Chapter 7

Survival Models

Our final chapter concerns models for the analysis of data which have three main characteristics: (1) the dependent variable or response is the waiting *time* until the occurrence of a well-defined event, (2) observations are *censored*, in the sense that for some units the event of interest has not occurred at the time the data are analyzed, and (3) there are predictors or *explanatory* variables whose effect on the waiting time we wish to assess or control. We start with some basic definitions.

7.1 The Hazard and Survival Functions

Let T be a non-negative random variable representing the waiting time until the occurrence of an event. For simplicity we will adopt the terminology of survival analysis, referring to the event of interest as 'death' and to the waiting time as 'survival' time, but the techniques to be studied have much wider applicability. They can be used, for example, to study age at marriage, the duration of marriage, the intervals between successive births to a woman, the duration of stay in a city (or in a job), and the length of life. The observant demographer will have noticed that these examples include the fields of fertility, mortality and migration.

7.1.1 The Survival Function

We will assume for now that T is a continuous random variable with probability density function (p.d.f.) $f(t)$ and cumulative distribution function (c.d.f.) $F(t) = \Pr\{T < t\}$, giving the probability that the event has occurred by duration t.

It will often be convenient to work with the complement of the c.d.f, the *survival* function

$$S(t) = \Pr\{T \geq t\} = 1 - F(t) = \int_t^\infty f(x)dx, \tag{7.1}$$

which gives the probability of being alive just before duration t, or more generally, the probability that the event of interest has not occurred by duration t.

7.1.2 The Hazard Function

An alternative characterization of the distribution of T is given by the *hazard* function, or instantaneous rate of occurrence of the event, defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt | T \geq t\}}{dt}. \tag{7.2}$$

The numerator of this expression is the conditional probability that the event will occur in the interval $[t, t + dt)$ given that it has not occurred before, and the denominator is the width of the interval. Dividing one by the other we obtain a rate of event occurrence per unit of time. Taking the limit as the width of the interval goes down to zero, we obtain an instantaneous rate of occurrence.

The conditional probability in the numerator may be written as the ratio of the joint probability that T is in the interval $[t, t + dt)$ *and* $T \geq t$ (which is, of course, the same as the probability that t is in the interval), to the probability of the condition $T \geq t$. The former may be written as $f(t)dt$ for small dt, while the latter is $S(t)$ by definition. Dividing by dt and passing to the limit gives the useful result

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{7.3}$$

which some authors give as a definition of the hazard function. In words, the rate of occurrence of the event at duration t equals the density of events at t, divided by the probability of surviving to that duration without experiencing the event.

Note from Equation 7.1 that $-f(t)$ is the derivative of $S(t)$. This suggests rewriting Equation 7.3 as

$$\lambda(t) = -\frac{d}{dt} \log S(t).$$

If we now integrate from 0 to t and introduce the boundary condition $S(0) = 1$ (since the event is sure not to have occurred by duration 0), we can solve the above expression to obtain a formula for the probability of surviving to duration t as a function of the hazard at all durations up to t:

$$S(t) = \exp\{-\int_0^t \lambda(x)dx\}. \tag{7.4}$$

This expression should be familiar to demographers. The integral in curly brackets in this equation is called the *cumulative hazard* (or cumulative risk) and is denoted

$$\Lambda(t) = \int_0^t \lambda(x)dx. \tag{7.5}$$

You may think of $\Lambda(t)$ as the sum of the risks you face going from duration 0 to t.

These results show that the survival and hazard functions provide alternative but equivalent characterizations of the distribution of T. Given the survival function, we can always differentiate to obtain the density and then calculate the hazard using Equation 7.3. Given the hazard, we can always integrate to obtain the cumulative hazard and then exponentiate to obtain the survival function using Equation 7.4. An example will help fix ideas.

Example: The simplest possible survival distribution is obtained by assuming a constant risk over time, so the hazard is

$$\lambda(t) = \lambda$$

for all t. The corresponding survival function is

$$S(t) = \exp\{-\lambda t\}.$$

This distribution is called the exponential distribution with parameter λ. The density may be obtained multiplying the survivor function by the hazard to obtain

$$f(t) = \lambda \exp\{-\lambda t\}.$$

The mean turns out to be $1/\lambda$. This distribution plays a central role in survival analysis, although it is probably too simple to be useful in applications in its own right.□

7.1.3 Expectation of Life

Let μ denote the mean or expected value of T. By definition, one would calculate μ multiplying t by the density $f(t)$ and integrating, so

$$\mu = \int_0^\infty tf(t)dt.$$

Integrating by parts, and making use of the fact that $-f(t)$ is the derivative of $S(t)$, which has limits or boundary conditions $S(0) = 1$ and $S(\infty) = 0$, one can show that

$$\mu = \int_0^\infty S(t)dt. \tag{7.6}$$

In words, the mean is simply the integral of the survival function.

7.1.4 A Note on Improper Random Variables*

So far we have assumed implicitly that the event of interest is bound to occur, so that $S(\infty) = 0$. In words, given enough time the proportion surviving goes down to zero. This condition implies that the cumulative hazard must diverge, i.e. we must have $\Lambda(\infty) = \infty$. Intuitively, the event will occur with certainty only if the cumulative risk over a long period is sufficiently high.

There are, however, many events of possible interest that are not bound to occur. Some men and women remain forever single, some birth intervals never close, and some people are happy enough at their jobs that they never leave. What can we do in these cases? There are two approaches one can take.

One approach is to note that we can still calculate the hazard and survival functions, which are well defined even if the event of interest is not bound to occur. For example we can study marriage in the entire population, which includes people who will never marry, and calculate marriage rates and proportions single. In this example $S(t)$ would represent the proportion still single at age t and $S(\infty)$ would represent the proportion who never marry.

One limitation of this approach is that if the event is not certain to occur, then the waiting time T could be undefined (or infinite) and thus not a proper random variable. Its density, which could be calculated from the hazard and survival, would be improper, i.e. it would fail to integrate to one. Obviously, the mean waiting time would not be defined. In terms of our example, we cannot calculate mean age at marriage for the entire population, simply because not everyone marries. But this limitation is of no great consequence if interest centers on the hazard and survivor functions, rather than the waiting time. In the marriage example we can even calculate a median age at marriage, provided we define it as the age by which half the population has married.

The alternative approach is to condition the analysis on the event actually occurring. In terms of our example, we could study marriage (perhaps retrospectively) for people who eventually marry, since for this group the

actual waiting time T is always well defined. In this case we can calculate not just the conditional hazard and survivor functions, but also the mean. In our marriage example, we could calculate the mean age at marriage for those who marry. We could even calculate a conventional median, defined as the age by which half the people who will eventually marry have done so.

It turns out that the conditional density, hazard and survivor function for those who experience the event are related to the unconditional density, hazard and survivor for the entire population. The conditional density is

$$f^*(t) = \frac{f(t)}{1 - S(\infty)},$$

and it integrates to one. The conditional survivor function is

$$S^*(t) = \frac{S(t) - S(\infty)}{1 - S(\infty)},$$

and goes down to zero as $t \to \infty$. Dividing the density by the survivor function, we find the conditional hazard to be

$$\lambda^*(t) = \frac{f^*(t)}{S^*(t)} = \frac{f(t)}{S(t) - S(\infty)}.$$

Derivation of the mean waiting time for those who experience the event is left as an exercise for the reader.

Whichever approach is adopted, care must be exercised to specify clearly which hazard or survival is being used. For example, the conditional hazard for those who eventually experience the event is always higher than the unconditional hazard for the entire population. Note also that in most cases all we observe is whether or not the event has occurred. If the event has not occurred, we may be unable to determine whether it will eventually occur. In this context, only the unconditional hazard may be estimated from data, but one can always translate the results into conditional expressions, if so desired, using the results given above.

7.2 Censoring and The Likelihood Function

The second distinguishing feature of the field of survival analysis is censoring: the fact that for some units the event of interest has occurred and therefore we know the exact waiting time, whereas for others it has not occurred, and all we know is that the waiting time exceeds the observation time.

7.2.1 Censoring Mechanisms

There are several mechanisms that can lead to censored data. Under censoring of *Type I*, a sample of n units is followed for a fixed time τ. The number of units experiencing the event, or the number of 'deaths', is random, but the total duration of the study is fixed. The fact that the duration is fixed may be an important practical advantage in designing a follow-up study.

In a simple generalization of this scheme, called *fixed censoring*, each unit has a potential maximum observation time τ_i for $i = 1, \ldots, n$ which may differ from one case to the next but is nevertheless fixed in advance. The probability that unit i will be alive at the end of her observation time is $S(\tau_i)$, and the total number of deaths is again random.

Under censoring of *Type II*, a sample of n units is followed as long as necessary until d units have experienced the event. In this design the number of deaths d, which determines the precision of the study, is fixed in advance and can be used as a design parameter. Unfortunately, the total duration of the study is then random and cannot be known with certainty in advance.

In a more general scheme called *random censoring*, each unit has associated with it a potential censoring time C_i and a potential lifetime T_i, which are assumed to the independent random variables. We observe $Y_i = \min\{C_i, T_i\}$, the minimum of the censoring and life times, and an indicator variable, often called d_i or δ_i, that tells us whether observation terminated by death or by censoring.

All these schemes have in common the fact that the censoring mechanism is *non-informative* and they all lead to essentially the same likelihood function. The weakest assumption required to obtain this common likelihood is that the censoring of an observation should not provide any information regarding the prospects of survival of that particular unit beyond the censoring time. In fact, the basic assumption that we will make is simply this: all we know for an observation censored at duration t is that the lifetime exceeds t.

7.2.2 The Likelihood Function for Censored Data

Suppose then that we have n units with lifetimes governed by a survivor function $S(t)$ with associated density $f(t)$ and hazard $\lambda(t)$. Suppose unit i is observed for a time t_i. If the unit died at t_i, its contribution to the likelihood function is the density at that duration, which can be written as the product of the survivor and hazard functions

$$L_i = f(t_i) = S(t_i)\lambda(t_i).$$

If the unit is still alive at t_i, all we know under non-informative censoring is that the lifetime exceeds t_i. The probability of this event is

$$L_i = S(t_i),$$

which becomes the contribution of a censored observation to the likelihood.

Note that both types of contribution share the survivor function $S(t_i)$, because in both cases the unit lived up to time t_i. A death multiplies this contribution by the hazard $\lambda(t_i)$, but a censored observation does not. We can write the two contributions in a single expression. To this end, let d_i be a death indicator, taking the value one if unit i died and the value zero otherwise. Then the likelihood function may be written as follows

$$L = \prod_{i=1}^{n} L_i = \prod_i \lambda(t_i)^{d_i} S(t_i).$$

Taking logs, and recalling the expression linking the survival function $S(t)$ to the cumulative hazard function $\Lambda(t)$, we obtain the log-likelihood function for censored survival data

$$\log L = \sum_{i=1}^{n} \{d_i \log \lambda(t_i) - \Lambda(t_i)\}. \tag{7.7}$$

We now consider an example to reinforce these ideas.

Example: Suppose we have a sample of n censored observations from an exponential distribution. Let t_i be the observation time and d_i the death indicator for unit i.

In the exponential distribution $\lambda(t) = \lambda$ for all t. The cumulative risk turns out to be the integral of a constant and is therefore $\Lambda(t) = \lambda t$. Using these two results on Equation 7.7 gives the log-likelihood function

$$\log L = \sum \{d_i \log \lambda - \lambda t_i\}.$$

Let $D = \sum d_i$ denote the total number of deaths, and let $T = \sum t_i$ denote the total observation (or exposure) time. Then we can rewrite the log-likelihood as a function of these totals to obtain

$$\log L = D \log \lambda - \lambda T. \tag{7.8}$$

Differentiating this expression with respect to λ we obtain the score function

$$u(\lambda) = \frac{D}{\lambda} - T,$$

and setting the score to zero gives the maximum likelihood estimator of the hazard

$$\hat{\lambda} = \frac{D}{T}, \tag{7.9}$$

the total number of deaths divided by the total exposure time. Demographers will recognize this expression as the general definition of a death rate. Note that the estimator is optimal (in a maximum likelihood sense) only if the risk is constant and does not depend on age.

We can also calculate the observed information by taking minus the second derivative of the score, which is

$$I(\lambda) = \frac{D}{\lambda^2}.$$

To obtain the expected information we need to calculate the expected number of deaths, but this depends on the censoring scheme. For example under Type I censoring with fixed duration τ, one would expect $n(1-S(\tau))$ deaths. Under Type II censoring the number of deaths would have been fixed in advance. Under some schemes calculation of the expectation may be fairly complicated if not impossible.

A simpler alternative is to use the observed information, estimated using the m.l.e. of λ given in Equation 7.9. Using this approach, the large sample variance of the m.l.e. of the hazard rate may be estimated as

$$\hat{\text{var}}(\hat{\lambda}) = \frac{D}{T^2},$$

a result that leads to large-sample tests of hypotheses and confidence intervals for λ.

If there are no censored cases, so that $d_i = 1$ for all i and $D = n$, then the results obtained here reduce to standard maximum likelihood estimation for the exponential distribution, and the m.l.e. of λ turns out to be the reciprocal of the sample mean.

It may be interesting to note in passing that the log-likelihood for censored exponential data given in Equation 7.8 coincides exactly (except for constants) with the log-likelihood that would be obtained by treating D as a Poisson random variable with mean λT. To see this point, you should write the Poisson log-likelihood when $D \sim P(\lambda T)$, and note that it differs from Equation 7.8 only in the presence of a term $D \log(T)$, which is a constant depending on the data but not on the parameter λ.

Thus, treating the deaths as Poisson conditional on exposure time leads to exactly the same estimates (and standard errors) as treating the exposure

times as censored observations from an exponential distribution. This result will be exploited below to link survival models to generalized linear models with Poisson error structure.

7.3 Approaches to Survival Modeling

Up to this point we have been concerned with a homogeneous population, where the lifetimes of all units are governed by the same survival function $S(t)$. We now introduce the third distinguishing characteristic of survival models—the presence of a vector of covariates or explanatory variables that may affect survival time—and consider the general problem of modeling these effects.

7.3.1 Accelerated Life Models*

Let T_i be a random variable representing the (possibly unobserved) survival time of the i-th unit. Since T_i must be non-negative, we might consider modeling its logarithm using a conventional linear model, say

$$\log T_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i,$$

where ϵ_i is a suitable error term, with a distribution to be specified. This model specifies the distribution of log-survival for the i-th unit as a simple *shift* of a standard or baseline distribution represented by the error term.

Exponentiating this equation, we obtain a model for the survival time itself

$$T_i = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}T_{0i},$$

where we have written T_{0i} for the exponentiated error term. It will also be convenient to use γ_i as shorthand for the multiplicative effect $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ of the covariates.

Interpretation of the parameters follows along standard lines. Consider, for example, a model with a constant and a dummy variable x representing a factor with two levels, say groups one and zero. Suppose the corresponding multiplicative effect is $\gamma = 2$, so the coefficient of x is $\beta = \log(2) = 0.6931$. Then we would conclude that people in group one live twice as long as people in group zero.

There is an interesting alternative interpretation that explains the name 'accelerated life' used for this model. Let $S_0(t)$ denote the survivor function in group zero, which will serve as a reference group, and let $S_1(t)$ denote the

survivor function in group one. Under this model,

$$S_1(t) = S_0(t/\gamma).$$

In words, the probability that a member of group one will be alive at age t is exactly the same as the probability that a member of group zero will be alive at age t/γ. For $\gamma = 2$, this would be half the age, so the probability that a member of group one would be alive at age 40 (or 60) would be the same as the probability that a member of group zero would be alive at age 20 (or 30). Thus, we may think of γ as affecting the passage of time. In our example, people in group zero age 'twice as fast'.

For the record, the corresponding hazard functions are related by

$$\lambda_1(t) = \lambda_0(t/\gamma)/\gamma,$$

so if $\gamma = 2$, at any given age people in group one would be exposed to half the risk of people in group zero half their age.

The name 'accelerated life' stems from industrial applications where items are put to test under substantially worse conditions than they are likely to encounter in real life, so that tests can be completed in a shorter time.

Different kinds of parametric models are obtained by assuming different distributions for the error term. If the ϵ_i are normally distributed, then we obtain a log-normal model for the T_i. Estimation of this model for censored data by maximum likelihood is known in the econometric literature as a Tobit model.

Alternatively, if the ϵ_i have an extreme value distribution with p.d.f.

$$f(\epsilon) = \exp\{\epsilon - \exp\{\epsilon\}\},$$

then T_{0i} has an exponential distribution, and we obtain the exponential regression model, where T_i is exponential with hazard λ_i satisfying the log-linear model

$$\log \lambda_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

An example of a demographic model that belongs to the family of accelerated life models is the Coale-McNeil model of first marriage frequencies, where the proportion ever married at age a in a given population is written as

$$F(a) = cF_0(\frac{a - a_0}{k}),$$

where F_0 is a model schedule of proportions married by age, among women who will ever marry, based on historical data from Sweden; c is the proportion who will eventually marry, a_0 is the age at which marriage starts, and k is the *pace* at which marriage proceeds relative to the Swedish standard.

Accelerated life models are essentially standard regression models applied to the log of survival time, and except for the fact that observations are censored, pose no new estimation problems. Once the distribution of the error term is chosen, estimation proceeds by maximizing the log-likelihood for censored data described in the previous subsection. For further details, see Kalbfleish and Prentice (1980).

7.3.2 Proportional Hazard Models

A large family of models introduced by Cox (1972) focuses directly on the hazard function. The simplest member of the family is the *proportional hazards* model, where the hazard at time t for an individual with covariates $\mathbf{x}_i$ (not including a constant) is assumed to be

$$\lambda_i(t|\mathbf{x}_i) = \lambda_0(t) \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}. \tag{7.10}$$

In this model $\lambda_0(t)$ is a baseline hazard function that describes the risk for individuals with $\mathbf{x}_i = \mathbf{0}$, who serve as a reference cell or pivot, and $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ is the relative risk, a proportionate increase or reduction in risk, associated with the set of characteristics $\mathbf{x}_i$. Note that the increase or reduction in risk is the same at all durations t.

To fix ideas consider a two-sample problem where we have a dummy variable x which serves to identify groups one and zero. Then the model is

$$\lambda_i(t|x) = \begin{cases} \lambda_0(t) & \text{if } x = 0, \\ \lambda_0(t)e^{\beta} & \text{if } x = 1. \end{cases}.$$

Thus, $\lambda_0(t)$ represents the risk at time t in group zero, and $\gamma = \exp\{\beta\}$ represents the ratio of the risk in group one relative to group zero at any time t. If $\gamma = 1$ (or $\beta = 0$) then the risks are the same in the two groups. If $\gamma = 2$ (or $\beta = 0.6931$), then the risk for an individual in group one at any given age is twice the risk of a member of group zero who has the same age.

Note that the model separates clearly the effect of time from the effect of the covariates. Taking logs, we find that the proportional hazards model is a simple additive model for the log of the hazard, with

$$\log \lambda_i(t|\mathbf{x}_i) = \alpha_0(t) + \mathbf{x}_i'\boldsymbol{\beta},$$

where $\alpha_0(t) = \log \lambda_0(t)$ is the log of the baseline hazard. As in all additive models, we assume that the effect of the covariates $\mathbf{x}$ is the same at all times or ages t. The similarity between this expression and a standard analysis of covariance model with parallel lines should not go unnoticed.

Returning to Equation 7.10, we can integrate both sides from 0 to t to obtain the cumulative hazards

$$\Lambda_i(t|\mathbf{x}_i) = \Lambda_0(t) \exp\{\mathbf{x}_i'\boldsymbol{\beta}\},$$

which are also proportional. Changing signs and exponentiating we obtain the survivor functions

$$S_i(t|\mathbf{x}_i) = S_0(t)^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}, \tag{7.11}$$

where $S_0(t) = \exp\{-\Lambda_0(t)\}$ is a baseline survival function. Thus, the effect of the covariate values $\mathbf{x}_i$ on the survivor function is to raise it to a power given by the relative risk $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$.

In our two-group example with a relative risk of $\gamma = 2$, the probability that a member of group one will be alive at any given age t is the square of the probability that a member of group zero would be alive at the same age.

7.3.3 The Exponential and Weibull Models

Different kinds of proportional hazard models may be obtained by making different assumptions about the baseline survival function, or equivalently, the baseline hazard function. For example if the baseline risk is constant over time, so $\lambda_0(t) = \lambda_0$, say, we obtain the exponential regression model, where

$$\lambda_i(t, \mathbf{x}_i) = \lambda_0 \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}.$$

Interestingly, the exponential regression model belongs to both the proportional hazards and the accelerated life families. If the baseline risk is a constant and you double or triple the risk, the new risk is still constant (just higher). Perhaps less obviously, if the baseline risk is constant and you imagine time flowing twice or three times as fast, the new risk is doubled or tripled but is still constant over time, so we remain in the exponential family.

You may be wondering whether there are other cases where the two models coincide. The answer is yes, but not many. In fact, there is only one distribution where they do, and it includes the exponential as a special case.

The one case where the two families coincide is the *Weibull* distribution, which has survival function

$$S(t) = \exp\{-(\lambda t)^p\}$$

and hazard function

$$\lambda(t) = p\lambda(\lambda t)^{p-1},$$

for parameters $\lambda > 0$ and $p > 0$. If $p = 1$, this model reduces to the exponential and has constant risk over time. If $p > 1$, then the risk increases over time. If $p < 1$, then the risk decreases over time. In fact, taking logs in the expression for the hazard function, we see that the log of the Weibull risk is a linear function of log time with slope $p - 1$.

If we pick the Weibull as a baseline risk and then multiply the hazard by a constant γ in a proportional hazards framework, the resulting distribution turns out to be still a Weibull, so the family is closed under proportionality of hazards. If we pick the Weibull as a baseline survival and then speed up the passage of time in an accelerated life framework, dividing time by a constant γ, the resulting distribution is still a Weibull, so the family is closed under acceleration of time.

For further details on this distribution see Cox and Oakes (1984) or Kalbfleish and Prentice (1980), who prove the equivalence of the two Weibull models.

7.3.4 Time-varying Covariates

So far we have considered explicitly only covariates that are fixed over time. The local nature of the proportional hazards model, however, lends itself easily to extensions that allows for covariates that change over time. Let us consider a few examples.

Suppose we are interested in the analysis of birth spacing, and study the interval from the birth of one child to the birth of the next. One of the possible predictors of interest is the mother's education, which in most cases can be taken to be fixed over time.

Suppose, however, that we want to introduce breastfeeding status of the child that begins the interval. Assuming the child is breastfed, this variable would take the value one ('yes') from birth until the child is weaned, at which time it would take the value zero ('no'). This is a simple example of a predictor that can change value only once.

A more elaborate analysis could rely on frequency of breastfeeding in a 24-hour period. This variable could change values from day to day. For example a sequence of values for one woman could be 4,6,5,6,5,4,...

Let $\mathbf{x}_i(t)$ denote the value of a vector of covariates for individual i at time or duration t. Then the proportional hazards model may be generalized to

$$\lambda_i(t, \mathbf{x}_i(t)) = \lambda_0(t) \exp\{\mathbf{x}_i(t)'\boldsymbol{\beta}\}. \tag{7.12}$$

The separation of duration and covariate effects is not so clear now, and on occasion it may be difficult to identify effects that are highly collinear with time. If all children were weaned when they are around six months old, for example, it would be difficult to identify effects of breastfeeding from general duration effects without additional information. In such cases one might still prefer a time-varying covariate, however, as a more meaningful predictor of risk than the mere passage of time.

Calculation of survival functions when we have time-varying covariates is a little bit more complicated, because we need to specify a path or trajectory for each variable. In the birth intervals example one could calculate a survival function for women who breastfeed for six months and then wean. This would be done by using the hazard corresponding to $x(t) = 0$ for months 0 to 6 and then the hazard corresponding to $x(t) = 1$ for months 6 onwards. Unfortunately, the simplicity of Equation 7.11 is lost; we can no longer simply raise the baseline survival function to a power.

Time-varying covariates can be introduced in the context of accelerated life models, but this is not so simple and has rarely been done in applications. See Cox and Oakes (1984, p.66) for more information.

7.3.5 Time-dependent Effects

The model may also be generalized to allow for *effects* that vary over time, and therefore are no longer proportional. It is quite possible, for example, that certain social characteristics might have a large impact on the hazard for children shortly after birth, but may have a relatively small impact later in life. To accommodate such models we may write

$$\lambda_i(t, \mathbf{x}_i) = \lambda_0(t) \exp\{\mathbf{x}_i'\boldsymbol{\beta}(t)\},$$

where the parameter $\boldsymbol{\beta}(t)$ is now a function of time.

This model allows for great generality. In the two-sample case, for example, the model may be written as

$$\lambda_i(t|x) = \begin{cases} \lambda_0(t) & \text{if } x = 0 \\ \lambda_0(t)e^{\beta}(t) & \text{if } x = 1 \end{cases},$$

which basically allows for two arbitrary hazard functions, one for each group. Thus, this is a form of saturated model.

Usually the form of time dependence of the effects must be specified parametrically in order to be able to identify the model and estimate the parameters. Obvious candidates are polynomials on duration, where $\beta(t)$ is a linear or quadratic function of time. Cox and Oakes (1984, p. 76) show how one can use quick-dampening exponentials to model transient effects.

Note that we have lost again the simple separation of time and covariate effects. Calculation of the survival function in this model is again somewhat complicated by the fact that the coefficients are now functions of time, so they don't fall out of the integral. The simple Equation 7.11 does not apply.

7.3.6 The General Hazard Rate Model

The foregoing extensions to time-varying covariates and time-dependent effects may be combined to give the most general version of the hazard rate model, as

$$\lambda_i(t, \mathbf{x}_i(t)) = \lambda_0(t) \exp\{\mathbf{x}_i(t)'\boldsymbol{\beta}(t)\},$$

where $\mathbf{x}_i(t)$ is a vector of time-varying covariates representing the characteristics of individual i at time t, and $\boldsymbol{\beta}(t)$ is a vector of time-dependent coefficients, representing the effect that those characteristics have at time or duration t.

The case of breastfeeding status and its effect on the length of birth intervals is a good example that combines the two effects. Breastfeeding status is itself a time-varying covariate $x(t)$, which takes the value one if the woman is breastfeeding her child t months after birth. The effect that breastfeeding may have in inhibiting ovulation and therefore reducing the risk of pregnancy is known to decline rapidly over time, so it should probably be modeled as a time dependent effect $\beta(t)$. Again, further progress would require specifying the form of this function of time.

7.3.7 Model Fitting

There are essentially three approaches to fitting survival models:

- The first and perhaps most straightforward is the *parametric* approach, where we assume a specific functional form for the baseline hazard $\lambda_0(t)$. Examples are models based on the exponential, Weibull, gamma and generalized F distributions.

- A second approach is what might be called a flexible or *semi-parametric* strategy, where we make mild assumptions about the baseline hazard

$\lambda_0(t)$. Specifically, we may subdivide time into reasonably small intervals and assume that the baseline hazard is constant in each interval, leading to a piece-wise exponential model.

- The third approach is a *non-parametric* strategy that focuses on estimation of the regression coefficients $\boldsymbol{\beta}$ leaving the baseline hazard $\lambda_0(t)$ completely unspecified. This approach relies on a partial likelihood function proposed by Cox (1972) in his original paper.

A complete discussion of these approaches in well beyond the scope of these notes. We will focus on the intermediate or semi-parametric approach because (1) it is sufficiently flexible to provide a useful tool with wide applicability, and (2) it is closely related to Poisson regression analysis.

7.4 The Piece-Wise Exponential Model

We will consider fitting a proportional hazards model of the usual form

$$\lambda_i(t|\mathbf{x}_i) = \lambda_0(t) \exp\{\mathbf{x}_i'\boldsymbol{\beta}\} \tag{7.13}$$

under relatively mild assumptions about the baseline hazard $\lambda_0(t)$.

7.4.1 A Piece-wise Constant Hazard

Consider partitioning duration into J intervals with cutpoints $0 = \tau_0 < \tau_1 < \ldots < \tau_J = \infty$. We will define the j-th interval as $[\tau_{j-1}, \tau_j)$, extending from the $(j-1)$-st boundary to the j-th and including the former but not the latter.

We will then assume that the baseline hazard is *constant* within each interval, so that

$$\lambda_0(t) = \lambda_j \quad \text{for } t \text{ in } \quad [\tau_{j-1}, \tau_j). \tag{7.14}$$

Thus, we model the baseline hazard $\lambda_0(t)$ using J parameters $\lambda_1, \ldots, \lambda_J$, each representing the risk for the reference group (or individual) in one particular interval. Since the risk is assumed to be piece-wise constant, the corresponding survival function is often called a piece-wise exponential.

Clearly, judicious choice of the cutpoints should allow us to approximate reasonably well almost any baseline hazard, using closely-spaced boundaries where the hazard varies rapidly and wider intervals where the hazard changes more slowly.

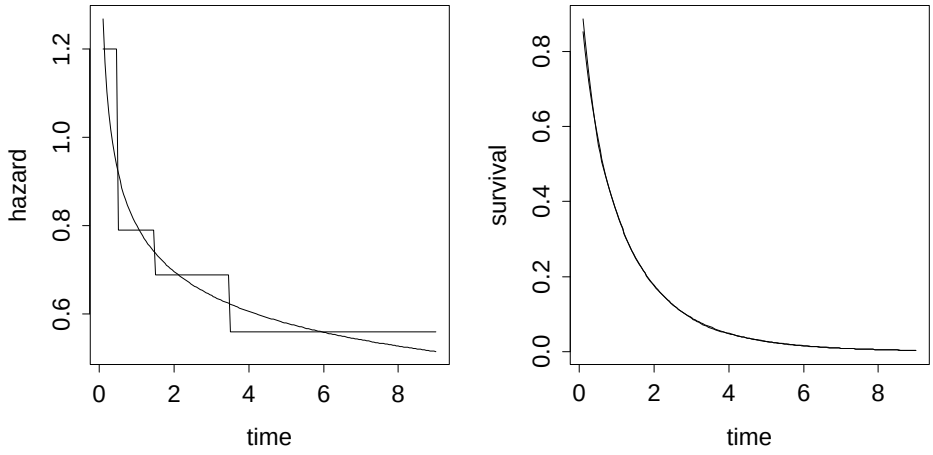

FIGURE 7.1: Approximating a Survival Curve Using a Piece-wise Constant Hazard Function

Figure 7.1 shows how a Weibull distribution with $\lambda = 1$ and $p = 0.8$ can be approximated using a piece-wise exponential distribution with boundaries at 0.5, 1.5 and 3.5. The left panel shows how the piece-wise constant hazard can follow only the broad outline of the smoothly declining Weibull hazard yet, as shown on the right panel, the corresponding survival curves are indistinguishable.

7.4.2 A Proportional Hazards Model

let us now introduce some covariates in the context of the proportional hazards model in Equation 7.13, assuming that the baseline hazard is piece-wise constant as in Equation 7.14. We will write the model as

$$\lambda_{ij} = \lambda_j \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}, \tag{7.15}$$

where λ_{ij} is the hazard corresponding to individual i in interval j, λ_j is the baseline hazard for interval j, and $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ is the relative risk for an individual with covariate values $\mathbf{x}_i$, compared to the baseline, at any given time.

Taking logs, we obtain the additive log-linear model

$$\log \lambda_{ij} = \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}, \tag{7.16}$$

where $\alpha_j = \log \lambda_j$ is the log of the baseline hazard. Note that the result is a standard log-linear model where the duration categories are treated as

a factor. Since we have not included an explicit constant, we do not have to impose restrictions on the α_j. If we wanted to introduce a constant representing the risk in the first interval then we would set $\alpha_1 = 0$, as usual.

The model can be extended to introduce time-varying covariates and time-dependent effects, but we will postpone discussing the details until we study estimation of the simpler proportional hazards model.

7.4.3 The Equivalent Poisson Model

Holford (1980) and Laird and Oliver (1981), in papers produced independently and published very close to each other, noted that the piece-wise proportional hazards model of the previous subsection was equivalent to a certain Poisson regression model. We first state the result and then sketch its proof.

Recall that we observe t_i, the total time lived by the i-th individual, and d_i, a death indicator that takes the value one if the individual died and zero otherwise. We will now define analogous measures for each interval that individual i goes through. You may think of this process as creating a bunch of pseudo-observations, one for each combination of individual and interval.

First we create measures of exposure. Let t_{ij} denote the time lived by the i-th individual in the j-th interval, that is, between τ_{j-1} and τ_j. If the individual lived beyond the end of the interval, so that $t_i > \tau_j$, then the time lived in the interval equals the width of the interval and $t_{ij} = \tau_j - \tau_{j-1}$. If the individual died or was censored in the interval, i.e. if $\tau_{j-1} < t_i < \tau_j$, then the timed lived in the interval is $t_{ij} = t_i - \tau_{j-1}$, the difference between the total time lived and the lower boundary of the interval. We only consider intervals actually visited, but obviously the time lived in an interval would be zero if the individual had died before the start of the interval and $t_i < \tau_{j-1}$.

Now we create death indicators. Let d_{ij} take the value one if individual i dies in interval j and zero otherwise. Let $j(i)$ indicate the interval where t_i falls, i.e. the interval where individual i died or was censored. We use functional notation to emphasize that this interval will vary from one individual to another. If t_i falls in interval $j(i)$, say, then d_{ij} must be zero for all $j < j(i)$ (i.e. all prior intervals) and will equal d_i for $j = j(i)$, (i.e. the interval where individual i was last seen).

Then, the piece-wise exponential model may be fitted to data by treating the death indicators d_{ij}'s as if they were independent Poisson observations with means

$$\mu_{ij} = t_{ij}\lambda_{ij},$$

where t_{ij} is the exposure time as defined above and λ_{ij} is the hazard for individual i in interval j. Taking logs in this expression, and recalling that the hazard rates satisfy the proportional hazards model in Equation 7.15, we obtain

$$\log \mu_{ij} = \log t_{ij} + \alpha_j + \mathbf{x}_i'\boldsymbol{\beta},$$

where $\alpha_j = \log \lambda_j$ as before.

Thus, the piece-wise exponential proportional hazards model is equivalent to a Poisson log-linear model for the pseudo observations, one for each combination of individual and interval, where the death indicator is the response and the log of exposure time enters as an offset.

It is important to note that we do not assume that the d_{ij} have independent Poisson distributions, because they clearly do not. If individual i died in interval $j(i)$, then it must have been alive in all prior intervals $j < j(i)$, so the indicators couldn't possibly be independent. Moreover, each indicator can only take the values one and zero, so it couldn't possibly have a Poisson distribution, which assigns some probability to values greater than one. The result is more subtle. It is the likelihood functions that coincide. Given a realization of a piece-wise exponential survival process, we can find a realization of a set of independent Poisson observations that happens to have the same likelihood, and therefore would lead to the same estimates and tests of hypotheses.

The proof is not hard. Recall from Section 7.2.2 that the contribution of the i-th individual to the log-likelihood function has the general form

$$\log L_i = d_i \log \lambda_i(t_i) - \Lambda_i(t_i),$$

where we have written $\lambda_i(t)$ for the hazard and $\Lambda_i(t)$ for the cumulative hazard that applies to the i-th individual at time t. Let $j(i)$ denote the interval where t_i falls, as before.

Under the piece-wise exponential model, the first term in the log-likelihood can be written as

$$d_i \log \lambda_i(t_i) = d_{ij(i)} \log \lambda_{ij(i)},$$

using the fact that the hazard is $\lambda_{ij(i)}$ when t_i is in interval $j(i)$, and that the death indicator d_i applies directly to the last interval visited by individual i, and therefore equals $d_{j(i)}$.

The cumulative hazard in the second term is an integral, and can be written as a sum as follows

$$\Lambda_i(t_i) = \int_0^{t_i} \lambda_i(t)dt = \sum_{j=1}^{j(i)} t_{ij}\lambda_{ij},$$

where t_{ij} is the amount of time spent by individual i in interval j. To see this point note that we need to integrate the hazard from 0 to t_i. We split this integral into a sum of integrals, one for each interval where the hazard is constant. If an individual lives through an interval, the contribution to the integral will be the hazard λ_{ij} multiplied by the width of the interval. If the individual dies or is censored in the interval, the contribution to the integral will be the hazard λ_{ij} multiplied by the time elapsed from the beginning of the interval to the death or censoring time, which is $t_i - \tau_{j-1}$. But this is precisely the definition of the exposure time t_{ij}.

One slight lack of symmetry in our results is that the hazard leads to *one* term on $d_{ij(i)} \log \lambda_{ij(i)}$, but the cumulative hazard leads to $j(i)$ terms, one for each interval from $j = 1$ to $j(i)$. However, we know that $d_{ij} = 0$ for all $j < j(i)$, so we can add terms on $d_{ij} \log \lambda_{ij}$ for all prior j's; as long as $d_{ij} = 0$ they will make no contribution to the log-likelihood. This trick allows us to write the contribution of the i-th individual to the log-likelihood as a sum of $j(i)$ contributions, one for each interval visited by the individual:

$$\log L_i = \sum_{j=1}^{j(i)} \{d_{ij} \log \lambda_{ij} - t_{ij}\lambda_{ij}\}.$$

The fact that the contribution of the individual to the log-likelihood is a *sum* of several terms (so the contribution to the likelihood is a product of several terms) means that we can treat each of the terms as representing an independent observation.

The final step is to identify the contribution of each pseudo-observation, and we note here that it agrees, except for a constant, with the likelihood one would obtain if d_{ij} had a Poisson distribution with mean $\mu_{ij} = t_{ij}\lambda_{ij}$. To see this point write the Poisson log-likelihood as

$$\log L_{ij} = d_{ij} \log \mu_{ij} - \mu_{ij} = d_{ij} \log(t_{ij}\lambda_{ij}) - t_{ij}\lambda_{ij}.$$

This expression agrees with the log-likelihood above except for the term $d_{ij} \log(t_{ij})$, but this is a constant depending on the data and not on the parameters, so it can be ignored from the point of view of estimation. This completes the proof.□

This result generalizes the observation made at the end of Section 7.2.2 noting the relationship between the likelihood for censored exponential data and the Poisson likelihood. The extension is that instead of having just one 'Poisson' death indicator for each individual, we have one for each interval visited by each individual.

Generating pseudo-observations can substantially increase the size of the dataset, perhaps to a point where analysis is impractical. Note, however, that the number of distinct covariate patterns may be modest even when the total number of pseudo-observations is large. In this case one can group observations, adding up the measures of exposure and the death indicators. In this more general setting, we can define d_{ij} as the number of deaths and t_{ij} as the total exposure time of individuals with characteristics $\mathbf{x}_i$ in interval j. As usual with Poisson aggregate models, the estimates, standard errors and likelihood ratio tests would be exactly the same as for individual data. Of course, the model deviances would be different, representing goodness of fit to the aggregate rather than individual data, but this may be a small price to pay for the convenience of working with a small number of units.

7.4.4 Time-varying Covariates

It should be obvious from the previous development that we can easily accommodate time-varying covariates provided they change values only at interval boundaries. In creating the pseudo-observations required to set-up a Poisson log-likelihood, one would normally replicate the vector of covariates $\mathbf{x}_i$, creating copies $\mathbf{x}_{ij}$, one for each interval. However, there is nothing in our development requiring these vectors to be equal. We can therefore redefine $\mathbf{x}_{ij}$ to represent the values of the covariates of individual i in interval j, and proceed as usual, rewriting the model as

$$\log \lambda_{ij} = \alpha_j + \mathbf{x}'_{ij}\boldsymbol{\beta}.$$

Requiring the covariates to change values only at interval boundaries may seem restrictive, but in practice the model is more flexible than it might seem at first, because we can always further split the pseudo observations. For example, if we wished to accommodate a change in a covariate for individual i half-way through interval j, we could split the pseudo-observation into two, one with the old and one with the new values of the covariates. Each half would get its own measure of exposure and its own death indicator, but both would be tagged as belonging to the same interval, so they would get the same baseline hazard. All steps in the above proof would still hold.

Of course, splitting observations further increases the size of the dataset, and there will usually be practical limitations on how far one can push this approach, even if one uses grouped data. An alternative is to use simpler indicators such as the mean value of a covariate in an interval, perhaps lagged to avoid predicting current hazards using future values of covariates.

7.4.5 Time-dependent Effects

It turns out that the piece-wise exponential scheme lends itself easily to the introduction of non-proportional hazards or time-varying effects, provided again that we let the effects vary only at interval boundaries.

To fix ideas, suppose we have a single predictor taking the value x_{ij} for individual i in interval j. Suppose further that this predictor is a dummy variable, so its possible values are one and zero. It doesn't matter for our current purpose whether the value is fixed for the individual or changes from one interval to the next.

In a proportional hazards model we would write

$$\log \lambda_{ij} = \alpha_j + \beta x_{ij},$$

where β represents the effect of the predictor on the log of the hazard at any given time. Exponentiating, we see that the hazard when $x = 1$ is $\exp\{\beta\}$ times the hazard when $x = 0$, and this effect is the same at all times. This is a simple additive model on duration and the predictor of interest.

To allow for a time-dependent effect of the predictor, we would write

$$\log \lambda_{ij} = \alpha_j + \beta_j x_{ij},$$

where β_j represents the effect of the predictor on the hazard during interval j. Exponentiating, we see that the hazard in interval j when $x = 1$ is $\exp\{\beta_j\}$ times the hazard in interval j when $x = 0$, so the effect may vary from one interval to the next. Since the effect of the predictor depends on the interval, we have a form of interaction between the predictor and duration, which might be more obvious if we wrote the model as

$$\log \lambda_{ij} = \alpha_j + \beta x_{ij} + (\alpha\beta)_j x_{ij}.$$

These models should remind you of the analysis of covariance models of Chapter 2. Here α plays the role of the intercept and β the role of the slope. The proportional hazards model has different intercepts and a common slope, so it's analogous to the parallel lines model. The model with a time-dependent effect has different intercepts and different slopes, and is analogous to the model with an interaction.

To sum up, we can accommodate non-proportionality of hazards simply by introducing interactions with duration. Obviously we can also test the assumption of proportionality of hazards by testing the significance of the interactions with duration. We are now ready for an example.

7.5 Infant and Child Mortality in Colombia

We will illustrate the use of piece-wise exponential survival models using data from an analysis of infant and child mortality in Colombia done by Somoza (1980). The data were collected in a 1976 survey conducted as part of the World Fertility Survey. The sample consisted of women between the ages of 15 and 49. The questionnaire included a maternity history, recording for each child ever born to each respondent the sex, date of birth, survival status as of the interview and (if applicable) age at death.

7.5.1 Calculating Events and Exposure

As if often the case with survival data, most of the work goes into preparing the data for analysis. In the present case we started from tables in Somoza's article showing living children classified by current age, and dead children classified by age at death. Both tabulations reported age using the groups shown in Table 7.1, using fine categories early in life, when the risk if high but declines rapidly, and wider categories at later ages. With these two bits of information we were able to tabulate deaths and calculate exposure time by age groups, assuming that children who died or were censored in an interval lived on the average half the length of the interval.

Table 7.1: Infant and Child Deaths and Exposure Time by Age of Child and Birth Cohort, Colombia 1976.

Exact Age	Birth Cohort 1941–59		1960–67		1968-76	
	deaths	exposure	deaths	exposure	deaths	exposure
0–1 m	168	278.4	197	403.2	195	495.3
1–3 m	48	538.8	48	786.0	55	956.7
3–6 m	63	794.4	62	1165.3	58	1381.4
6–12 m	89	1550.8	81	2294.8	85	2604.5
1–2 y	102	3006.0	97	4500.5	87	4618.5
2–5 y	81	8743.5	103	13201.5	70	9814.5
5–10 y	40	14270.0	39	19525.0	10	5802.5

Table 7.1 shows the results of these calculations in terms of the number of deaths and the total number of person-years of exposure to risk between birth and age ten, by categories of age of child, for three groups of children (or cohorts) born in 1941–59, 1960–67 and 1968–76. The purpose of our

analysis will be to assess the magnitude of the expected decline in infant and child mortality across these cohorts, and to study whether mortality has declined uniformly at all ages or more rapidly in certain age groups.

7.5.2 Fitting The Poisson Models

Let y_{ij} denote the number of deaths for cohort i (with $i = 1, 2, 3$) in age group j (for $j = 1, 2, \ldots, 7$). In view of the results of the previous section, we treat the y_{ij} as realizations of Poisson random variables with means μ_{ij} satisfying

$$\mu_{ij} = \lambda_{ij} t_{ij},$$

where λ_{ij} is the hazard rate and t_{ij} is the total exposure time for group i at age j. In words, the expected number of deaths is the product of the death rate by exposure time.

A word of caution about units of measurement: the hazard rates must be interpreted in the same units of time that we have used to measure exposure. In our example we measure time in years and therefore the λ_{ij} represent rates per person-year of exposure. If we had measured time in months the λ_{ij} would represent rates per person-month of exposure, and would be exactly one twelfth the size of the rates per person-year.

To model the rates we use a log link, so that the linear predictor becomes

$$\eta_{ij} = \log \mu_{ij} = \log \lambda_{ij} + \log t_{ij},$$

the sum of two parts, $\log t_{ij}$, an *offset* or known part of the linear predictor, and $\log \lambda_{ij}$, the log of the hazard rates of interest.

Finally, we introduce a log-linear model for the hazard rates, of the usual form

$$\log \lambda_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta},$$

where $\mathbf{x}_{ij}$ is a vector of covariates. In case you are wondering what happened to the baseline hazard, we have folded it into the vector of parameters $\boldsymbol{\beta}$. The vector of covariates $\mathbf{x}_{ij}$ may include a constant, a set of dummy variables representing the age groups (i.e. the shape of the hazard by age), a set of dummy variables representing the birth cohorts (i.e. the change in the hazard over time) and even a set of cross-product dummies representing combinations of ages and birth cohorts (i.e. interaction effects).

Table 7.2 shows the deviance for the five possible models of interest, including the null model, the two one-factor models, the two-factor additive model, and the two-factor model with an interaction, which is saturated for these data.

TABLE 7.2: Deviances for Various Models Fitted to Infant and Child Mortality Data From Colombia

Model	Name	$\log \lambda_{ij}$	Deviance	d.f.
ϕ	Null	η	4239.8	20
A	Age	$\eta + \alpha_i$	72.7	14
C	Cohort	$\eta + \beta_j$	4190.7	18
$A + C$	Additive	$\eta + \alpha_i + \beta_j$	6.2	12
AC	Saturated	$\eta + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	0	0

7.5.3 The Equivalent Survival Models

The null model assumes that the hazard is a constant from birth to age ten and that this constant is the same for all cohorts. It therefore corresponds to an *exponential survival model with no covariates*. This model obviously does not fit the data, the deviance of 4239.8 on 20 d.f. is simply astronomical. The m.l.e. of η is -3.996 with a standard error of 0.0237. Exponentiating we obtain an estimated hazard rate of 0.0184. Thus, we expect about 18 deaths per thousand person-years of exposure. You may want to verify that 0.0184 is simply the ratio of the total number of deaths to the total exposure time. Multiplying 0.0184 by the amount of exposure in each cell of the table we obtain the expected number of deaths. The deviance quoted above is simply twice the sum of observed times the log of observed over expected deaths.

The age model allows the hazard to change from one age group to another, but assumes that the risk at any given age is the same for all cohorts. It is therefore equivalent to *a piece-wise exponential survival model with no covariates*. The reduction in deviance from the null model is 4167.1 on 6 d.f., and is extremely significant. The risk of death varies substantially with age over the first few months of life. In other words the hazard is clearly not constant. Note that with a deviance of 72.7 on 14 d.f., this model does not fit the data. Thus, the assumption that all cohorts are subject to the same risks does not seem tenable.

Table 7.3 shows parameter estimates for the one-factor models A and C and for the additive model $A + C$ in a format reminiscent of multiple classification analysis. Although the A model does not fit the data, it is instructive to comment briefly on the estimates. The constant, shown in parentheses, corresponds to a rate of $\exp\{-0.7427\} = 0.4758$, or nearly half a death per person-year of exposure, in the first month of life. The estimate for ages 1–3 months corresponds to a multiplicative effect of $\exp\{-1.973\} =$

0.1391, amounting to an 86 percent reduction in the hazard after surviving the first month of life. This downward trend continues up to ages 5–10 years, where the multiplicative effect is $\exp\{-5.355\} = 0.0047$, indicating that the hazard at these ages is only half-a-percent what it was in the first month of life. You may wish to verify that the m.l.e.'s of the age effects can be calculated directly from the total number of deaths and the total exposure time in each age group. Can you calculate the deviance by hand?

Let us now consider the model involving only birth cohort, which assumes that the hazard is constant from birth to age ten, but varies from one birth cohort to another. This model is equivalent to *three exponential survival models*, one for each birth cohort. As we would expect, it is hopelessly inadequate, with a deviance in the thousands, because it fails to take into account the substantial age effects that we have just discussed. It may of of interest, however, to note the parameter estimates in Table 7.3. As a first approximation, the overall mortality rate for the older cohort was $\exp\{-3.899\} = 0.0203$ or around 20 deaths per thousand person-years of exposure. The multiplicative effect for the cohort born in 1960–67 is $\exp\{-0.3020\} = 0.7393$, indicating a 26 percent reduction in overall mortality. However, the multiplicative effect for the youngest cohort is $\exp\{0.0742\} = 1.077$, suggesting an eight percent *increase* in overall mortality. Can you think of an explanation for this apparent anomaly? We will consider the answer after we discuss the next model.

TABLE 7.3: Parameter Estimates for Age, Cohort and Age+Cohort Models of Infant and Child Mortality in Colombia

Factor	Category	Gross Effect	Net Effect
Baseline			−0.4485
Age	0–1 m	(−0.7427)	–
	1–3 m	−1.973	−1.973
	3–6 m	−2.162	−2.163
	6–12 m	−2.488	−2.492
	1–2 y	−3.004	−3.014
	2–5 y	−4.086	−4.115
	5–10 y	−5.355	−5.436
Cohort	1941–59	(−3.899)	–
	1960–67	−0.3020	−0.3243
	1968–76	0.0742	−0.4784

Consider now the additive model with effects of both age and cohort, where the hazard rate is allowed to vary with age and may differ from one cohort to another, but the age (or cohort) effect is assumed to be the same for each cohort (or age). This model is equivalent to a *proportional hazards model*, where we assume a common shape of the hazard by age, and let cohort affect the hazard proportionately at all ages. Comparing the proportional hazards model with the age model we note a reduction in deviance of 66.5 on two d.f., which is highly significant. Thus, we have strong evidence of cohort effects net of age. On the other hand, the attained deviance of 6.2 on 12 d.f. is clearly not significant, indicating that the proportional hazards model provides an adequate description of the patterns of mortality by age and cohort in Colombia. In other words, the assumption of proportionality of hazards is quite reasonable, implying that the decline in mortality in Colombia has been the same at all ages.

Let us examine the parameter estimates on the right-most column of Table 7.3. The constant is the baseline hazard at ages 0–1 months for the earliest cohort, those born in 1941–59. The age parameters representing the baseline hazard are practically unchanged from the model with age only, and trace the dramatic decline in mortality from birth to age ten, with half the reduction concentrated in the first year of life. The cohort affects adjusted for age provide a more reasonable picture of the decline in mortality over time. The multiplicative effects for the cohorts born in 1960–67 and 1068–76 are $\exp\{-0.3243\} = 0.7233$ and $\exp\{-0.4784\} = 0.6120$, corresponding to mortality declines of 28 and 38 percent at every age, compared to the cohort born in 1941–59. This is a remarkable decline in infant and child mortality, which appears to have been the same at all ages. In other words, neonatal, post-neonatal, infant and toddler mortality have all declined by approximately 38 percent across these cohorts.

The fact that the gross effect for the youngest cohort was positive but the net effect is substantially negative can be explained as follows. Because the survey took place in 1976, children born between 1968 and 76 have been exposed mostly to mortality at younger ages, where the rates are substantially higher than at older ages. For example a child born in 1975 would have been exposed only to mortality in the first year of life. The gross effect ignores this fact and thus overestimates the mortality of this group at ages zero to ten. The net effect adjusts correctly for the increased risk at younger ages, essentially comparing the mortality of this cohort to the mortality of earlier cohorts when they had the same ages, and can therefore unmask the actual decline.

A final caveat on interpretation: the data are based on retrospective re-

ports of mothers who were between the ages of 15 and 49 at the time of the interview. These women provide a representative sample of both mothers and births for recent periods, but a somewhat biased sample for older periods. The sample excludes mothers who have died before the interview, but also women who were older at the time of birth of the child. For example births from 1976, 1966 and 1956 come from mothers who were under 50, under 40 and under 30 at the time of birth of the child. A more careful analysis of the data would include age of mother at birth of the child as an additional control variable.

7.5.4 Estimating Survival Probabilities

So far we have focused attention on the hazard or mortality rate, but of course, once the hazard has been calculated it becomes an easy task to calculate cumulative hazards and therefore survival probabilities. Table 7.4 shows the results of just such an exercise, using the parameter estimates for the proportional hazards model in Table 7.3.

TABLE 7.4: Calculation of Survival Probabilities for Three Cohorts Based on the Proportional Hazards Model

Age Group	Width	Baseline			Survival for Cohort		
		Log-haz	Hazard	Cum.Haz	<1960	1960–67	1968–76
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0–1 m	1/12	−0.4485	0.6386	0.0532	0.9482	0.9623	0.9676
1–3 m	2/12	−2.4215	0.0888	0.0680	0.9342	0.9520	0.9587
3–6 m	3/12	−2.6115	0.0734	0.0864	0.9173	0.9395	0.9479
6–12 m	1/2	−2.9405	0.0528	0.1128	0.8933	0.9217	0.9325
1–2 y	1	−3.4625	0.0314	0.1441	0.8658	0.9010	0.9145
2–5 y	3	−4.5635	0.0104	0.1754	0.8391	0.8809	0.8970
5–10 y	5	−5.8845	0.0028	0.1893	0.8275	0.8721	0.8893

Consider first the baseline group, namely the cohort of children born before 1960. To obtain the log-hazard for each age group we must add the constant and the age effect, for example the log-hazard for ages 1–3 months is $-0.4485 - 1.973 = -2.4215$. This gives the numbers in column (3) of Table 7.3. Next we exponentiate to obtain the hazard rates in column (4), for example the rate for ages 1–3 months is $\exp\{-2.4215\} = 0.0888$. Next we calculate the cumulative hazard, multiply the hazard by the width of the interval and summing across intervals. In this step it is crucial to express the width of the interval in the same units used to calculate exposure, in

this case years. Thus, the cumulative hazard at then end of ages 1–3 months is $0.6386 \times 1/12 + 0.0888 \times 2/12 = 0.0680$. Finally, we change sign and exponentiate to calculate the survival function. For example the baseline survival function at 3 months is $\exp\{-0.0680\} = 0.9342$.

To calculate the survival functions shown in columns (7) and (8) for the other two cohorts we could multiply the baseline hazards by $\exp\{-0.3242\}$ and $\exp\{-0.4874\}$ to obtain the hazards for cohorts 1960–67 and 1968–76, respectively, and then repeat the steps described above to obtain the survival functions. This approach would be necessary if we had time-varying effects, but in the present case we can take advantage of a simplification that obtains for proportional hazard models. Namely, the survival functions for the two younger cohorts can be calculated as the baseline survival function *raised* to the relative risks $\exp\{-0.3242\}$ and $\exp\{-0.4874\}$, respectively. For example the probability of surviving to age three months was calculated as 0.9342 for the baseline group, and turns out to be $0.9342^{\exp\{-0.3242\}} = 0.9520$ for the cohort born in 1960–67, and $0.9342^{\exp\{-0.4874\}} = 0.9587$ for the cohort born in 1968–76.

Note that the probability of dying in the first year of life has declined from 106.7 per thousand for children born before 1960 to 78.3 per thousand for children born in 1960–67 and finally to 67.5 per thousand for the most recent cohort. Results presented in terms of probabilities are often more accessible to a wider audience than results presented in terms of hazard rates. (Unfortunately, demographers are used to calling the probability of dying in the first year of life the 'infant mortality rate'. This is incorrect because the quantity quoted is a probability, not a rate. In our example the rate varies substantially within the first year of life. If the probability of dying in the first year of life is q, say, then the average rate is approximately $-\log(1-q)$, which is not too different from q for small q.)

By focusing on events and exposure, we have been able to combine infant and child mortality in the same analysis and use all available information. An alternative approach could focus on infant mortality (deaths in the first year of life), and solve the censoring problem by looking only at children born at least one year before the survey, for whom the survival status at age one is know. One could then analyze the probability of surviving to age one using ordinary logit models. A complementary analysis could then look at survival from age one to five, say, working with children born at least five years before the survey who survived to age one, and then analyzing whether or not they further survive to age five, using again a logit model. While simple, this approach does not make full use of the information, relying on cases with complete (uncensored) data. Cox and Oakes (1980) show that

this so-called reduced sample approach can lead to inconsistencies. Another disadvantage of this approach is that it focuses on survival to key ages, but cannot examine the shape of the hazard in the intervening period.

7.6 Discrete Time Models

We discuss briefly two extensions of the proportional hazards model to discrete time, starting with a definition of the hazard and survival functions in discrete time and then proceeding to models based on the logit and the complementary log-log transformations.

7.6.1 Discrete Hazard and Survival

Let T be a discrete random variable that takes the values $t_1 < t_2 < \ldots$ with probabilities

$$f(t_j) = f_j = \Pr\{T = t_j\}.$$

We define the survivor function at time t_j as the probability that the survival time T is at least t_j

$$S(t_j) = S_j = \Pr\{T \geq t_j\} = \sum_{k=j}^{\infty} f_j.$$

Next, we define the hazard at time t_j as the conditional probability of dying at that time given that one has survived to that point, so that

$$\lambda(t_j) = \lambda_j = \Pr\{T = t_j | T \geq t_j\} = \frac{f_j}{S_j}. \tag{7.17}$$

Note that in discrete time the hazard is a conditional probability rather than a rate. However, the general result expressing the hazard as a ratio of the density to the survival function is still valid.

A further result of interest in discrete time is that the survival function at time t_j can be written in terms of the hazard at all prior times $t_1, \ldots, t_{j-1}$, as

$$S_j = (1 - \lambda_1)(1 - \lambda_2) \ldots (1 - \lambda_{j-1}). \tag{7.18}$$

In words, this result states that in order to survive to time t_j one must first survive t_1, then one must survive t_2 given that one survived t_1, and so on, finally surviving t_{j-1} given survival up to that point. This result is analogous to the result linking the survival function in continuous time to the integrated or cumulative hazard at all previous times.

An example of a survival process that takes place in discrete time is time to conception measured in menstrual cycles. In this case the possible values of T are the positive integers, f_j is the probability of conceiving in the j-th cycle, S_j is the probability of conceiving in the j-th cycle or later, and λ_j is the conditional probability of conceiving in the j-th cycle given that conception had not occurred earlier. The result relating the survival function to the hazard states that in order to get to the j-th cycle without conceiving, one has to fail in the first cycle, then fail in the second given that one didn't succeed in the first, and so on, finally failing in the $(j-1)$-st cycle given that one hadn't succeeded yet.

7.6.2 Discrete Survival and Logistic Regression

Cox (1972) proposed an extension of the proportional hazards model to discrete time by working with the conditional odds of dying at each time t_j given survival up to that point. Specifically, he proposed the model

$$\frac{\lambda(t_j|\mathbf{x}_i)}{1-\lambda(t_j|\mathbf{x}_i)} = \frac{\lambda_0(t_j)}{1-\lambda_0(t_j)} \exp\{\mathbf{x}_i'\boldsymbol{\beta}\},$$

where $\lambda(t_j|\mathbf{x}_i)$ is the hazard at time t_j for an individual with covariate values $\mathbf{x}_i$, $\lambda_0(t_j)$ is the baseline hazard at time t_j, and $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ is the relative risk associated with covariate values $\mathbf{x}_i$.

Taking logs, we obtain a model on the *logit* of the hazard or conditional probability of dying at t_j given survival up to that time,

$$\text{logit}\lambda(t_j|\mathbf{x}_i) = \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}, \tag{7.19}$$

where $\alpha_j = \text{logit}\lambda_0(t_j)$ is the logit of the baseline hazard and $\mathbf{x}_i'\boldsymbol{\beta}$ is the effect of the covariates on the logit of the hazard. Note that the model essentially treats time as a discrete factor by introducing one parameter α_j for each possible time of death t_j. Interpretation of the parameters $\boldsymbol{\beta}$ associated with the other covariates follows along the same lines as in logistic regression.

In fact, the analogy with logistic regression goes further: we can fit the discrete-time proportional-hazards model by running a logistic regression on a set of pseudo observations generated as follows. Suppose individual i dies or is censored at time point $t_{j(i)}$. We generate death indicators d_{ij} that take the value one if individual i died at time j and zero otherwise, generating one for each discrete time from t_1 to $t_{j(i)}$. To each of these indicators we associate a copy of the covariate vector $\mathbf{x}_i$ and a label j identifying the time point. The proportional hazards model 7.19 can then be fit by treating

the d_{ij} as independent Bernoulli observations with probability given by the hazard λ_{ij} for individual i at time point t_j.

More generally, we can group pseudo-observations with identical covariate values. Let d_{ij} denote the number of deaths and n_{ij} the total number of individuals with covariate values $\mathbf{x}_i$ observed at time point t_j. Then we can treat d_{ij} as binomial with parameters n_{ij} and λ_{ij}, where the latter satisfies the proportional hazards model.

The proof of this result runs along the same lines as the proof of the equivalence of the Poisson likelihood and the likelihood for piece-wise exponential survival data under non-informative censoring in Section 7.4.3, and relies on Equation 7.18, which writes the probability of surviving to time t_j as a product of the conditional hazards at all previous times. It is important to note that we do not assume that the pseudo-observations are independent and have a Bernoulli or binomial distribution. Rather, we note that the likelihood function for the discrete-time survival model under non-informative censoring coincides with the binomial likelihood that would be obtained by treating the death indicators as independent Bernoulli or binomial.

Time-varying covariates and time-dependent effects can be introduced in this model along the same lines as before. In the case of time-varying covariates, note that only the values of the covariates at the discrete times $t_1 < t_2 < \ldots$ are relevant. Time-dependent effects are introduced as interactions between the covariates and the discrete factor (or set of dummy variables) representing time.

7.6.3 Discrete Survival and the C-Log-Log Link

An alternative extension of the proportional hazards model to discrete time starts from the survival function, which in a proportional hazards framework can be written as

$$S(t_j|\mathbf{x}_i) = S_0(t_j)^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}},$$

where $S(t_j|\mathbf{x}_i)$ is the probability that an individual with covariate values $\mathbf{x}_i$ will survive up to time point t_j, and $S_0(t_j)$ is the baseline survival function. Recalling Equation 7.18 for the discrete survival function, we obtain a similar relationship for the complement of the hazard function, namely

$$1 - \lambda(t_j|\mathbf{x}_i) = [1 - \lambda_0(t_j)]^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}},$$

so that solving for the hazard for individual i at time point t_j we obtain the model

$$\lambda(t_j|\mathbf{x}_i) = 1 - [1 - \lambda_0(t_j)]^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}.$$

The transformation that makes the right hand side a linear function of the parameters is the complementary log-log. Applying this transformation we obtain the model

$$\log(-\log(1-\lambda(t_j|\mathbf{x}_i))) = \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}, \tag{7.20}$$

where $\alpha_j = \log(-\log(1-\lambda_0(t_j)))$ is the complementary log-log transformation of the baseline hazard.

This model can be fitted to discrete survival data by generating pseudo-observations as before and fitting a generalized linear model with binomial error structure and complementary log-log link. In other words, the equivalence between the binomial likelihood and the discrete-time survival likelihood under non-informative censoring holds both for the logit and complementary log-log links.

It is interesting to note that this model can be obtained by grouping time in the continuous-time proportional-hazards model. To see this point let us assume that time is continuous and we are really interested in the standard proportional hazards model

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}.$$

Suppose, however, that time is grouped into intervals with boundaries $0 = \tau_0 < \tau_1 < \ldots < \tau_J = \infty$, and that all we observe is whether an individual survives or dies in an interval. Note that this construction imposes some constraints on censoring. If an individual is censored at some point inside an interval, we do not know whether it would have survived the interval or not. Therefore we must censor it at the end of the previous interval, which is the last point for which we have complete information. Unlike the piecewise exponential set-up, here we can not use information about exposure to part of an interval. On the other hand, it turns out that we do not need to assume that the hazard is constant in each interval.

Let λ_{ij} denote the discrete hazard or conditional probability that individual i will die in interval j given that it was alive at the start of the interval. This probability is the same as the complement of the conditional probability of surviving the interval given that one was alive at the start, and can be written as

$$\begin{aligned}
\lambda_{ij} &= 1 - \Pr\{T_i > \tau_j | T_i > \tau_{j-1}\} \\
&= 1 - \exp\{-\int_{\tau_{j-1}}^{\tau_j} \lambda(t|\mathbf{x}_i)dt\} \\
&= 1 - \exp\{-\int_{\tau_{j-1}}^{\tau_j} \lambda_0(t)dt\}^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}
\end{aligned}$$

$$= \quad 1-(1-\lambda_j)^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}},$$

where λ_j is the baseline probability of dying in interval j given survival to the start of the interval. The second line follows from Equation 7.4 relating the survival function to the integrated hazard, the third line follows from the proportional hazards assumption, and the last line defines λ_j.

As noted by Kalbfleish and Prentice (1980, p. 37), "this discrete model is then the uniquely appropriate one for grouped data from the continuous proportional hazards model". In practice, however, the model with a logit link is used much more often than the model with a c-log-log link, probably because logistic regression is better known that generalized linear models with c-log-log links, and because software for the former is more widely available than for the latter. In fact, the logit model is often used in cases where the piece-wise exponential model would be more appropriate, probably because logistic regression is better known than Poisson regression.

In closing, it may be useful to provide some suggestions regarding the choice of approach to survival analysis using generalized linear models:

- If time is truly discrete, then one should probably use the discrete model with a logit link, which has a direct interpretation in terms of conditional odds, and is easily implemented using standard software for logistic regression.

- If time is continuous but one only observes it in grouped form, then the complementary log-log link would seem more appropriate. In particular, results based on the c-log-log link should be more robust to the choice of categories than results based on the logit link. However, one cannot take into account partial exposure in a discrete time context, no matter which link is used.

- If time is continuous and one is willing to assume that the hazard is constant in each interval, then the piecewise exponential approach based on the Poisson likelihood is preferable. This approach is reasonably robust to the choice of categories and is unique in allowing the use of information from cases that have partial exposure.

Finally, if time is truly continuous and one wishes to estimate the effects of the covariates without making any assumptions about the baseline hazard, then Cox's (1972) partial likelihood is a very attractive approach.