CHAPTER 13

Models for survival data

13.1 Introduction

This chapter deals with models for the analysis of data in which
the response variate is the lifetime of a component or the survival
time of a patient. Survival data usually refers to medical trials, but
the ideas are useful also in industrial reliability experiments where
the emphasis is on failure times rather than survival times.

Survival data are distinguished from most other types by the
widespread occurrence of *censoring*. Censoring occurs when the
outcome of a particular unit (patient or component) is unknown
at the end of the study. Thus we may know only that a particular
patient was still alive six months into the study, but the exact
failure time is unknown either because the patient withdrew from
the study or because the study ended while the patient was
still alive. Censoring is so common in medical experiments that
estimation methods must allow for it if they are to be generally
useful.

A second characteristic of survival data is the frequent occur-
rence of *time-dependent covariates*. These arise when the status of
a subject changes during a trial. Such a covariate x, say, cannot
be represented by a single value x_i for patient i, but takes values
that may change with time.

13.1.1 *Survival functions and hazard functions*

Let the survival time, T, for individuals in a population have a
density function $f(t)$. (In practice $f(\cdot)$ usually depends on other
parameters, but for the moment we omit reference to these.) The
corresponding distribution function

$$F(t) = \int_{-\infty}^{t} f(s)\,ds$$

419

is the fraction of the population dying by time t. The complementary function $1 - F(t)$, often called the *survivor function*, is the fraction still surviving at time t. The *hazard function* $h(t)$ measures the instantaneous risk, in that $h(t)\,\delta t$ is the probability of dying in the next small interval δt given survival to time t. From the relation

$$\text{pr(survival to } t + \delta t)$$
$$= \text{pr(survival to } t)\,\text{pr(survival for } \delta t \,|\, \text{survival to } t)$$

we have

$$1 - F(t + \delta t) = \{1 - F(t)\}\{1 - h(t)\,\delta t\},$$

whence

$$\delta t\, F'(t) = \{1 - F(t)\}\, h(t)\,\delta t,$$

so that the hazard function is given by

$$h(t) = f(t)/\{1 - F(t)\}.$$

A distribution for survival times must have a hazard function with suitable properties. Thus for large t a hazard function should not decrease, because beyond a certain point the chance of breakdown or death does not ordinarily decrease with time. For small t various forms can be justified, including one that initially declines with t, for such a distribution could describe the behaviour of a machine part with a settling-in period, where reliability increases once the initial period is over.

The simplest hazard function, a constant, implies an exponential distribution of survival times and hence a Poisson process. For if T has the density

$$f(t) = \lambda e^{-\lambda t}; \qquad t \geq 0,$$

then

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

and so

$$h(t) = \lambda.$$

Other forms of hazard function appear in later sections.

13.2 Proportional-hazards models

The hazard function depends in general both on time and on a set of covariates, some of which may be time-dependent. The proportional-hazards model separates these components by specifying that the hazard at time t for an individual whose covariate vector is $\mathbf{x}$ is given by

$$h(t; \mathbf{x}) = \lambda(t) \exp\{G(\mathbf{x}; \boldsymbol{\beta})\},$$

where the second term is written in exponential form because it must be positive. This model implies that the ratio of the hazards for two individuals is constant over time provided that their covariates do not change. It is conventional, but not necessary (Oakes, 1981) to assume that the effects of the covariates on the hazard are also multiplicative. This additional assumption leads to models that may be written in the form

$$h(t; \mathbf{x}) = \lambda(t) \exp(\boldsymbol{\beta}^T \mathbf{x}), \qquad (13.1)$$

where $\boldsymbol{\eta} = \boldsymbol{\beta}^T \mathbf{x}$ is the linear predictor. The model thus implies that the ratio of hazards for two individuals depends on the difference between their linear predictors at any time, and so, with no time-dependent covariates, is a constant independent of time. This is a strong assumption that clearly needs checking in applications. Various assumptions may be made about the $\lambda(t)$ function. If a continuous survival distribution is assumed, $\lambda(t)$ is a smooth function of t, defined for all $t \geq 0$. Cox's model (Cox, 1972a) treats $\lambda(t)$ as analogous to the block factor in a blocked experiment, defined only at points where deaths occur, thus making no assumptions about the trend with time. In practice it frequently makes surprisingly little difference to estimates and inferences whether we put a structure on the base-line hazard function $\lambda(t)$ or not. We consider first estimation of $\boldsymbol{\beta}$ in the linear predictor with an explicit survival distribution, following closely the development in Aitkin and Clayton (1980).

13.3   Estimation with a specified survival distribution

We begin with the proportional-hazards model (13.1), and develop the likelihood for the data, some of which may be censored; for this we need both the density function and the survivor function.

From the definition of the hazard function we have

$$h(t) = F'(t)/\{1 - F(t)\} = \lambda(t)e^{\eta},$$

so that

$$-\log\{1 - F(t)\} = \Lambda(t)e^{\eta},$$

where

$$\Lambda(t) = \int_{-\infty}^{t} \lambda(u)\,du$$

is known as the cumulative hazard function. Thus the survivor function is given by

$$S(t) = 1 - F(t) = \exp\{-\Lambda(t)e^{\eta}\},$$

and the density function by minus its derivative, i.e.

$$f(t) = \lambda(t)\exp\{\eta - \Lambda(t)e^{\eta}\}.$$

At the end of the study an individual who died at time t contributes a factor $f(t)$ to the likelihood, while one censored at time t contributes $S(t)$. Suppose now that we define w as a variate taking the value 1 for an uncensored observation and value 0 for a censored one, and let there be n uncensored and m censored observations. Then the log likelihood takes the form

$$
\begin{aligned}
l &= \sum_{i=1}^{n+m}\{w_i \log f(t_i) + (1 - w_i)\log S(t_i)\}\\
&= \sum_i \{w_i\{\log\lambda(t_i) + \eta_i\} - \Lambda(t_i)e^{\eta_i}\}\\
&= \sum_i \left\{ w_i\{\log\Lambda(t_i) + \eta_i\} - \Lambda(t_i)e^{\eta_i} + w_i \log\left(\frac{\lambda(t_i)}{\Lambda(t_i)}\right) \right\}.
\end{aligned}
$$

Now if we write $\mu_i = \Lambda(t_i)e^{\eta_i}$, l becomes

$$\sum_i (w_i \log\mu_i - \mu_i) + \sum_i w_i \log\left(\frac{\lambda(t_i)}{\Lambda(t_i)}\right).$$

The first term is identical to the kernel of the likelihood function for $(n + m)$ independent Poisson variates w_i with means μ_i, while the second term does not depend on the unknown βs. Thus, given $\Lambda(t)$, we can obtain estimates of the βs by treating the censoring indicator variate w_i as Poisson distributed with mean $\mu_i = \Lambda(t_i)e^{\eta_i}$. The link function is the same as for log-linear models except that there is a fixed intercept $\log \Lambda(t_i)$ to be included in the linear predictor. Such a quantity is known in GLIM terminology as an *offset*.

The estimation process is less straightforward if the offset contains parameters of the survival density whose values are not known in advance. First, however, we deal with the exponential distribution for which no such difficulties arise.

13.3.1 *The exponential distribution*

For this distribution $\lambda(t)$ is the constant λ, so that the cumulative hazard function is

$$\Lambda(t) = \int_0^t \lambda(s)\, ds = \lambda t.$$

Thus $\lambda(t)/\Lambda(t) = 1/t$ and no extra parameters are involved. It follows that

$$\log \mu_i = \log t_i + \eta_i,$$

so that the offset is just $\log t_i$ and the log-linear model can be fitted directly. Two other distributions give particularly simple forms for $\Lambda(t)$ and these we now consider.

13.3.2 *The Weibull distribution*

By setting $\Lambda(t) = t^\alpha$, $\alpha > 0$, we obtain a hazard function proportional to $\alpha t^{\alpha-1}$ and a corresponding density $f(t)$ of the Weibull form:

$$f(t) = \alpha t^{\alpha-1}\{\exp(\eta - t^\alpha e^\eta)\}; \qquad t \geq 0.$$

Now $\lambda(t)/\Lambda(t) = \alpha/t$ and depends on the unknown parameter α, which must be jointly estimated with the βs. The kernel of the log-likelihood function is

$$n \log \alpha + \sum_i (w_i \log \mu_i - \mu_i),$$

Table 13.1   *Times of remission (weeks) of leukaemia patients, treated with drug (sample 1) and placebo (sample 2)*

Sample 1	(6)	6	6	6	7	(9)	(10)
	10	(11)	13	16	(17)	(19)	(20)
	22	23	(25)	(32)	(32)	(34)	(35)
Sample 2	1	1	2	2	3	4	4
	5	5	8	8	8	8	11
	11	12	12	15	17	22	23

Data from Freireich *et al.* (1963).

Figures in parentheses denote censored observations.

and, given α, the likelihood equations for the βs are the same as those for a log-linear model with offset $\alpha \log t_i$. The equation for α given the βs takes the form

$$n/\hat{\alpha} = \sum_i (\hat{\mu}_i - w_i) \log t_i. \tag{13.2}$$

The estimation procedure begins with $\alpha = 1$ (the exponential distribution), uses the log-linear model algorithm to fit the βs, then estimates α from (13.2), oscillating between the two stages until convergence is attained.

Note that the log likelihood for the model differs from that of the log-linear model by the inclusion of the extra term $n \log \hat{\alpha}$, so that the deviance requires adjustment by a term $-2n \log \hat{\alpha}$.

13.3.3   *The extreme-value distribution*

For this distribution $\Lambda(t) = e^{\alpha t}$, giving a hazard function proportional to $\alpha e^{\alpha t}$ and a density $f(t)$ in the extreme-value form

$$f(t) = \alpha e^{\alpha t} \exp(\eta - e^{\alpha t + \eta}). \tag{13.3}$$

Note that the transformation $u = \exp(t)$ transforms the distribution to the Weibull form. It follows that we need only replace t by u in the estimating procedure for the Weibull to obtain the corresponding one for this distribution.

13.4 Example: remission times for leukaemia

The data in Table 13.1 from Freireich *et al.* (1963) have been analysed by Gehan (1965), Aitkin and Clayton (1980) and others. There are two samples of 21 patients each, sample 1 having been given an experimental drug and sample 2 a placebo. The times of remission are given in weeks and figures in parentheses denote censored observations. To fit a survival model of the type discussed in section 13.3, we set up a pseudo-Poisson variable taking values 0 for the censored and 1 for the uncensored observations. To fit the exponential distribution we apply an offset of $\log t$, and models with a single mean and with separate sample means (S) give deviances as follows:

Model	Deviance	d.f.
1	54.50	41
S	38.02	40

The more general Weibull distribution yields for model S the value $\hat{\alpha} = 1.366$ with a deviance of 34.13 on 39 d.f. The reduction in deviance of 3.89 is thus marginally significant at the 5% level. Separate fits of α to the two samples give similar estimates of 1.35 and 1.37; the estimate of the sample difference for $\hat{\alpha} = 1.366$ is $b_1 = 1.731$, corresponding to a hazard ratio of $\exp(1.731) = 5.65$ for sample 2 as compared with sample 1. The standard error of b_1, for α with a fixed prior value of 1.366, is ± 0.398; to adjust this for the simultaneous fitting of α, we must border the information matrix for the two parameters in the linear predictor with the second derivatives that include α and then invert the expanded matrix. The details for this example are given in Aitkin and Clayton (1980) and give $b_1 = 1.73 \pm 0.41$; the validity of this SE may be checked by plotting the deviance for fixed values of β_1 in the neighbourhood of the estimate 1.73. The resulting curve rises slightly more steeply on the lower than on the upper side, but the effect at the $2 \times$ SE distance is quite small. Thus our 95% limits for the log hazard difference are $1.73 \pm (1.96 \times 0.41) = (0.93, 2.53)$ corresponding to hazard ratios of $(2.52, 12.6)$.

The data of this example have been analysed by Whitehead (1980) using Cox's model and treating ties by both Peto's and Cox's

Table 13.2 *Comparison of estimators for the leukaemia data*

Model (treatment of ties)	b_1	SE
Exponential	1.53	0.40
Weibull	1.73	0.41
Cox (Peto)	1.51	0.41
Cox (Cox)	1.63	0.43

methods. His results (after correcting observation 6 in sample 1, which was censored), together with those obtained above using parametric survival functions, are summarized in Table 13.2. The estimates all fall within a range of about half a standard error, and the increase in standard error from the Cox model as against the parametric survival functions is quite small. Efron (1977) and Oakes (1977) discuss this phenomenon from a theoretical viewpoint.

13.5 Cox's proportional-hazards model

Cox's (1972a) version of the proportional-hazards model is only partially parametric in the sense that the baseline hazard function $\lambda(t)$ is not modelled as a smooth function of t. Instead, $\lambda(t)$ is permitted to take arbitrary values and is irrelevant in the sense that it does not enter into the estimating equations derived from Cox's partial likelihood (Cox, 1975).

13.5.1 *Partial likelihood*

The argument used to derive the partial likelihood function is as follows. First observe that we need only consider times at which failures occur because, in principle at least, the hazard could be zero over intervals that are free of failures and no contribution to the likelihood would be made by these intervals. Let $t_1 < t_2 < \ldots$ be the distinct failure times and suppose for simplicity that there are no tied failure times. The risk set immediately prior to the jth failure, $R(t_j)$, is the set of individuals any of whom may be found to fail at time t_j. Thus, individuals who have previously failed or who have been censored are excluded from $R(t_j)$. Given that one failure is to occur in the interval $(t_j - \delta t, t_j)$, the relative probabilities of failure for the individuals in $R(t_j)$ are proportional

to the values of their hazard functions. Let $\mathbf{x}_j$ be the value of the covariate vector for the failed individual. The probability under the proportional-hazards model that the individual who fails at time t_j is the one actually observed is

$$\frac{\lambda(t)\exp(\boldsymbol{\beta}^T\mathbf{x}_j)}{\Sigma\lambda(t)\exp(\boldsymbol{\beta}^T\mathbf{x})} = \frac{\exp(\boldsymbol{\beta}^T\mathbf{x}_j)}{\sum\exp(\boldsymbol{\beta}^T\mathbf{x})}, \qquad (13.4)$$

where summation extends over the risk set $R(t_j)$.

This conditional probability is the probability of observing $\mathbf{x}_j$ in sampling from the finite population corresponding to the covariate vectors in $R(t_j)$, where the selection probabilities are proportional to $\exp(\boldsymbol{\beta}^T\mathbf{x})$. This is a generalization of the non-central hypergeometric distribution (section 7.3.2). This argument effectively reverses the roles of random failure times and fixed covariates to fixed failure times and covariates selected according to the probability distribution described above.

The partial likelihood for $\boldsymbol{\beta}$ is the product over the failure times of the conditional probabilities (13.4), and so independent of the baseline hazard function $\lambda(t)$. These conditional probabilities have the form of a linear exponential-family model so that $\boldsymbol{\beta}$ can be estimated by equating the vector sum of the covariates of the failed individuals to the sum of their conditional means. Note, however, that the conditioning event changes from one failure time to the next as individuals are removed from the risk set either through failure or through censoring.

13.5.2 *The treatment of ties*

The occurrence of ties among the failure times complicates the analysis, and several techniques have been proposed for dealing with this complication. One method due to Cox (1972a) is as follows. Suppose for definiteness that two failures occur at time t and that the vector sum of the covariates of these two failed individuals is $\mathbf{s}_j$. The factor corresponding to (13.4) is then defined to be

$$\exp(\boldsymbol{\beta}^T\mathbf{s}_j)\Big/\sum\exp(\boldsymbol{\beta}^T\mathbf{s}), \qquad (13.5)$$

where the sum in the denominator extends over all distinct pairs of individuals in $R(t_j)$. In other words we construct the finite

population consisting of sums of the covariate vectors for all distinct pairs of individuals in the risk set at time t_j. The probability under an exponentially weighted sampling scheme that the failures were those of the pair actually observed is given by (13.5), which again has the exponential-family form. Note however that the number of terms in the denominator of (13.5) quickly becomes exceedingly large for even a moderate number of ties at any failure time.

Any reasonable method for dealing with ties is likely to be satisfactory if the number of failed individuals constitutes only a small fraction of the risk set. In fact the likelihood contribution (13.5) is exact only if failures are thought of as occurring in discrete time. In practice, however, ties occur principally because of grouping. With grouped data the appropriate likelihood (Peto, 1972) involves the sum over all permutations of the failed individuals consistent with the ties observed. Suppose, for example, that two failures are tied and that the failed individuals have covariate vectors x_1 and x_2. The probability for the sequence in time (x_1, x_2) or (x_2, x_1), either of which is possible given the tie, is

$$\frac{\exp(\boldsymbol{\beta}^T x_1)}{\sum_R \exp(\boldsymbol{\beta}^T x)} \frac{\exp(\boldsymbol{\beta}^T x_2)}{\sum_{R_1} \exp(\boldsymbol{\beta}^T x)} + \frac{\exp(\boldsymbol{\beta}^T x_2)}{\sum_R \exp(\boldsymbol{\beta}^T x)} \frac{\exp(\boldsymbol{\beta}^T x_1)}{\sum_{R_2} \exp(\boldsymbol{\beta}^T x)},$$

$$(13.6)$$

where R_j is the risk set excluding $x_j (j = 1, 2)$. Clearly the likelihood contribution becomes increasingly cumbersome as the number of ties becomes appreciable.

Expressions (13.5) and (13.6) for the contribution to the likelihood can both be derived by arguments involving exponentially weighted sampling from a finite population without replacement. If the number of ties is small we may use the simpler expression

$$\frac{\exp(\boldsymbol{\beta}^T s)}{\left\{\sum_R \exp(\boldsymbol{\beta}^T x)\right\}^m}, \qquad (13.7)$$

where s is the sum of the covariate vectors of the m tied individuals (Peto, 1972). This term corresponds to sampling with replacement.

13.5.3 *Numerical methods*

The likelihood formed by taking the product over failure times of the conditional probabilities (13.4) can, in principle, be maximized directly using the weighted least-squares method discussed in Chapters 2 and 8. Alternatively we can regard the covariate vector of the failed individuals as the response and condition on the set of covariates of all individuals in the risk set at each failure time, these being regarded as fixed. If we write $\mathbf{y}$ for the covariate vector of the failed individual the log likelihood for one failure time takes the form

$$\boldsymbol{\beta}^T \mathbf{y} - \log\left\{\sum \exp(\boldsymbol{\beta}^T \mathbf{x})\right\},$$

with summation over the risk set. This has the form of an exponential family model with canonical parameter $\boldsymbol{\beta}$ and $b(\theta)$ (in the notation of section 2.2) equal to $\log\{\sum \exp(\boldsymbol{\beta}^T \mathbf{x})\}$. The (conditional) mean is then given by $b'(\theta)$ and the variance by $b''(\theta)$. However, this formulation is unhelpful computationally because there is no explicit expression for the quadratic weight (here equal to the variance function) as a function of the mean.

The computational difficulty can be avoided by a device similar to that used in section 13.4. Suppose that k_j individuals are at risk immediately prior to t_j and that just one individual is about to fail. If we regard the observation on the failed individual as a multinomial observation with k_j categories, taking the value 1 for the failed observation and 0 for the remainder, then the contribution to the likelihood is again of the form (13.4), but now interpreted as a log-linear model for the cell probabilities. Thus the numerical methods of Chapter 5 may be used provided that the algorithm allows variable numbers of categories for the multinomial observations.

Alternatively (Whitehead, 1980) a Poisson log likelihood may be used provided that a blocking factor associated with failure times is included. The idea here is that at each failure time each individual in the risk set contributes an artificial Poisson response of 1 for failure and 0 for survival. The mean of this response is $\exp(\alpha + \boldsymbol{\beta}^T \mathbf{x})$ for an individual whose covariate value is $\mathbf{x}$ and α represents the blocking factor associated with failure times. Because of the equivalence of the Poisson and multinomial likelihoods discussed in section 6.4, the estimate of $\boldsymbol{\beta}$ and the estimate of its precision are identical to those obtained from the multinomial likelihood and

hence to the partial likelihood.

The computations can be simplified if the number of distinct covariate vectors is small so that individuals in the risk set may be grouped into sets of constant hazard. The adjustment for ties is simple for the third method described above (often called Peto's method). In the multinomial log likelihood we set the multinomial total equal to the observed number of tied failures at that time. No adjustment to the algorithm is required. The corresponding Poisson log likelihood is equivalent to Peto's version of the partial likelihood.

Whitehead (1980) describes the adjustments to the Poisson likelihood required to maximize the likelihood corresponding to Cox's method for dealing with ties.

13.6   Bibliographic notes

The recent literature on the analysis of survival data includes books by Cox and Oakes (1984), Elandt-Johnson and Johnson (1980), Gross and Clark (1975), Lawless (1982), Lee (1980), Kalbfleisch and Prentice (1980) and Miller (1981).

Cox's model was proposed by Cox (1972a), and fitting via GLIM discussed by Whitehead (1980); the pseudo-Poisson model for parametric survival functions was proposed by Aitkin and Clayton (1980), who also discuss the definition of residuals and the necessary adaptation of standard graphical techniques (see also Crowley and Hu 1977). For a comparison of Cox and Weibull models, see Byar (1983).

13.7   Further results and exercises 13

13.1   In medical trials the recruitment of patients frequently continues over a prolonged period, spanning perhaps the entire trial. Consider such a trial to test a new drug that is claimed to benefit patients suffering from angina by reducing the incidence of coronary disease. The protocol specifies eligible patients to be those aged 55–75, showing symptoms of angina who have no previous record of heart attack and are taking no other medication. After being judged eligible and consent has been obtained, a patient

is randomized to one of two groups, either the new drug or the standard treatment.

Discuss how you might analyse the data that have accumulated after two years in such a trial. Consider in particular the following points.

1. What are appropriate definitions of failure:
   deaths from all causes;
   deaths from coronary disease only;
   all heart attacks whether fatal or not.
2. Choice of origin for the time scale:
   calendar time from the beginning of the study;
   time from individual patient randomization;
   time from first appearance of patient's angina symptoms.
3. Non-compliance because of non-fatal side-effects:
4. Who to include in the risk set:
   all known survivors among those randomized;
   all survivors excluding those no longer complying.

13.2 Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be an ordered sample of $i.i.d.$ exponential random variables of unit mean. Define the normalized differences

$$Y_1 = nX_{(1)}, \qquad Y_i = (n - i + 1)(X_{(i)} - X_{(i-1)}), \quad i = 2, \ldots, n.$$

Show that $Y_1, \ldots, Y_n$ are $i.i.d.$ exponential random variables of unit mean.