MAS3301 / MAS8311 Biostatistics
Part II: Survival

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 2, 2009-10

13 The Cox proportional hazards model

13.1 Introduction

In the Weibull proportional hazards model (Section 12.1) we assumed

- that the hazard function $h_i(t) = \phi_i h_0(t)$ where $h_0(t)$ is the baseline hazard and ϕ_i depends on the covariate values for individual i but not on t, i.e. that we have proportional hazards.
- that the baseline hazard $h_0(t) = \lambda\gamma t^{\gamma-1}$, i.e. that we have a Weibull distribution (given any particular set of covariate values).

Now we will consider a model which retains the proportional hazards assumption but makes no assumption about the form of the baseline hazard. Because we are not assuming a particular form for the lifetime distribution but we are assuming that

$$h_i(t) = \phi_i h_0(t)$$

where

$$\phi_i = \exp(\eta_i) \quad \text{and} \quad \eta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$

this is described as a *semi-parametric* model.

The model was first introduced by Cox (1972) and has come to be known as the 'Cox regression model.' (It's also sometimes described simply as 'proportional hazards regression.') It's probably the most widespread model used for survival analysis.

13.2 Likelihood with no censoring

Note that what we call the "likelihood" in what follows is really a *partial likelihood* because we condition on the observed death times.

We now have a model with parameters $\beta_1, \ldots, \beta_p$ and the proportional hazards assumption. In order to estimate the model parameters we need to write down a likelihood function. Let's assume for the present that there is no censoring, and that *there are no ties* in the death times. (This last assumption is important and we will return to it.) If we assume the death times are ordered (as in our usual notation) we therefore have a list:

$$t_1 < t_2 < \cdots < t_n.$$

- The set of covariates for individual i is denoted $\underline{x}_i$.
- The set of indices of death times is D.
  With no censoring $D = \{1, \ldots, n\}$.
- The size of D is n_D.
  (With no censoring, $n_D = n$).
- The set of individuals alive and uncensored the instant before t_i is denoted A_i.
- Following our usual notation there are N_i individuals in the set A_i.
- With the individuals ordered according to their times and no censoring, we have $A_i = \{i, i+1, \ldots, n\}$.

Since we are saying nothing about the form of the baseline hazard, we condition on the fact that one death occurred at time t_i. We therefore consider the probability L_i that the i-th individual dies at t_i, conditional on t_i being one of the death times:

$$
\begin{aligned}
L_i &= \Pr(\text{individual with covariates } \underline{x}_i \text{ dies at } t_i \mid t_i \text{ is a death time}) \\
&= \frac{\Pr(\text{individual with covariates } \underline{x}_i \text{ dies at } t_i)}{\Pr(t_i \text{ is a death time})} \\
&= \frac{\Pr(\text{individual with covariates } \underline{x}_i \text{ dies at } t_i)}{\sum_{j \in A_i} \Pr(\text{individual with covariates } \underline{x}_j \text{ dies at } t_i)}
\end{aligned}
$$

The summation in the denominator uses the assumption that there are no ties: we're evaluating the probability that any of the individuals $j \in A_i$ die at t_i by assuming deaths are independent. Now, the two lines in the equations above aren't quite right technically: the probability of an individual dying at any given instant is zero (if we're dealing with continuous probability distributions). To make sense of this we work on an interval $[t_i, t_i + \delta t)$ for small δt and consider the limit as $\delta t \to 0$:

$$
\begin{aligned}
L_i &= \lim_{\delta t \to 0} \frac{\Pr(\text{individual with covariates } \underline{x}_i \text{ dies on } [t_i, t_i + \delta_t))}{\sum_{j \in A_i} \Pr(\text{individual with covariates } \underline{x}_j \text{ dies on } [t_i, t_i + \delta_t))} \\
&= \frac{h_i(t_i)}{\sum_{j \in A_i} h_j(t_i)} \\
&= \frac{\phi_i}{\sum_{j \in A_i} \phi_j} \\
&= \frac{\exp(\underline{\beta}^T \underline{x}_i)}{\sum_{j \in A_i} \exp(\underline{\beta}^T \underline{x}_j)}.
\end{aligned}
$$

The (partial) likelihood is just the product of these terms:

$$
L(\underline{\beta}) = \prod_{i \in D} \frac{\exp(\underline{\beta}^T \underline{x}_i)}{\sum_{j \in A_i} \exp(\underline{\beta}^T \underline{x}_j)}.
$$

You can see this equation does not depend on how we ordered the individuals (it's just that if we drop the ordering assumption, A_i doesn't have such a tidy form). Above all, however, notice that the formula does not depend on the exact death times, just the order in which they occur.

Because we have had to condition on the death times, which, in turn, is because we have not said anything about the form of $h_0(t)$, the expression above is not a true likelihood but is described as a *partial likelihood*. The partial likelihood L is maximised to estimate the parameters $\beta_1, \ldots, \beta_p$ given the data in the covariates $\underline{x}_i$, $i = 1, \ldots, n$.

13.3 Right censoring

If an individual's survival time is right-censored at time t_i^* then that individual appears in the risk set A_j at all death times less than t_i^* but, because we do not give a form for $h_0(t)$, it does not make any other contribution to the partial likelihood. So when we have a set of death times indexed by the set D and right-censored times indexed by R with $D \cup R = \{1, 2, \ldots, n\}$ then the likelihood formula is:

$$L(\underline{\beta}) = \prod_{i \in D} \frac{\exp(\underline{\beta}^T \underline{x}_i)}{\sum_{j \in A_i} \exp(\underline{\beta}^T \underline{x}_j)}.$$

In other words, it's exactly the same as above (it's just that D is no longer the whole set of times). This can also be written as:

$$L(\underline{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\underline{\beta}^T \underline{x}_i)}{\sum_{j \in A_i} \exp(\underline{\beta}^T \underline{x}_j)} \right)^{\delta_i}$$

where δ_i is an indicator variable, $\delta_i = 1$ when $i \in D$ and zero otherwise.

13.4 Ties

Although, in theory, ties should not occur (when we work with continuous life time random variables), they do occur in practice because time is not usually recorded sufficiently precisely to distinguish death times which are close together (e.g. same day).

Ties of censored observations do not cause problems. Censored observations which occur at the "same time" as a death are assumed to occur "just after" the death so that the censored individuals appear in the risk set.

When two or more death times are tied, this is more of a problem. Various formulae have been proposed. Most of these use approximations to the exact conditional probability that all of the tied deaths occur before censored times of the same value and before the next death time. For further comments see the book by Collett for example. The most common method used is that due to Breslow: the idea is you average over all the possible sequences of the tied death times. However, R uses the method due to Efron. This can be computationally intensive, and if you find your models are taking a long time to fit, it might be worth changing to the Breslow method: look at `help(coxph)` to see how to do it.

13.5 Fitting Cox models in R

This is carried out using the `coxph` function. It's very similar to using the `survreg` function, except you get the estimates of the model coefficients out directly without having to do any additional calculations. For example:

```
> s <- coxph(formula =
    Surv(time, cens)~ascites+logbilirubin, data=my.data)

> s

              coef exp(coef) se(coef)    z       p
ascites      1.244      3.47    0.285 4.36 1.3e-05
logbilirubin 0.994      2.70    0.112 8.84 0.0e+00

Likelihood ratio test=119  on 2 df, p=0  n= 250
> s$loglik
[1] -504.1208 -444.7277
```

The maximized log partial likelihood is the second term in the vector in the last line above. You can use it to carry out hypothesis tests just as we described in section 12.3. The numbers in the column labelled `coeff` are β_1 and β_2.

13.6 Writing down likelihood functions

In the theory above the notation we adopted assumed that the individuals were ordered according to their death times. When we drop this assumption the formulae for the (partial) likelihood remain the same, but the sets A_i will change. Recall that A_i is the set of indices of individuals who are alive and uncensored at time t_i so that

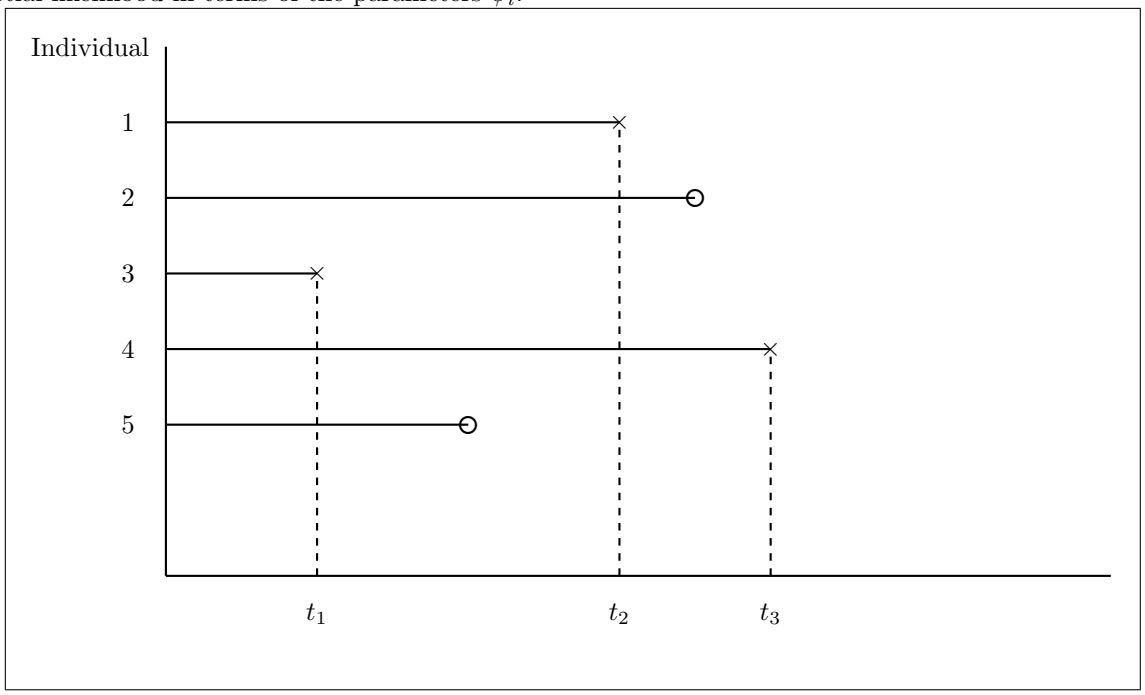
$$A_i = \{j : t_j \geq t_i\}.$$

The following example shows how to write down the likelihood when the death times are not ordered.

Example: Suppose we have survival data on just five individuals as follows.

Individual i	1	2	3	4	5
Survival time	6	7*	2	8	4*

where * indicates a right censored time. Also suppose we model these data using a Cox proportional hazards model, so that the hazard function for individual i is $h_i(t) = \phi_i h_0(t)$. Write down the Cox partial likelihood in terms of the parameters ϕ_i.



At t_1 the risk set contains all five subjects.
At t_2 the risk set contains subjects 1, 2, 4.
At t_3 the risk set contains only subject 4
The partial likelihood is therefore

$$\frac{\phi_3}{\phi_1+\phi_2+\phi_3+\phi_4+\phi_5} \times \frac{\phi_1}{\phi_1+\phi_2+\phi_4} \times \frac{\phi_4}{\phi_4}.$$

14 Confidence intervals for regression coefficients and hazard ratios

The maximum likelihood estimates of the coefficients $\underline{\beta}$ can be taken to be approximately normally distributed (under a large sample assumption). R gives a standard error for each parameter in a Cox model. (These errors are harder to obtain in R for parametric models). Look back at the last lecture where we fitted a Cox model to some data in R. The standard error of the coefficient for ascites is 0.285, and the standard error for the coefficient for log bilirubin is 0.112. Suppose the estimate of the coeffient β of log bilirubin is $\hat{\beta}$ and its standard error is s. Then an approximate 95% confidence interval for β is $\hat{\beta} \pm 1.96 \times s$. Using the data from the last lecture this becomes $0.994 \pm 1.96 \times 0.112$, ie. $[0.77, 1.21]$.

Confidence intervals for hazard ratios can be obtained in an analogous fashion. The hazard ratio for two individuals A and B is:

$$\frac{h_A}{h_B} = \frac{\exp\left(\underline{\beta}^T \underline{x}_A\right)}{\exp\left(\underline{\beta}^T \underline{x}_B\right)}$$

If the individuals are identical apart from in one covariate x, this ratio becomes

$$\frac{h_A}{h_B} = \frac{\exp(\beta x_A)}{\exp(\beta x_B)} = \exp(\beta(x_A - x_B)).$$

A 95% confidence interval for this ratio is therefore

$$\exp\left((\hat{\beta} - 1.96 \times s)(x_A - x_B)\right), \exp\left((\hat{\beta} + 1.96 \times s)(x_A - x_B)\right).$$

Using the data above, suppose we want to calculate a confidence interval for the hazard ratio between two individuals whose log bilirubin differs by 0.5, but who are identical in all other ways. The confidence interval is

$$[\exp(0.5 \times (0.994 - 1.96 \times 0.112)), \exp(0.5 \times (0.994 + 1.96 \times 0.112))].$$

The interval is $[1.47, 1.83]$, so the difference of 0.5 in log bilirubin gives a hazard ratio significantly different from unity.

For hazard ratios involving more than one parameter (eg. compare two individuals with different bpd and different smoking status) the *covariances* of the β parameters need to be taken into account, and this lies beyond the scope of these lectures.

Example

A Cox proportional hazards model was used to model the survival times of cancer patients. Tumour size (in mm) was included as a covariate, with coefficient β. The maximum likelihood estimate of β was $\hat{\beta} = 0.0176$ with standard error 0.004. Find an estimate of the hazard ratio between two individuals with tumours measuring 46mm and 37mm who are identical in other ways. Construct a 95% CI for the hazard ratio.

Hazard ratio:

$$\frac{e^{46\beta}}{e^{37\beta}} = e^{9\beta}.$$

Estimate:

$$e^{9\hat{\beta}} = e^{0.162} = 1.176.$$

95% confidence interval for β : $0.0176 \pm 1.96 \times 0.004$
That is: $0.01016 < \beta < 0.02584$.
95% confidence interval for hazard ratio:

$$1.096 < e^{9\beta} < 1.262$$

Miscellaneous proportional hazards question

Suppose the life times for n individuals are modelled in such a way that individual i is assumed to have survivor function

$$S_i(t) = \frac{1}{(1+t)^{\alpha_i}}$$

where $\alpha_1, \ldots, \alpha_n$ are positive constants. Show that this constitutes a proportional hazards model.

Distribution function:

$$F_i(t) = 1 - S_i(t) = 1 - (1+t)^{-\alpha_i}$$

Probability density function:

$$f_i(t) = \frac{d}{dt}F_i(t) = \alpha_i(1+t)^{-(\alpha_i+1)}$$

Hazard:

$$h_i(t) = \frac{f_i(t)}{S_i(t)} = \frac{\alpha_i(1+t)^{-(\alpha_i+1)}}{(1+t)^{-\alpha_i}} = \frac{\alpha_i}{1+t}$$

Similarly

$$h_j(t) = \frac{\alpha_j}{1+t}$$

Hence

$$\frac{h_i(t)}{h_j(t)} = \frac{\alpha_i}{\alpha_j} = \psi_{ij}$$

which does not depend on t.

15 Accelerated life models

15.1 Introduction

Proportional hazards models are not the only way to relate survival to covariates. The most important alternative to proportional hazards is an *accelerated life model* (sometimes called an accelerated failure time model). As in proportional hazards, in an accelerated life model there is a different survivor function for each individual. Instead of scaling the hazard function, however, accelerated life models scale *time* in the following way. We assume some underlying baseline survivor function $S_0(t)$ and assume that the survivor function for individual i is of the form

$$S_i(t) = S_0(\phi_i t)$$

where ϕ_i is a positive constant called the *acceleration factor* for individual i. As we did for proportional hazards models, we can make the constants ϕ_i depend on the covariates for each individual. Together the baseline S_0 and constants ϕ_i specify the model.

In an accelerated life model time effectively 'runs faster' for some individuals compared to others. Consider the following example. Suppose that we have two groups and for group 1 we take $\phi_1 = 1$ while for group 2 we take $\phi_2 = 2$. It follows that

$$S_2(t) = S_1(2t)$$

or in other words 'time runs twice as fast for group 2 as group 1'. Note that if the median survival time for group 1 is t_1, then the median for group 2 is $t_2 = \frac{1}{2}t_1$.

Suppose that $t_1(\alpha)$ and $t_2(\alpha)$ are the 100α percentiles for two individuals with acceleration factors ϕ_1, ϕ_2. Then

$$S_1[t_1(\alpha)] = S_2[t_2(\alpha)] = 1 - \alpha$$

so

$$S_0[\phi_1 t_1(\alpha)] = S_0[\phi_2 t_2(\alpha)].$$

The two arguments are therefore equal:

$$\phi_1 t_1(\alpha) = \phi_2 t_2(\alpha)$$

and so

$$t_2(\alpha) = (\phi_1/\phi_2)t_1(\alpha).$$

We can relate the distribution function, density function, and hazard function for each individual i to the baseline quantities.

Since $S_i(t) = S_0(\phi_i t)$, it follows that

$$F_i(t) = F_0(\phi_i t)$$

where $F_0(t) = 1 - S_0(t)$ is the baseline distribution function. Now,

$$f_i(t) = \frac{d}{dt}F_i(t) = \frac{d}{dt}F_0(\phi_i t) = \phi_i F_0'(\phi_i t) = \phi_i f_0(\phi_i t).$$

The hazard function for individual i is

$$h_i(t) = f_i(t)/S_i(t) = \phi_i f_0(\phi_i t)/S_0(\phi_i t) = \phi_i h_0(\phi_i t).$$

You might find it useful to back-track in your notes at this point, and look at the equivalent relationships for proportional hazards models.

Typical question:

Suppose that we have a model in which the hazard function for individual i is assumed to be of the form

$$h_i(t) = \frac{\rho_i}{1+\rho_i t}.$$

Show that this constitutes an accelerated life model.

We obtain:

$$H_i(t) = \int_0^t \frac{\rho_i}{1+\rho_i s} ds = \log(1+\rho_i t)$$

so

$$S_i(t) = \exp[-H_i(t)] = \frac{1}{1+\rho_i t}$$

It follows that

$$S_i(t) = S_0(\rho_i t) \text{where } S_0(t) = \frac{1}{1+t}.$$

This is an accelerated life model.

15.2 Covariates

Covariates are incorporated in exactly the same way as for proportional hazards models, namely by making the constants ϕ_i depend on the covariates in the following way:

$$\phi_i = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$$

where $x_{i1}, \ldots, x_{ip}$ are the covariates for individual i and $\beta_1, \ldots, \beta_p$ are coefficients that must be estimated.

The next two subsections explain how the Weibull and log-logistic distributions can be used to construct parametric accelerated life models.

15.3 The Weibull accelerated life model

A life time random variable with Weibull distribution has survivor function

$$S(t) = \exp(-\lambda t^\gamma).$$

An accelerated life model is defined by taking

$$S_i(t) = \exp[-\lambda(\phi_i t)^\gamma]$$

for fixed λ and γ (which specify the baseline) and some positive constants $\phi_1, \ldots, \phi_n$. The baseline survivor function is $S_0(t) = \exp(-\lambda t^\gamma)$. Then

$$S_i(t) = \exp[-\lambda \phi_i^\gamma t^\gamma]$$

and this is a Weibull survivor function for each i.

Now if this all seems very similar to the Weibull proportional hazards model, that's because it's the same up to re-parameterization. Let ϕ_i, β_j denote the linear predictor and coefficients in the accelerated life model, and let $\tilde{\phi}_i, \tilde{\beta}_j$ be the same objects in the proportional hazards model. In the Weibull proportional hazards model we had

$$h_i(t) = \tilde{\phi}_i h_0(t) = \tilde{\phi}_i \lambda \gamma t^{\gamma-1}$$

so that

$$S_i(t) = \exp[-\tilde{\phi}_i \lambda t^\gamma].$$

If we take

$$\phi_i^\gamma = \tilde{\phi}_i$$

then the proportional hazards and accelerated life models are identical. The coefficients are then related by

$$\phi_i^\gamma = \exp[\gamma(\beta_1 x_{i1} + \ldots + \beta_p x_{ip})] = \exp[\tilde{\beta}_1 x_{i1} + \ldots + \tilde{\beta}_p x_{ip}] = \tilde{\phi}_i$$

so that

$$\gamma\beta_j = \tilde{\beta}_j$$

for $j = 1, \ldots, p$. This shows that the two models are identical up to a reparameterization of the coefficients. No distribution except the Weibull distribution (and hence also the exponential distribution) has this property.

15.4 The log-logistic accelerated life model

For the log-logistic distribution

$$S(t) = \frac{1}{1 + (\rho t)^\gamma}.$$

If we take $\rho_i = \phi_i \rho$ then this defines an accelerated life model with baseline

$$S_0(t) = \frac{1}{1 + (\rho t)^\gamma}.$$

and $S_i(t) = S(\phi_i t)$.

The baseline hazard is

$$h_0(t) = \frac{\gamma \rho^\gamma t^{\gamma-1}}{1 + (\rho t)^\gamma},$$

so the relation $h_i(t) = \phi_i h_0(\phi_i t)$ gives

$$h_i(t) = \frac{\gamma \rho^\gamma \phi_i^\gamma t^{\gamma-1}}{1 + (\rho \phi_i t)^\gamma}.$$

We do not have $h_i(t) = \text{const} \times h_0(t)$ so the model is *not* a proportional hazards model.

Refering back to Section 8.4 of the notes, the log linear representation for the log-logistic distribution is

$$\begin{aligned} \log T_i &= -\log(\rho_i) + \frac{1}{\gamma} \log E_i \\ &= -\log(\rho) - (\beta_1 x_{i1} + \cdots + \beta_p x_{ip}) + \frac{1}{\gamma} \log E_i \end{aligned}$$

where the E_i are (standardized) error terms. R fits the log-logistic accelerated life model using this representation: a call to `survreg` specifying `dist="loglogistic"` returns the scale $= 1/\gamma$, intercept $= -\log \rho$, and minus the β coefficients.

15.5 Q-Q plots

Suppose we have data from two groups (and no other covariates). We know from earlier that, if an accelerated life model is appropriate, $t_2(\alpha) = (\phi_1/\phi_2) t_1(\alpha)$ where $t_i(\alpha)$ is the 100α percentile for group i.

Now suppose that we obtain estimates $\hat{t}_i(\alpha_k)$ for a number of percentiles $(100\alpha_1, 100\alpha_2, \ldots)$ for each group, using e.g., Kaplan-Meier. Then, if we plot $\hat{t}_2(\alpha_k)$ against $\hat{t}_1(\alpha_k)$ for $k = 1, 2, \ldots$ we should obtain, approximately, a straight line, passing through the origin, with gradient ϕ_1/ϕ_2. This can be used to identify cases where an accelerated life model might be appropriate.

The plot is called a "percentile-percentile plot," a "quantile-quantile plot" or simply a "Q-Q plot."

15.6 More examples

1. Suppose that we have a model in which the hazard function for individual i is assumed to be of the form

$$h_i(t) = \lambda_i + \mu\lambda_i^2 t$$

for some positive constant μ and positive constants $\lambda_1, \ldots, \lambda_n$. Show that this constitutes an accelerated life model.

Cumulative hazard:

$$\begin{aligned} H_i(t) &= \int_0^t \lambda_i + \mu\lambda_i^2 u \; du \\ &= \lambda_i t + \mu\lambda_i^2 t^2/2 = (\lambda_i t) + \mu(\lambda_i t)^2/2 \end{aligned}$$

Survivor function:

$$\begin{aligned} S_i(t) &= \exp\{-H_i(t)\} \\ &= \exp\{-(\lambda_i t) - \mu(\lambda_i t)^2/2\} \\ &= S_0(\lambda_i t) \end{aligned}$$

where

$$S_0(t) = \exp\{-t - \mu t^2/2\}.$$

2. Suppose that $h_1(t)$ and $h_2(t)$ are the hazard functions for two individuals and that the corresponding survivor functions are $S_1(t)$ and $S_2(t)$. Show that, if $h_1(t)/h_2(t) = \psi \geq 1$, then $S_1(t) \leq S_2(t)$ for all t.

Hazard: $h_1(t) = \psi h_2(t)$
Cumulative hazard: $H_1(t) = \int_0^t \psi h_2(u) \; du = \psi H_2(t)$
Survivor function: $S_1(t) = \exp\{-H_1(t)\} = \exp\{-\psi H_2(t)\} = [S_2(t)]^\psi$
At $t = 0$, $S_1(t) = S_2(t) = 1$.
At $t > 0$, $S_2(t) \leq 1$ (only equal to 1 if $h(u) = 0$ for $u < t$) so $S_1(t) \leq S_2(t)$ for $\psi \geq 1$.

Revision examples

1. A large number of individuals were enrolled in a study and were followed for 30 years to assess the age at which a disease symptom first appeared. For ten selected individuals described below, state the types of censoring they represent.

   (a) The first individual, enrolled in the study at age 45, entered the study with the symptom already present.

   (b) The next two individuals enrolled in the study at ages 35 and 40 and never had any symptoms.

   (c) The next two individuals, enrolled in the study at ages 35 and 40, did not exhibit the symptom when examined by a doctor 6 years after enrollment, but did exhibit the symptom when examined 8 years after enrollment.

   (d) The next two individuals, enrolled at ages 47 and 50, died of causes unrelated to the disease (with no symptoms of the disease) at ages 61 and 65 respectively.

   (e) The last three individuals, enrolled in the study at ages 36, 42, and 50, moved away and were lost to follow-up at ages 40, 55, and 60 respectively, having never shown any symptoms.

2. (a) A life time distribution has hazard function

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

   Derive the corresponding survivor and density functions, and write down an expression for the median survival time.

   (b) A second life time distribution has hazard function

$$h(t) = \frac{\mu + t}{1 + t}$$

   where μ is a positive constant. Derive the corresponding survivor function.

3. An experiment was carried out over a period of years to measure the time between recurrence of oral herpes (cold sores). 76 patients recovering from an outbreak were subsequently asked by phone call whether they had suffered a recurrence. Calls were made at varoius time intervals for 10 years. Some patients didn't reply and so were lost to follow-up. From the data below use the actuarial method to estimate the survivor function and hazard for the time till recurrence.

Months	Suffered recurrence	Lost to follow-up
$0 \leq t < 12$	5	0
$12 \leq t < 24$	7	3
$24 \leq t < 36$	12	5
$36 \leq t < 48$	10	7
$48 \leq t < 60$	8	7
$60 \leq t < 84$	6	2
$84 \leq t < 120$	3	1

4. An experiment leads to the following set of survival data:

   6*, 29, 31, 38*, 42, 42, 43*, 44, 53, 58*, 63*, 85

   where a * symbol indicates a right-censored time. Calculate the Kaplan Meier estimate of the survivor function and hazard function. Construct a 95% confidence interval for $S(50)$. Obtain an estimate of the median survival time together with a 95% confidence interval.

5. The survival times of two groups of patients with a disease were recorded. The two groups received different drug treatments and the survival times in months were as follows:

   Drug A: 12, 18*, 22, 22, 23*, 27, 31, 36*
   Drug B: 24, 25*, 27, 36*, 37, 48

   Carry out a log-rank test of the null hypothesis that there is no difference in survival distribution for the two groups. Note that

$$V = \sum_j \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} = 1.55.$$

6. A disease has three identifiable stages, and survival times of patients suffering from various stages of the the disease were recorded. Some times were right-censored (indicated by a '*'). The times are given below.

   Stage 1: 20, 27, 36*, 42*, 42*

   Stage 2: 12, 15, 19, 21*, 25*, 27, 42*

   Stage 3: 6, 8, 10*, 13, 15, 30*

   Carry out the simplified version of the log-rank test to test the null hypothesis that the distribution of survival time is independent of stage. Carry out a log-rank test for trend with weights $3, 2, 1$ to test same hypothesis but where the alternative consists of a trend with stage. The test statistic is

$$U_T = \sum_k w_k (O_k - E_k)$$

   which has variance

$$V_T = \sum_k (w_k - \bar{w})^2 E_k$$

   where

$$\bar{w} = \frac{\sum w_k E_k}{\sum E_k} = 2.23.$$

7. The survival times for patients in a study are given below. Those marked $*$ were right-censored.

   Female, non-smokers: 19, 22, 25*, 27, 30*, 31*

   Female, smokers: 12*, 14, 16*, 18, 22*, 24*

   Male, non-smokers: 16, 17*, 20, 22*, 23*, 28*

   Male, smokers: 8, 9*, 14, 15, 16*, 19*

   Carry out a stratified log-rank test to test the hypothesis that males and females have the same life time distribution, using smoking status to define the strata.

8. A survival distribution has survivor function

$$S(t) = \frac{1}{(1 + \lambda^2 t^2)^{\frac{1}{2}}}.$$

   Suppose we are given a set of survival times $t_1, \ldots, t_n$, some of which are right-censored. The sets D and R are defined such that, if $i \in D$ then the failure time t_i is observed and, if $i \in R$ then the observation is right censored. There are n_D observed failures and n_R right-censored observations, so that $n_D + n_R = n$. Write down the likelihood function the model defined above given these data. Obtain and simplify the log likelihood function.

9. Suppose that in a survival analysis of n individuals we model the survivor function for individual i as

$$S_i(t) = S_0(t)^{\alpha_i}$$

   where $S_0(t)$ is a baseline survivor function and α_i are positive constants $i = 1, \ldots, n$. Show that this constitutes a proportional hazards model.

10. A Cox proportional hazards model was used to model the survival times of lymphoma patients. White blood cell count was included as a covariate, with coefficient β. The maximum likelihood estimate of β was $\hat{\beta} = 0.453$ with standard error 0.104. Find an estimate of the hazard ratio between two individuals with white blood cell counts 15 and 10 who are indentical in other ways. Construct a 95% CI for the hazard ratio.

11. A survival analysis was carried out as part of a study of time to re-offence for released prisoners in the UK. The age and sex of each individual in the study was recorded and used to define the following covariates. x is a binary variable taking value 1 if the individual is female and zero otherwise. a is the age (in years) of the individual. A Weibull parametric proportional hazards model was used to analyse the data. A null model, with no covariates, and models with each of the covariates alone and with both together were fitted by maximum likelihood. The resulting values of $\log L$, where L is the maximised likelihood, were as follows.

Model	$\log L$
null	-140.350
x	-139.094
a	-137.843
x, a	-137.577

Use suitable tests to compare these models and state your conclusions about the evidence for the effects of the covariates.

12. Suppose we have survival data on six individuals as follows.

Individual i	1	2	3	4	5	6
Survival time	25	12*	19	28	21*	35*

where * indicates a right censored time. Also suppose we model these data using a Cox proportional hazards model, so that the hazard function for individual i is $h_i(t) = \phi_i h_0(t)$. Write down the Cox partial likelihood in terms of the parameters ϕ_i and simplify the expression.

13. Suppose that we have a model in which the hazard function for individual i is assumed to be of the form

$$h_i(t) = \frac{\mu_i t}{1 + \mu_i t^2}$$

for some positive constant μ and positive constants $\mu_1, \ldots, \mu_n$. Show that this constitutes an accelerated life model.