u)} (1+a) r_0(u)\, du \\ &= \frac{a}{1+a}, \end{aligned}$$

where the equality in the last line follows upon observing that the integrand in the next to last line is the density of the random variable T^*.

Third, the censoring indicator D and the censored duration time T^* are independent. This follows from the same reasoning as above, in reverse order:

$$\begin{aligned} \Pr(D = 0) \Pr(T^* > t) &= \frac{a}{1+a} e^{-(1+a)H_0(t)} \\ &= \frac{a}{1+a} \int_t^\infty (1+a) r_0(u) e^{-(1+a)H_0(u)}\, du \\ &= \int_t^\infty e^{-H_0(u)} a r_0(u) e^{-aH_0(u)}\, du \\ &= \int_t^\infty \Pr(T > u) h_C(u)\, du \\ &= \Pr(T > C > t) = \Pr(D = 0, T^* > t). \end{aligned}$$

It can also be shown that the independence of the censoring indicator and the censored durations is sufficient for the Koziol–Green model to hold.

If the Koziol–Green model holds, it is possible to simulate censored observations by independently simulating the censoring indicator D and the censored times T^*. This allows for a simple control over censoring proportions in simulations. Moreover, some awkward computations

in the evaluation of the performance of estimators are considerably reduced. The Koziol–Green model of censoring has therefore become a convenient starting point for the evaluation of censored data models, both practically—through simulations—and theoretically.

5 Estimation

In the presence of censored observations there is no unified method for the construction of estimators with good properties. Of the many proposals, we have already mentioned two non–parametric estimators of survivor functions and the Buckley–James regression estimator. Three other construction methods that are especially useful in the context of regression models are treated next.

5.1 Maximum likelihood

Suppose first that a fully specified model for both the distribution and the covariate effect are given. Denote the parameter(s) of the distribution by θ, the parameters of the covariate effect model by β, and the resulting conditional density by $f(t|x;\theta,\beta)$. In the case of uncensored observations from independent replications of $T|x$, the joint density of n observations is given by

$$\prod_{i=1}^{n} f(t_i|x_i;\theta,\beta).$$

This may also be seen as a function of θ,β for given $(t_i,x_i), i=1,\ldots,n$, in which case it is called the *likelihood function*

$$L(\theta,\beta)=\prod_{i=1}^{n} f(t_i|x_i;\theta,\beta). \tag{85}$$

One may define estimators as those values of θ,β that maximize the function L,

$$(\hat{\theta},\hat{\beta})=\arg\max_{\theta,\beta} L(\theta,\beta), \tag{86}$$

the *maximum likelihood* estimator. To be of use in the analysis of durations, censoring must be included in the definition. Using the independent random censoring model from section 2.2, the data are now

(t_i,d_i,x_i). Their density involves the survivor function K and the density k of of the censoring variable C and is given by

$$\prod_{i=1}^{n} \left(f(t_i|x_i;\theta,\beta)K(t_i)\right)^{d_i} \left(G(t_i|x_i;\theta,\beta)k(t_i)\right)^{1-d_i}. \tag{87}$$

The contribution of n uncensored observation ($d_i=1$) to the likelihood is the density of the duration, $f(t_i|x_i;\theta,\beta)$, times the probability of a censoring time C after the observed duration, $K(t_i)$. The contribution of a censored observation ($1-d_i=1$) is the probability of a duration larger than t_i time the density of a censoring time at t_i.

If the censoring distribution does not contain information on (θ,β), the likelihood function is up to a multiplicative constant (terms not depending on θ or β)

$$L(\theta,\beta)=\prod_{i=1}^{n}(f(t_i|x_i;\theta,\beta))^{d_i}(G(t_i|x_i;\theta,\beta))^{1-d_i}. \tag{88}$$

That is, the likelihood is the product of the densities of the uncensored observations times the survivor functions of the censored observations.

Because of the product structure of the likelihood function it is advantageous to use the logarithm of the likelihood, the *log–likelihood* $\ell(\theta,\beta)=\ln L(\theta,\beta)$ as the function to be maximized. It is the sum of the logarithms of the densities or the survivor function respectively.

As an example, suppose $T|x$ is exponential with hazard rate $e^{x\beta}$. The density is $e^{x\beta}\exp(-e^{x\beta}t)$ and the survivor function is $\exp(-e^{x\beta}t)$, so that the log–likelihood function is

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{n} d_i\left(x_i\beta-e^{x_i\beta}t_i\right)+(1-d_i)(-e^{x_i\beta}t_i)\\ &= \sum_{i=1}^{n} d_i x_i\beta-\sum_{i=1}^{n} e^{x_i\beta}t_i.\end{aligned}$$

If the covariate vector contains only a constant, the maximum likelihood estimator can be given explicitly, since then

$$\ell(\beta)=\sum_{i=1}^{n} d_i\beta-\sum_{i=1}^{n} e^{\beta}t_i.$$

The derivative of the log–likelihood function, the *score function* $U(\beta)$ is

$$U(\beta) = \frac{\partial}{\partial\beta}\ell(\beta) = \sum_{i=1}^{n} d_i - \sum_{i=1}^{n} e^{\beta} t_i.$$

Setting this to 0 results in

$$\hat{\beta} = \ln \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} t_i}.$$

In general, the score function of the i th observation has expectation 0 and its covariance, the *information matrix*, can also be expressed as

$$\mathcal{I}_i(\beta) = \mathbb{E}_\beta\left(U_i(\beta)U_i'(\beta)\right) = -\mathbb{E}_\beta\left(\frac{\partial^2\ell_i(\beta)}{\partial\beta^2}\right). \tag{89}$$

In the exponential example, $\mathcal{I} = \mathbb{E}_\beta(e^\beta T^*)$. This equals 1 in the absence of censoring. The large sample theory of regular models suggests that the inverse of the information matrix is the asymptotic variance of $\hat{\beta}$, so that it can be used for the computation of confidence intervals and test statistics. However, the expectation in the definition of the information will normally depend on the censoring distribution so that it cannot be evaluated without strong assumptions. In our example, $\mathbb{E}_\beta(e^\beta T^*) = e^\beta \int \exp(-e^\beta u)K(u)\,du$.

In practice, the information is therefore replaced by the *observed information*, the negative of the sum of the second derivatives of the log–likelihood function. In the case of the exponential,

$$\mathcal{I}_{\text{obs}}(\beta) = -\sum_{i=1}^{n}\left(\frac{\partial^2\ell_i(\beta)}{\partial\beta^2}\right) = e^\beta\sum_{i=1}^{n} t_i \tag{90}$$

In the context of tests, $\hat{\beta}$ is substituted for β in $\mathcal{I}_{\text{obs}}(\beta)$. In the example, $\mathcal{I}_{\text{obs}}(\beta) = \sum_i d_i$. The Wald test then uses $(\hat{\beta}-\beta_0)\mathcal{I}_{\text{obs}}(\hat{\beta})^{-1}(\hat{\beta}-\beta_0)$ as a test statistic of the hypotheses β_0. It should be born in mind that this procedure is not invariant under re-parameterization, such as when the exponential distribution in the example is reexpressed by $a = e^\beta$. Moreover, in regression contexts the procedure may lead to unreliable results if the absolute value of some regression coefficients $\hat{\beta}$ gets large.

The method of maximum likelihood is applicable in most situations where the censored likelihood (88) can be written down and where the factoring of the likelihood (87) is judged appropriate. It provides a general method of estimation in many situations and is algorithmically simple. It may fail to produce reliable results, however, in situations with threshold parameters, for models containing many parameters, and in the presence of forms of incomplete data other than random censoring.

5.2 EM and the missing information principle

A much more flexible approach to incomplete data follows from the missing information principle that was already encountered when discussing the self-consistency property of the Kaplan–Meier estimator and the Buckley–James estimator. In both cases, some standard estimators, the empirical distribution function in the case of the Kaplan–Meier estimator and the least squares estimator in the case of the Buckley–James estimator were generalized to allow for censored data by replacing the unknown quantities by their expectation given the available data. The same principle can be used within the context of maximum likelihood estimation. The starting point in this case is the log–likelihood function. If there are incomplete observations, the full data log–likelihood terms are replaced by

$$\mathbb{E}_{\theta,\beta}\left(\ell_i(\theta,\beta;T,x)|T^* = t\right), \tag{91}$$

where the expectation depends on the current parameter values (θ,β) and T^* are the incomplete data ($\min(T,C), D$ in the case of censoring). The resulting log–likelihood function is then maximized with respect to the parameters, and the procedure is iterated.

In the case of the Buckley–James procedure the complete data score function is $U(\beta) = x'(Y - x\beta)$ from the normal linear regression model (27). The expectation satisfies

$$\mathbb{E}_\beta\left(U(\beta;Y,x)\right) = 0\,. \tag{92}$$

and the root $\hat{\beta}$ of

$$\sum_i U(\hat{\beta};y_i,x_i) = 0$$

is the maximum likelihood estimator. Even if the distribution is not normal — so that the root of the score function $\hat{\beta}$ need no longer be a maximum likelihood estimator — it is often consistent and efficient.

When the variables are censored with variables Z, d, x, then the censored normal score function can be expressed as

$$U^*(\beta; Z, d, x) = \mathbb{E}(U(\beta; Y, x) \mid Z, d, x)\,, \tag{93}$$

the conditional expectation of the score function with complete observations given the incomplete observations. This suggests using an empirical version

$$\mathbb{E}_{\hat{\beta}}\,(U(\hat{\beta}, Y, x \mid Z, d, x)) = 0$$

for estimation, and this is just (31). It remains to consider the computation of the conditional expectation. From the perspective of the normal linear regression model one might try to use the normal distribution. However, one can only expect the good properties of the estimators even outside the normal distribution to extend to censored data situations if the conditional expectation is computed from a non–parametric estimator. In the case of right censored observations, this amounts to using the Kaplan–Meier estimator (which is the non–parametric maximum likelihood estimator) as in the Buckley–James procedure.

5.3 Partial likelihood

Another extension of maximum likelihood ideas is the partial likelihood that allows estimation of proportional covariate effects without specifying a parameterized baseline distribution. Consider the proportional model

$$f(t, x; \beta) = e^{x\beta} r(t) e^{-\int_0^t e^{x\beta} r(u)\, du}\,. \tag{94}$$

Let $t_{(1)} < t_{(2)} \ldots < t_{(n)}$ be n ordered event times, all assumed to be uncensored. Let I_j be the label of an observation with an event at $t_{(j)}$ and $\mathcal{R}(t_{(j)})$ be the set of observations without an event before $t_{(j)}$. $\mathcal{R}(t_{(j)})$ is called the risk set at the event time $t_{(j)}$. Note that $R(t_{(j)})$ from (15) is the number of elements in $\mathcal{R}(t_{(j)})$. As an example consider figure 5.3. Here, $\mathcal{R}(t_{(1)}) = \{1, 2, 3, 4\}, \mathcal{R}(t_{(2)}) = \{1, 2, 4\}, \mathcal{R}(t_{(3)}) = \{1, 2\}$, and $\mathcal{R}(t_{(4)}) = \{2\}$. The set of indices I_j, the ordered event times $t_{(j)}$, and the covariates $x_{(i)}$ are jointly equivalent to the original data. If nothing is known about the hazard function r, the $t_{(j)}$ will contain little information about β. On the other hand, the distribution of I_j can be computed without

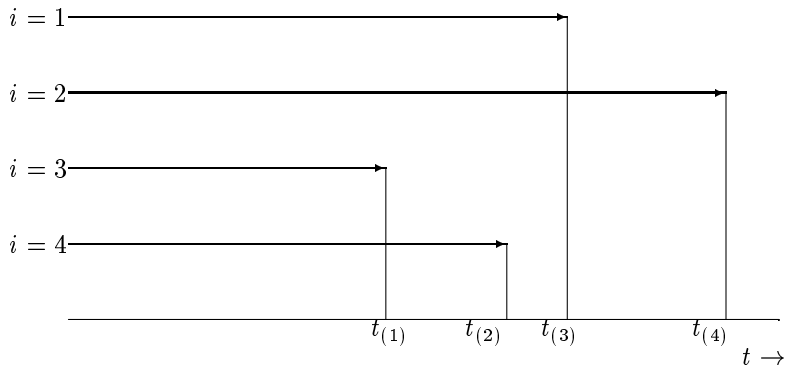

Figure 8: Risk Sets

knowledge of r. The conditional probability of an event for observation i at the j th event time given the history up to the j th event is

$$\begin{aligned}
&\Pr(I_j = i \mid (t_{(k)}, x_{(k)})_{k=1\ldots j}, (I_k)_{k=1\ldots j-1}) \\
&= \frac{r(t_{(i)}) e^{x_{(i)}\beta}}{\sum_{k \in \mathcal{R}(t_{(i)})} r(t_{(k)}) e^{x_{(k)}\beta}} \\
&= \frac{e^{x_{(i)}\beta}}{\sum_{k \in \mathcal{R}(t_{(i)})} e^{x_{(k)}\beta}}
\end{aligned} \tag{95}$$

Because of the proportional covariate effect, this conditional probability does not depend on the hazard rate r. Neither does it depend on the event times $t_{(j)}$. Therefore, the joint distribution of the indices $\{I_1, \ldots, I_n\}$ is the product of the above conditional probabilities

$$\Pr(I_1, I_2, \cdots, I_n) = \prod_{j=1}^{n} \frac{e^{x_{I_j}\beta}}{\sum_{k \in \mathcal{R}(t_{(j)})} e^{x_{(k)}\beta}} \tag{96}$$

If some observations are censored, a similar expression results in which all possible event times of the censored observations are considered. If $\mathcal{D}$

is the set of distinct uncensored observations and $\mathcal{R}_i$ the risk set corresponding to the event time of the i th observation, the *partial likelihood* can be written as

$$\mathrm{PL}(\beta) = \prod_{i \in \mathcal{D}} \frac{e^{x_i \beta}}{\sum_{k \in \mathcal{R}_i} e^{x_k \beta}}, \tag{97}$$

where the product is taken with respect to uncensored observations only. The partial likelihood depends only on the order of events, not on their timing. It is therefore invariant with respect to monotone transforms of the time scale.

The derivation of the partial likelihood included only the probabilities of the indices and information from the covariates. But from (95) alone, one cannot reconstruct the probability of the sample. Thus, no fully specified probability distribution is used, in contrast to the derivation of maximum likelihood estimators. Hence the name partial likelihood.

Though the maximizer of the partial likelihood, the *partial likelihood estimator*, is not in general equivalent to a maximum likelihood estimator, it shares a lot of the properties of the maximum likelihood estimators. Specifically, the second derivatives of the log partial likelihood behave like the observed information and can be used for the construction of tests and confidence intervals.

The score function of the partial likelihood is

$$\begin{aligned} \frac{\partial}{\partial \beta} \ln \mathrm{PL}(\beta) &= \sum U(\beta; t_i, d_i, x_i) \\ &= \sum_{i \in \mathcal{D}} \left(x_i' - \frac{\sum_{k \in \mathcal{R}_i} x_k' e^{x_k \beta}}{\sum_{k \in \mathcal{R}_i} e^{x_k \beta}} \right) \\ &= \sum_{i \in \mathcal{D}} (x_i' - A_i(\beta)'). \end{aligned} \tag{98}$$

The term $A_i(\beta)$ may be interpreted as the expectation of the covariates x in the i th risk set if the x_i are sampled proportional to $e^{x_i \beta}$ from the risk set. Similarly, the negative of the second derivative of the partial likelihood is the sum of covariance matrices of covariates from the risk sets. It follows that it is non negative definite if the moment matrices in

the risk sets are non singular. The partial likelihood is therefore concave and function maximizing algorithms generally converge rapidly.

6 References

6.1 Text Books

P.K. Andersen/Ø. Borgan/R.D. Gill/N. Keiding: Statistical Models Based on Counting Processes; Springer 1993

H.-P. Blossfeld/G. Rohwer: Techniques of Event History Modeling; Lawrence Erlbaum 1995

D.R. Cox/D. Oakes: Analysis of Survival Data; Chapman and Hall 1984

T.J. Hastie/R.J. Tibshirani: Generalized Additive Models; Chapman and Hall 1990

T.R. Fleming/D.P. Harrington: Counting Processes and Survival Analysis; Wiley 1991

J.D. Kalbfleisch/R.L. Prentice: The Statistical Analysis of Failure Time Data; Wiley 1980

N.B. Tuma/M. Hannan: Social Dynamics; Academic Press 1984

6.2 Articles

P.K. Andersen: Survival analysis 1982-1991: The second decade of the proportional hazards regression model; Statist. Med., 10, 1991, 1931–1941

E. Arjas: Survival models and martingal dynamics (with discussion); Scand. J. Statist., 16, 1989, 177–225

W.E. Barlow/R.L. Prentice: Residuals for relative risk regression; Biometrika, 75, 1988, 65–74

J. Buckley/I. James: Linear regression with censored data; Biometrika, 66, 1979, 429–436

K.A. Doksum/M. Gasko: On a correspondence between models in binary regression analysis and in survival analysis; Int. Statist. Rev., 58, 1990, 243–252

B. Efron: Logistic regression, survival analysis, and the Kaplan-Meier curve; J. Am. Statist. Assoc., 83, 1988, 414–425

R.D. Gill: Understanding Cox's regression model: A martingale approach; J. Am. Statist. Assoc., 79, 1984, 441–447

N.L. Hjort: On inference in parametric survival data models; Int. Statist. Rev., 60, 1992, 355–387

J. Hobcroft/M. Murphy: Demographic event history analysis: A selective review; Population Index, 52, 1986, 3–27

O. Intrator/C. Kooperberg: Trees and splines in survival analysis; Statist. Med. Res., 4, 1995, 237–261

N.M. Kiefer: Economic duration data and hazard functions; J. Ec. Literature, 26, 1988, 646–679

L. Le Cam: Maximum likelihood: an introduction; Int. Statist. Rev., 58, 1990, 153–171

R.G. Miller: What price Kaplan-Meier?; Biometrics, 39, 1983, 1077–1081

D. Oakes: Multiple time scales in survival analysis; Lifetime Data Anal., 1, 1995, 7–18

T. Petersen: Recent advances in longitudinal methodology; Ann. Rev. Sociol., 19, 1993, 425–454

J.D. Teachman/M.D. Hayward: Interpreting hazard rate models; Soc. Meth. Res., 21, 1993, 340–371

L.J. Wei: The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis (with discussion); Statist. Med., 11, 1992, 1871–1885

Index

γ–odds model, 31

Accelerated failure time model, 26

Buckley–James estimator, 19

Censoring
    Koziol–Green model, 55
Censoring time, 10

Distribution function, 3
Distributions
    defective, 45
    exponential, 37
    frailty model, 46
    gamma, 44
    log–logistic, 42
    Pareto, 47
    piecewise exponential, 52
    proportional mixture, 46
    scale mixture, 50
    Weibull, 39

Equivalent effects, 32

Greenwood's formula, 15

Hazard function
    time dependence, 37, 39
Hazard rate, 6
Hazard transform, 38
Heterogeneity, 45

Independent random censoring, 10
Integrated hazard rate, 8

Kaplan–Meier estimator, 14

Lack of memory property, 37
Laplace transform, 46
Likelihood function, 57
Log–odds model, 31

Mixture, 46
Mover–stayer model, 45

Nelson–Aalen estimator, 13

Partial likelihood, 63
Partly additive model, 34
Proportional hazards model, 27

Regression tree, 35

Scale model, 26
Self consistency, 15
Survivor function, 3

Ties, 15
Time dependent covariates
    ancillary, 52
    defined, 52
    evolutionary, 54