MATH3085/6143

Survival Models

Contents

Chapter 1

Introduction

1.1 Survival analysis

The aim of statistical modelling is to investigate the relationship between an observed response (usually denoted by y) and k explanatory variables (denoted by $\mathbf{x} = (x_1, \ldots, x_k)$). In MATH3085/6143, the response is always the **time from origin until an event of interest occurs**. Because the response is a time, we denote the observation t and the corresponding random variable T.

Survival analysis refers to a set of special statistical methods required to analyse time-to-event data.

The object of survival modelling is to learn about the variability in T in a population of interest and (often) how this is associated with other potentially explanatory variables (covariates).

1.2 Applications of survival analysis and alternative terminology

Historically, survival analysis originated from medical applications where the event was death and the time to event was called the survival time. However, survival analysis now has applications in many areas beyond medicine including the following.

- Demography - age at 'milestone' such as
  - death
  - birth of first child
- Engineering
  - failure time of component
- Criminology
  - time to recidivism
- Medicine
  - survival after heart attack
  - time from cancer treatment to relapse
- Psychology
  - response time to stimulus
- Economics
  - duration of unemployment

Important actuarial examples where we are required to model survival data include:

- time between pensionable age and death;
- time between taking out a life insurance policy and death;
- time between taking out a critical illness insurance policy and onset of illness;
- failure time for a product with warranty insurance.

Due to the different application areas, you may encounter different terminology. The following table summarises alternative terminology you may encounter.

survival analysis	**origin**	**event of interest**	**time to event**
event history analysis	initial event	death	survival time
duration analysis	initiating event	failure	failure time
hazard modelling	starting event	endpoint	response time
reliability analysis	time origin	outcome	waiting time
		terminating event	duration
		target event	spell
			episode

1.3 Why is Survival Analysis 'special'?

Why does survival analysis need its own module? Answers include the following.

- Models for nonnegative random variable T. In standard normal linear modelling, the response has a normal distribution which allows for negative responses.
- Data are often not well-described by standard probability distributions, e.g. the normal distribution.
- Data are typically censored (see Section 1.5).
- Model parameters often *not* of primary interest. Often interest in the whole survival distribution, e.g. all quantiles, not just measures of location such as mean or median.
- Time-dependent covariates, i.e. they are not fixed like in standard regression modelling.
- Truncation: some cases may be missing (see Section 1.6).

1.4 Types of time to event

The time to event can materialise in different ways.

- Continuous time
  In theory the value of T can be recorded to arbitrary precision.
  In practice the value of T is rounded to convenient level of precision.
  Tied data values cannot occur in theory but do in practice.
- Grouped continuous time
  Imprecise time measurements, only reporting interval in which observation lies.
  e.g. time in completed years, months, weeks or days
  common in population mortality studies (report age in completed years at death)
- Discrete time
  True discrete-time scale, $T = 1, 2, 3, \cdots$
  Examples include number of operations of a machine to first failure, number of attempts to pass a test ...

1.5 Censoring

A common feature of survival data is censoring. *Censoring* occurs when we do not know all times-to-event T exactly, but only have bounds on some of the survival times.

Special methods are needed for censored data because:

- censored observations provide information;
- exclusion of censored data leads to bias;
- discarding censored data would be inefficient.

Observations of T may be:

- observed precisely;
- right censored;
- left censored;

- interval censored.

In statistical modelling, censoring has to be taken into account to avoid bias.

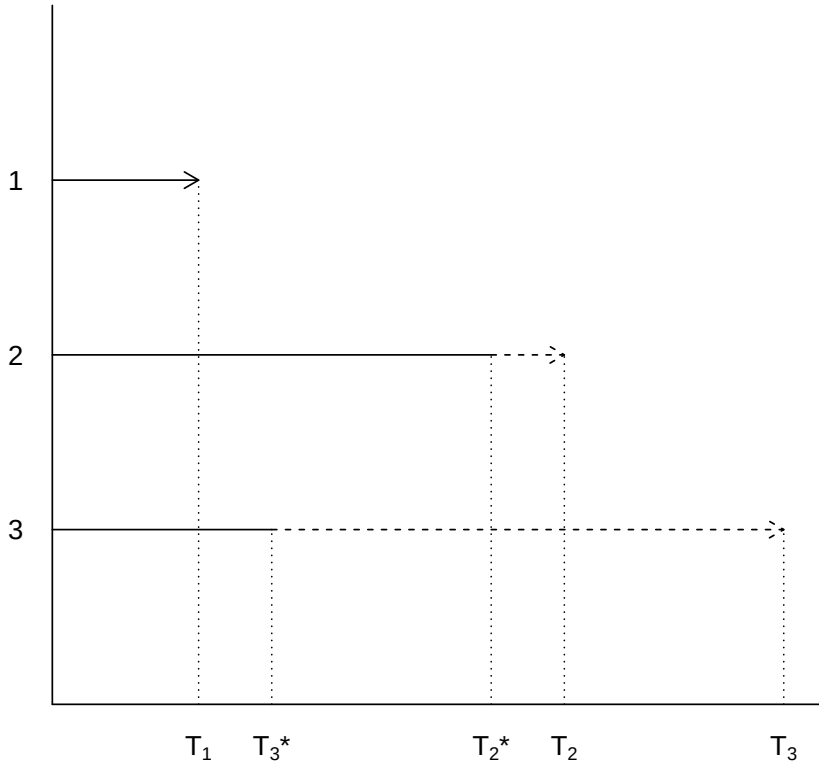
1.5.1 Right censoring

An observation of T is *right censored* if we only observe a lower bound for T, i.e. we know that T was greater than some value, e.g. 18, but not the exact value itself.

A censored value with be denoted by T^* or T^+, so we know

$$T \geq T^* \qquad \text{equivalently} \qquad T \in (T^*, \infty)\,.$$

In the picture below, observation 1 is not censored and we would observe the response T_1. For observations 2 and 3 are right censored: we only know that the true responses T_2 and T_3 are greater than T_2^* and T_3^*, respectively.



Reasons for right censoring include:

- event (*e.g.* death) has not been observed before the end of study;
- individual lost-to-followup;
- individual withdrawal from study.

In most survival analyses (particularly involving mortality) we should expect right censoring.

1.5.2 Left and interval censoring

An observation of T is *left censored* if we only observe an upper bound for T, i.e. we know that T was less than some value, e.g. 18, but not the exact value itself. Left censoring is not as common as right censoring.

An observation of T is *interval censored* if we only observe an interval for T, i.e. we know that T is between two values, e.g. $(18, 20)$, but not the exact value itself. Note that left and right censored observations are actually interval censored where the upper or lower limit of the interval is $-\infty$ or ∞, respectively.

1.5.3 Informative and non-informative censoring

Censoring is *informative* if the event of censoring conveys extra information about T (other than it is in a particular range). Otherwise, censoring is *non-informative*.

Let C be the variable governing the time at which an observation is (right) censored. If $T > C$ then we observe $T = (C, \infty]$, and if $C > T$ then we observe T precisely.

A sufficient condition for non-informative censoring is that C and T are independent variables. For example C is a (non-random) time fixed in advance of the study.

Examples of causes of informative censoring include:

- patients lost to follow-up because of good prognosis;
- withdrawals of patients due to ill health related to outcome of interest.

Informative censoring causes complications for statistical modelling because we need a joint model for T and C. We will assume non-informative censoring throughout.

1.6 Truncation

Left truncation of survival data occurs when cases whose survival times T are shorter than a given time, either fixed or random, are *not* observed *e.g.* a study of heart attack survival which excludes individuals who died before reaching hospital.

Right truncation of survival data occurs when cases which haven't experienced the event are not observed, *e.g.* data obtained from death certificates

Truncation is different to censoring in that with truncation we do not observe anything associated with a truncated value but with censoring we know the observation existed we just do not know its exact value.

In statistical modelling, truncation has to be taken into account to avoid bias.

1.7 Goals of survival analysis

Some basic goals of survival analysis include the following.

- Describe how survival in the sample depends on time.
- Inference to the population of interest.
- Compare whole survival distributions for groups.
- Explain survival differentials using explanatory variables:
- Predict future survival.

Chapter 2

Statistical Models

2.1 Introduction

Statistical analysis (or inference) involves drawing conclusions, and making predictions and decisions, using the evidence provided to us by observed data.

To do this, we use *statistical models*, where we simulate the process by which the observed data were generated through a probability distribution.

- The form of the model helps us to understand the real-world process by which the data were generated.
- If the model explains the observed data well, then it should also inform us about future (or unobserved) data, and hence help us to make predictions (and decisions contingent on unobserved data).
- The use of statistical models, together with a carefully constructed methodology for their analysis also allows us to quantify the uncertainty associated with any conclusions, predictions or decisions we make.

We rarely believe in our models, but regard them as temporary constructs subject to improvement.

2.2 Example: Leukaemia

Survival times are given for 33 patients who died from acute myelogenous leukaemia. In `R`, this can be found in the `leuk` object in the `MASS` package.

```
> library(MASS)
> head(leuk)
```

```
    wbc      ag time
1  2300 present   65
2   750 present  156
3  4300 present  100
4  2600 present  134
5  6000 present   16
6 10500 present  108
```

```
> tail(leuk)
```

```
      wbc     ag time
28  31000 absent    8
29  26000 absent    4
30  21000 absent    3
31  79000 absent   30
32 100000 absent    4
33 100000 absent   43
```

For the moment, ignore the columns for `wbc` and `ag`.

2.3 Notation

Suppose that we have n data observations, then we use

$$t_1, t_2, \ldots, t_n$$

to denote these observed event (failure, death, ...) or censoring times.

For the leukaemia survival times in Section 2.2, $n = 33$ and $t_1 = 65$, $t_2 = 156$, ..., $t_{33} = 43$.

We denote the complete data by the vector $\mathbf{t} = (t_1, t_2, \ldots, t_n)$.

In a *statistical model*, we consider $t_1, t_2, \ldots, t_n$ to be observations of random variables (denoted with the corresponding capital letters)

$$T_1, T_2, \ldots, T_n$$

We also use the vector notation $\mathbf{T} = (T_1, T_2, \ldots, T_n)$.

This is the same as MATH2010, but with t and T, instead of y and Y, respectively.

2.4 Statistical models

A statistical model specifies a probability distribution for the random variables $\mathbf{T}$ corresponding to the data observations $\mathbf{t}$.

Providing a specification for the distribution of n jointly varying random variables is made much easier if we can make some *simplifying assumptions*, such as

1. $T_1, T_2, \ldots, T_n$ are *independent* random variables
2. $T_1, T_2, \ldots, T_n$ have the same probability distribution (so $t_1, t_2, \ldots, t_n$ are observations of a single random variable T)

Assumption 1 is very common, even in quite complex examples.

Assumption 2 is not always appropriate, but may be reasonable when we are modelling a homogeneous population, without other information. Making assumption 2 means we cannot include explanatory variables so sometimes we will not make this assumption.

When we make assumptions 1 and 2, we say that $T_1, T_2, \ldots, T_n$ are *independent and identically distributed* (i.i.d.)

2.5 A fully specified model

Sometimes a model completely specifies the probability distribution of $T_1, T_2, \ldots, T_n$.

For example, for the leukaemia survival times in Section 2.2, we might assume the model

$$T_1, T_2, \ldots, T_n \overset{iid}{\sim} \text{lognormal}(\mu, \sigma^2)$$

where $\mu = 3$ and $\sigma^2 = 4$. Note that $T_1, T_2, \ldots, T_n \overset{iid}{\sim} \text{lognormal}(\mu, \sigma^2)$ is equivalent to

$$\log T_1, \log T_2, \ldots, \log T_n \overset{iid}{\sim} \text{N}(\mu, \sigma^2),$$

where log is the natural logarithm.

The key is that we have assumed exact values for μ and σ^2. This would be appropriate when there is some external (to the data) theory as to why the model (in particular the values of $\mu = 3$ and $\sigma^2 = 4$) is appropriate.

The data can then be used to assess the plausibility of the model. Do the data support the model or not?

We rarely have external theory to specify a model so precisely.

2.6 A parametric statistical model

For the leukaemia survival times in Section 2.2, a more common model would be

$$T_1, T_2, \ldots, T_n \overset{iid}{\sim} \text{lognormal}(\mu, \sigma^2)$$

where μ and σ^2 are unspecified.

This is called a *parametric statistical model* as it completely specifies the probability distribution which generated the data, *apart from a (small) number of constants (parameters)*, in this case the values of μ and σ^2.

The data are then used to *estimate* the unknown parameters (μ and σ^2 here), and to assess the plausibility of other assumptions (lognormal distribution, independence, etc.)

2.7 A nonparametric statistical model

Sometimes, it is not appropriate, or we want to avoid, making a precise specification for the distribution which generated $T_1, T_2, \ldots, T_n$. Then, we might propose the model

$T_1, T_2, \ldots, T_n$ are i.i.d. random variables.

This is sometimes described as a nonparametric (or distribution-free) specification. It imposes limitations on the kind of statistical inferences we can obtain, but still allows us to do some interesting things, such as the following.

- Use exploratory (graphical) techniques, such as box plots and histograms to learn about the distribution of the (common) random variable T which generated the data
- Estimate features of the distribution of T, such as its expectation $E(T)$, variance $Var(T)$, $P(T > t_0)$ for some specified t_0, etc.
- Estimate the distribution or survivor function of T (to be covered in this module).

2.8 Regression models

Often, we model survival data to learn about the relationship between survival time T and other potentially explanatory variables $x_1, x_2, \ldots$.

The leukaemia survival times in Section 2.2 include values of two such explanatory variables:

- `ag`, taking values in the set $\{\text{present}, \text{absent}\}$ (a test result)
- `wbc` (white blood cell count), a numerical variable

In a regression model, we assume that $T_1, T_2, \ldots, T_n$ are independent random variables but they are not identically distributed (we make assumption 1 but not assumption 2).

Instead, we assume the differences between their distributions is explained by a *regression function* of the values of the explanatory variables.

For example, for $i = 1, \ldots, n$ we assume $T_i \sim \text{lognormal}(\mu_i, \sigma^2)$ independently with

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

where x_{i1} is the ith value of `wbc` and

$$x_{i2} = \begin{cases} 1 & \text{if } \texttt{ag} \text{ is present;} \\ 0 & \text{if } \texttt{ag} \text{ is absent;} \end{cases}$$

i.e. x_{i2} is a dummy variable.

2.9 The data analysis process

The data analysis can be summarised by the following diagram.

Obtain data

↓

Examine (visualise) data

↓

Specify models

↓

Estimate model parameters

↓

Compare models

↓

Assess chosen model

↓

Base predictions/decisions on chosen model

In MATH3085/6143 we focus on models and methods which have been specifically developed for survival data.

Chapter 3

The Survival Distribution

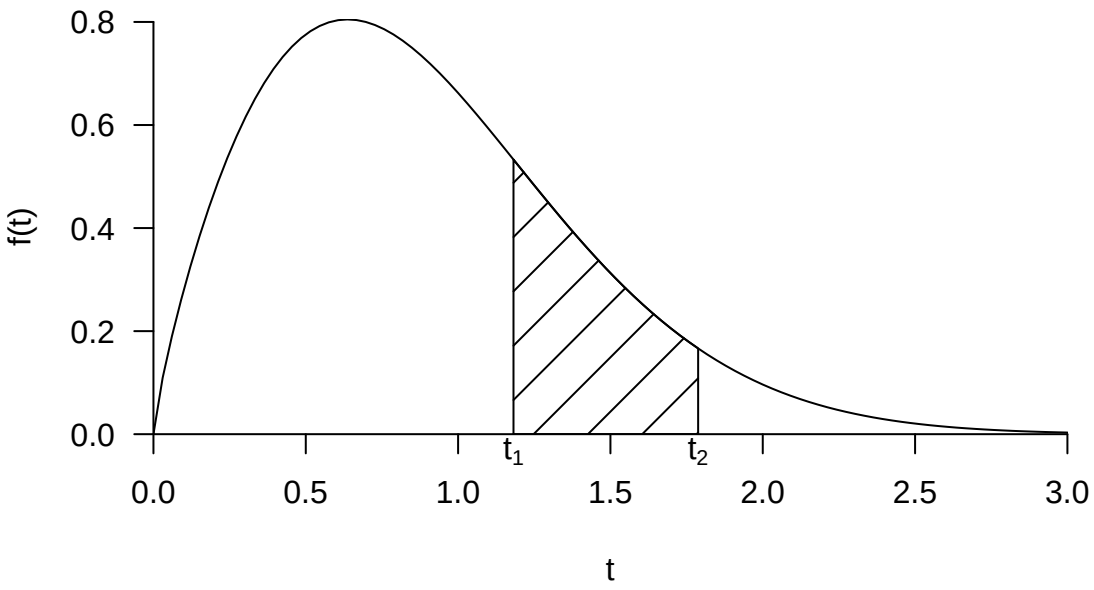
3.1 The density function

We model survival time using a random variable T.

If T is a continuous random variable (as we shall generally assume throughout this module) then its distribution is defined by its probability density function (p.d.f.) $f_T(t)$, or $f(t)$ for short.

Recall that, for any values t_1 and t_2, we have

$$P(t_1 \leq T \leq t_2) = \int_{t_1}^{t_2} f_T(t)\mathrm{d}t$$



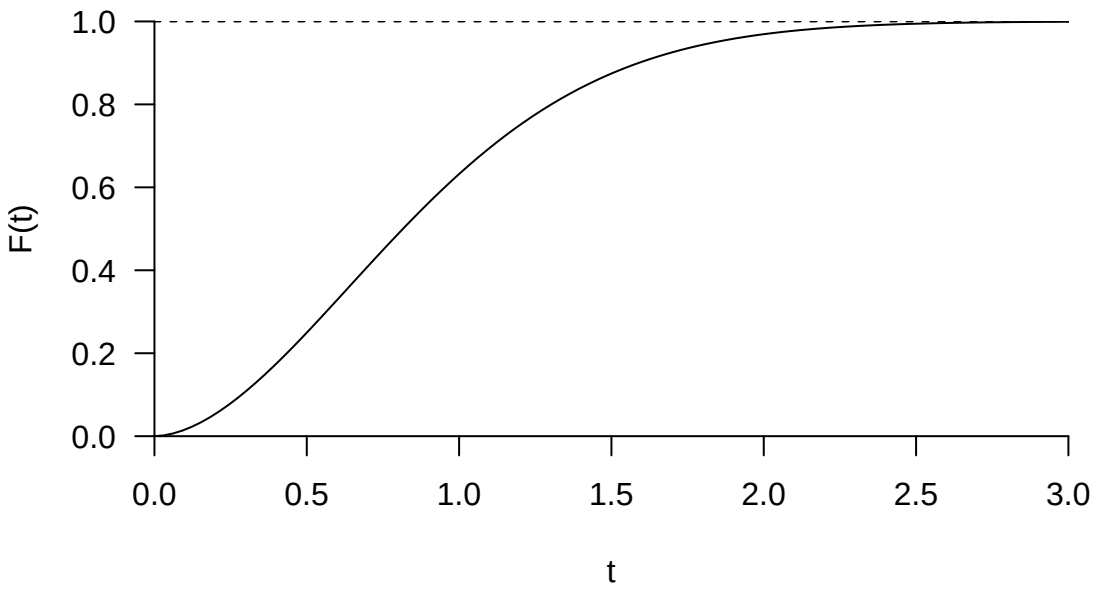
3.2 The distribution function

Recall that the distribution function $F_T(t)$, for a random variable T, is defined as

$$F_T(t) = P(T \leq t)$$

and that for a continuous survival variable with p.d.f. f_T, we have

$$F_T(t) = \int_0^t f_T(s)\mathrm{d}s$$



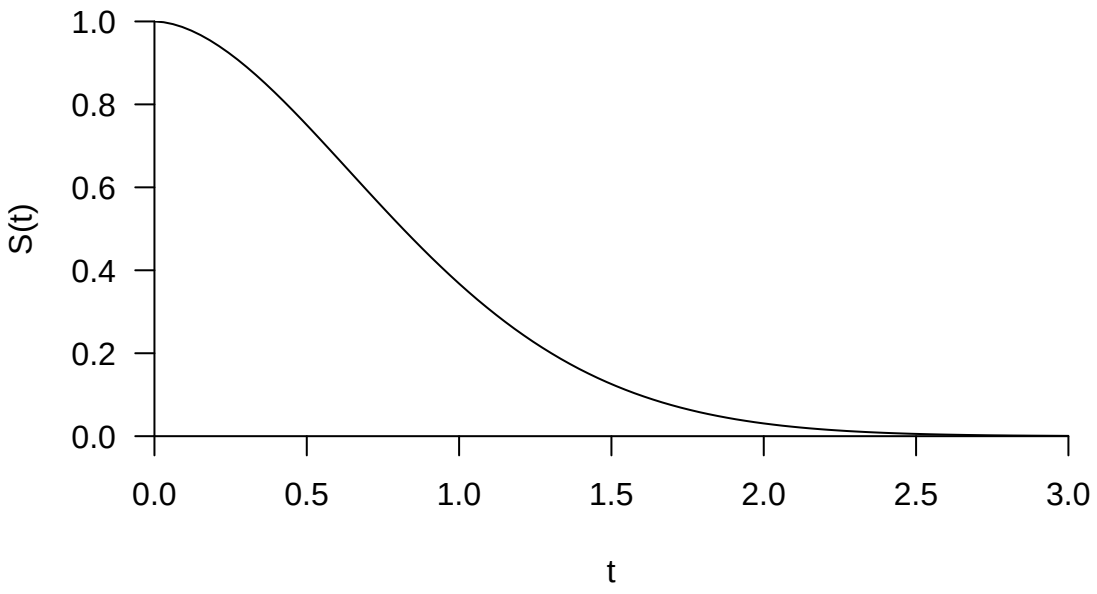
3.3 The survival function

In survival analysis, we usually prefer to focus on the *Survival Function* (sometimes called the *Survivor Function*), $S_T(t)$, which is defined as

$$S_T(t) = P(T > t)$$

so for a continuous survival variable with p.d.f. f_T, we have

$$S_T(t) = \int_t^\infty f_T(s)\mathrm{d}s$$



3.3.1 The survival function: properties

- The survival function is a decreasing function of t with $S_T(0) = 1$ and $S_T(t) \to 0$ as $t \to \infty$.

- The survival function is related to the distribution function through

$$S_T(t) + F_T(t) = 1 \qquad \Rightarrow \qquad S_T(t) = 1 - F_T(t)$$

- The expected lifetime is given, in terms of the survival function by

$$\begin{aligned} E(T) &= \int_0^\infty t f_T(u)\mathrm{d}u \\ &= [-uS_T(u)]_0^\infty + \int_0^\infty S_T(u)\mathrm{d}u \\ &= \int_0^\infty S_T(t)dt, \end{aligned}$$

the area under the survival function curve.

3.3.2 The residual survival function

The *residual survival function* is defined for $t \geq 0$ as

$$S_T(t|x) = P(T > t + x|T > x)$$

for some specified x. It gives the probability of surviving beyond $t + x$ given that you have survived past x.

It defines the distribution of future (beyond x) survival for the subpopulation of survivors up to time x.

Using the standard definition of a conditional probability, we have

$$\begin{aligned} S_T(t|x) &= \frac{P(T > t+x \quad \text{and} \quad T > x)}{P(T > x)} \\ &= \frac{P(T > t+x)}{P(T > x)} \\ &= \frac{S_T(t+x)}{S_T(x)} \end{aligned}$$

Similarly the residual distribution function is $F_T(t|x) = 1 - S_T(t|x)$.

3.4 The hazard function

The *hazard function* is defined for $t \geq 0$ as

$$h_T(t) = \lim_{\delta t \to 0} \frac{P(T \leq t + \delta t | T > t)}{\delta t}$$

The numerator $P(T \leq t + \delta t | T > t)$ can be thought of as the expected number of future events (failures) in the time interval $(t, t + \delta t]$ (in a residual population of size 1 at time t).

Hence, the ratio

$$\frac{P(T \leq t + \delta t | T > t)}{\delta t}$$

is the event (failure) *rate* over the time interval $(t, t+\delta t]$.

The hazard function is the limiting (or instantaneous) event rate at time t.

3.4.1 The hazard function: properties

- The hazard function tells you how likely the event is to occur at (or around) a particular time t, *given that it has not occurred before then.*

- The *hazard function* is given by

$$\begin{aligned} h_T(t) &= \lim_{\delta t \to 0} \frac{P(T \leq t + \delta t | T > t)}{\delta t} \\ &= \lim_{\delta t \to 0} \frac{P(T \leq t + \delta t \quad \text{and} \quad T > t)}{\delta t P(T > t)} \\ &= \lim_{\delta t \to 0} \frac{P(t < T \leq t + \delta t)}{\delta t P(T > t)} \\ &= \lim_{\delta t \to 0} \frac{f_T(t)\delta t}{\delta t P(T > t)} \\ &= \frac{f_T(t)}{S_T(t)}. \end{aligned}$$

- The only constraint on the hazard function $h_T(t)$ is that it must be non-negative for $t \in (0, \infty)$.

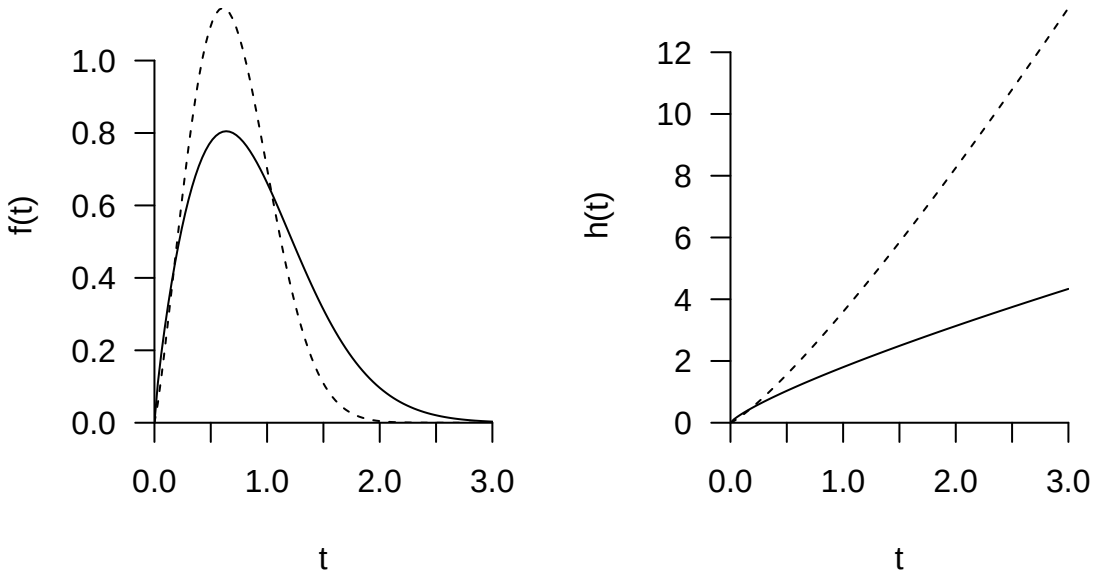
- There is no need to construct a 'residual hazard function', because $h_T(t)$ already explicitly conditions on $T > t$.

3.4.2 Hazard v. density

The hazard function compares the relative probabilities of events occurring at different times t, conditional on $T > t$ and hence accounts for the size of the population at risk at t.

The p.d.f. simply compares the relative probabilities of events occurring at different times in the population at large.

A time t with a high value for $h(t)$ is not necessarily a likely failure time, as there may be a small probability that any individual survives until then.



3.5 The integrated hazard function

The *integrated hazard function* (cumulative hazard function) is defined by

$$H_T(t) = \int_0^t h_T(u)du.$$

The cumulative hazard describes the "total exposure to risk" for a survivor up to $T = t$.

Note that

$$H_T(t) = \int_0^t h_T(u)du = \int_0^t \frac{f_T(u)}{S_T(u)}\mathrm{d}u.$$

Using the substitution, $v = 1/S_T(t)$, then

$$\frac{\mathrm{d}v}{\mathrm{d}u} = \frac{f_T(u)}{S_T(u)^2},$$

and

$$\mathrm{d}u = \frac{\mathrm{d}v S_T(u)^2}{f_T(u)}.$$

Then

$$\begin{aligned} H_T(t) &= \int_1^{1/S_T(t)} \frac{1}{v}\mathrm{d}v \\ &= [\log v]_1^{1/S_T(t)} \\ &= -\log S_T(t). \end{aligned}$$

3.6 Relationships

Only one of the functions f_T, F_T, S_T, h_T or H_T needs to be specified to completely determine the distribution of T.

In survival analysis, interest is usually focussed on the survivor function $S(t)$ and/or the hazard function $h(t)$.

The others can then be calculated using the relationships presented in the following table.

	f_T	S_T	h_T
$f_T(t) =$		$-\frac{\mathrm{d}}{\mathrm{d}t}S_T(t)$	$h_T(t)\exp[-\int_0^t h_T(s)\mathrm{d}s]$
$S_T(t) =$	$\int_t^\infty f_T(s)\mathrm{d}s$		$\exp[-\int_0^t h_T(s)\mathrm{d}s]$
$h_T(t) =$	$\frac{f_T(t)}{\int_t^\infty f_T(s)\mathrm{d}s}$	$-\frac{\mathrm{d}}{\mathrm{d}t}\log S_T(t)$	

Chapter 4

Distributions for Survival Modelling

We now introduce some distributions which are commonly used in survival models. In each case, we present a family of distributions which depend on one or more parameters.

Each of these distributions has sample space $(0, \infty)$ so is appropriate as a model for a survival time T.

In each case, we shall present the density function $f_T(t)$, the survival function $S_T(t)$ and the hazard function $h_T(t)$.

4.1 The exponential distribution

The exponential (or negative exponential) distribution has a single parameter, the rate (scale) $\beta > 0$, is denoted $\exp(\beta)$ and has p.d.f.

$$f_T(t) = \beta \exp(-\beta t)$$

survival function

$$S_T(t) = \exp(-\beta t),$$

and hazard function

$$h_T(t) = \beta.$$

If $T \sim \exp(\beta)$ then

$$E(T) = \frac{1}{\beta} \qquad \text{and} \qquad Var(T) = \frac{1}{\beta^2}.$$

Note that the p.d.f. is sometimes parameterised as $\frac{1}{\gamma} \exp\left(-\frac{t}{\gamma}\right)$, i.e. $\gamma = 1/\beta$.

The survival function is given by

$$\begin{aligned} S_T(t) &= \int_t^\infty f_T(u)\mathrm{d}u \\ &= \int_t^\infty \beta \exp\left(-\beta u\right) \mathrm{d}u \\ &= \left[-\exp\left(-\beta u\right)\right]_t^\infty \\ &= \exp\left(-\beta t\right). \end{aligned}$$

The hazard function is

$$\begin{aligned} h_T(t) &= \frac{f_T(t)}{S_T(t)} \\ &= \frac{\beta \exp\left(-\beta t\right)}{\exp\left(-\beta t\right)} \\ &= \beta. \end{aligned}$$

Using integration by parts, the expectation is

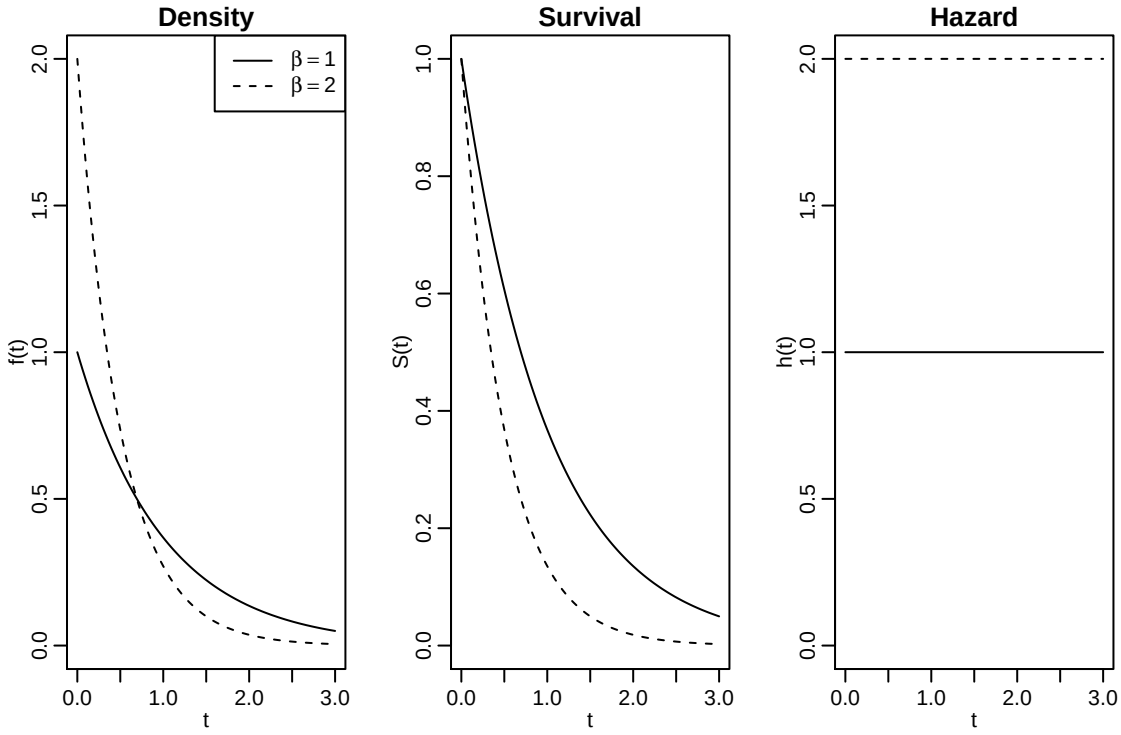
$$\begin{aligned} E(T) &= \int_0^\infty \beta t \exp(-\beta t)\, \mathrm{d}t \\ &= [-t \exp(-\beta t)]_0^\infty + \int_0^\infty \exp(-\beta t)\, \mathrm{d}t \\ &= \left[-\frac{1}{\beta} \exp(-\beta t)\right]_0^\infty \\ &= \frac{1}{\beta}. \end{aligned}$$

The variance can also be calculated via $Var(T) = E(T^2) - E(T)^2$, where

$$E(T^2) = \int_0^\infty \beta t^2 \exp(-\beta t)\, \mathrm{d}t$$

is found using integration by parts (twice).

The plots below show the density, survival and hazard for two different exponential distributions with $\beta = 1$ and $\beta = 2$.



4.2 The Weibull distribution

The Weibull distribution has two parameters, the shape $\alpha > 0$ and the scale $\beta > 0$, is denoted Weibull(α, β) and has p.d.f.

$$f_T(t) = \alpha\beta\,(\beta t)^{\alpha-1} \exp\left\{-(\beta t)^{\alpha}\right\}$$

survival function

$$S_T(t) = \exp\left\{-(\beta t)^{\alpha}\right\},$$

and hazard function

$$h_T(t) = \alpha\beta\,(\beta t)^{\alpha-1}.$$

If $T \sim$ Weibull(α, β) then

$$\begin{aligned} E(T) &= \frac{1}{\beta}\Gamma\left(1+\frac{1}{\alpha}\right) \\ Var(T) &= \frac{1}{\beta^2}\left\{\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma\left(1+\frac{1}{\alpha}\right)^2\right\} \end{aligned}$$

where $\Gamma(r) = \int_0^\infty x^{r-1}e^{-x}dx$ is the Gamma function.

The survival function is given by

$$\begin{aligned} S_T(t) &= \int_t^\infty f_T(u)\mathrm{d}u \\ &= \int_t^\infty \alpha\beta\left(\beta u\right)^{\alpha-1}\exp\left\{-\left(\beta u\right)^\alpha\right\}\mathrm{d}u. \end{aligned}$$

Using the substitution $v = \beta^\alpha u^\alpha$,

$$\frac{\mathrm{d}v}{\mathrm{d}u} = \alpha\beta^\alpha u^{\alpha-1} \qquad \text{and} \qquad \mathrm{d}u = \alpha^{-1}\beta^{-\alpha}u^{1-\alpha}\mathrm{d}v.$$

Then

$$\begin{aligned} S_T(t) &= \int_{\beta^\alpha t^\alpha}^\infty \exp\left(-v\right)\mathrm{d}v \\ &= \left[-\exp\left(-v\right)\right]_{\beta^\alpha t^\alpha}^\infty \\ &= \exp\left\{-\left(\beta t\right)^\alpha\right\}. \end{aligned}$$

The hazard function is

$$\begin{aligned} h_T(t) &= \frac{f_T(t)}{S_T(t)} \\ &= \frac{\alpha\beta\,(\beta t)^{\alpha-1}\exp\{-(\beta t)^{\alpha}\}}{\exp\{-(\beta t)^{\alpha}\}} \\ &= \alpha\beta\,(\beta t)^{\alpha-1}. \end{aligned}$$

The expectation is

$$\begin{aligned} E(T) &= \int_0^{\infty} t f_T(t)\mathrm{d}t \\ &= \int_0^{\infty} \alpha\,(\beta t)^{\alpha}\exp\{-(\beta t)^{\alpha}\}\,\mathrm{d}t. \end{aligned}$$

Using the substitution $u = (\beta t)^{\alpha}$, (i.e. $t = u^{1/\alpha}/\beta$), then

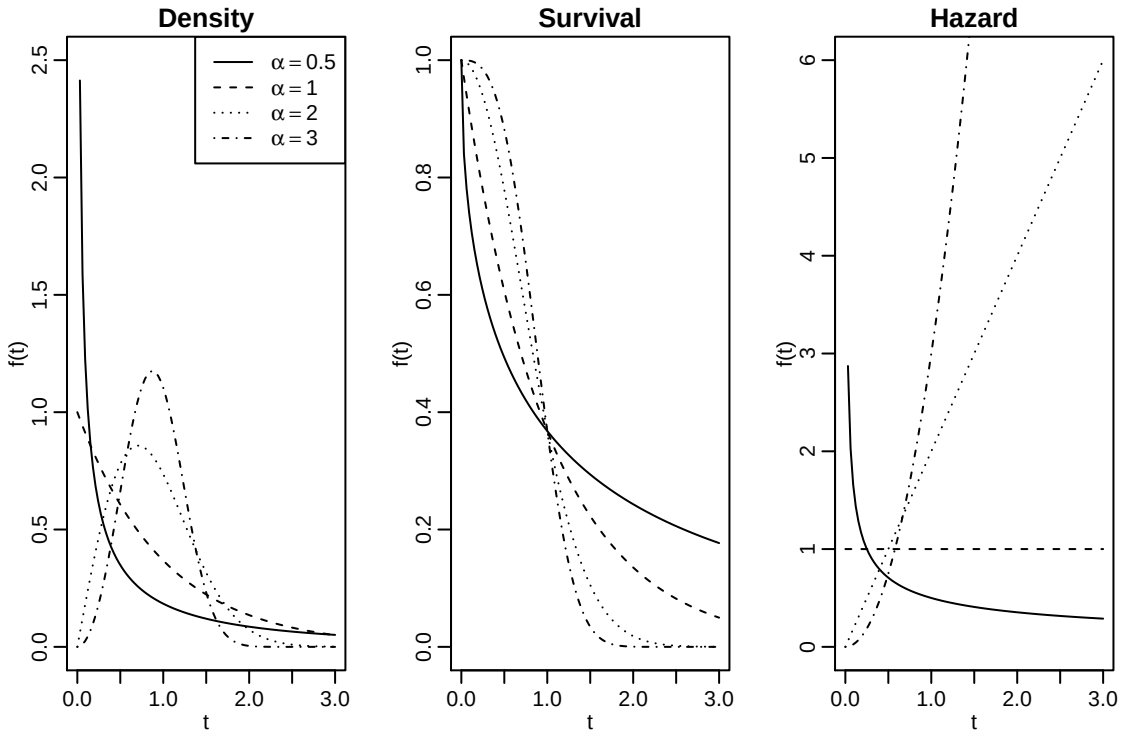
$$\frac{\mathrm{d}u}{\mathrm{d}t} = \alpha\beta^{\alpha}t^{\alpha-1} \qquad \text{and} \qquad \mathrm{d}t = \alpha^{-1}\beta^{-\alpha}t^{1-\alpha}\mathrm{d}u.$$

Then

$$\begin{aligned} E(T) &= \frac{1}{\beta}\int_0^{\infty} u^{1/\alpha}\exp(-u)\mathrm{d}u \\ &= \frac{1}{\beta}\Gamma\left(1+\frac{1}{\alpha}\right). \end{aligned}$$

The expression for the variance can be proved in a similar way.

The plots below show the density, survival and hazard for four different Weibull distributions given by different values of α but all with $\beta = 1$.



4.2.1 Weibull distribution properties

- Weibull$(1, \beta) = \exp(\beta)$

  So the Weibull with shape parameter $\alpha = 1$ is exponential.

- If $X \sim$ Weibull(α, β), then

$$T = bX \sim \text{Weibull}(\alpha, \beta/b).$$

- If $X \sim \exp(1)$, then

$$T = \frac{X^{1/\alpha}}{\beta} \sim \text{Weibull}(\alpha, \beta).$$

4.3 The log-logistic distribution

The log-logistic distribution has two parameters, the shape $\alpha > 0$ and the scale $\beta > 0$, is denoted loglogistic(α, β) and has p.d.f.

$$f_T(t) = \frac{\alpha\beta(\beta t)^{\alpha-1}}{[1 + (\beta t)^\alpha]^2}$$

survival function

$$S_T(t) = \frac{1}{1 + (\beta t)^\alpha},$$

and hazard function

$$h_T(t) = \frac{\alpha\beta(\beta t)^{\alpha-1}}{1 + (\beta t)^\alpha}.$$

If $T \sim$ loglogistic(α, β) then

$$E(T) = \frac{\pi}{\alpha\beta \sin(\pi/\alpha)}$$

The survival function is given by

$$\begin{aligned} S_T(t) &= \int_t^\infty f_T(u)\mathrm{d}u \\ &= \int_t^\infty \frac{\alpha\beta(\beta u)^{\alpha-1}}{[1 + (\beta u)^\alpha]^2}\mathrm{d}u. \end{aligned}$$

Using the substitution $v = 1 + (\beta u)^\alpha$, then

$$\frac{\mathrm{d}v}{\mathrm{d}u} = \alpha\beta^\alpha u^{\alpha-1} \qquad \text{and} \qquad \mathrm{d}u = \frac{\mathrm{d}v}{\alpha\beta(\beta u)^{\alpha-1}}.$$

Then

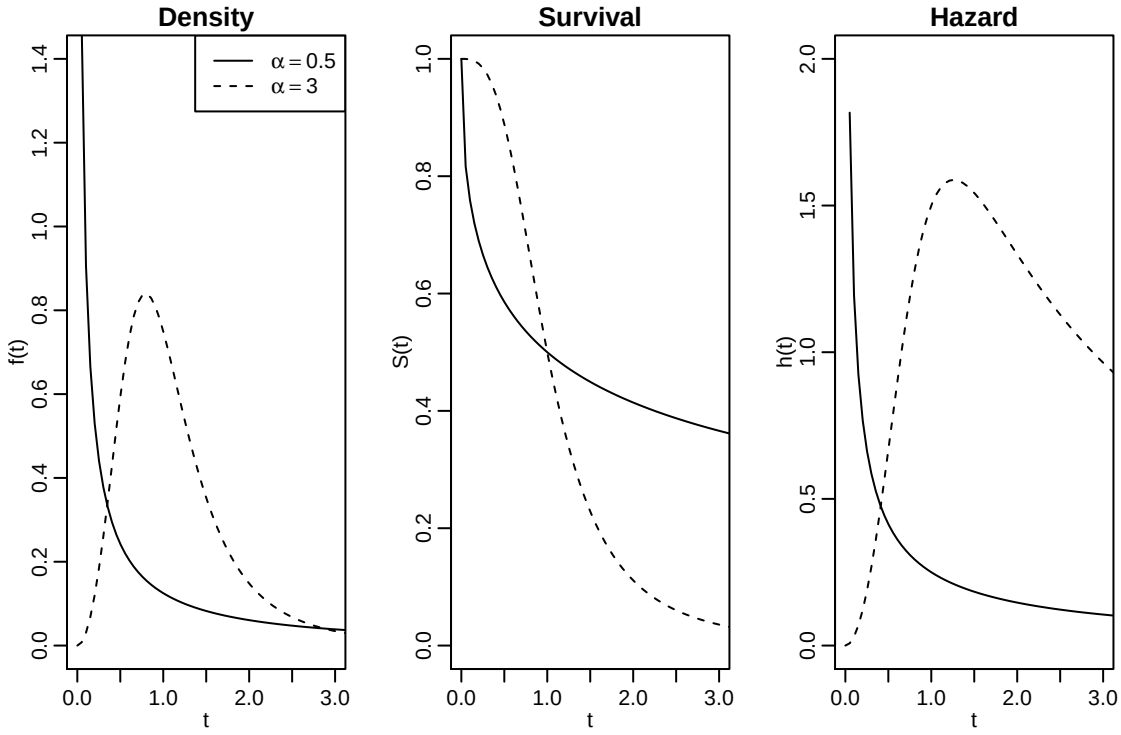
$$\begin{aligned} S_T(t) &= \int_{1+(\beta t)^\alpha}^{\infty} \frac{1}{v^2}\mathrm{d}v \\ &= \left[-\frac{1}{v}\right]_{1+(\beta t)^\alpha}^{\infty} \\ &= \frac{1}{1+(\beta t)^\alpha}. \end{aligned}$$

The hazard function is

$$\begin{aligned} h_T(t) &= \frac{f_T(t)}{S_T(t)} \\ &= \frac{\frac{\alpha\beta(\beta t)^{\alpha-1}}{[1+(\beta t)^\alpha]^2}}{\frac{1}{1+(\beta t)^\alpha}} \\ &= \frac{\alpha\beta(\beta t)^{\alpha-1}}{1+(\beta t)^\alpha}. \end{aligned}$$

The expectation will not be proven here!

The plots below show the density, survival and hazard for two different log-logistic distributions given by different values of α but all with $\beta = 1$.



4.4 The lognormal distribution

The lognormal distribution has two parameters, μ and $\sigma^2 > 0$, is denoted lognormal(μ, σ^2), is the distribution of $\exp X$ where $X \sim N(\mu, \sigma^2)$ and has p.d.f.

$$f_T(t) = \frac{1}{(2\pi)^{1/2}\sigma t} \exp\left(-\frac{1}{2\sigma^2}[\log t - \mu]^2\right) = \frac{1}{\sigma t}\phi\left(\frac{\log t - \mu}{\sigma}\right)$$

survival function

$$S_T(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right),$$

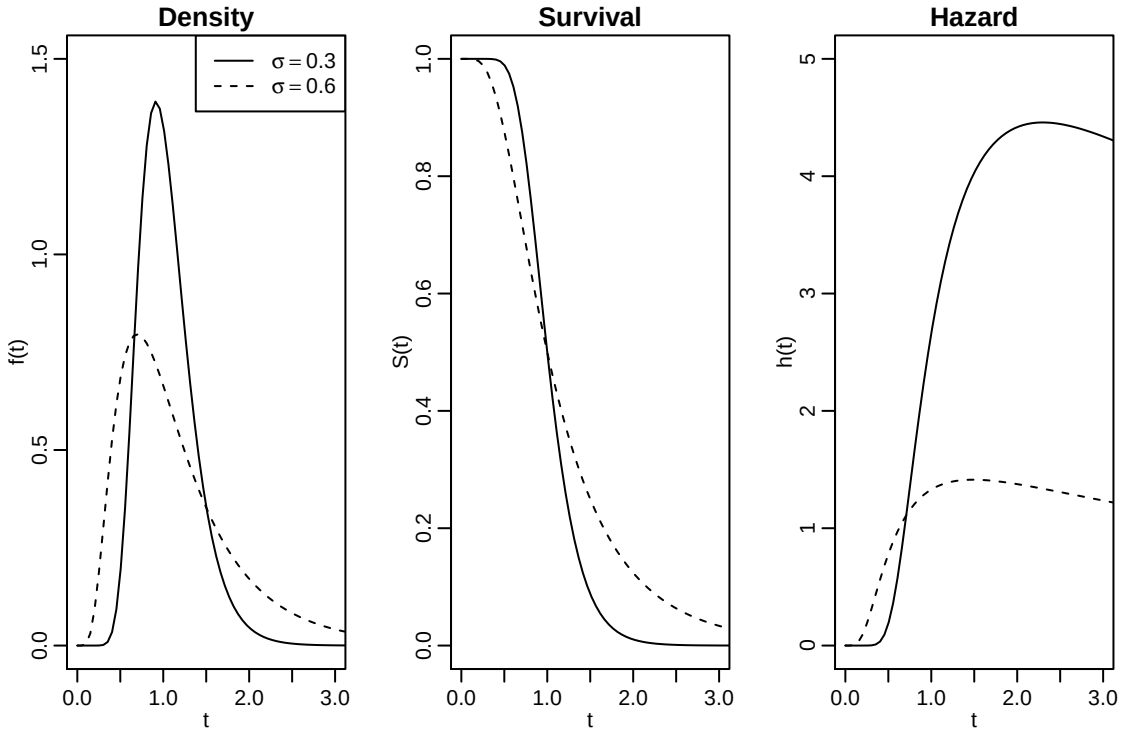
where ϕ and Φ are the standard normal density and distribution functions. The lognormal hazard function is $h_T(t) = f_T(t)/S_T(t)$.

If $T \sim$ lognormal(μ, σ^2) then

$$E(T) = \exp(\mu + \sigma^2/2) \quad \text{and} \quad Var(T) = \exp(2\mu + \sigma^2)[\exp \sigma^2 - 1].$$

We do not derive the survival and hazard function here.

The plots below show the density, survival and hazard for two different lognormal distributions given by different values of σ but all with $\mu = 0$.



4.5 The Gompertz distribution

The Gompertz distribution has two parameters, the shape $\alpha > 0$ and the scale $\beta > 0$, is denoted Gompertz(α, β) and has p.d.f.

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

survival function

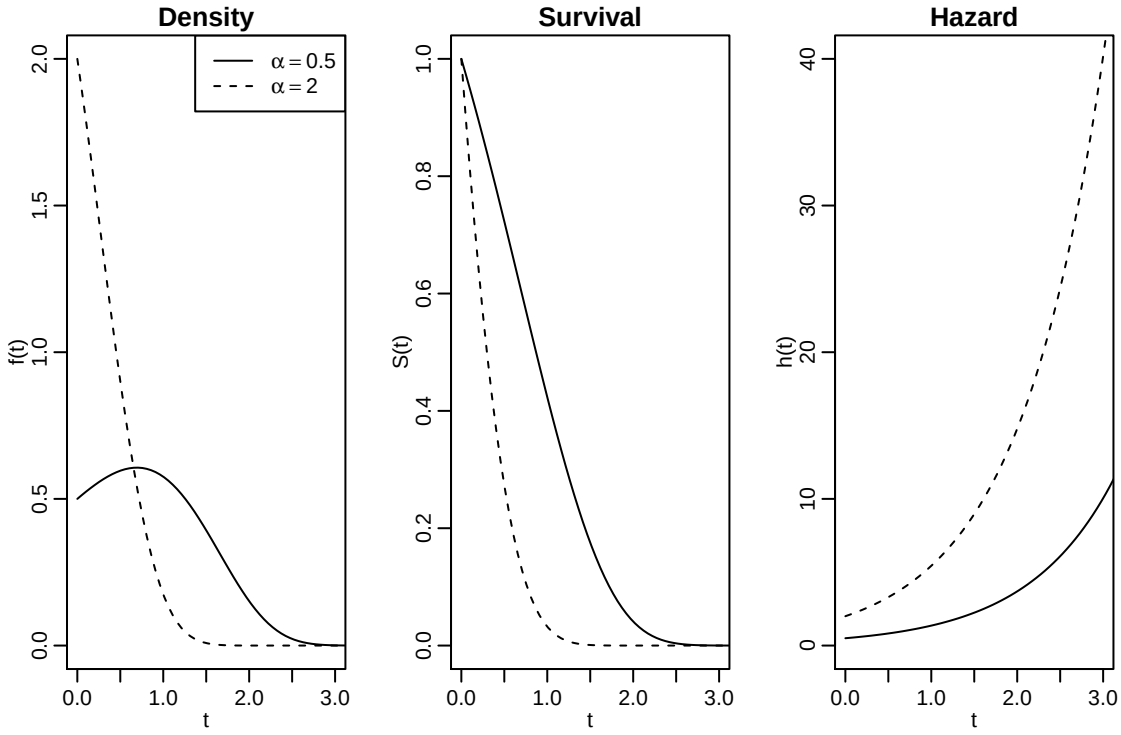
$$S_T(t) = \exp\left(-\frac{\alpha}{\beta}(e^{\beta t} - 1)\right),$$

and hazard function

$$h_T(t) = \alpha \exp(\beta t)$$

See Exercise 5 on Worksheet 1 for derivations of the survival and hazard function.

The plots below show the density, survival and hazard for two different Gompertz distributions given by different values of α but all with $\beta = 1$.



Exponential hazard implies that log-hazard is linear in t which can be a plausible model for human lifetimes from middle age onwards.

Chapter 5

Survival models: parameter estimation

We start by considering models for a *homogenous* population of survival times, from which we have observed a sample $\mathbf{t} = (t_1, \ldots t_n)$.

Some of the observations may be censored.

- Here, we will only consider right censoring (most common) only.
- Extension to other forms of censoring is possible.

Let $d_1, \ldots, d_n$ be a set of censoring indicators, with

$$d_i = \begin{cases} 0 & \text{unit } i \text{ was censored at } t_i \text{ so actual failure time } > t_i \\ 1 & \text{failure of unit } i \text{ was observed at } t_i \end{cases}$$

Therefore $t_1, \ldots t_n$ are (possibly censored) observations of i.i.d. random variables $\mathbf{T} = (T_1, \ldots, T_n) \equiv T$.

5.1 Parametric models

A parametric model for T specifies the p.d.f f_T, apart from the values of a (small) number of unknown parameters, which we denote by $\boldsymbol{\theta}$.

For example, for a Weibull model, $\boldsymbol{\theta}$ comprises the shape and scale parameters α and β, i.e. $\boldsymbol{\theta} = (\alpha, \beta)$.

We write $f_T(t; \boldsymbol{\theta})$ to recognise the dependence of the p.d.f. f_T on $\boldsymbol{\theta}$.

Similarly, we explicitly recognise the dependence of the survivor $S_T(t; \boldsymbol{\theta})$ and hazard $h_T(t; \boldsymbol{\theta})$ functions on $\boldsymbol{\theta}$.

Estimating the distribution of T then simply involves estimating $\boldsymbol{\theta}$.

Once we have an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, we can obtain the corresponding estimates $S_T(t;\hat{\boldsymbol{\theta}})$ and $h_T(t;\hat{\boldsymbol{\theta}})$ of the survival and hazard functions.

Typically, we use *maximum likelihood estimation* to obtain $\hat{\boldsymbol{\theta}}$ in a parametric model.

5.2 Maximum likelihood estimation

Maximum likelihood estimation is a general method for parameter estimation with many good properties (see MATH3044 for more).

Initially consider a scalar parameter θ.

The maximum likelihood estimator (m.l.e.) $\hat{\theta}$ maximises the *likelihood*, $L(\theta)$, which is simply the joint probability (density) of the observed data, treated as a function of the unknown θ.

Assuming i.i.d. observations with no censoring

$$L(\theta) = \prod_{i=1}^{n} f_T(t_i;\theta)$$

as the joint p.d.f. of independent variables is just the product of their individual (marginal) p.d.f.s

5.3 Censored data likelihood

For a right censored observation t_i is not an observed value of T_i, but an interval (t_i, ∞) for T_i.

Hence, the appropriate contribution to the likelihood for a censored t_i is not $f_T(t_i;\theta)$ but rather $P(T_i > t_i) = S_T(t_i)$.

Assuming i.i.d. observations with (right) censoring indicators $d_1, \ldots, d_n$,

$$L(\theta) = \prod_{i:d_i=1} f_T(t_i;\theta) \prod_{i:d_i=0} S_T(t_i;\theta)$$

This is a product of two terms: one for the exact observations and one for the censored observations.

5.4 Maximum likelihood properties

Usually, it is easier to maximise the log-likelihood $\ell(\theta) = \log L(\theta)$

For large samples, we have the asymptotic approximation

$$\hat{\theta} \overset{approx}{\sim} N(\theta, I(\theta)^{-1})$$

where

$$I(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right]$$

is called the Fisher information. A multi-parameter extension of this result exists but is outside the scope of this module

In large samples the m.l.e. is approximately unbiased, and we can construct $100(1-\alpha)\%$ confidence intervals as

$$\hat{\theta} \pm z_{1-\alpha/2} s.e.(\hat{\theta})$$

where $s.e.(\hat{\theta})$, the standard error given by

$$s.e.(\hat{\theta}) = \left[I(\hat{\theta})^{-1} \right]^{1/2}$$

and $z_{1-\alpha/2}$ is the $1-\alpha/2$th quantile of the standard normal. This can be numerically calculated using the `R` function `qnorm`. For example, for $\alpha = 0.1$, $\alpha = 0.05$ and $\alpha = 0.01$

```
> qnorm(0.95)

[1] 1.644854

> qnorm(0.975)

[1] 1.959964

> qnorm(0.995)

[1] 2.575829
```

which would be needed for 90%, 95% and 99% confidence intervals, respectively.

5.5 Example: exponential model likelihood

Suppose that we want to fit an exponential model to our data, so

$$\begin{aligned} L(\beta) &= \prod_{i:d_i=1} f_T(t_i;\beta) \prod_{i:d_i=0} S_T(t_i;\beta) \\ &= \prod_{i:d_i=1} \beta \exp(-\beta t_i) \prod_{i:d_i=0} \exp(-\beta t_i) \\ &= \beta^{d_+} \exp\left(-\beta \sum_{i=1}^{n} t_i\right) \end{aligned}$$

where $d_+ = \sum_{i=1}^{n} d_i$ is the number of uncensored observations.

Therefore the log-likelihood is

$$\ell(\beta) = d_+ \log \beta - \beta \sum_{i=1}^{n} t_i.$$

The m.l.e. for β in the exponential model maximises $\ell(\beta)$ and therefore solves

$$0 = \frac{\partial}{\partial \beta} \ell(\beta) = \frac{d_+}{\beta} - \sum_{i=1}^{n} t_i.$$

Therefore

$$\frac{d_+}{\hat{\beta}} = \sum_{i=1}^{n} t_i \qquad \Rightarrow \qquad \hat{\beta} = \frac{d_+}{\sum_{i=1}^{n} t_i}.$$

so $\hat{\theta}$ is the number of uncensored observations divided by the sum of the (censored and observed) survival times.

The second derivative of the log-likelihood is

$$\frac{\partial^2}{\partial \beta^2}\ell(\beta) = -\frac{d_+}{\beta^2}.$$

The Fisher information is

$$\begin{aligned} I(\beta) &= E\left(-\frac{\partial^2}{\partial \beta^2}\ell(\beta)\right) \\ &= E\left(\frac{d_+}{\beta^2}\right) \\ &= \frac{d_+}{\beta^2}. \end{aligned}$$

So now

$$\begin{aligned} s.e.(\hat{\beta}) &= \left[I(\hat{\beta})^{-1}\right]^{1/2} \\ &= \frac{\hat{\beta}}{d_+^{1/2}}. \end{aligned}$$

5.5.1 Gehan data

Remission times (in weeks) from a clinical trial of 42 leukaemia patients. Patients matched in pairs and randomised to 6-mercaptopurine or control. in R, data is found in the MASS package in the gehan.

```
> library(MASS)
> head(gehan)
```

```
  pair time cens   treat
1    1    1    1 control
2    1   10    1    6-MP
3    2   22    1 control
4    2    7    1    6-MP
5    3    3    1 control
6    3   32    0    6-MP
```

```
> tail(gehan)
```

```
   pair time cens   treat
37   19    4    1 control
38   19    9    0    6-MP
39   20    1    1 control
40   20    6    0    6-MP
41   21    8    1 control
42   21   10    0    6-MP
```

We fit exponential models separately to the treatment group and the control group.

For the 6-MP treatment group, $n = 21$, $d_+ = 9$ and $\sum_{i=1}^{n} t_i = 359$, so

$$\hat{\beta} = \frac{9}{359} = 0.025 \qquad \text{and} \qquad s.e.(\hat{\theta}) = \frac{0.025}{9^{1/2}} = 0.008$$

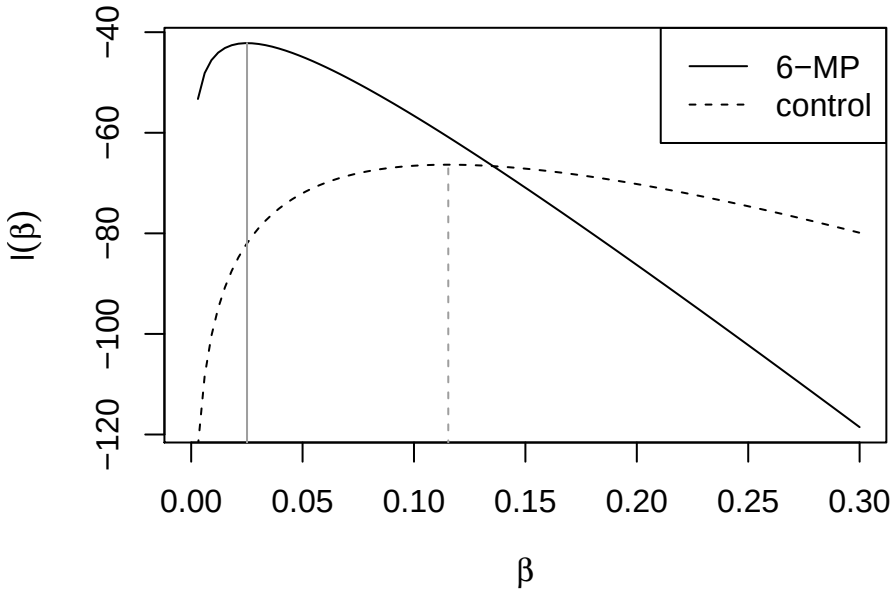
For the `control` treatment group, $n = 21$, $d_+ = 21$ and $\sum_{i=1}^{n} t_i = 182$, so

$$\hat{\beta} = \frac{21}{182} = 0.115 \qquad \text{and} \qquad s.e.(\hat{\beta}) = \frac{0.115}{21^{1/2}} = 0.025$$

Assuming this model, approximate 95% confidence intervals for β are

- $[0.009, 0.041]$ (6-MP group)
- $[0.066, 0.165]$ (control group)

The plot below shows the log-likelihood curves for both groups.



5.6 Example: Weibull model likelihood

For a Weibull(α, β) model,

$$\begin{aligned} L(\alpha,\beta) &= \prod_{i:d_i=1} f_T(t_i;\alpha,\beta) \prod_{i:d_i=0} S_T(t_i;\alpha,\beta) \\ &= \prod_{i:d_i=1} \alpha\beta\,(\beta t)^{\alpha-1} \exp\left\{-(\beta t)^{\alpha}\right\} \prod_{i:d_i=0} \exp\left\{-(\beta t)^{\alpha}\right\} \\ &= \alpha^{d_+}\beta^{\alpha d_+} \left(\prod_{i:d_i=1} t_i\right)^{\alpha-1} \exp\left(-\sum_{i=1}^{n} (\beta t_i)^{\alpha}\right). \end{aligned}$$

Therefore

$$\ell(\alpha,\beta) = d_+ \log\alpha + \alpha d_+ \log\beta + (\alpha-1)\sum_{i:d_i=1} \log t_i - \beta^{\alpha}\sum_{i=1}^{n} t_i^{\alpha}.$$

Requires numerical solution of $\frac{\partial \ell}{\partial \alpha} = 0$, $\frac{\partial \ell}{\partial \beta} = 0$ (use `R` or similar).

Chapter 6

Non-parametric Survival Estimation

Frequently, we want to estimate the distribution of T for the (i.i.d.) homogeneous model based on observations $t_1, \ldots, t_n$ with (right) censoring indicators $d_1, \ldots, d_n$, *without* assuming a particular parametric family for f_T.

The likelihood is given by

$$L = \prod_{i:d_i=1} f_T(t_i) \prod_{i:d_i=0} S_T(t_i)$$

and this can be made infinitely large by concentrating f_T in infinitesimally narrow regions around each observed t_i (i.e. those for which $d_i = 1$).

Hence the maximum likelihood estimate for the survival distribution is a discrete distribution on the observed failure times $\{t_i : d_i = 1\}$.

6.1 Discrete survival distributions and likelihood

A discrete survival variable T takes values in a sample space $0 < t'_1 < t'_2 < \cdots < t'_m < \infty$ and is defined by any of

- probability function (p.f.)

$$f_i = P(T = t'_i),$$

for $i = 1, \ldots, m$;

- survival function

$$\begin{aligned} S(t) = P(T > t) &= 1 - \sum_{j:t'_j \le t} f_j \\ &\equiv s_i \qquad t \in [t'_i, t'_{i+1}), \quad i = 0, \ldots, m \end{aligned}$$

where $t'_0 \equiv 0$ and $t'_{m+1} \equiv \infty$.

- hazard function

$$\begin{aligned} h_i &= P(T = t'_i | T \ge t'_i) \\ &= \frac{P(T = t'_i \quad \text{and} \quad T \ge t'_i)}{P(T \ge t'_i)} \\ &= \frac{P(T = t'_i)}{P(T \ge t'_i)} \\ &= \frac{f_i}{P(T > t'_{i-1})} \\ &= \frac{f_i}{s_{i-1}} \end{aligned}$$

We have $s_0 = 1$ and

$$\begin{aligned} s_i &= P(T > t'_i) \\ &= 1 - P(T \leq t'_i) \\ &= 1 - \big(P(T \leq t'_{i-1}) + f_i\big) \\ &= P(T > t'_{i-1}) - f_i \\ &= s_{i-1} - f_i \\ &= s_{i-1} - h_i s_{i-1} \\ &= s_{i-1}(1 - h_i) \end{aligned}$$

Therefore

$$s_i = \prod_{j=1}^{i}(1 - h_j) \qquad \text{and} \qquad f_i = h_i \prod_{j=1}^{i-1}(1 - h_j)$$

It is best to describe the survival distributions through $\{h_i\}$ which are simply constrained to be in $[0, 1]$, rather than $\{f_i\}$ or $\{s_i\}$ which have more complex constraints.

- Let $0 < t'_1 < t'_2 < \cdots < t'_m < \infty$ denote the *distinct* observed failure times observed in our sample $\{t_1, \ldots, t_n\}$
- $t'_0 = 0$ and $t'_{m+1} = \infty$

- d'_i is the number of failures observed at t'_i, $(i = 1, \ldots, m)$, so

$$\sum_{i=1}^{m} d'_i = \sum_{i=1}^{n} d_i = d_+$$

- c_i $(i = 0, \ldots, m)$ is the number of censored observations with censoring times between t'_i and t'_{i+1}

The likelihood is given by

$$\begin{aligned}
L = \prod_{i=1}^{m} f_i^{d'_i} \prod_{i=0}^{m} s_i^{c_i} &= \prod_{i=1}^{m} \left[h_i \prod_{j=1}^{i-1} (1-h_j) \right]^{d'_i} \prod_{i=1}^{m} \left[\prod_{j=1}^{i} (1-h_j) \right]^{c_i} \\
&= \prod_{i=1}^{m} \left[\left(\frac{h_i}{1-h_i} \right)^{d'_i} \prod_{j=1}^{i} (1-h_j)^{d'_i + c_i} \right] \\
&= \left[\prod_{i=1}^{m} \left(\frac{h_i}{1-h_i} \right)^{d'_i} \right] \prod_{i=1}^{m} \prod_{j=1}^{i} (1-h_j)^{d'_i + c_i} \\
&= \left[\prod_{i=1}^{m} \left(\frac{h_i}{1-h_i} \right)^{d'_i} \right] \prod_{i=1}^{m} \prod_{j=i}^{m} (1-h_i)^{d'_j + c_j} \\
&= \left[\prod_{i=1}^{m} \left(\frac{h_i}{1-h_i} \right)^{d'_i} \right] \prod_{i=1}^{m} (1-h_i)^{\sum_{j=i}^{m} (d'_j + c_j)}
\end{aligned}$$

The log-likelihood is given by

$$\ell = \sum_{i=1}^{m} \left[d'_i \log h_i - d'_i \log(1-h_i) + \log(1-h_i) \sum_{j=i}^{m} (d'_j + c_j) \right]$$

so

$$\frac{\partial \ell}{\partial h_i} = \frac{d'_i}{h_i} + \frac{d'_i}{1-h_i} - \frac{\sum_{j=i}^{m}(d'_j + c_j)}{1-h_i}$$

and

$$\frac{d'_i}{\hat{h}_i} + \frac{d'_i - \sum_{j=i}^{m}(d'_j + c_j)}{1-\hat{h}_i} = 0$$

$$\Rightarrow \qquad \hat{h}_i = \frac{d'_i}{\sum_{j=i}^{m}(d'_j + c_j)}, \qquad i = 1, \ldots, m$$

6.2 The Kaplan-Meier estimator

The discrete hazard m.l.e. is

$$\hat{h}_i = \frac{d'_i}{r_i}, \qquad i = 1, \ldots,$$

where

$$r_i = \sum_{j=i}^{m}(d'_j + c_j)$$

is called the *number at risk* (of failure) at t'_i.

The hazard at each t'_i is therefore estimated by the observed number of failures at t'_i as a proportion of the number at risk at t'_i.

The corresponding estimator of the survival function

$$\hat{S}(t) = \prod_{j=1}^{i}(1 - \hat{h}_j) = \prod_{j=1}^{i}\left(1 - \frac{d'_j}{r_j}\right) \qquad t \in [t'_i, t'_{i+1}), \quad i = 0, \ldots, m$$

is called the *Kaplan-Meier* (or product-limit) estimator.

6.2.1 Notes on the Kaplan-Meier estimator

- Although the Kaplan-Meier estimator represents a discrete survival distribution, we use it to estimate $S_T(t)$ for continuous T.

- For censored times tied with uncensored times, we treat the censored times as infinitessimally larger than the uncensored ones with which they are tied (so c_i includes observations censored at t'_i).

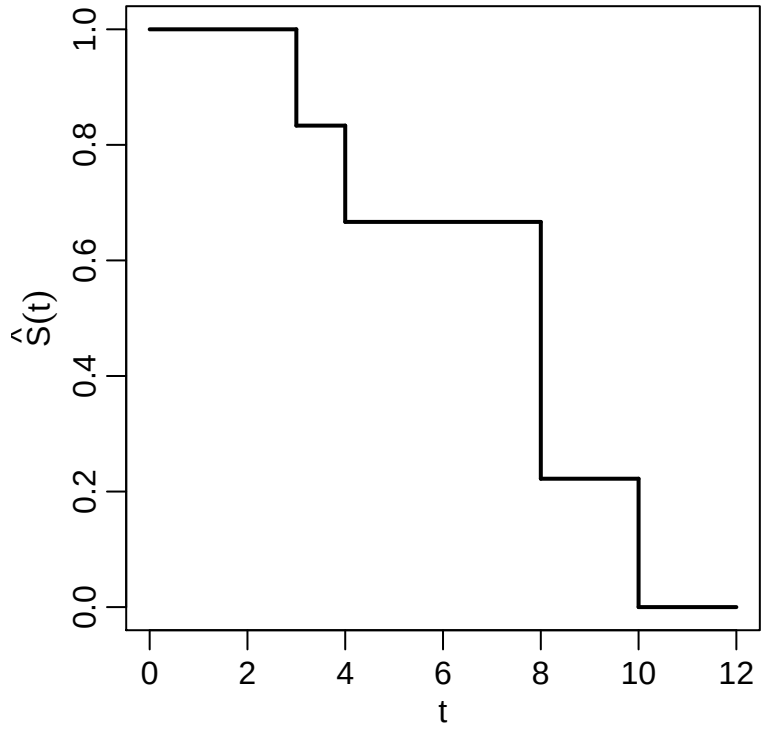
- If the largest time is censored (so $c_m > 0$), then the Kaplan-Meier estimator is positive at the largest uncensored time and therefore the estimated survivor function is nowhere zero.

- Sometimes, it is informative to plot $\log S(t)$, recalling that the hazard function $h(t)$ is equal to $-\frac{\mathrm{d}}{\mathrm{d}t}\log S(t)$. Hence $\log S(t)$ is linear for exponentially distributed data (constant hazard).

6.2.2 Example of calculating the Kaplan-Meier estimate by hand

Consider the following survival times $\mathbf{t} = (3, 4, 6^*, 8, 8, 10)$. There are $m = 4$ distinct observed survival times.

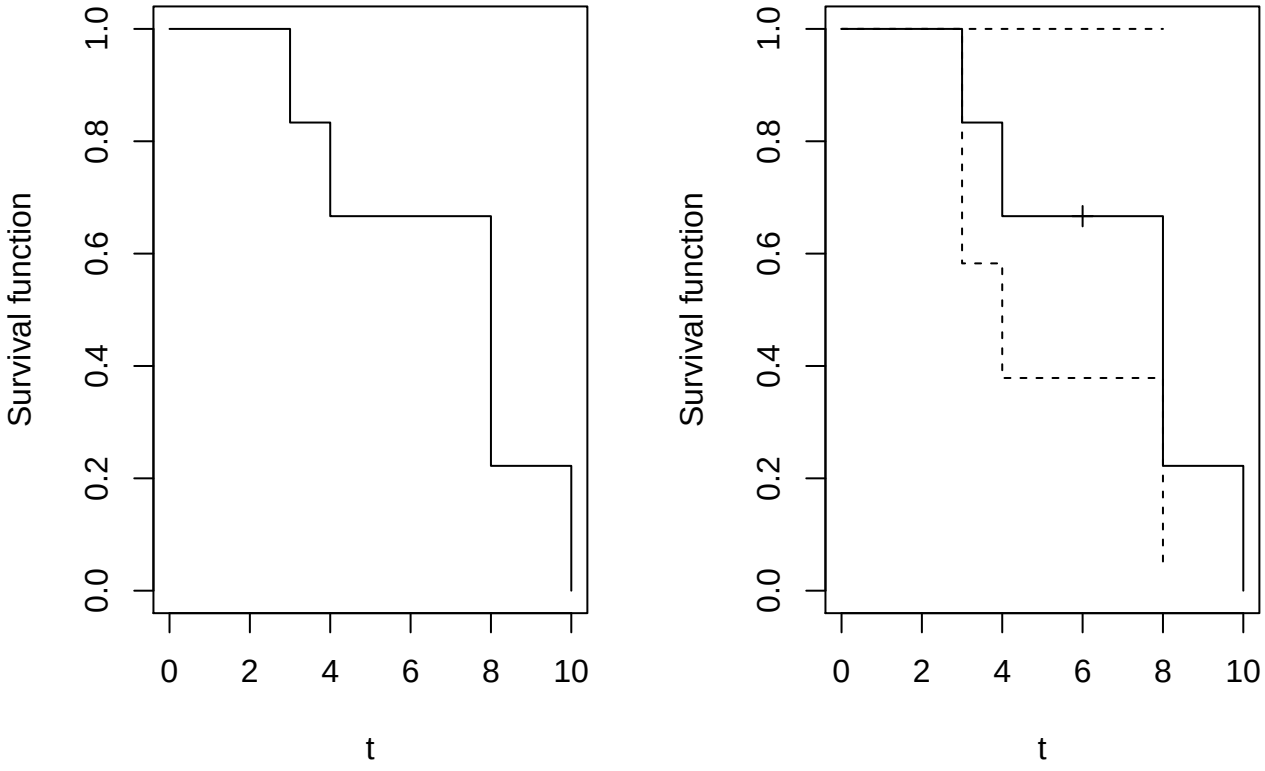
i	t_i	t_{i+1}	r_i	d'_i	c_i	$\hat{S}(t)$
0	0	3	6	0	0	1
1	3	4	6	1	0	5/6
2	4	8	5	1	1	2/3
3	8	10	3	2	0	2/9
4	10	∞	1	1	0	0



6.2.3 Calculating the Kaplan-Meier estimate using R

The R code below calculates the Kaplan-Meier curve for the times in Section 6.2.2.

```
> time<-c(3, 4, 6, 8, 8, 10)
> cens<-c(1,1,0,1,1,1)
>
> km <- survfit(Surv(time, cens)~ 1)
>
> par(mfrow=c(1,2))
> plot(km, conf.int=FALSE, xlab="t", ylab="Survival function")
> plot(km, mark.time=TRUE, xlab="t", ylab="Survival function")
```



By default the plot will include confidence intervals. In the left hand plot these have been turned off by `conf.int = TRUE`. In the right hand plot, `mark.time=TRUE` will indicate censored times with a cross.

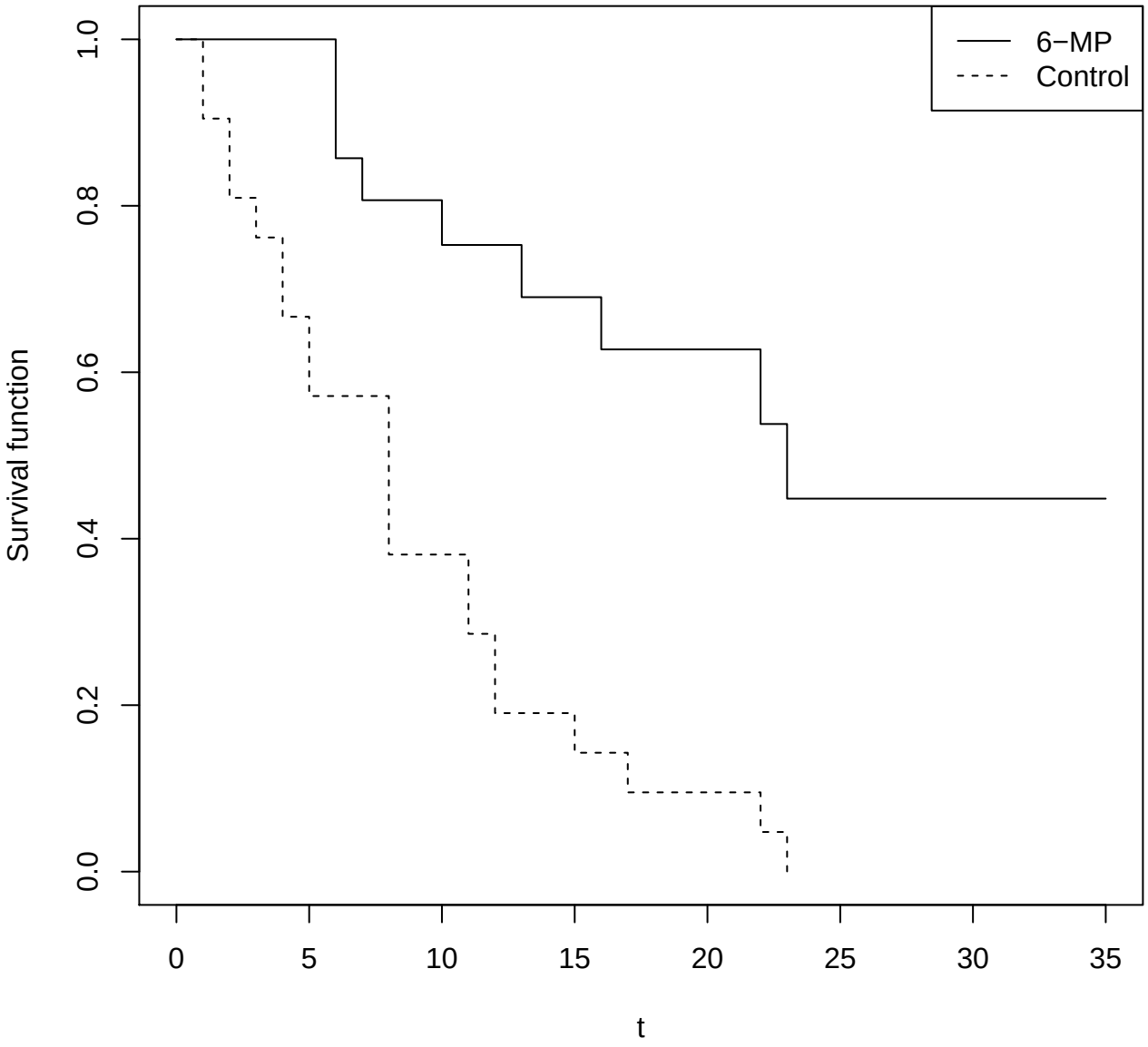
The below code plots survival curves for the two groups (control and 6-MP) in the `gehan` data.

```
> library(MASS)
>
> head(gehan)
```

```
  pair time cens   treat
1    1    1    1 control
2    1   10    1    6-MP
3    2   22    1 control
4    2    7    1    6-MP
5    3    3    1 control
6    3   32    0    6-MP
```

```
> gehan.km <- survfit(Surv(time, cens)~ treat, data = gehan)
```

```
> plot(gehan.km, xlab="t", ylab="Survival function", lty=c(1,2))
> legend("topright", lty=c(1,2), legend = c("6-MP","Control"))
```



6.2.4 Standard errors and confidence intervals

The standard error of the Kaplan-Meier estimator is given by *Greenwood's formula*

$$s.e.[\hat{S}(t)]^2 = \hat{S}(t)^2 \sum_{j=1}^{i} \frac{d'_j}{r_j(r_j - d'_j)} \qquad t \in [t'_i, t'_{i+1}), \quad i = 0, \ldots, m$$

Using a normal approximation for the sampling distribution of $\hat{S}(t)$, we can obtain endpoints of an approximate $(1-\alpha)\%$ confidence interval for $S(t)$ as

$$\hat{S}(t) \pm z_{1-\alpha/2} s.e.[\hat{S}(t)]$$

Better confidence intervals are often obtained on the log scale, using

$$s.e.[\log \hat{S}(t)]^2 = \sum_{j=1}^{i} \frac{d'_j}{r_j(r_j - d'_j)} \qquad t \in [t'_i, t'_{i+1}), \quad i = 0, \ldots, m$$

to give endpoints of an alternative 95% confidence interval as

$$\hat{S}(t) \exp(\pm z_{1-\alpha/2} s.e.[\log \hat{S}(t)])$$

6.3 The Nelson-Aalen estimator

The *Nelson-Aalen* estimator estimates the cumulative hazard function $H(t)$ by integrating (summing) the discrete hazard estimators $\hat{h}_i = d'_i/r_i$, and is given by

$$\hat{H}(t) = \sum_{j=1}^{i} \hat{h}_j = \sum_{j=1}^{i} \frac{d'_j}{r_j} \qquad t \in [t'_i, t'_{i+1}), \quad i = 0, \ldots, m$$

As $S(t) = \exp\{-H(t)\}$, the Nelson-Aalen estimator provides an estimator of $S(t)$ as

$$\hat{S}(t) = \exp\left(-\sum_{j=1}^{i} \frac{d'_j}{r_j}\right) = \prod_{j=1}^{i} \exp\left(-\frac{d'_j}{r_j}\right) \qquad t \in [t'_i, t'_{i+1}), \quad i = 0, \ldots, m$$

which is sometimes called the *Fleming-Harrington* estimator.

6.3.1 Notes on the Nelson-Aalen estimator

Plotting $\log \hat{H}(t)$ against $\log t$ checks whether a Weibull model might fit the data, as for the Weibull,

$$H(t) = (\beta t)^{\alpha} \qquad \Rightarrow \qquad \log H(t) = \alpha \log \beta + \alpha \log t$$

so for Weibull data we would expect $\log \hat{H}(t)$ to be linear in $\log t$, with gradient α.

If d_i'/r_i is small (typical for smaller values of t) then the Kaplan-Meier and Fleming-Harrington estimators of $S(t)$ will be close, as

$$\begin{aligned} exp\left(-\frac{d_j'}{r_j}\right) &= 1 - \frac{d_j'}{r_j} + \frac{1}{2}\left(\frac{d_j'}{r_j}\right)^2 - \cdots \\ &\approx 1 - \frac{d_j'}{r_j} \end{aligned}$$

Chapter 7

Survival Regression Models

Commonly, when we observe (possibly censored) survival times $t_1, \ldots, t_n$, we also observe the values of k other variables, $x_1, \ldots, x_k$, for each of the n units of observation.

Then, we drop the assumption that the survival time variables $T_1, \ldots, T_n$ are identically distributed, and investigate how their distribution depends on the *explanatory* variables (or *covariates*) $x_1, \ldots, x_k$.

In a *regression* model, we assume that the dependence of the distribution of T_i on the values of $x_1, \ldots, x_k$ is through a regression function, which is typically assumed to have linear structure, as

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \sum_{j=1}^{k} x_{ij}\beta_j = \mathbf{x}_i^T \boldsymbol{\beta}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T$ is the k-vector containing the values of $x_1, \ldots, x_k$ for unit i and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)^T$ is a k-vector of regression parameters.

7.1 Example: `leuk` data

Survival times for $n = 33$ patients who died from acute myelogenous leukaemia.

```
> library(MASS)
> head(leuk)
```

```
    wbc      ag time
1  2300 present   65
2   750 present  156
3  4300 present  100
4  2600 present  134
5  6000 present   16
6 10500 present  108
```

Here, we have two potential explanatory variables, *wbc* and *ag*. As *ag* is a factor (a non-numerical variable) we transform it in the regression function to an indicator (dummy) variable or variables, for example

$$I(\text{ag} = \text{“present”}) = \begin{cases} 1 & \text{if ag} = \text{“present”} \\ 0 & \text{if ag} = \text{“absent”} \end{cases}$$

Similarly further explanatory variables may be created from the numeric covariate *wbc*, e.g. wbc^2 to investigate possible quadratic dependence.

7.2 Proportional hazards

As was the case with homogeneous survival models in Chapter 6, we frequently want to investigate the dependence of T_i on $x_1, \ldots, x_k$ *without* assuming a particular parametric family for f_{T_i}.

A *proportional hazards* model (or Cox regression model) assumes that the hazard function for the survival variable corresponding to the ith unit is

$$h_{T_i}(t) = \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right) h_0(t)$$

where h_0 is called the *baseline hazard function* and is not assumed to have a particular mathematical form.

Hence, we model the effect of the explanatory variables on the distribution of T_i as multiplying the baseline hazard by the exponentiated regression function, $\exp \eta_i$.

Equivalently, using the table from page 23

$$\begin{aligned} S_{T_i}(t) &= \exp\left[-\int_0^t h_{T_i}(s)\mathrm{d}s\right] \\ &= \exp\left[-\int_0^t \exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right) h_0(s)\mathrm{d}s\right] \\ &= \exp\left[-\int_0^t h_0(s)\mathrm{d}s\right]^{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)} \\ &= S_0(t)^{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)} \end{aligned}$$

where

$$S_0(t) = \exp\left[-\int_0^t h_0(s)\mathrm{d}s\right]$$

is the baseline survivor function.

7.3 Partial likelihood

We require to estimate the regression parameters $\boldsymbol{\beta}$ without specifying (or even estimating) the baseline hazard h_0 (so we cannot write down a likelihood, as f_{T_i} and S_{T_i} are not fully specified).

The *partial likelihood function* is the joint probability of the observed data *conditional on the observed failure times* $\{t_i : d_i = 1\}$.

Conditioning on $\{t_i : d_i = 1\}$, the partial likelihood reduces to the probability of the failures occurring in the order in which they were observed.

Let R_i be the *risk set* at t_i, that is

$$R_i = \{j : t_j \geq t_i\}$$

the units with failure or censoring times at, or after, t_i. Then,

$$P(T_j \in [t_i, t_i + \delta t] \,|\, j \in R_i) = h_{T_j}(t_i)\delta t = \exp\left(\mathbf{x}_j^T \boldsymbol{\beta}\right) h_0(t_i)\delta t$$

for some (infinitesimally small) δt.

Therefore

$$P(\text{some } j \in R_i \text{ fails in } [t_i, t_i + \delta t]) = \sum_{j \in R_i} \exp\left(\mathbf{x}_j^T \boldsymbol{\beta}\right) h_0(t_i)\delta t + O(\delta t^2)$$

Hence,

$$
\begin{aligned}
&P(T_i \in [t_i, t_i + \delta t] \,|\, \text{some } j \in R_i \text{ fails in } [t_i, t_i + \delta t]) \\
&\qquad = \frac{\exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right) h_0(t_i)\delta t}{\sum_{j \in R_i} \exp\left(\mathbf{x}_j^T \boldsymbol{\beta}\right) h_0(t_i)\delta t + O(\delta t^2)} \\
&\qquad = \frac{\exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right)}{\sum_{j \in R_i} \exp\left(\mathbf{x}_j^T \boldsymbol{\beta}\right) + O(\delta t)}
\end{aligned}
$$

and in the limit $\delta t \to 0$,

$$
P(i \text{ fails at } t_i \,|\, \text{some } j \in R_i \text{ fails at } t_i) = \frac{\exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right)}{\sum_{j \in R_i} \exp\left(\mathbf{x}_j^T \boldsymbol{\beta}\right)}
$$

The partial likelihood is therefore given by

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= P(\text{failures occurred in observed order}) \\
&= \prod_{i:d_i=1} P(i \text{ fails at } t_i \,|\, \text{some } j \in R_i \text{ fails at } t_i) \\
&= \prod_{i:d_i=1} \frac{\exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right)}{\sum_{j \in R_i} \exp\left(\mathbf{x}_j^T \boldsymbol{\beta}\right)}
\end{aligned}
$$

which does not depend on the baseline hazard h_0.

7.4 Tied failure times

Suppose two observations (for example $i = 1$ and $i = 2$) have the same recorded failure time. Our definition of $R_i = \{j : t_j \geq t_i\}$ includes $1 \in R_2$ and $2 \in R_1$, both of which cannot be true

Different partial likelihoods arise from assuming

- 1 failed just before 2 (so $2 \in R_1$ but $1 \notin R_2$)
- 2 failed just before 1 (so $2 \notin R_1$ but $1 \in R_2$).

The correct action is to average over these two assumptions, so we replace the contributions of units $i = 1$ and $i = 2$ to the partial likelihood with

$$\frac{\exp\left(\mathbf{x}_1^T\boldsymbol{\beta}\right)}{\sum_{j\in R_1}\exp\left(\mathbf{x}_j^T\boldsymbol{\beta}\right)}\frac{\exp\left(\mathbf{x}_2^T\boldsymbol{\beta}\right)}{\sum_{j\in R_2'}\exp\left(\mathbf{x}_j^T\boldsymbol{\beta}\right)} + \frac{\exp\left(\mathbf{x}_2^T\boldsymbol{\beta}\right)}{\sum_{j\in R_2}\exp\left(\mathbf{x}_j^T\boldsymbol{\beta}\right)}\frac{\exp\left(\mathbf{x}_1^T\boldsymbol{\beta}\right)}{\sum_{j\in R_1'}\exp\left(\mathbf{x}_j^T\boldsymbol{\beta}\right)}$$

where $R_1' = R_1 \setminus \{2\}$ and $R_2' = R_2 \setminus \{1\}$.

The exact partial likelihood for tied failures times can be complicated if we have $m > 2$ tied failure times ($m!$ terms in the average). There exist approximations to the likelihood (e.g. Breslow and Efron) in the case of tied failure times.

7.5 Estimation

We estimate the regression parameters $\boldsymbol{\beta}$ using the values $\hat{\boldsymbol{\beta}}$ which maximise the partial likelihood $L(\boldsymbol{\beta})$, or partial log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i:d_i=1} \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i:d_i=1} \log \left[\sum_{j \in R_i} \exp\left(\mathbf{x}_j^T \boldsymbol{\beta}\right) \right]$$

Maximisation is performed by solving (numerically) the simultaneous equations

$$\frac{\partial}{\partial \beta_i} \ell(\boldsymbol{\beta}) = 0, \qquad i = 1, \ldots, k.$$

We also obtain

$$Var(\hat{\beta}_i) \approx [I(\boldsymbol{\beta})^{-1}]_{ii}$$

where $I(\boldsymbol{\beta})$ is the *observed information matrix* defined by

$$I(\boldsymbol{\beta})_{ij} = -\frac{\partial^2}{\partial \beta_i \partial \beta_j} \ell(\boldsymbol{\beta}).$$

In practice, we rely on computer packages to compute estimates.

7.6 Confidence intervals

As with maximum likelihood, maximum partial likelihood estimators have an asymptotic (as $n \to \infty$) normal distribution, so

$$\frac{\hat{\beta}_i - \beta_i}{s.e.(\hat{\beta}_i)} \sim N(0,1)$$

can be used as an approximation in moderately large samples, where

$$s.e.(\hat{\beta}_i)^2 = [I(\hat{\boldsymbol{\beta}})^{-1}]_{ii}$$

is an estimate of the (asymptotic) standard deviation of $\hat{\beta}_i$.

Hence, a $(1-\alpha)\%$ confidence interval for β_i is

$$[\hat{\beta}_i - z_{1-\alpha/2} s.e.(\hat{\beta}_i), \hat{\beta}_i + z_{1-\alpha/2} s.e.(\hat{\beta}_i)]$$

In practice, we rely on computer packages to compute standard errors.

7.7 Hypothesis testing

In regression models, the hypothesis $H_0 : \beta_j = 0$ is equivalent to a model where the explanatory variable x_j has no effect on survival (as its values are omitted from the regression function $\mathbf{x}_i^T \boldsymbol{\beta}$).

We test this hypothesis, by noting that, under $H_0 : \beta_j = 0$, we have the approximation

$$\frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \sim N(0,1).$$

So $\frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$ is a test statistic whose values can be calibrated against a standard normal distribution. For a

$\alpha\%$ significance level, we reject H_0 when

$$\left| \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \right| > z_{1-\alpha/2}.$$

We can also calculate a p-value using

$$\text{p-value} = 2\left(1 - \text{P}\left(Z \leq \left| \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \right|\right)\right).$$

This is the *Wald* test.

7.8 Estimating the baseline $h_0(t)$, $H_0(t)$ and $S_0(t)$

It is not necessary to estimate h_0 to answer interesting questions about which covariates affect survival time, and how they do so.

However, as in the homogeneous model, we can estimate the complete survival distribution nonparametrically, by estimating the hazard function as if the underlying process was discrete.

An estimate of the baseline hazard is given by

$$\hat{h}_0(t_i) = \frac{d'_i}{\sum_{j \in R_i} \exp(\mathbf{x}_j^T \hat{\boldsymbol{\beta}})},$$

where d'_i is the total number of failures observed at t_i.

As in the homogeneous case in Chapter 6, the discrete hazard estimates can be transformed into an estimate of the cumulative hazard or survival functions. Again, we let $t'_1, \ldots, t'_m$ be the m ordered distinct failure times with corresponding numbers of failures $d'_1, \ldots, d'_m$. Then

$$\hat{H}_0(t) = \sum_{j=1}^{i} \hat{h}_0(t'_j) = \sum_{j=1}^{i} \frac{d'_j}{\sum_{k:t_k \geq t'_j} \exp(\mathbf{x}_k^T \hat{\boldsymbol{\beta}})} \quad t \in [t'_i, t'_{i+1}), \quad i = 0, \ldots, m$$

(like the Nelson-Aalen estimator in the homogeneous case) and

$$\hat{S}_0(t) = \prod_{j=1}^{i} \exp\left(-\frac{d'_j}{\sum_{k:t_k \geq t'_j} \exp(\mathbf{x}_k^T \hat{\boldsymbol{\beta}})}\right) \qquad t \in [t'_i, t'_{i+1}), \quad i = 0, \ldots, m.$$

which implies

$$\hat{S}_{T_i}(t) = \hat{S}_0(t)^{\exp\left(\mathbf{x}_i^T \hat{\beta}\right)}.$$

7.9 Fitting Cox models using R

7.9.1 gehan data

The R code below fits a Cox model with a dummy variable for treatment.

```
> gehan.cox <- coxph(Surv(time,event=cens)~treat,data=gehan)
> summary(gehan.cox)


Call:
coxph(formula = Surv(time, event = cens) ~ treat, data = gehan)

  n= 42, number of events= 30

               coef exp(coef)
treatcontrol 1.5721    4.8169
             se(coef)     z
treatcontrol   0.4124 3.812
             Pr(>|z|)
treatcontrol 0.000138 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*'
  0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef)
treatcontrol     4.817     0.2076
             lower .95 upper .95
treatcontrol     2.147     10.81
```

```
Concordance= 0.69  (se = 0.041 )
Likelihood ratio test= 16.35  on 1 df,   p=5e-05
Wald test            = 14.53  on 1 df,   p=1e-04
Score (logrank) test = 17.25  on 1 df,   p=3e-05
```

The p-value associated with the treatment dummy variable is 1.4×10^{-4}, i.e. the treatment significantly affects survival time.

The estimated coefficient for treatment is 1.57213. This means the control treatment is estimated to increase the hazard by a factor of

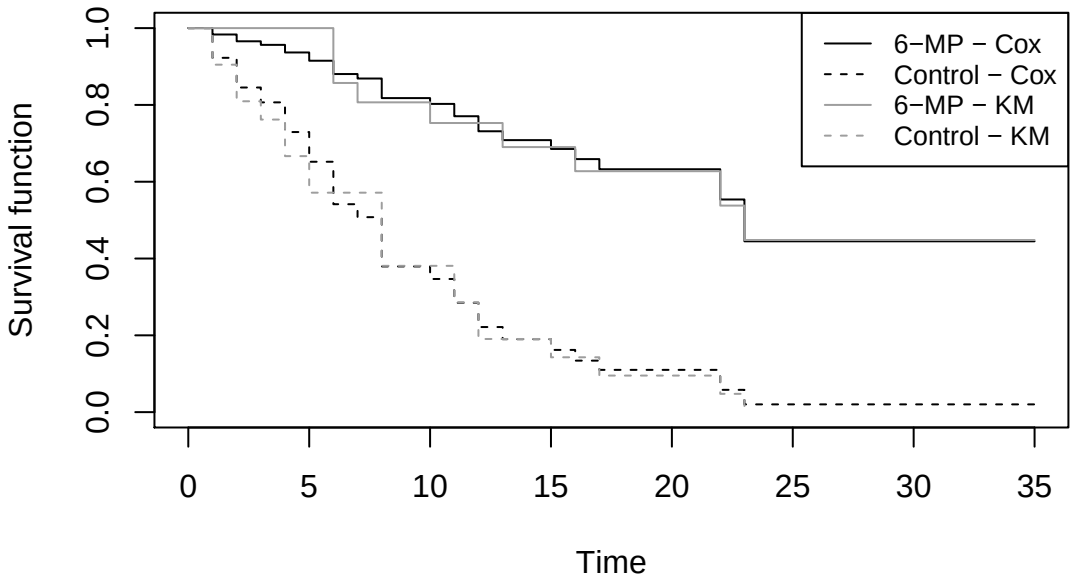
$$\exp(1.57213) = 4.81687$$

relative to the 6-MP treatment. This estimate can be seen in the R output above (along with a 95% confidence interval).

The R code below will create survival curves. The argument `newdata` allows us to specify values of the explanatory variables. In this case there is one explanatory variable which can take two values. So we compute two survival curves for those two values. For comparison, the Kaplan Meier survival curves are also plotted in grey.

```
> gehan.S <- survfit(gehan.cox,
+ newdata = data.frame(treat = c("control","6-MP")))
> plot(gehan.S, ylab= "Survival function", xlab="Time", lty = c(2,1))
> lines(gehan.km, lty=c(1,2), col=8)
> legend("topright", lty = c(1,2,1,2), col = c(1,1,8,8),
```

```
+ legend = c("6-MP - Cox","Control - Cox","6-MP - KM","Control - KM"), cex=0.8)
```



7.9.2 `leuk` data

The R code below fits a Cox model with two explanatory variables: a dummy variable for *ag* and a variable given by the natural logarithm of *wbc*.

```
> leuk.cox <- coxph(Surv(time)~ag+log(wbc),data=leuk)
> summary(leuk.cox)


Call:
coxph(formula = Surv(time) ~ ag + log(wbc), data = leuk)

  n= 33, number of events= 33

             coef exp(coef)
agpresent -1.0691    0.3433
log(wbc)   0.3677    1.4444
          se(coef)      z
agpresent   0.4293 -2.490
log(wbc)    0.1360  2.703
```

```
          Pr(>|z|)
agpresent  0.01276 *
log(wbc)   0.00687 **
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*'
  0.05 '.' 0.1 ' ' 1

          exp(coef) exp(-coef)
agpresent    0.3433     2.9126
log(wbc)     1.4444     0.6923
          lower .95 upper .95
agpresent     0.148    0.7964
log(wbc)      1.106    1.8857

Concordance= 0.726  (se = 0.047 )
Likelihood ratio test= 15.64  on 2 df,   p=4e-04
Wald test            = 15.06  on 2 df,   p=5e-04
Score (logrank) test = 16.49  on 2 df,   p=3e-04
```

The significance of both explanatory variables is immediately obvious from the two p-values 0.01276 and 0.00687. Strictly we should formal model selection using either backwards or forwards selection, or using information criteria (see MATH2010).

The estimated coefficient for ag is -1.06905. This means ag being present is estimated to increase the hazard by a factor of

$$\exp(-1.06905) = 0.34333,$$

relative to ag being absent, i.e. ag being present reduces the hazard.

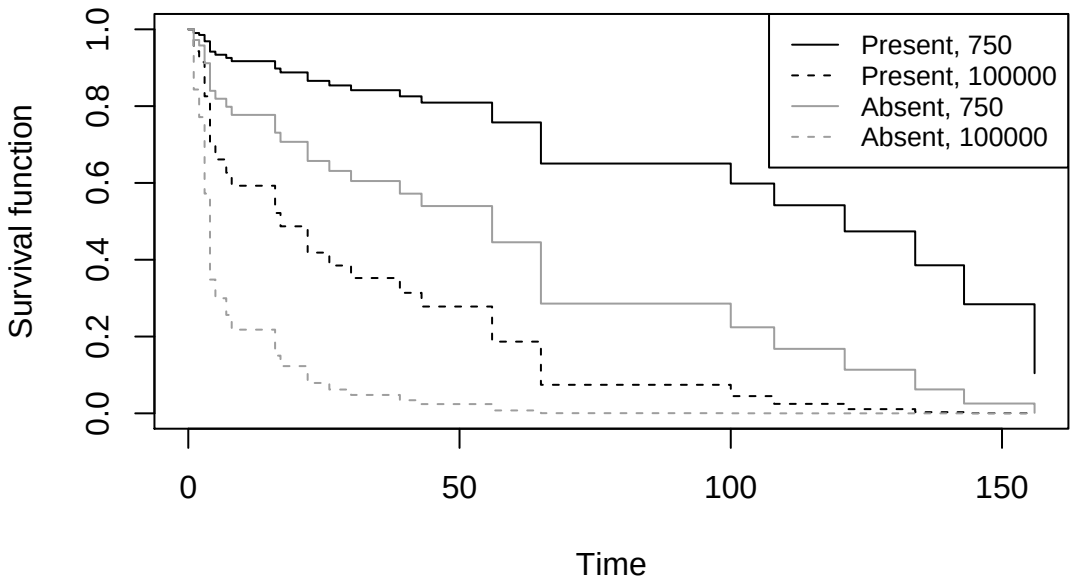
The estimated coefficient for $log(wbc)$ is 0.3677. This $\log(wbc)$ increasing by one unit is estimated to

increase the hazard by a factor of

$$\exp(0.3677) = 1.44441,$$

i.e. the higher the *wbc*, the higher the hazard.

```
> leuk.S <- survfit(leuk.cox,
+ newdata = data.frame(ag = c("present","present","absent","absent"),
+ wbc = c(750,100000,750,100000)))
> plot(leuk.S, ylab= "Survival function", xlab="Time", lty = c(1,2,1,2),col=c(1,1,8,8))
> legend("topright", lty = c(1,2,1,2), col = c(1,1,8,8),
+ legend = c("Present, 750", "Present, 100000","Absent, 750", "Absent, 100000"), cex=0.8)
```



7.10 Checking proportional hazards

The assumption that covariates affect the response through proportional hazards is a strong one and should be checked where possible.

Recall that the survival function for T_i can be written as

$$S_{T_i}(t) = S_0(t)^{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}.$$

Taking the complementary log-log of both sides

$$\begin{aligned} \log\left(-\log S_{T_i}(t)\right) &= \log\left(-\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\log S_0(t)\right) \\ &= \mathbf{x}_i^T\boldsymbol{\beta} + \log\left(-\log S_0(t)\right). \end{aligned}$$

Consider units $i = 1$ and $i = 2$ who differ only in the jth covariate. Then

$$\begin{aligned} \mathbf{x}_1^T\boldsymbol{\beta} &= \sum_{l=1}^{k} x_{1l}\beta_l \\ &= x_{1j}\beta_j + \sum_{l\neq j} x_{1l}\beta_l \\ \mathbf{x}_2^T\boldsymbol{\beta} &= x_{2j}\beta_j + \sum_{l\neq j} x_{2l}\beta_l \end{aligned}$$

and

$$\begin{aligned} \mathbf{x}_1^T\boldsymbol{\beta} - \mathbf{x}_2^T\boldsymbol{\beta} &= x_{1j}\beta_j + \sum_{l\neq j} x_{1l}\beta_l - x_{2j}\beta_j - \sum_{l\neq j} x_{2l}\beta_l \\ &= (x_{1j} - x_{2j})\beta_j \end{aligned}$$

since $\sum_{l \neq j} x_{1l}\beta_l = \sum_{l \neq j} x_{2l}\beta_l$.

Now

$$\log\left(-\log S_{T_1}(t)\right) - \log\left(-\log S_{T_2}(t)\right) = (x_{1j} - x_{2j})\beta_j$$

which does not depend on t. That means if we were to plot $\log\left(-\log S_{T_1}(t)\right)$ and $\log\left(-\log S_{T_2}(t)\right)$ against t we should get two parallel lines (if the assumption of proportional hazards is correct).

Therefore, a strategy for checking proportional hazards for covariate x_j is

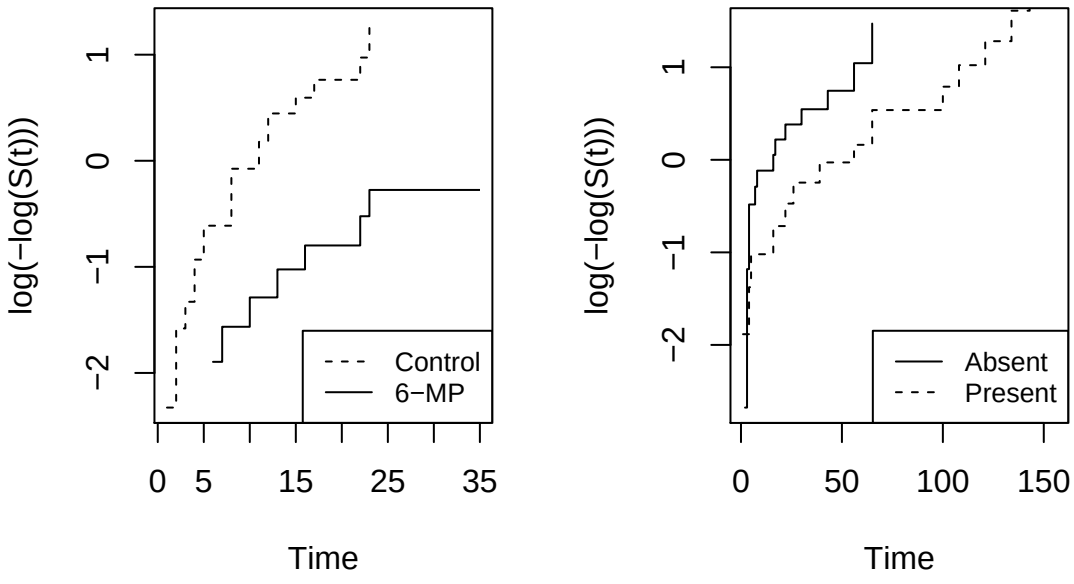
- partition the data according to each distinct x_j value observed
- fit a separate proportional hazards model (including all other covariates) to each subset
- the estimates of $\log\left(-\log S_T(t)\right)$ for those separate models will be parallel if the proportional hazards assumption holds for x_j.

Clearly this is only practical if x_j has few distinct values, each appearing several times (typically if x_j is a factor).

The following `R` code produces these plots for the `gehan` and `leuk` datasets, where x_j corresponds to `treat` and `ag`, respectively.

```
> logmlog<-function(x){
+ log(-log(x))}
>
```

```
> par(mfrow = c(1,2))
>
> gehan.cox1 <- coxph(Surv(time,event=cens)~1, data = gehan[gehan$treat=="control", ])
> gehan.cox2 <- coxph(Surv(time,event=cens)~1, data = gehan[gehan$treat=="6-MP", ])
>
> gehan.S1<- survfit(gehan.cox1)
> gehan.S2<- survfit(gehan.cox2)
> plot(gehan.S1, ylab= "log(-log(S(t)))", xlab = "Time", fun = logmlog,
+ conf.int = FALSE, xlim = range(gehan$time), lty = 2)
> lines(gehan.S2, fun=logmlog, conf.int = FALSE, lty = 1)
> legend("bottomright", lty = c(2,1), legend = c("Control","6-MP"), cex = 0.8)
>
> leuk.cox1 <- coxph(Surv(time) ~ log(wbc), data = leuk[leuk$ag == "absent", ])
> leuk.cox2 <- coxph(Surv(time) ~ log(wbc), data = leuk[leuk$ag == "present", ])
>
> leuk.S1<- survfit(leuk.cox1, newdata = data.frame(wbc = mean(leuk$wbc)))
> leuk.S2<- survfit(leuk.cox2, newdata = data.frame(wbc = mean(leuk$wbc)))
>
> plot(leuk.S1, ylab= "log(-log(S(t)))", xlab = "Time", fun = logmlog,
+ conf.int = FALSE, xlim = range(leuk$time), lty = 1)
> lines(leuk.S2, fun=logmlog, conf.int = FALSE, lty = 2)
> legend("bottomright", lty = c(1,2), legend = c("Absent","Present"), cex = 0.8)
```



7.11 Accelerated failure and parametric models

We have focused on *semiparametric* models, where the parameters $\boldsymbol{\beta}$ do not completely specify the survival distribution.

Fully parametric survival regression models do exist. For example, consider a Weibull model, with $T_i \sim \text{Weibull}(\alpha, \theta_i)$. (We have changed the scale parameter to θ to stop confusion with the $\boldsymbol{\beta}$ regression parameters.) The shape parameter α is the same for all units but the scale parameter θ_i is unit-specific depending on the explanatory variables as follows

$$\begin{aligned}
\theta_i &= 1/\exp(\eta_i) \\
&= \exp(-\eta_i) \\
&= \exp\left(-\beta_0 - \mathbf{x}_i^T\boldsymbol{\beta}\right) \\
&= \exp\left(-\beta_0 - \sum_{j=1} x_{ij}\beta_j\right).
\end{aligned}$$

Now suppose $T_{0i} \sim \text{Weibull}(\alpha, 1)$, then

$$T_i = \exp(\eta_i)\, T_{0i}.$$

which can be shown using the properties in Section 4.2.1. The survival functions for T_{0i} and T_i are

$$S_{T_{0i}}(t) = \exp(-t^\alpha) \qquad S_{T_i}(t) = \exp\left(-\left(\exp(-\eta_i)\, t\right)^\alpha\right).$$

We see that the effect on the distribution of T_i of the explanatory variables, through η_i is to shrink (or stretch) the time axis, resulting in correspondingly longer (shorter) failure times. ["one dog year = seven human years"]

This is called an *accelerated failure time* (AFT) model (or accelerated life model), and is generally expressed through the survival function

$$S_{T_i}(t) = S_0\left(t \exp(-\eta_i)\right)$$

for some baseline survival function S_0. Although it is possible to consider semiparametric models (with S_0 not from a standard family) it is more usual to use fully parametric accelerated failure models.

7.11.1 Families of accelerated failure models

$$T_i = \exp(\eta_i) T_{0i} \qquad \text{where} \qquad \eta_i = \beta_0 + \mathbf{x}_i^T \beta$$

T_{0i}	T_i
$\exp(1)$	$\exp(\exp[-\eta_i])$
$\text{Weibull}(\alpha, 1)$	$\text{Weibull}(\alpha, \exp[-\eta_i])$
$\text{loglogistic}(\alpha, 1)$	$\text{loglogistic}(\alpha, \exp[-\eta_i])$
$\text{lognormal}(0, \sigma^2)$	$\text{lognormal}(\eta_i, \sigma^2)$

7.12 Estimating accelerated failure models

For any parametric AFT model, we know the form of S_{T_i} and f_{T_i}, so we can write down the likelihood

$$L(\theta) = \prod_{i:d_i=1} f_{T_i}(t_i;\theta) \prod_{i:d_i=0} S_{T_i}(t_i;\theta)$$

where θ comprises the regression parameters β_0 and $\boldsymbol{\beta}$ which enter into the computation of η_i and any other parameters, such as α.

Estimation is performed by maximum likelihood, to obtain $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$ etc. with standard errors computed in the usual way.

Maximisation must be done numerically (e.g. using R)

7.12.1 Example: Weibull ART model for the gehan data

The R code below fits a Weibull ART model with a dummy variable for treatment.

```
> gehan.wb <- survreg(Surv(time,event=cens)~treat,data=gehan)
> summary(gehan.wb)



Call:
survreg(formula = Surv(time, event = cens) ~ treat, data = gehan)
              Value Std. Error
(Intercept)   3.516      0.252
treatcontrol -1.267      0.311
Log(scale)   -0.312      0.147
                 z       p
(Intercept)  13.96 < 2e-16
treatcontrol -4.08 4.5e-05
Log(scale)   -2.12   0.034

Scale= 0.732

Weibull distribution
Loglik(model)= -106.6   Loglik(intercept only)= -116.4
        Chisq= 19.65 on 1 degrees of freedom, p= 9.3e-06
```

```
Number of Newton-Raphson Iterations: 5
n= 42
```

Confusingly, R uses the following parameterisation for the shape parameter:

$$\hat{\alpha} = \frac{1}{\texttt{scale}}.$$

The estimates are

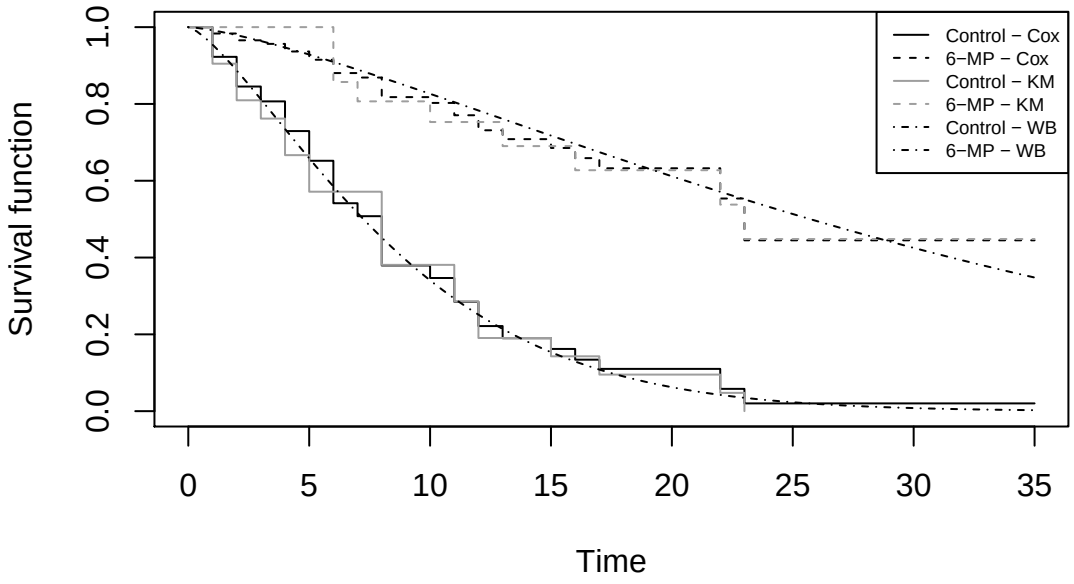
$$\begin{aligned} \hat{\beta}_0 &= 3.516 \\ \hat{\beta}_1 &= -1.267 \\ \hat{\alpha} &= 1.366. \end{aligned}$$

The estimate of β_1 is negative meaning the treatment control will increase the scale parameter θ and increase the hazard (all relative to the 6-MP treatment).

The following R code plots the estimated Kaplan-Meier and Cox survival curves. It thens adds estimated survival curves from the Weibull model using the curve function.

```
> plot(gehan.S, ylab= "Survival function", xlab="Time", lty = c(1,2))
> lines(gehan.km, lty=c(2,1), col=8)
>
> alpha<-1/gehan.wb$scale
```

```
> theta.control<-exp(-gehan.wb$coefficients[1]-gehan.wb$coefficients[2])
> theta.6MP<-exp(-gehan.wb$coefficients[1])
>
> curve(expr = exp(-(theta.control*x)^alpha), from = 0, to = 35, add = TRUE, lty=4)
> curve(expr = exp(-(theta.6MP*x)^alpha), from = 0, to = 35, add = TRUE, lty = 4)
>
> legend("topright", lty = c(1,2,1,2,4,4), col = c(1,1,8,8,1,1),
+ legend = c("Control - Cox","6-MP - Cox","Control - KM",
+ "6-MP - KM","Control - WB","6-MP - WB"),
+ cex = 0.6)
```



7.12.2 Example: Weibull ART model for the `leuk` data

The `R` code below fits a Weibull ART model with a dummy variable for treatment.

```
> leuk.wb <- survreg(Surv(time)~ag + log(wbc), data = leuk)
> summary(leuk.wb)
```

```
Call:
survreg(formula = Surv(time) ~ ag + log(wbc), data = leuk)
```

```
              Value Std. Error
(Intercept)  5.8524     1.3227
agpresent    1.0206     0.3781
log(wbc)    -0.3103     0.1313
Log(scale)   0.0399     0.1392
                z       p
(Intercept)  4.42 9.7e-06
agpresent    2.70  0.0069
log(wbc)    -2.36  0.0181
Log(scale)   0.29  0.7745

Scale= 1.04

Weibull distribution
Loglik(model)= -146.5   Loglik(intercept only)= -153.6
        Chisq= 14.18 on 2 degrees of freedom, p= 0.00084
Number of Newton-Raphson Iterations: 6
n= 33
```

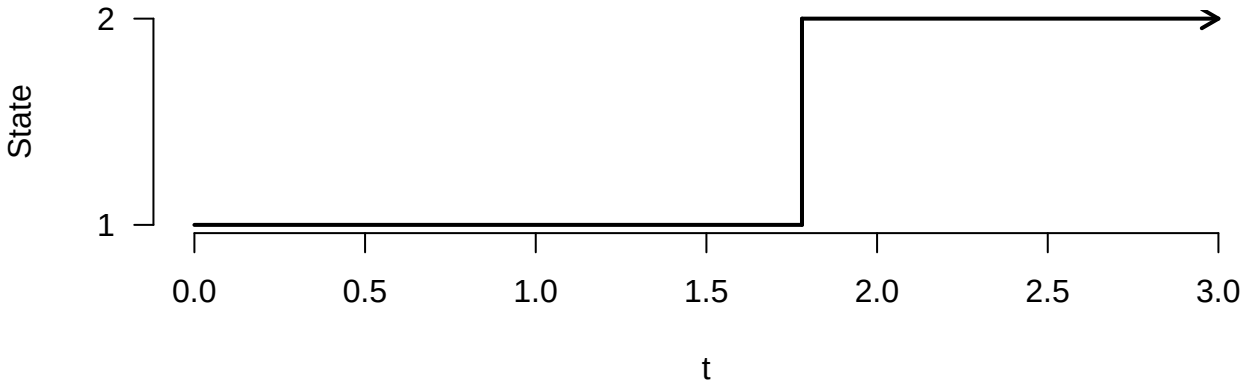
Chapter 8

Multistate Survival Models

Our survival models in Chapters 1 to 7 have assumed that survival (time-to-failure) of each unit is an observation of a random variable T, and our survival models have been stochastic (probability) models describing the distribution of T.

An alternative approach is to model the variable Y_t representing the status (alive or dead) of a unit at time t.

A probability model for a time-indexed variable (one which changes over time) is a *stochastic process*. Survival is a simple stochastic process where, at time $t = 0$, $Y_0 = 1$ (alive) and then the process remains in state 1 unless at some value of t a transition to $Y_t = 2$ (dead) is made after which the process remains in state 2.



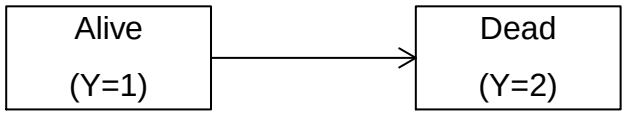
8.1 State space

The set of possible values that a stochastic process Y_t can take over time t is called its state space, S.

In the simple survival model $S = \{\text{alive}, \text{dead}\}$ or $S = \{1, 2\}$.

This is an example of a discrete state space, and in this module we will restrict consideration to discrete state spaces with a small number of possible states.

The simple survival state space can be represented as:



The arrow indicates that transition is possible only in one direction.

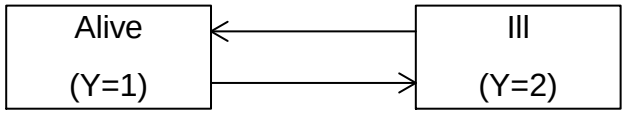
In this case the state dead is called an *absorbing* state.

8.2 Multi-state models

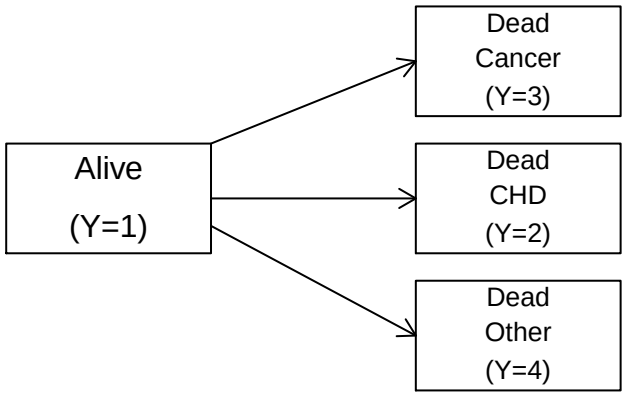
Stochastic processes allow us to model richer survival time processes than simply two-state alive/dead processes with an absorbing (dead) state.

Two such examples are:

General two-state model (no absorbing states)



Multiple decrement (absorbing state) model



8.3 Markov processes

A *Markov process* model for Y_t is one where the future is conditionally independent of the history of the process given the current state. For example, tomorrow's weather only depends on today's weather, and is independent on all weather before today.

Hence, a Markov process can be specified by transition probabilities

$$P(Y_{x+t} = j \,|\, Y_x = i) \;\equiv\; p_{ij}(x,t), \quad j \in S$$

for any present time x, future time $x+t$ and present state i. This makes explicit that the probability, at time x, of future realisations of the process depends only on the current state Y_x, and not on the *history* of Y (its values in the period $[0,t)$).

For a *time-homogeneous* Markov process, the transition probabilities do not depend on the value of x, so we can write

$$P(Y_{x+t} = j \,|\, Y_x = i) \;=\; p_{ij}(t).$$

8.4 Transition intensity

The transition intensity function $\mu_{ij}(x)$ is defined as

$$\mu_{ij}(x) = \lim_{\delta t \to 0} \frac{p_{ij}(x, \delta t)}{\delta t} \qquad i \neq j.$$

It follows that

$$p_{ij}(x, \delta t) = \mu_{ij}(x)\delta t + o(\delta t) \qquad i \neq j.$$

We define

$$\mu_{ii}(x) = -\sum_{j \neq i} \mu_{ij}(x)$$

so that

$$p_{ii}(x, \delta t) = 1 + \mu_{ii}(x)\delta t + o(\delta t) \qquad i \neq j.$$

For any transition $i \to j$ which is prohibited (e.g. dead $\to$ alive), $\mu_{ij} = 0$.

For a homogenous process, transition intensity $\mu_{ij}(x)$ does not depend on x, i.e. it is written μ_{ij}.

8.5 Chapman-Kolmogorov equations: solving a Markov process for $p_{ij}(x, t)$

The transition intensity function $\mu_{ij}(x)$ (or just μ_{ij} for a time-homogeneous process) provides an efficient representation of the process.

How are the transition probabilities $p_{ij}(x, t)$ obtained from the transition intensities $\mu_{ij}(x)$?

We need to derive the *Kolmogorov forward equations*, the solution to which are the required transition probabilities $p_{ij}(x, t)$.

For any times $x, s, t \geq 0$ and states $i, j, k \in S$, we have

$$\begin{aligned} & P(Y_{x+s+t} = j, Y_{x+s} = k \,|\, Y_x = i) \\ & = P(Y_{x+s+t} = j \,|\, Y_{x+s} = k, Y_x = i) P(Y_{x+s} = k \,|\, Y_x = i) \\ & = P(Y_{x+s+t} = j \,|\, Y_{x+s} = k) P(Y_{x+s} = k \,|\, Y_x = i) \end{aligned}$$

Hence

$$\begin{aligned} & P(Y_{x+s+t} = j \,|\, Y_x = i) \\ & \qquad = \sum_{k \in S} P(Y_{x+s+t} = j, Y_{x+s} = k \,|\, Y_x = i) \\ & \qquad = \sum_{k \in S} P(Y_{x+s+t} = j \,|\, Y_{x+s} = k) P(Y_{x+s} = k \,|\, Y_x = i) \end{aligned}$$

or

$$p_{ij}(x, s+t) = \sum_{k \in S} p_{ik}(x, s) p_{kj}(x+s, t) \qquad i, j \in S$$

These are the *Chapman-Kolmogorov equations*.

Suppose that there are m states in S, labelled $1, \ldots, m$.

For a Markov process, we define the transition matrix between times x and $x+t$ to be the matrix

$$P(x,t) = \begin{pmatrix} p_{11}(x,t) & p_{12}(x,t) & \cdots & p_{1m}(x,t) \\ p_{21}(x,t) & p_{22}(x,t) & \cdots & p_{2m}(x,t) \\ \vdots & \vdots & & \vdots \\ p_{m1}(x,t) & p_{m2}(x,t) & \cdots & p_{mm}(x,t) \end{pmatrix}$$

The Chapman-Kolmogorov equations can be written as

$$P(x, s+t) = P(x, s)P(x+s, t)$$

For a time-homogeneous process, we can drop the first argument x in the above.

The Chapman-Kolmogorov equations give

$$p_{ij}(x, t+\delta t) = \sum_{k \in S} p_{ik}(x, t) p_{kj}(x+t, \delta t) \qquad i, j \in S$$

which implies that

$$\begin{aligned}
&\frac{p_{ij}(x, t+\delta t) - p_{ij}(x, t)}{\delta t} \\
&\qquad = \sum_{k \in S} p_{ik}(x, t) \frac{p_{kj}(x+t, \delta t)}{\delta t} - \frac{p_{ij}(x, t)}{\delta t} \\
&\qquad = \sum_{k \neq j} p_{ik}(x, t) \frac{p_{kj}(x+t, \delta t)}{\delta t} - p_{ij}(x, t) \frac{1 - p_{jj}(x+t, \delta t)}{\delta t}
\end{aligned}$$

In the limit $\delta t \to 0$, we have

$$\begin{aligned} \frac{d}{dt}p_{ij}(x,t) &= \sum_{k\neq j} p_{ik}(x,t)\mu_{kj}(x+t) + p_{ij}(x,t)\mu_{jj}(x+t) \\ &= \sum_{k\in S} p_{ik}(x,t)\mu_{kj}(x+t) \qquad i,j \in S \end{aligned}$$

These are the Kolmogorov forward equations, which are a set of differential equations for $p_{ij}(x,t)$.

They can be written in matrix form as $\frac{d}{dt}P(x,t) = P(x,t)M(x+t)$ where $M(x)$ is the transition intensity matrix

$$M(x) = \begin{pmatrix} \mu_{11}(x) & \mu_{12}(x) & \cdots & \mu_{1m}(x) \\ \mu_{21}(x) & \mu_{22}(x) & \cdots & \mu_{2m}(x) \\ \vdots & \vdots & & \vdots \\ \mu_{m1}(x) & \mu_{m2}(x) & \cdots & \mu_{mm}(x) \end{pmatrix}$$

8.6 The holding time distribution

Consider a unit in state i at time x. The holding time T_{xi} is the random variable indicating the length of

the time interval between x and the next transition to any other state. We have

$$\begin{aligned} P(T_{xi} \leq t + \delta t \,|\, T_{xi} \geq t) &= P(Y_{x+t+\delta t} \neq i \,|\, Y_{[x,x+t]} = i) \\ &= 1 - P(Y_{x+t+\delta t} = i \,|\, Y_{x+t} = i) \end{aligned}$$

and therefore

$$\frac{P(T_{xi} \leq t + \delta t \,|\, T_{xi} \geq t)}{\delta t} = \frac{1 - P(Y_{x+t+\delta t} = i \,|\, Y_{x+t} = i)}{\delta t}$$

Taking the limit as $\delta t \to 0$, we have

$$h_{xi}(t) = -\mu_{ii}(x+t) = \sum_{i \neq j} \mu_{ij}(x+t).$$

where h_{xi} is the hazard function for the holding time variable T_{xi}.

The hazard function $h_{xi}(t)$ can be transformed to a survival function in the usual way, so we have

$$\begin{aligned} S_{xi}(t) = \exp\left(-\int_0^t h_{xi}(t)\mathrm{d}t\right) &= \exp\left(\int_0^t \mu_{ii}(x+t)\mathrm{d}t\right) \\ &= \exp\left(-\int_0^t \sum_{i \neq j} \mu_{ij}(x+t)\mathrm{d}t\right) \end{aligned}$$

For a time-homogeneous Markov process, all $\mu_{ij}(x)$ are constant (independent of x) so we have $h_{xi}(t) =$

$-\mu_{ii} = \sum_{j \neq i} \mu_{ij}$, a constant hazard. Hence,

$$S_{xi}(t) = \exp\left(\mu_{ii} t\right) = \exp\left(-t \sum_{j \neq i} \mu_{ij}\right)$$

and the holding time distribution is exponential, with rate $\sum_{i \neq j} \mu_{ij}$.

8.7 Conditional transition probability

Conditional on a transition being made from state i at time $x + t$, what is the probability that transition is made to state j for each $j \neq i$?

$$P(\text{transition from } i \text{ to } j \text{ at } x+t \mid \text{transition from } i \text{ at } x+t)$$

$$= \lim_{\delta t \to 0} P(Y_{x+t+\delta t} = j \mid Y_{x+t+\delta t} \neq i \text{ and } Y_{x+t} = i)$$

$$= \lim_{\delta t \to 0} \frac{P(Y_{x+t+\delta t} = j \mid Y_{x+t} = i)}{P(Y_{x+t+\delta t} \neq i \mid Y_{x+t} = i)}$$

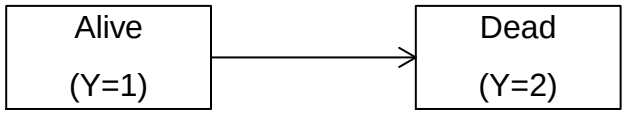
$$= \lim_{\delta t \to 0} \frac{\mu_{ij}(x+t)\delta t + o(\delta t)}{\sum_{k \neq i} \mu_{ik}(x+t)\delta t + o(\delta t)}$$

$$= \frac{\mu_{ij}(x+t)}{\sum_{k \neq i} \mu_{ik}(x+t)}$$

For a time-homogeneous Markov process, all $\mu_{ij}(x)$ are constant (independent of x) so this becomes $\frac{\mu_{ij}}{\sum_{k \neq i} \mu_{ik}}$.

8.8 Examples

8.8.1 Two-state model (absorbing state)



Any two-state process is determined by the transition intensities between the two states, $\mu_{12}(x)$ and $\mu_{21}(x)$.

As state 2 is absorbing,

$$p_{21}(x,t) = 0 \quad \text{for all } x,t \qquad \Rightarrow \qquad \mu_{21}(x) = 0 \quad \text{for all } x$$

Hence the transition intensity matrix is given by

$$M(x) = \begin{pmatrix} -\mu_{12}(x) & \mu_{12}(x) \\ 0 & 0 \end{pmatrix}$$

We have

$$p_{21}(x,t) = 0 \quad \text{for all } x,t \qquad \Rightarrow \qquad p_{22}(x,t) = 1 \quad \text{for all } x,t$$

To completely determine the transition structure for the process, all that remains is to solve the Kolmogorov forward equation for either $p_{11}(x,t)$ or $p_{12}(x,t)$, and then obtain the other using

$$p_{11}(x,t) + p_{12}(x,t) = 1$$

For completeness, we shall solve both and show that they give equivalent results.

The Kolmogorov forward equation for $p_{11}(x,t)$ is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} p_{11}(x,t) &= p_{11}(x,t)\mu_{11}(x+t) + p_{12}(x,t)\mu_{21}(x+t) \\ &= -p_{11}(x,t)\mu_{12}(x+t) \end{aligned}$$

So

$$\int \frac{\mathrm{d}p_{11}(x,t)}{p_{11}(x,t)} = -\int_0^t \mu_{12}(x+t)\mathrm{d}t$$

$$\Rightarrow \quad \log p_{11}(x,t) = -\int_0^t \mu_{12}(x+t)\mathrm{d}t + C$$

$$\Rightarrow \quad p_{11}(x,t) = A\exp\left(-\int_0^t \mu_{12}(x+t)\mathrm{d}t\right)$$

$$\Rightarrow \quad p_{11}(x,t) = \exp\left(-\int_0^t \mu_{12}(x+t)\mathrm{d}t\right)$$

because we have the boundary condition $p_{11}(x,0) = 1$.

The Kolmogorov forward equation for $p_{12}(x,t)$ is

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}p_{12}(x,t) &= p_{11}(x,t)\mu_{12}(x+t) + p_{12}(x,t)\mu_{22}(x+t) \\
&= p_{11}(x,t)\mu_{12}(x+t) \\
&= [1 - p_{12}(x,t)]\mu_{12}(x+t)
\end{aligned}$$

So

$$\int \frac{\mathrm{d}p_{12}(x,t)}{1-p_{12}(x,t)} = \int_0^t \mu_{12}(x+t)\mathrm{d}t$$

$$\Rightarrow \quad -\log[1-p_{12}(x,t)] = \int_0^t \mu_{12}(x+t)\mathrm{d}t + C$$

$$\Rightarrow \quad 1-p_{12}(x,t) = A\exp\left(-\int_0^t \mu_{12}(x+t)\mathrm{d}t\right)$$

$$\Rightarrow \quad p_{12}(x,t) = 1-\exp\left(-\int_0^t \mu_{12}(x+t)\mathrm{d}t\right)$$

because we have the boundary condition $p_{12}(x,0)=0$.

Clearly, the two solutions are equivalent, as we have

$$p_{11}(x,t) = \exp\left(-\int_0^t \mu_{12}(x+t)\mathrm{d}t\right) = 1-p_{12}(x,t).$$

We also note that if we define T_x to be the random variable representing the time (from x) in the alive state (1) before death (transition to state 2), then

$$S_{T_x}(t) = p_{11}(x,t)$$

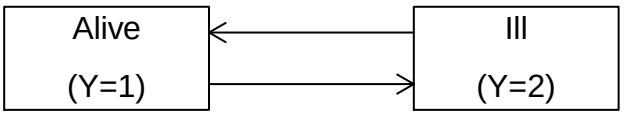
because for this state space, being in state 1 at time $x+t$ is equivalent to holding in state 1 throughout the interval $[x, x+t]$.

It follows from the solution for $p_{11}(x,t)$ above that

$$h_{T_x}(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\log S_{T_x}(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\log p_{11}(x,t) = \mu_{12}(x+t)$$

so the transition intensity plays the role of the hazard function for the variable representing the time spent in the alive state (1).

8.8.2 Two-state model (no absorbing state)



As in the first example, this is a two-state process, so is determined by the transition intensities between the two states, $\mu_{12}(x)$ and $\mu_{21}(x)$.

However, neither state is absorbing, so the transition intensity matrix is given by

$$M(x) = \begin{pmatrix} -\mu_{12}(x) & \mu_{12}(x) \\ \mu_{21}(x) & -\mu_{21}(x) \end{pmatrix}$$

To completely determine the transition structure for the process, we need to solve for $p_{11}(x,t)$ (or $p_{12}(x,t)$) and $p_{21}(x,t)$ (or $p_{22}(x,t)$)

The Kolmogorov forward equation for $p_{11}(x,t)$ is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}p_{11}(x,t) &= p_{11}(x,t)\mu_{11}(x+t) + p_{12}(x,t)\mu_{21}(x+t) \\ &= -p_{11}(x,t)\mu_{12}(x+t) + [1 - p_{11}(x,t)]\mu_{21}(x+t) \\ &= -p_{11}(x,t)[\mu_{12}(x+t) + \mu_{21}(x+t)] + \mu_{21}(x+t) \end{aligned}$$

Similarly, the Kolmogorov forward equation for $p_{22}(x,t)$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}p_{22}(x,t) = -p_{22}(x,t)[\mu_{12}(x+t) + \mu_{21}(x+t)] + \mu_{12}(x+t)$$

The existence of an analytic solution to these equations depends on the form of the transition intensity functions $\mu_{12}(x)$ and $\mu_{21}(x)$.

We restrict attention to the time-homogeneous case where

$$\mu_{12}(x) = \mu_{12} \qquad \text{and} \qquad \mu_{21}(x) = \mu_{21}$$

independent of x.

The Kolmogorov forward equation for $p_{11}(x,t)$ is now

$$\begin{aligned}
& \frac{\mathrm{d}}{\mathrm{d}t} p_{11}(x,t) = -p_{11}(x,t)[\mu_{12} + \mu_{21}] + \mu_{21} \\
\Rightarrow \quad & \int \frac{\mathrm{d}p_{11}(x,t)}{p_{11}(x,t)[\mu_{12} + \mu_{21}] - \mu_{21}} = -\int \mathrm{d}t \\
\Rightarrow \quad & \frac{\log\{p_{11}(x,t)[\mu_{12} + \mu_{21}] - \mu_{21}\}}{\mu_{12} + \mu_{21}} = -t + C \\
\Rightarrow \quad & p_{11}(x,t)[\mu_{12} + \mu_{21}] - \mu_{21} = A \exp\left(-[\mu_{12} + \mu_{21}]t\right) \\
\Rightarrow \quad & p_{11}(x,t) = \frac{A \exp\left(-[\mu_{12} + \mu_{21}]t\right) + \mu_{21}}{\mu_{12} + \mu_{21}} \\
\Rightarrow \quad & p_{11}(x,t) = \frac{\mu_{12} \exp\left(-[\mu_{12} + \mu_{21}]t\right) + \mu_{21}}{\mu_{12} + \mu_{21}}
\end{aligned}$$

because we have the boundary condition $p_{11}(x,0) = 1$.

Similarly, the Kolmogorov forward equation for $p_{22}(x,t)$ is solved by

$$p_{22}(x,t) = \frac{\mu_{21}\exp\left(-[\mu_{12}+\mu_{21}]t\right)+\mu_{12}}{\mu_{21}+\mu_{12}}$$

and we also obtain, using the identities $p_{11}(x,t)+p_{12}(x,t)=1$ and $p_{21}(x,t)+p_{22}(x,t)=1$,

$$p_{12}(x,t) = \frac{\mu_{12}\{1-\exp\left(-[\mu_{12}+\mu_{21}]t\right)\}}{\mu_{12}+\mu_{21}}$$

and

$$p_{21}(x,t) = \frac{\mu_{21}\{1-\exp\left(-[\mu_{12}+\mu_{21}]t\right)\}}{\mu_{21}+\mu_{12}}$$

Note that, as $t\to\infty$

$$p_{i1}(x,t)\to\frac{\mu_{21}}{\mu_{21}+\mu_{12}} \qquad \text{and} \qquad p_{i2}(x,t)\to\frac{\mu_{12}}{\mu_{12}+\mu_{21}}$$

for $i=1,2$, so the final state becomes independent of the initial state.

Note that, in this model, the survivor function for the holding time random variables T_{x1} (state 1) and T_{x2} (state 2) are

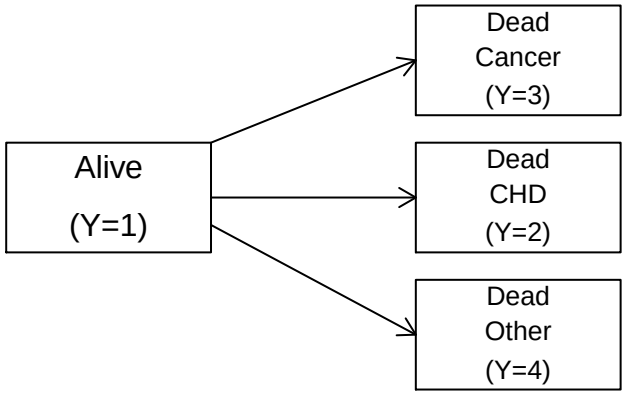
$$S_{x1}(t)=\exp(-\mu_{12}t) \qquad \text{and} \qquad S_{x2}(t)=\exp(-\mu_{21}t)$$

and that

$$S_{x1}(t) \neq p_{11}(x,t) \qquad \text{and} \qquad S_{x2}(t) \neq p_{22}(x,t)$$

because this model allows transition out of a state and then return at a later time.

8.8.3 Four-state process



This four-state process is determined by the transition intensities, $\mu_{ij}(x)$ $i \neq j$.

As states 2,3,4 are absorbing,

$$p_{ij}(x,t) = 0 \text{ for all } x,t,i > 1, j \neq i \quad \Rightarrow \quad \mu_{ij}(x) = 0 \text{ for all } x, i > 1, j \neq i$$

Hence the transition intensity matrix is given by

$$M(x) = \begin{pmatrix} -\mu_{12}(x) - \mu_{13}(x) - \mu_{14}(x) & \mu_{12}(x) & \mu_{13}(x) & \mu_{14}(x) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

To completely determine the transition structure for the process, we need to solve the Kolmogorov forward equation for three of $\{p_{11}(x,t), p_{12}(x,t), p_{13}(x,t), p_{14}(x,t)\}$, and then obtain the other using

$$p_{11}(x,t) + p_{12}(x,t) + p_{13}(x,t) + p_{14}(x,t) = 1.$$

The Kolmogorov forward equation for $p_{11}(x,t)$ is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} p_{11}(x,t) &= p_{11}(x,t)\mu_{11}(x+t) + p_{12}(x,t)\mu_{21}(x+t) \\ &\qquad + p_{13}(x,t)\mu_{31}(x+t) + p_{14}(x,t)\mu_{41}(x+t) \\ &= p_{11}(x,t)\mu_{11}(x+t) \end{aligned}$$

So

$$\int \frac{\mathrm{d}p_{11}(x,t)}{p_{11}(x,t)} = -\int_0^t [\mu_{12}(x+t)+\mu_{13}(x+t)+\mu_{14}(x+t)]\mathrm{d}t$$

$$\Rightarrow \quad \log p_{11}(x,t) = -\int_0^t [\mu_{12}(x+t)+\mu_{13}(x+t)+\mu_{14}(x+t)]\mathrm{d}t + C$$

$$\Rightarrow \quad p_{11}(x,t) = A\exp\left(-\int_0^t [\mu_{12}(x+t)+\mu_{13}(x+t)+\mu_{14}(x+t)]\mathrm{d}t\right)$$

$$\Rightarrow \quad p_{11}(x,t) = \exp\left(-\int_0^t [\mu_{12}(x+t)+\mu_{13}(x+t)+\mu_{14}(x+t)]\mathrm{d}t\right)$$

because we have the boundary condition $p_{11}(x,0)=1$.

The Kolmogorov forward equation for $p_{12}(x,t)$ is

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}p_{12}(x,t) &= p_{11}(x,t)\mu_{12}(x+t) + p_{12}(x,t)\mu_{22}(x+t) \\
&\qquad + p_{13}(x,t)\mu_{32}(x+t) + p_{14}(x,t)\mu_{42}(x+t) \\
&= p_{11}(x,t)\mu_{12}(x+t) \\
&= \mu_{12}(x+t)\times \\
&\qquad \exp\left(-\int_0^t [\mu_{12}(x+t)+\mu_{13}(x+t)+\mu_{14}(x+t)]\mathrm{d}t\right)
\end{aligned}$$

Similarly, for $j = 2, 3, 4$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}p_{1j}(x,t) &= \mu_{1j}(x+t) \times \\ &\qquad \exp\left(-\int_0^t [\mu_{12}(x+t) + \mu_{13}(x+t) + \mu_{14}(x+t)]\mathrm{d}t\right)\end{aligned}$$

The existence of an analytic solution to the differential equations for $p_{2j}(x,t)$, $p_{3j}(x,t)$ and $p_{4j}(x,t)$, depends on the form of the intensity functions $\mu_{12}(x)$ and $\mu_{13}(x)$ and $\mu_{14}(x)$

We restrict attention to the time-homogeneous case where

$$\mu_{12}(x) = \mu_{12} \qquad \mu_{13}(x) = \mu_{13} \qquad \mu_{14}(x) = \mu_{214}$$

independent of x.

It immediately follows that

$$\begin{aligned}p_{11}(x,t) &= \exp\left(-\int_0^t [\mu_{12}(x+t) + \mu_{13}(x+t) + \mu_{14}(x+t)]\mathrm{d}t\right) \\ &= \exp\left(-t[\mu_{12} + \mu_{13} + \mu_{14}]\right)\end{aligned}$$

Then, for $j > 1$,

$$\frac{\mathrm{d}}{\mathrm{d}t}p_{1j}(x,t) = \mu_{1j}\exp\left(-t[\mu_{12}+\mu_{13}+\mu_{14}]\right)$$

$$\Rightarrow \quad p_{1j}(x,t) = C - \frac{\mu_{1j}}{\mu_{12}+\mu_{13}+\mu_{14}}\exp\left(-t[\mu_{12}+\mu_{13}+\mu_{14}]\right)$$

$$\Rightarrow \quad p_{1j}(x,t) = \frac{\mu_{1j}}{\mu_{12}+\mu_{13}+\mu_{14}}\left\{1-\exp\left(-t[\mu_{12}+\mu_{13}+\mu_{14}]\right)\right\}$$

because we have the boundary condition $p_{1j}(x,0) = 0$, $j > 1$.

It is easy to verify that

$$p_{11}(x,t) + p_{12}(x,t) + p_{13}(x,t) + p_{14}(x,t) = 1.$$

Chapter 9

Inference for Multistate Models

In Chapter 8 we have introduced multistate models and their properties, and presented several examples of models with possible applications to survival data analysis.

A key part of the statistical analysis process is making inference about the parameters of the model, on the basis of observed data.

For a multistate model, with states $1, \ldots, m$ the parameters are the transition intensity functions

$$\mu_{k\ell}(x), \quad i, j = 1, \ldots, m, \quad k \neq \ell$$

We will only consider time-homogenous models, so the parameters to be estimated are the transition intensities

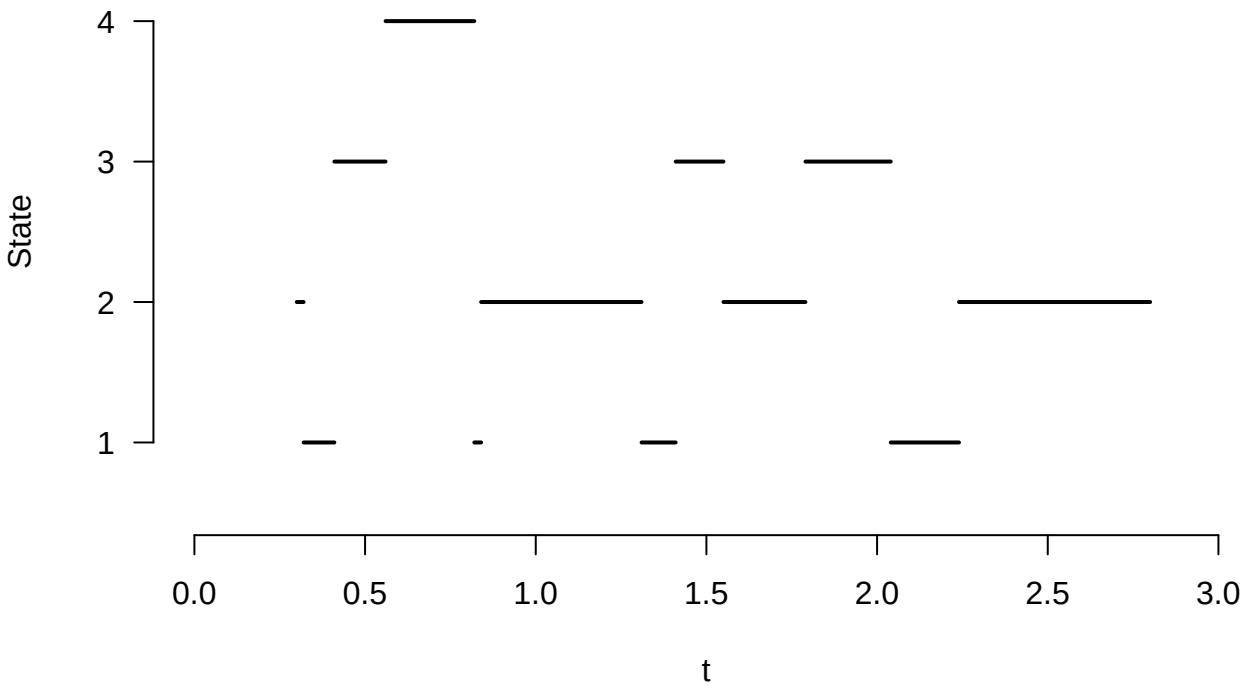
$$\mu_{k\ell}, \quad i, j = 1, \ldots, m, \quad k \neq \ell$$

9.1 Data

For a multistate model, the data will be the transition histories of a set of individuals $i = 1, \ldots, n$. We start by considering the case $n = 1$

In this case, we might observe something like:

Transition	Start	1	2	3	4	5	6	7	8	9	10	11	End
Time	0.3	0.32	0.41	0.56	0.82	0.84	1.31	1.41	1.55	1.79	2.04	2.24	2.8
State	2	1	3	4	1	2	1	3	2	3	1	2	2



9.2 Data structure

We will denote the data observations using

Transition (j)	0	1	2	$\cdots$	j	$\cdots$	J	$J+1$ (end)
Time (t_j)	t_0	t_1	t_2	$\cdots$	t_j	$\cdots$	t_J	t_{J+1}
State (y_j)	y_0	y_1	y_2	$\cdots$	y_j	$\cdots$	y_J	

The likelihood for the transition intensity parameters $\boldsymbol{\mu} = \{\mu_{k\ell}, k \neq \ell\}$ is the probability of the observed

sequence of holding times and transitions, given $\boldsymbol{\mu}$, that is

$$\begin{aligned} L(\boldsymbol{\mu}) &= f_{y_0}(t_1 - t_0)p(y_1|y_0) \times f_{y_1}(t_2 - t_1)p(y_2|y_1) \times \cdots \\ &\quad \times f_{y_{J-1}}(t_J - t_{J-1})p(y_J|y_{J-1}) \times S_{y_J}(t_{J+1} - t_J) \end{aligned}$$

where

- $f_k(t)$ is the exponential p.d.f. of the holding time t in state k
- $p(\ell|k)$ is the conditional probability of transition from state k to state ℓ given that a transition has happened

9.3 Likelihood

We have

$$\begin{aligned} L(\boldsymbol{\mu}) &= f_{y_0}(t_1 - t_0)p(y_1|y_0) \times f_{y_1}(t_2 - t_1)p(y_2|y_1) \times \cdots \\ &\quad \times f_{y_{J-1}}(t_J - t_{J-1})p(y_J|y_{J-1}) \times S_{y_J}(t_{J+1} - t_J) \\ &= S_{y_J}(t_{J+1} - t_J) \prod_{j=1}^{J} f_{y_{j-1}}(t_j - t_{j-1})p(y_j|y_{j-1}) \end{aligned}$$

where

$$f_k(t) = \sum_{\ell \neq k} \mu_{k\ell} \exp\left(-t \sum_{\ell \neq k} \mu_{k\ell}\right)$$

$$S_k(t) = \exp\left(-t \sum_{\ell \neq k} \mu_{k\ell}\right)$$

$$p(\ell|k) = \frac{\mu_{k\ell}}{\sum_{\ell \neq k} \mu_{k\ell}}$$

Hence

$$\begin{aligned}
L(\boldsymbol{\mu}) &= S_{y_J}(t_{J+1} - t_J) \prod_{j=1}^{J} f_{y_{j-1}}(t_j - t_{j-1}) p(y_j|y_{j-1}) \\
&= \exp\left(-(t_{J+1} - t_J) \sum_{\ell \neq y_J} \mu_{y_J \ell}\right) \\
&\qquad \times \prod_{j=1}^{J} \exp\left(-(t_j - t_{j-1}) \sum_{\ell \neq y_{j-1}} \mu_{y_{j-1}\ell}\right) \mu_{y_{j-1}y_j} \\
&= \prod_{j=1}^{J+1} \exp\left(-(t_j - t_{j-1}) \sum_{\ell \neq y_{j-1}} \mu_{y_{j-1}\ell}\right) \prod_{j=1}^{J} \mu_{y_{j-1}y_j} \\
&= \prod_{j=1}^{J+1} \prod_{\ell \neq y_{j-1}} \exp\left\{-(t_j - t_{j-1}) \mu_{y_{j-1}\ell}\right\} \prod_{j=1}^{J} \mu_{y_{j-1}y_j}
\end{aligned}$$

So the likelihood factorises into a series of terms, each containing a single transition intensity $\mu_{k\ell}$.

Gathering together all the terms containing a particular $\mu_{k\ell}$, we obtain

$$\begin{aligned} L(\mu_{k\ell}) &= \mu_{k\ell}^{n_{k\ell}} \prod_{j:y_{j-1}=k}^{J+1} \exp\left\{-(t_j - t_{j-1})\mu_{k\ell}\right\} \\ &= \mu_{k\ell}^{n_{k\ell}} \exp\left(-t_k^+ \mu_{k\ell}\right) \end{aligned}$$

where

- t_k^+ is the total observed holding time in state k
- $n_{k\ell}$ is the total observed number of transitions from state k to state ℓ

are called *sufficient statistics*.

Therefore the likelihood for the full set of transition intensity parameters $\boldsymbol{\mu} = \{\mu_{k\ell}, k \neq \ell\}$ for a single individual transition history is

$$L(\boldsymbol{\mu}) = \prod_{k,\ell:k\neq\ell} \mu_{k\ell}^{n_{k\ell}} \exp\left(-t_k^+ \mu_{k\ell}\right)$$

9.4 Maximum likelihood estimation for $\boldsymbol{\mu}$

The log likelihood is

$$\ell(\boldsymbol{\mu}) = \sum_{k,\ell:k\neq\ell} n_{k\ell} \log \mu_{k\ell} - \sum_{k,\ell:k\neq\ell} t_k^+ \mu_{k\ell}$$

and therefore

$$\frac{\partial}{\partial \mu_{k\ell}} \ell(\boldsymbol{\mu}) = \frac{n_{k\ell}}{\mu_{k\ell}} - t_k^+$$

so the m.l.e. solves

$$\frac{n_{k\ell}}{\hat{\mu}_{k\ell}} - t_k^+ = 0 \qquad \Rightarrow \qquad \hat{\mu}_{k\ell} = \frac{n_{k\ell}}{t_k^+}$$

9.5 Standard errors for $\hat{\boldsymbol{\mu}}$

The second derivatives of the log likelihood are

$$\frac{\partial^2}{\partial \mu_{k\ell}^2} \ell(\boldsymbol{\mu}) = -\frac{n_{k\ell}}{\mu_{k\ell}^2}$$

and all other second derivatives, $\frac{\partial^2}{\partial \mu_{k\ell} \partial \mu_{ij}} \ell(\boldsymbol{\mu})$, are zero.

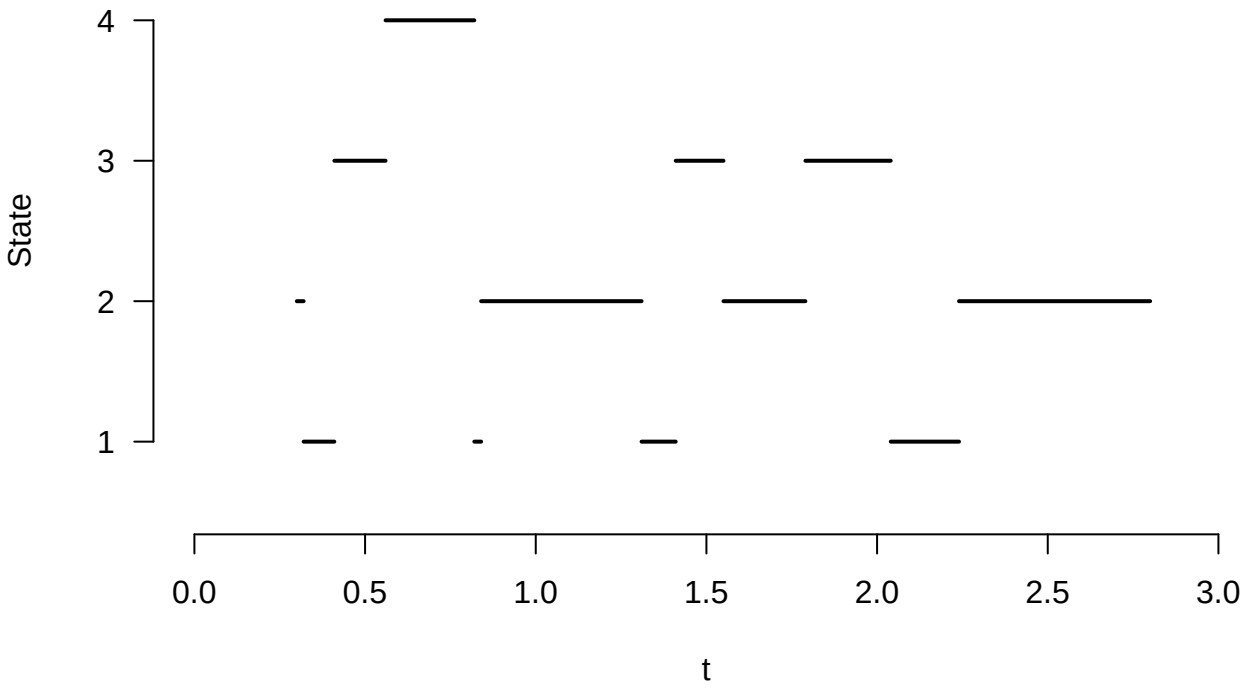
Therefore, we have that the m.l.e.s are approximately independent (in large samples) and

$$Var\left(\hat{\mu}_{k\ell}\right) = \frac{\mu_{k\ell}^2}{n_{k\ell}} \qquad \Rightarrow \qquad s.e.\left(\hat{\mu}_{k\ell}\right) = \frac{\hat{\mu}_{k\ell}}{n_{k\ell}^{1/2}}$$

so confidence intervals for the transition intensities can be easily constructed, based on the large sample normal distribution of the m.l.e.

9.6 Example

Transition	Start	1	2	3	4	5	6	7	8	9	10	11	End
Time	0.3	0.32	0.41	0.56	0.82	0.84	1.31	1.41	1.55	1.79	2.04	2.24	2.8
State	2	1	3	4	1	2	1	3	2	3	1	2	2



$t_1^+ = 0.41 \quad t_2^+ = 1.29 \quad t_3^+ = 0.54 \quad t_4^+ = 0.26$

$$[n_{k\ell}] = \begin{pmatrix} - & 2 & 2 & 0 \\ 2 & - & 1 & 0 \\ 1 & 1 & - & 1 \\ 1 & 0 & 0 & - \end{pmatrix}$$

$$[\hat{\mu}_{k\ell}] = \begin{pmatrix} - & \frac{2}{0.41} & \frac{2}{0.41} & 0 \\ \frac{2}{1.29} & - & \frac{1}{1.29} & 0 \\ \frac{1}{0.54} & \frac{1}{0.54} & - & \frac{1}{0.54} \\ \frac{1}{0.26} & 0 & 0 & - \end{pmatrix} = \begin{pmatrix} - & 4.88 & 4.88 & 0 \\ 1.55 & - & 0.78 & 0 \\ 1.85 & 1.85 & - & 1.85 \\ 3.85 & 0 & 0 & - \end{pmatrix}$$

$$[s.e.(\hat{\mu}_{k\ell})]) = \begin{pmatrix} - & 3.45 & 3.45 & 0 \\ 1.10 & - & 0.78 & 0 \\ 1.85 & 1.85 & - & 1.85 \\ 3.85 & 0 & 0 & - \end{pmatrix}$$

9.7 Multiple sequences ($n > 1$)

In practice, we will observe multiple histories of transitions corresponding to individuals $i = 1, \ldots, n$ in our data sample, with corresponding transition frequencies $[n_{ik\ell}]$ and holding times $\{t_{ik}^{+}\}$.

We make the standard assumption that the n observed histories are realizations of *independent* random processes, so the full likelihood (joint probability) is obtained by multiplying the individual likelihoods:

$$\begin{aligned} L(\boldsymbol{\mu}) &= \prod_{i=1}^{n} \prod_{k,\ell:k\neq\ell} \mu_{k\ell}^{n_{ik\ell}} \exp\left(-t_{ik}^{+}\mu_{k\ell}\right) \\ &= \prod_{k,\ell:k\neq\ell} \mu_{k\ell}^{n_{k\ell}} \exp\left(-t_{k}^{+}\mu_{k\ell}\right) \end{aligned}$$

where now

- t_k^+ is the total observed holding time in state k *across all individuals*
- $n_{k\ell}$ is the total observed number of transitions from state k to state ℓ *across all individuals*

Inference proceeds exactly as Sections 9.4 and 9.5 with these new sufficient statistics.

Chapter 10

Modelling Human Lifetime

For the rest of this module our focus will be on modeling human life length, that is the time (from birth or some other specified time origin) to death in a specified human population.

Models for human human lifetimes have a huge importance, for example

- Billions of pounds are invested in pension funds. Calculation of the liabilities requires us to be able to predict lifetimes of current and future pensioners.
- Planning public services requires requires us to predict age-structured populations, sometimes within a small geographical area. This requires us to be able to forecast mortality (along with fertility and migration)

10.1 Special features of human lifetime

There are a number of practical reasons why human lifetime modelling requires special treatment:

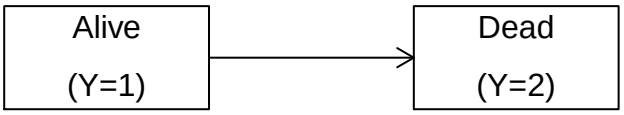
- The long time scales involved (much longer than the horizon of a standard statistical study)
- The requirement to use secondary data (collected for other purposes, such as census and death registration data)
- Data sets are often large (good!) but the data can be coarse (bad!) For example, ages may only be provided in whole years.
- Standard distributions tend not to provide a good fit for human lifetimes.
- The distribution of human lifetime is changing (lifetimes are getting longer)

10.2 Models

As previously, we use T to denote the time from a specified origin (usually birth) until death.

Then we know that a survival model can be specified by the survival function $S_T(t)$ or the hazard function $h_T(t)$ for T.

Alternatively, we can think of human survival as a two-state process for $Y_x \in \{1, 2\}$ (alive/dead) with an absorbing state



Then the process is specified by the transition intensity function $\mu_{12}(x)$ at time x or alternatively the transition probabilities $p_{11}(x, t) = 1 - p_{12}(x, t)$.

10.3 Models are equivalent

Recall that

$$\begin{aligned} p_{11}(x,t) &= P(Y_{x+t}=1 \,|\, Y_x=1) &&= P(T > x+t \,|\, T > x) \\ &&&= S_T(t|x) \\ &&&= \frac{S_T(x+t)}{S_T(x)} \end{aligned}$$

where $S_T(t|x)$ is the residual survival function (see Section 3.3.2). So $p_{11}(0,t) = S_T(t)$

Similarly

$$\mu_{12}(x) = \lim_{\delta t \to 0} \frac{p_{12}(x,\delta t)}{\delta t} = \lim_{\delta t \to 0} \frac{P(T \leq x+\delta t \,|\, T > x)}{\delta t} = h_T(x)$$

So the models are equivalently specified.

10.4 Alternative notation

Because of the importance of human lifetime modelling in demography and actuarial science, a specific notation has been developed:

$$p_{11}(x,t) = S_T(t|x) \equiv {}_tp_x$$

and

$$p_{12}(x,t) = 1 - S_T(t|x) = 1 - {}_tp_x \equiv {}_tq_x$$

so

- ${}_tq_x$ is the probability of death in $(x, x+t]$ given survival to age x
- ${}_tp_x$ is probability of probability of survival to $x+t$ given survival to age x.

10.5 p_x and q_x notation

When $t = 1$, the first subscript is usually omitted, so

- q_x is the probability of death in $(x, x+1]$ given survival to age x
- $p_x = 1 - q_x$ is probability of probability of survival to $x+1$ given survival to age x.

This is most commonly used when x is an exact integer number of years.

Note that if t is an integer then

$$\begin{aligned} {}_tp_x &= \frac{S_T(x+t)}{S_T(x)} \\ &= \frac{S_T(x+1)}{S_T(x)}\frac{S_T(x+2)}{S_T(x+1)}\cdots\frac{S_T(x+t)}{S_T(x+t-1)} \\ &= p_x p_{x+1}\cdots p_{x+t-1} \end{aligned}$$

and so

$${}_tq_x = 1-(1-q_x)(1-q_{x+1})\cdots(1-q_{x+t-1}).$$

10.6 Force of mortality

We also have:

$$\mu_{12}(x) = h_T(x) \equiv \mu_x$$

which is called (in human lifetime applications) the *force of mortality* at age x.

The force of mortality μ_x is the limiting death rate at age x, and as such is only constrained by $\mu_x \geq 0$.

In practice, if we are measuring time in years, $\mu_x < 1$ for all except *very* extreme ages x.

The functions ${}_tp_x$, ${}_tq_x$ and μ_x are related by

$$ {}_tp_x \;=\; 1 - {}_tq_x \;=\; \exp\left(-\int_x^{x+t} \mu_x \mathrm{d}x\right) $$

Chapter 11

The Life Table and Life Expectancy

11.1 Life tables

The distribution of lifetime for a population is often summarised in a *life table*.

Excerpt for males:

Age	Males							
x	m_x	q_x	l_x	d_x	L_x	T_x	μ_x	e_x
0	0.004757	0.004746	100000	475	99576.3	7896837		78.97
1	0.000306	0.000306	99525	30	99510.2	7797072	0.000369	78.34
2	0.000207	0.000207	99495	21	99484.6	7697562	0.000246	77.37
3	0.000147	0.000147	99474	14	99467.0	7598078	0.000172	76.38
4	0.000115	0.000115	99460	12	99453.9	7498612	0.000128	75.39
109	0.676172	0.491440	8	4	5.5	11	0.661588	1.43
110	0.701065	0.503943	4	2	2.8	5	0.685990	1.39
111	0.725677	0.516003	2	1	1.4	3	0.709972	1.34
112	0.750015	0.528125	1	1	0.7	1	0.733841	1.30

Excerpt for females:

Age	Females							
x	m_x	q_x	l_x	d_x	L_x	T_x	μ_x	e_x
0	0.003818	0.003811	100000	381	99660.7	8279504		82.80
1	0.000238	0.000238	99619	24	99607.0	8179692	0.000276	82.11
2	0.000176	0.000176	99595	17	99586.4	8080086	0.000202	81.13
3	0.000133	0.000133	99578	14	99571.0	7980500	0.000152	80.14
4	0.000107	0.000107	99564	10	99559.1	7880929	0.000118	79.15
109	0.668760	0.487656	27	13	20.0	39	0.652973	1.44
110	0.697334	0.502089	14	7	10.1	19	0.681051	1.39
111	0.724789	0.515573	7	4	5.0	9	0.708358	1.34
112	0.751040	0.528125	3	1	2.4	4	0.734218	1.30
113	0.776042	0.539776	2	1	1.1	2	0.758668	1.26
114	0.799785	0.550574	1	1	0.5	1	0.781684	1.23

This is ELT17, the latest decennial life table for England, available at:

`http://www.ons.gov.uk/ons/rel/lifetables/decennial-life-tables/`
`    english-life-tables--no-17--2010-12/stb-elt17.html`

11.2 The life table quantities ℓ_x, q_x and d_x

The life table summarises the distribution of a lifetime variable by presenting the function ℓ_x for a set of discrete values x.

The number ℓ_0 is called the radix for the table, and

$$\begin{aligned} \ell_x &= E[\text{number of survivors to exact age } x \\ &\qquad\qquad \text{from an initial population of size } \ell_0] \\ &= \ell_0 P(T > x) \\ &= \ell_0 S_T(x) \end{aligned}$$

So ℓ_x/ℓ_0 is simply the survival function $S_T(x)$ or ${}_xp_0$.

- The radix is usually a power of 10, such as 10 000 or 100 000
- The x values are usually evenly spaced but need not be every year, for example every 5 years or every 10 years.

The life table usually contains a column with q_x values.

If the gap between the ages x is 1 year, then q_x is exactly as defined in Section 10.5, that is

$$\begin{aligned} q_x = P(T \leq x+1 | T > x) &= 1 - S_T(1|x) \\ &= 1 - \frac{S_T(x+1)}{S_T(x)} \end{aligned}$$

It follows that there is a simple relationship between $\{\ell_x\}$ and $\{q_x\}$, as

$$q_x = 1 - \frac{S_T(x+1)}{S_T(x)} = 1 - \frac{\ell_{x+1}}{\ell_x}$$

and

$$\ell_{x+1} = \ell_x(1-q_x) = \ell_0 \prod_{k=0}^{x}(1-q_k)$$

If the gap between the x values is $g > 1$ year, the the q_x should be replaced by ${}_gq_x$ so that these relationships still hold.

The life table usually contains a column with d_x values.

$$\begin{aligned} d_x &= E[\text{number of deaths in } [x, x+1) \\ &\qquad \text{from an initial population of size } \ell_0] \\ &= \ell_0 P(x \le T < x+1) \\ &= \ell_0[S_T(x) - S_T(x+1)] \\ &= \ell_x - \ell_{x+1} \end{aligned}$$

11.3 Life expectancy

The complete life expectancy at age x, $\mathring{e}_x$ is defined as the expected future lifetime given survival to age x,

$$\mathring{e}_x = E(T - x \,|\, T > x)$$

The *curtate* life expectancy at age x, e_x is defined as the expected future lifetime in completed years given survival to age x (usually an integer),

$$\begin{aligned} e_x &= E([T - x] \,|\, T > x) \\ &= \sum_{k=0}^{\infty} kP(x + k \leq T < x + k + 1 \,|\, T > x) \\ &= \sum_{k=0}^{\infty} k[S_T(k|x) - S_T(k + 1|x)] \\ &= \sum_{k=0}^{\infty} k\frac{S_T(x + k) - S_T(x + k + 1)}{S_T(x)} \end{aligned}$$

11.3.1 Obtaining e_x from the life table

If x is an integer, then

$$\begin{aligned} e_x &= \sum_{k=0}^{\infty} k \frac{S_T(x+k) - S_T(x+k+1)}{S_T(x)} \\ &= \sum_{k=0}^{\infty} k \frac{\ell_{x+k} - \ell_{x+k+1}}{\ell_x} \\ &= \frac{1}{\ell_x} \sum_{k=0}^{\infty} k(\ell_{x+k} - \ell_{x+k+1}) \\ &= \frac{1}{\ell_x} \sum_{k=1}^{\infty} \ell_{x+k} \end{aligned}$$

In practice, we only need to sum up until $\ell_{x+k} \approx 0$ (typically around $x + k = 120$).

11.3.2 Obtaining $\mathring{e}_x$ from e_x

The complete expectation of life $\mathring{e}_x$ includes fractional parts of a year between an individuals last birthday (the death age for the purposes of calculation of e_x) and the actual date of death. Hence

$$e_x \leq \mathring{e}_x < e_x + 1.$$

If we make the assumption that, in every age interval $[x, x+1)$, deaths are uniformly distributed, then we

have

$$\mathring{e}_x = e_x + \frac{1}{2}.$$

This gives a reasonable approximation to $\mathring{e}_x$ which can then be computed using life table data.

However, in cases where the x values are not evenly spaced, or when the assumption that deaths are evenly distributed over $[x, x+1)$ is unreasonable (for example in the first year of life) then an alternative approach is required to compute $\mathring{e}_x$.

11.3.3 Obtaining $\mathring{e}_x$ from an irregular life table

We have

$$\begin{aligned}
\mathring{e}_x = E(T - x | T > x) &= \int_0^\infty S_T(t|x)\mathrm{d}t \\
&= \int_0^\infty \frac{S_T(x+t)}{S_T(x)}\mathrm{d}t \\
&= \frac{1}{\ell_x}\int_0^\infty \ell_0 S_T(x+t)\mathrm{d}t \\
&= \frac{1}{\ell_x}\int_x^\infty \ell_0 S_T(t)\mathrm{d}t \\
&= \frac{1}{\ell_x}\sum_{i=1}\int_{x_{i-1}}^{x_i} \ell_0 S_T(t)\mathrm{d}t
\end{aligned}$$

where $x = x_0 < x_1 < x_2 \cdots$ are the ages $\geq x$ represented in the life table.

11.3.4 Expectation of life in an interval

Define ${}_sL_x$ to be the expected total years lived between ages x and $x+s$ by an initial population of size ℓ_0. Then

$$ {}_sL_x \;=\; {}_\infty L_x - {}_\infty L_{x+s} $$

where

$$ {}_\infty L_x \;=\; \ell_0 P(T > x)\mathring{e}_x \;=\; \ell_0 \frac{\ell_x}{\ell_0}\frac{1}{\ell_x}\int_x^\infty \ell_0 S_T(t)\mathrm{d}t \;=\; \int_x^\infty \ell_0 S_T(t)\mathrm{d}t. $$

Therefore,

$$ {}_sL_x \;=\; \int_x^{x+s} \ell_0 S_T(t)\mathrm{d}t $$

The quantity ${}_\infty L_x$, the expected total years lived after age x by an initial population of size ℓ_0, is usually denoted by T_x.

11.3.5 Life expectancy using L_x and/or T_x

It follows from Section 11.3.3 and 11.3.4 that life expectancy at birth is

$$\mathring{e}_x = \frac{1}{\ell_x}\int_x^\infty \ell_0 S_T(t)\mathrm{d}t = \frac{T_x}{\ell_x}$$

$$= \frac{1}{\ell_x}\sum_{i=1}\int_{x_{i-1}}^{x_i} \ell_0 S_T(t)\mathrm{d}t = \frac{1}{\ell_x}\sum_{i=1} {}_{x_i-x_{i-1}}L_{x_{i-1}}$$

When $t = 1$, ${}_tL_x$ is simply denoted by L_x.

Therefore, for a life table with annual spacing, $x_0 = x$, $x_1 = x + 1, \ldots$, we have

$$\mathring{e}_x = \frac{T_x}{\ell_x} = \frac{1}{\ell_x}\sum_{t=0} L_{x+t}$$

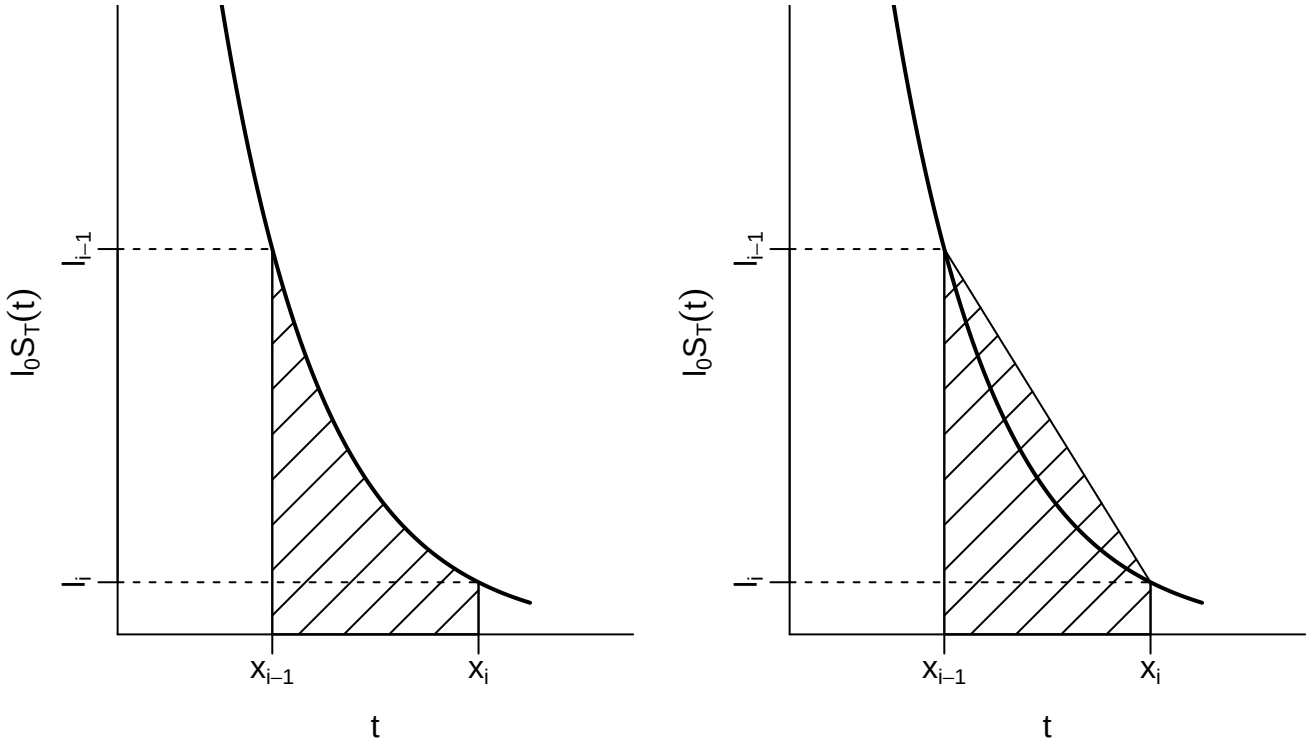
Life tables often present values of T_x and/or L_x.

11.3.6 The trapezium rule approximation

To obtain $\mathring{e}_x$, we need to approximate ${}_{x_i-x_{i-1}}L_{x_{i-1}} = \int_{x_{i-1}}^{x_i} \ell_0 S_T(t)\mathrm{d}t$ where the only relevant values of the integrand that we know are at the endpoints:

$$\ell_0 S_T(x_{i-1}) = \ell_{x_{i-1}} \qquad \text{and} \qquad \ell_0 S_T(x_i) = \ell_{x_i}$$

We use the trapezium rule approximation:



$$ {}_{x_i-x_{i-1}}L_{x_{i-1}} \quad = \quad \int_{x_{i-1}}^{x_i} \ell_0 S_T(t)\mathrm{d}t \quad \approx \quad (x_i - x_{i-1})\frac{\ell_{x_i} + \ell_{x_{i-1}}}{2} $$

so, when $x_i - x_{i-1} = 1$ for all i, we have $L_x \approx (\ell_{x+1} + \ell_x)/2$ which gives the expression in Section 11.3.2

when substituted into $\mathring{e}_x = \frac{1}{\ell_x}\sum_{t=0} L_{x+t}$.

11.4 Concluding remarks

11.4.1 Period life tables

The life table ELT17 in Section 11.1 is a *period life table*.

That means that the table was constructed by estimating q_x from mortality data over a particular period in time (here 2010-2012).

- q_{21} is estimated using individuals born around 1990
- q_{81} is estimated using individuals born around 1930

etc.

So the table does not actually describe the distribution of lifetime of any actual individual or group of individuals.

- Individuals born around 1990 will expect to experience a different (lower?) q_{81} in 2071
- Individuals born around 1930 experienced a higher q_{21} in 1951

11.4.2 Cohort life tables

A life table describing the mortality experience of a particular population, as it develops over time is called a *cohort life table*.

A *cohort* is the name given to a population all born at or around the same time (typically in the same calendar year).

The cohort life table estimates q_x based on the mortality experience of the cohort at that age.

The problem with a cohort life table is that it cannot be completed until the cohort have died out, by which point it is only really of interest to historians!

Chapter 12

Interpolating a Life Table

The life table only includes values of ℓ_x and q_x and hence $S_T(x)$ for certain, usually evenly spaced integer, values of x.

We can 'fill in the gaps' by interpolating, but this requires us to make assumptions about the distribution of lifetime T in the intervals between ages represented in the life table.

We shall initially assume that the rows of the life table represent $x = 0, 1, 2, 3, \ldots$. The generalisation to other intervals is straightforward.

At age $x + t$, where $t \in [0, 1)$ we have

$$\ell_{x+t} \;=\; \ell_0 S_T(x+t) \;=\; \ell_0 P(T > x+t | T > x) P(T > x) \;=\; {}_tp_x\, \ell_x$$

So completion of the life table requires us to specify our belief about ${}_tp_x$ (or equivalently ${}_tq_x$) for $t \in [0, 1)$.

12.1 Uniform distribution of deaths

Perhaps the easiest assumption is to assume that the deaths in $[x, x+1)$ are uniformly distributed. Then

$$P(T \le x+t | x \le T < x+1) \;=\; t \qquad t \in [0, 1)$$

Hence

$$\begin{aligned} {}_tq_x &= P(T \le x+t | T \ge x) \\ &= P(T \le x+t | x \le T < x+1) P(T < x+1 | T \ge x) \\ &= tq_x \end{aligned}$$

So

$$\ell_{x+t} = \ell_x(1 - tq_x) \qquad t \in [0,1)$$

which is the linear interpolation which makes the trapezium rule approximation, used in the calculation of life expectancy, exact.

12.2 Constant force of mortality

An alternative assumption is to assume that the force of mortality (hazard of death) μ_x takes a constant value μ in $[x, x+1)$. Then

$${}_tq_x = P(T \le x+t | T \ge x) = 1 - \exp(-\mu t) \qquad t \in [0,1)$$

Hence

$$q_x = 1 - \exp(-\mu) \qquad \Rightarrow \qquad \mu = -\log(1 - q_x)$$

and therefore

$$ {}_tq_x \;=\; 1 - (1 - q_x)^t \qquad t \in [0, 1)$$

and

$$\ell_{x+t} \;=\; \ell_x(1 - q_x)^t \qquad t \in [0, 1)$$

12.3 Balducci assumption

The final possible interpolation that we will consider is called the Balducci assumption, which is expressed as

$$ {}_{1-t}q_{x+t} \;=\; (1 - t)q_x \qquad t \in [0, 1)$$

meaning that the probability of death before time $x + 1$ having survived to $x + t$ is proportional to the length of time remaining $(1 - t)$ with constant of proportionality equal to q_x (so that the assumption is correct for $t = 0$). Then we have

$$1 - q_x \;=\; p_x \;=\; {}_tp_x \, {}_{1-t}p_{x+t} \;=\; (1 - {}_tq_x)[1 - (1 - t)q_x]$$

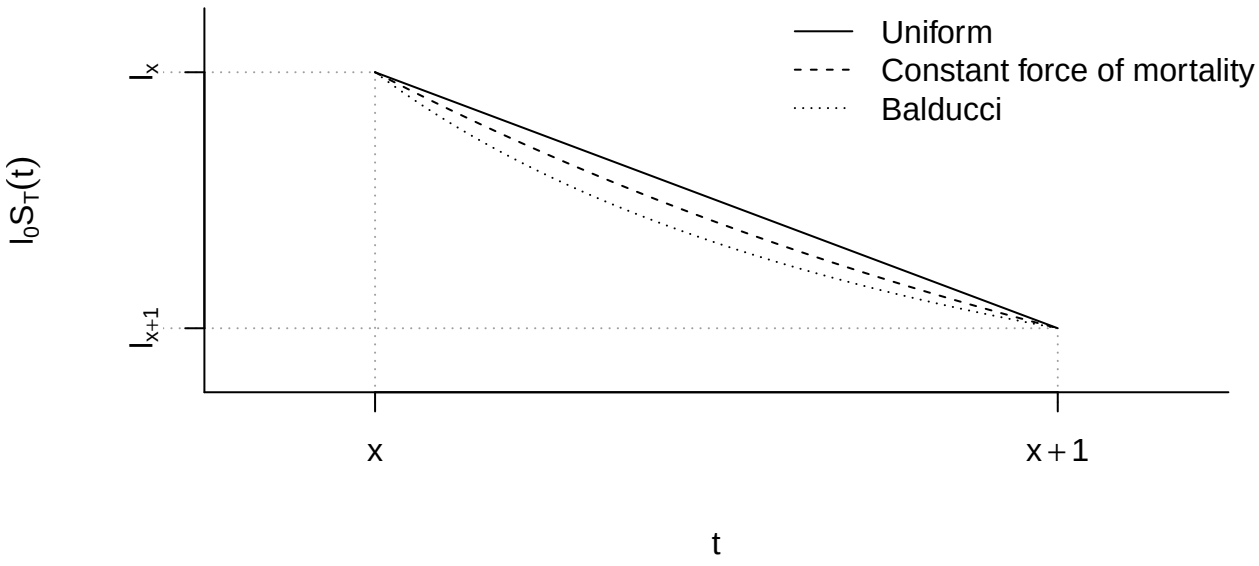
using the Balducci assumption. So

$$ {}_tq_x = 1 - \frac{1-q_x}{1-(1-t)q_x} \qquad t \in [0,1) $$

and

$$ \ell_{x+t} \;=\; \ell_x \frac{1-q_x}{1-(1-t)q_x} \qquad t \in [0,1) $$

12.4 Shape of the survival function



12.5 Implied force of mortality

The uniform distribution of deaths and Balducci assumptions have implications for the force of mortality over the region $[x, x+1)$ to which they are applied.

Based on the known relationship between hazard and survival function, we have

$$\mu_{x+t} = -\frac{\mathrm{d}}{\mathrm{d}t} \log \ell_{x+t}$$

For the uniform distribution of deaths this gives

$$\mu_{x+t} = \frac{q_x}{1 - tq_x} \qquad t \in [0, 1)$$

an increasing function of t and for the Balducci assumption, we get

$$\mu_{x+t} = \frac{q_x}{1 - (1-t)q_x} \qquad t \in [0, 1)$$

a decreasing function of t.

12.6 Other intervals

Where a life table contains an interval $g \neq 1$, so that ${}_gq_x$ is presented for (at least) one x, we have the straightforward generalisations:

$${}_tq_x = \frac{t}{g}{}_gq_x \qquad t \in [0, g) \qquad \text{(Uniform)}$$

$$
\begin{aligned}
{}_tq_x &= 1-(1-{}_gq_x)^{t/g} \qquad t\in[0,g) \qquad \text{(Constant force)} \\
{}_tq_x &= 1-\frac{1-{}_gq_x}{1-\left(1-\frac{t}{g}\right){}_gq_x} \qquad t\in[0,g) \qquad \text{(Balducci)}
\end{aligned}
$$

12.7 The life table quantity μ_x

Sometimes a life table will include a column of μ_x values which are the forces of mortality at the exact ages x in the life table. We know that

$$
\mu_{x+t} = -\frac{\mathrm{d}}{\mathrm{d}t}\log \ell_{x+t} = -\frac{1}{\ell_{x+t}}\frac{\mathrm{d}\ell_{x+t}}{\mathrm{d}t}
$$

This can be estimated using the life table quantities ℓ_x if we approximate ℓ_{x+t} around x by a polynomial.

If ℓ_{x+t} is well approximated by a quadratic, we have

$$
\ell_{x+t} = at^2+bt+c \qquad \Rightarrow \qquad \begin{cases} \ell_{x+1} = a+b+c \\ \ell_{x-1} = a-b+c \end{cases}
$$

and therefore at $t=0$

$$
\mu_x = -\frac{1}{\ell_x}\frac{\mathrm{d}\ell_x}{\mathrm{d}t} = -\frac{b}{\ell_x} = \frac{\ell_{x-1}-\ell_{x+1}}{2\ell_x}
$$

An alternative quartic approximation gives

$$\mu_x = \frac{8(\ell_{x-1} - \ell_{x+1}) - (\ell_{x-2} - \ell_{x+2})}{12\ell_x}$$

Chapter 13

Life Table Models

The life table is a summary of the distribution of a variable representing the length of a (usually human) life time.

Life tables are produced by estimating the distribution, using observed data on the relevant population.

The link between the data and the distribution is the statistical model for the data.

In this section, we introduce some different statistical models which can be used to construct a life table (and to quantify the uncertainty about life table quantities) based on observed data.

We will assume throughout this section that the life table intervals are whole years (although the results generalise easily).

13.1 The binomial model

In its simplest form, the binomial model assumes that we observe individuals through whole years of age, so either the sample is completely closed (no one enters or is censored) or individuals can only enter or leave the sample under observation at an exact age x.

Then the data can be presented as

Age	$\cdots$	$x-1$	x	$x+1$	$\cdots$
Initial exposed to risk	$\cdots$	E^0_{x-1}	E^0_x	E^0_{x+1}	$\cdots$
Deaths in $[x, x+1)$	$\cdots$	D_{x-1}	D_x	D_{x+1}	$\cdots$

where the *initial exposed to risk*, E_x^0, at age x is the number of individuals alive and in our observation sample on their xth birthday.

In a completely closed sample $E_{x+1}^0 = E_x^0 - D_x$.

13.1.1 The binomial likelihood

The D_x are observations of independent binomial(E_x^0, q_x) random variables, so

$$L(q_x) \;=\; \binom{E_x^0}{D_x} q_x^{D_x} (1-q_x)^{E_x^0 - D_x}$$

The log-likelihood is then

$$\ell(q_x) = \log \binom{E_x^0}{D_x} + D_x \log q_x + (E_x^0 - D_x) \log(1-q_x).$$

Taking first derivative with respect to q_x:

$$\frac{\mathrm{d}\ell(q_x)}{\mathrm{d}q_x} = \frac{D_x}{q_x} - \frac{E_x^0 - D_x}{1-q_x}.$$

Setting equal to 0 and solving

$$0 = \frac{D_x}{\hat{q}_x} - \frac{E_x^0 - D_x}{1 - \hat{q}_x}$$

$$\Rightarrow \qquad \hat{q}_x = \frac{D_x}{E_x^0}$$

The second derivative of $\ell(q_x)$ with respect to q_x is

$$\frac{\mathrm{d}^2\ell(q_x)}{\mathrm{d}q_x^2} = -\frac{D_x}{q_x^2} - \frac{E_x^0 - D_x}{(1-q_x)^2}.$$

The Fisher information is

$$\begin{aligned} I(q_x) &= \mathrm{E}\left[-\frac{\mathrm{d}^2\ell(q_x)}{\mathrm{d}q_x^2}\right] \\ &= \frac{\mathrm{E}\left[D_x\right]}{q_x^2} + \frac{E_x^0 - \mathrm{E}\left[D_x\right]}{(1-q_x)^2} \\ &= \frac{E_x^0}{q_x} + \frac{E_x^0}{(1-q_x)} \\ &= \frac{E_x^0}{q_x(1-q_x)} \end{aligned}$$

Therefore

$$s.e.(\hat{q}_x) = \left(\frac{\hat{q}_x(1-\hat{q}_x)}{E_x^0}\right)^{1/2}$$

Although this approach is feasible for a cohort study, where the same sample is followed from birth, it is

limited for more general studies (for example where a population is observed for a fixed period of time) where individuals enter and exit at ages other than exact birthdays.

13.1.2 Data example

Suppose that we are constructing a period life table, based on deaths observed in the year 2013, and that our data on 80-year olds ($n = 6$ individuals) is as follows:

Life	Date of 80^{th} birthday	Date of death (if died during 2013)	Other information
1	1 June 2012	1 March 2013	
2	1 November 2012		
3	1 January 2013		
4	1 March 2013		Lost to follow-up on 1 May 2013
5	1 March 2013	1 September 2013	
6	1 September 2013		

Note that only individual 3 was observed through the whole of $[80, 81)$.

How do we estimate q_{80}?

13.1.3 Likelihood for ${}_{b-a}q_{x+a}$

In the example in Section 13.1.2, we have each life i observed over a period $[x + a_i, x + b_i]$ where $0 \leq a_i < b_i \leq 1$.

Where observation ends in death, then b_i is the time at which *observation would have ended* if death had not happened. Hence we have

Life	1	2	3	4	5	6
a_i	7/12	2/12	0	0	0	0
b_i	1	1	1	2/12	10/12	4/12
y_i	1	0	0	0	1	0

The likelihood for a sample of size n is

$$L = \prod_{i=1}^{n} {}_{b_i-a_i}q_{x+a_i}^{y_i}(1 - {}_{b_i-a_i}q_{x+a_i})^{1-y_i}$$

so here

$$L = {}_{0.42}q_{80.58}\ {}_{0.83}q_{80}(1 - {}_{0.83}q_{80.17})(1 - q_{80})(1 - {}_{0.17}q_{80})(1 - {}_{0.33}q_{80})$$

13.1.4 Approximate likelihood for q_x

To estimate q_x we need to re-write the likelihood in Section 13.1.3 in terms of q_x by using one of the interpolation formulae in Chapter 12.

Generally, we have

$$\begin{aligned} {}_{b-a}q_{x+a} &= 1 - {}_{b-a}p_{x+a} &= 1 - S_T(b-a|x+a) \\ &= 1 - \frac{S_T(x+b)}{S_T(x+a)} \\ &= 1 - \frac{S_T(b|x)}{S_T(a|x)} \\ &= 1 - \frac{{}_bp_x}{{}_ap_x} \\ &= 1 - \frac{1 - {}_bq_x}{1 - {}_aq_x} \end{aligned}$$

Therefore under the assumption of uniform distribution of deaths (UDD; ${}_tq_x = tq_x$) we have

$${}_{b-a}q_{x+a} = \frac{(b-a)q_x}{1 - aq_x}.$$

Under constant force of mortality (${}_tq_x = 1 - (1-q_x)^t$) we have

$${}_{b-a}q_{x+a} = 1 - (1-q_x)^{b-a}.$$

Under Balducci (${}_tq_x = 1 - \frac{1-q_x}{1-(1-t)q_x}$) we have

$$ {}_{b-a}q_{x+a} = \frac{(b-a)q_x}{1-(1-b)q_x}. $$

All of these can be written as

$$ \begin{aligned} {}_{b-a}q_{x+a} &= (b-a)q_x + O(q_x^2) \\ &\approx (b-a)q_x \end{aligned} $$

if q_x is small.

13.1.5 Likelihood under UDD

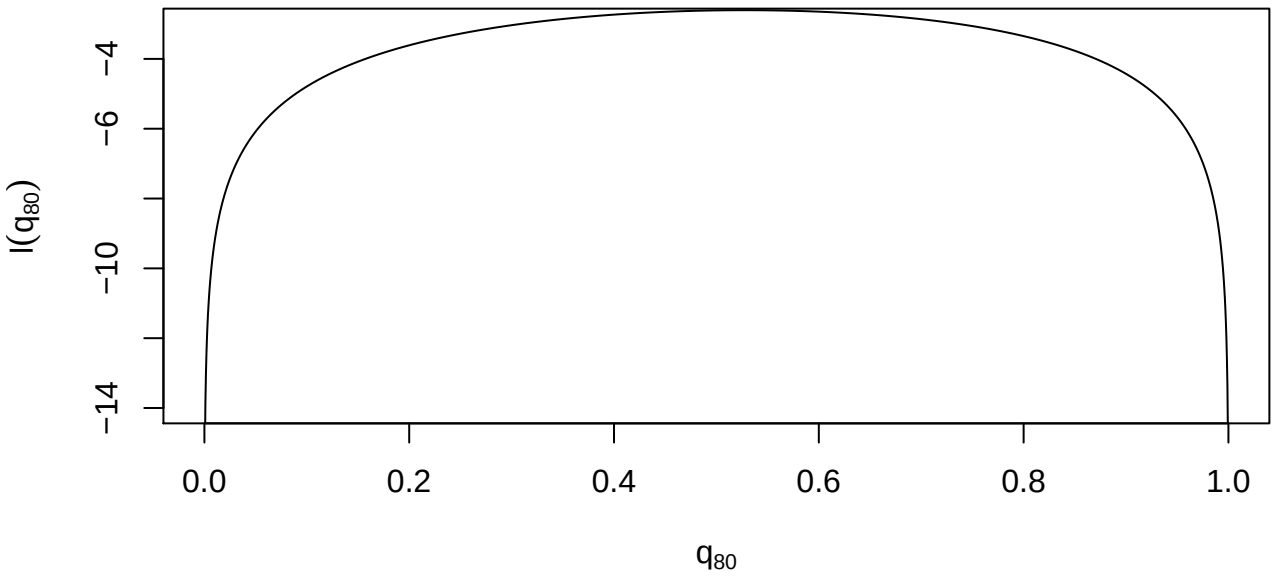
Under the uniform distribution of deaths

$$ \begin{aligned} L(q_x) &= \prod_{i=1}^{n} \left(\frac{(b_i - a_i)q_x}{1 - a_i q_x} \right)^{y_i} \left(1 - \frac{(b_i - a_i)q_x}{1 - a_i q_x} \right)^{1-y_i} \\ \Rightarrow \quad \ell(q_x) &= C + \sum_{i=1}^{n} y_i \log q_x + \sum_{i=1}^{n} (1 - y_i) \log(1 - b_i q_x) \\ &\quad - \sum_{i=1}^{n} \log(1 - a_i q_x) \end{aligned} $$

13.1.6 Log-likelihood example

For the data in Section 13.1.2,

$$\ell(q_{80}) = C + 2\log q_{80} + 2\log(1-q_{80}) + \log\left(1-\frac{4}{12}q_{80}\right) - \log\left(1-\frac{7}{12}q_{80}\right)$$



Numerical maximisation is necessary, giving

$$\hat{q}_{80} = 0.528 \qquad s.e.(\hat{q}_{80}) = 0.253$$

13.1.7 A simple estimator for q_x

Under the approximation

$$ {}_{b-a}q_{x+a} = (b-a)q_x $$

then, denoting by Y_i the random variable representing death of the ith individual, we have

$$ E(Y_i) = P(Y_i = 1) = (b_i - a_i)q_x \quad \Rightarrow \quad E\left(\sum_{i=1}^{n} Y_i\right) = q_x \sum_{i=1}^{n} (b_i - a_i) $$

so a (method-of-moments) estimator of q_x is

$$ \tilde{q}_x \;=\; \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(b_i - a_i)} \;=\; \frac{D_x}{\sum_{i=1}^{n}(b_i - a_i)} \;=\; \frac{D_x}{E_x^0} $$

where $E_x^0 = \sum_{i=1}^{n}(b_i - a_i)$ is the *initial exposed to risk* (in this more general situation).

For small values of q_x

$$ s.e.(\tilde{q}_x) \;\approx\; \left(\frac{\hat{q}_x}{E_x^0}\right)^{1/2} $$

so for the data on 14.4-14.5, we have $E_x^0 = \frac{43}{12}$, so

$$ \tilde{q}_{80} = \frac{2}{43/12} = 0.558 \qquad s.e.(\tilde{q}_{80}) = 0.395 $$

The estimate is close to the m.l.e. (0.528) but the standard error approximation is not accurate here, as q_{80} is not small.

13.1.8 Central exposed to risk

The central exposed to risk at age x, E_x^C is defined as the total time of observation in the age range $[x, x+1)$ of all individuals under study, where observation ends either with death, censoring, or the end of the study period.

$$E_x^C = \sum_{i=1}^{n} (b_i' - a_i)$$

where

$$b_i' = \begin{cases} b_i & \text{if } y_i = 0 \text{ (death not observed)} \\ t_i & \text{if } y_i = 1 \text{ (death observed at age } x + t_i,\ 0 \le t_i < 1) \end{cases}$$

Therefore

$$E_x^C = E_x^0 - \sum_{i=1}^{n} y_i (b_i - t_i)$$

13.1.9 The actuarial estimator

As we shall see in Chapter 14, it is usually easier to compute E_x^C from the available data than it is to

compute E_x^0.

Hence it is common to approximate E_x^0 using

$$E_x^C \;=\; E_x^0 - \sum_{i=1}^{n} y_i(b_i - t_i)$$

and two further assumptions/approximations:

1. Observed deaths (if they hadn't died) would have been observed through to $b_i = 1$

2. Observed death times $t_i \in [0, 1)$ are uniformly distributed through the interval

The consequence of 1 and 2 is that

$$E_x^C \;\approx\; E_x^0 - \frac{1}{2}D_x$$

Using this approximation, the estimator in 14.10 can be expressed as

$$\tilde{q}_x \;=\; \frac{D_x}{E_x^0} \;=\; \frac{D_x}{E_x^C + \frac{1}{2}D_x}$$

which is sometimes called the *actuarial estimator* for q_x.

For the data in Section 13.1.2, we have

Life	1	2	3	4	5	6
a_i	7/12	2/12	0	0	0	0
b_i	1	1	1	2/12	10/12	4/12
b_i'	9/12	1	1	2/12	6/12	4/12
y_i	1	0	0	0	1	0

so $E_x^C = \frac{36}{12}$ and $\tilde{q}_{80} = \frac{2}{36/12+2/2} = 0.5 \quad (s.e.(\tilde{q}_{80}) = 0.354)$.

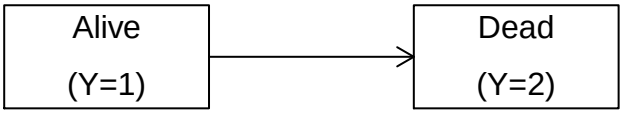
13.2 Force of mortality

An alternative approach to estimating the life table is to focus estimation on the force of mortality (hazard) μ_x rather than the death probabilities (survival function) q_x.

13.2.1 The two-state model

Whereas a binomial model is natural when considering q_x, a two-state model is much more natural for μ_x.

The two-state model is represented by



and parameterised by the transition intensity (force of mortality) $\mu_{12}(x) \equiv \mu_x$ from state 1 (alive) to state 2 (dead).

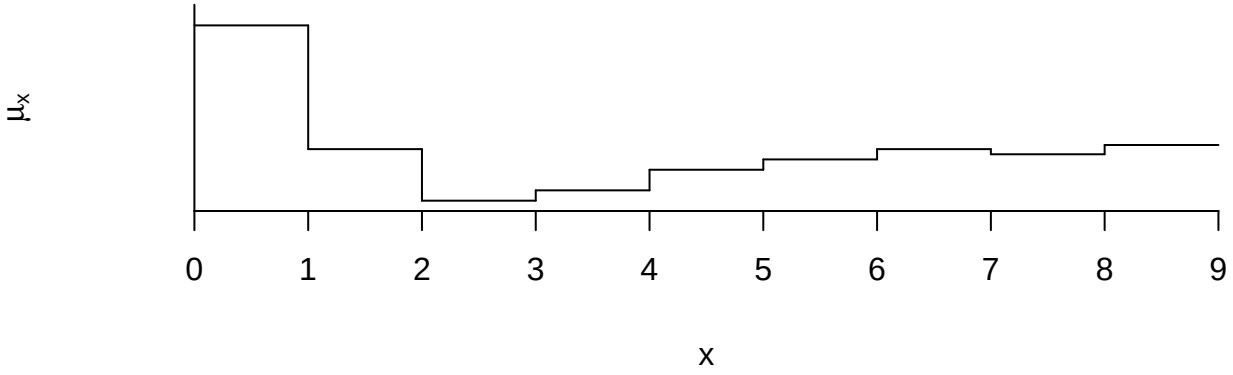
13.2.2 Estimating μ_x

We recall that maximum likelihood estimation of transition intensities in a multi-state model is easy in the time-homogenous case $\mu_x = \mu$, for all x.

For human lifetime models it is unrealistic to assume $\mu_x = \mu$, for all $x > 0$, but it may be reasonable to assume, for $x = 0, 1, 2, \ldots$, that

$$\mu_{x+t} = \mu_{x+\frac{1}{2}}, \quad t \in [0, 1)$$

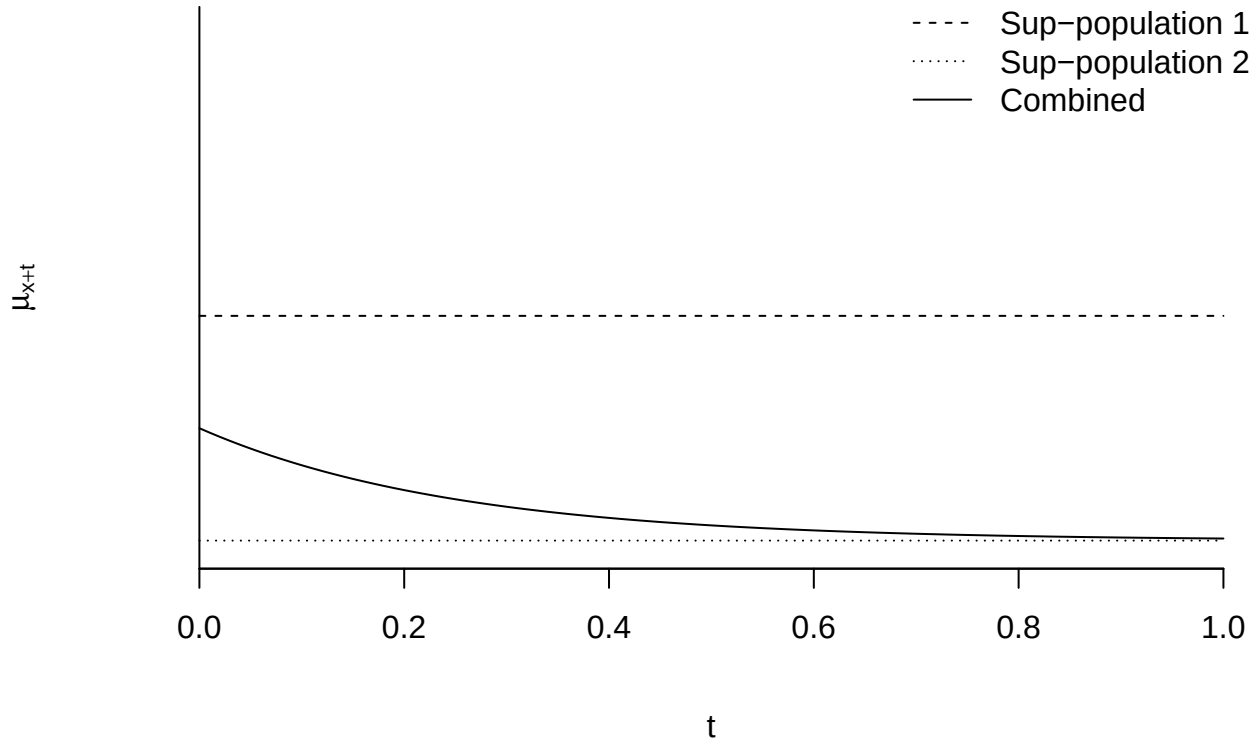
so the force of mortality is modelled as a step function:



13.2.3 Homogeneity assumption

The assumption that $\mu_{x+t} = \mu_{x+\frac{1}{2}},\ t \in [0, 1)$ is usually reasonable for human mortality, because, except in very early life, μ_x varies slowly with x.

However it is also necessary for the population being modelled to be as homogeneous as possible, so ideally separate models should be used for males and females, and for any other variables which are recorded and which may influence mortality (for example smokers and non-smokers).



13.2.4 Maximum likelihood for μ_x

Recall from Chapter 9 that the m.l.e. for a transition intensity μ_x in a multi-state model is

$$\begin{aligned}\hat{\mu}_{x+\frac{1}{2}} &= \frac{\text{Observed number of transitions to death state in } [x, x+1)}{\text{Total observed holding time in alive state in } [x, x+1)} \\ &= \frac{D_x}{\sum_{i=1}^{n}(b'_i - a_i)} = \frac{D_x}{E_x^C}\end{aligned}$$

with

$$s.e.\left(\hat{\mu}_{x+\frac{1}{2}}\right) = \frac{\hat{\mu}_{x+\frac{1}{2}}}{d_x^{1/2}}$$

For the data on 14.4, we have

$$\hat{\mu}_{80.5} = \frac{2}{\frac{36}{12}} = 0.667 \qquad s.e.(\hat{\mu}_{80.5}) = \frac{0.667}{2^{1/2}} = 0.471$$

13.2.5 Mortality rates

The estimator $\hat{\mu}_{x+\frac{1}{2}} = D_x/E_x^C$ is called the observed (or crude) *central mortality rate* and denoted by $\hat{m}_x$.

Hence

$$\hat{m}_x = \frac{D_x}{E_x^C}$$

Similarly, the estimator

$$\tilde{q}_x = \frac{D_x}{E_x^0}$$

is sometimes called the observed *initial mortality rate* (though this is a bit of a misnomer as this estimates a probability and not a hazard rate).

The crude central mortality rate $\hat{m}_x$ is an estimator of the population central mortality rate, defined as

$$m_x = \frac{E(D_x)}{E_x^C}$$

the ratio of the expected number of deaths in $[x, x+1)$ to the central exposed to risk, E_x^C. So $E(\hat{m}_x) = m_x$.

13.2.6 Central mortality rate

In a 'closed' population, where exits (except through death) are balanced by entries, we have

$$E(D_x) = E_x^0 q_x \qquad \text{and} \qquad E_x^C = L_x = E_x^0 \int_0^1 {}_tp_x \, \mathrm{d}t$$

Hence, the population central mortality rate is

$$m_x = \frac{E(D_x)}{E_x^C} = \frac{E_x^0 q_x}{E_x^0 \int_0^1 {}_tp_x \, \mathrm{d}t} = \frac{q_x}{\int_0^1 {}_tp_x \, \mathrm{d}t}$$

This is an alternative definition of m_x which can be used more generally. Under the assumption of constant force of mortality $\mu_{x+t} = \mu_{x+\frac{1}{2}}$ over $[x, x+1)$,

$$m_x = \frac{q_x}{\int_0^1 {}_tp_x \mathrm{d}t} = \frac{1 - \exp(-\mu_{x+\frac{1}{2}})}{\int_0^1 \exp(-t\mu_{x+\frac{1}{2}})\mathrm{d}t} = \mu_{x+\frac{1}{2}}$$

which is consistent with the use of the crude central mortality rate being the m.l.e. of $\mu_{x+\frac{1}{2}}$ under this assumption.

13.2.7 Estimating q_x from $\hat{m}_x$

Under the assumption of a constant force of mortality over $[x, x+1)$, we have $m_x = -\log(1 - q_x)$ and hence, we estimate q_x using

$$\hat{q}_x = 1 - \exp(-\hat{m}_x) \tag{13.1}$$

Alternatively, under the assumption of a uniform distribution of deaths over $[x, x+1)$, we have

$$m_x = \frac{q_x}{1 - \frac{1}{2}q_x}$$

leading to the actuarial estimator

$$\tilde{q}_x = \frac{\hat{m}_x}{1+\frac{1}{2}\hat{m}_x} = \frac{\frac{D_x}{E_x^C}}{1+\frac{1}{2}\frac{D_x}{E_x^C}} = \frac{D_x}{E_x^C+\frac{1}{2}D_x} \tag{13.2}$$

Typically life tables are computed by obtaining $\hat{m}_x$ and then transforming to an estimate of q_x using (1) or (2) above.

13.2.8 Similarity of $\hat{q}_x$ and $\tilde{q}_x$

$$\begin{aligned}\hat{q}_x = 1-p_x &= 1-\exp(-\hat{m}_x)\\ &= 1-\left(1-\hat{m}_x+\frac{1}{2}\hat{m}_x^2-\frac{1}{6}\hat{m}_x^3+\ldots\right)\\ &= \hat{m}_x-\frac{1}{2}\hat{m}_x^2+O(\hat{m}_x^3)\end{aligned}$$

$$\begin{aligned}\tilde{q}_x &= \frac{\hat{m}_x}{1+\frac{1}{2}\hat{m}_x}\\ &= \hat{m}_x\left(1-\frac{1}{2}\hat{m}_x+\frac{1}{4}\hat{m}_x^2+\ldots\right)\\ &= \hat{m}_x-\frac{1}{2}\hat{m}_x^2+O(\hat{m}_x^3)\end{aligned}$$

So the two estimators of q_x are typically very close. For the data on 14.4, we have $\hat{q}_{80} = 0.487$ $\tilde{q}_{80} = 0.5$

13.3 The Poisson model

The Poisson model assumes that the D_x are observations of independent Poisson$(m_x E_x^C)$ random variables.

Hence,

$$\begin{aligned} & L(m_x) &=& \frac{\exp\left(-m_x E_x^C\right)\left(m_x E_x^C\right)^{D_x}}{D_x!} \\ \Rightarrow \quad & \ell(m_x) &=& C - m_x E_x^C + D_x \log m_x \\ \Rightarrow \quad & \hat{m}_x &=& \frac{D_x}{E_x^C} \end{aligned}$$

the same estimator as for the two state model in Section 13.2.5 (with the same standard error).

13.3.1 Poisson model v. multi-state model

As models for estimation, the Poisson and two-state models lead to identical inferences.

Generally, the two-state interpretation is preferred because:

- The model is an exact description of the process, whereas the Poisson is approximate
  - a Poisson$\left(\mu_{x+\frac{1}{2}} E_x^C\right)$ distribution actually allows the number of deaths D_x to exceed n the number of lives under observation.
  - the Poisson model treats E_x^C as fixed and known in advance (it is not)
- Although the two-stage model is most easily estimated in the case of constant μ_{x+t} for $t \in [0, 1)$, it can be estimated more generally (using Kaplan-Meier etc). The Poisson mode requires the assumption of constant force of mortality.
- Both models extend to examples with multiple decrements (absorbing states) but only the multi-state approach allows transitions to non-absorbing states.

13.4 Binomial model v. multi-state model

Generally, the two-state model is preferred because:

- We can perform maximum likelihood estimation exactly in the two-state model, whereas the Binomial model usually requires additional assumptions, and is usually more complicated.
- The multi-state model extends to examples with multiple decrements (absorbing states) and increments (transitions to non-absorbing states) but the binomial model is only valid for the two-state (alive/dead).
- The two-state model uses the exact times of deaths, whereas the Binomial model only uses the number of deaths.

Chapter 14

Exposure to Risk

As defined in Chapter 13:

The central exposed to risk at age x, E_x^C is the total time of observation in the age range $[x, x+1)$ of all individuals under study, where observation ends either with death, censoring, or the end of the study period.

$$E_x^C = \sum_{i=1}^{n} (b_i' - a_i) \tag{14.1}$$

where

- a_i is the earliest age in $[x, x+1)$ at which individual i was observed within the study period
- b_i' is the latest age in $[x, x+1)$ at which individual i was alive and observed within the study period

14.1 Calculating E_X^C

The central exposed to risk at age x, E_x^C can be calculated using (1) if we have records, at individual level, of precise ages at entry into the study and at death or other exit from the study.

However, in mortality studies, records are often secondary data sources where such precise information is not available. For example, it is common to have records of:

- the number of deaths in the age range $[x, x+1)$ within the study period
- the total number of individuals within particular age ranges,such as $[x, x+1)$, at set dates (often January 1 each year)
- nothing else.

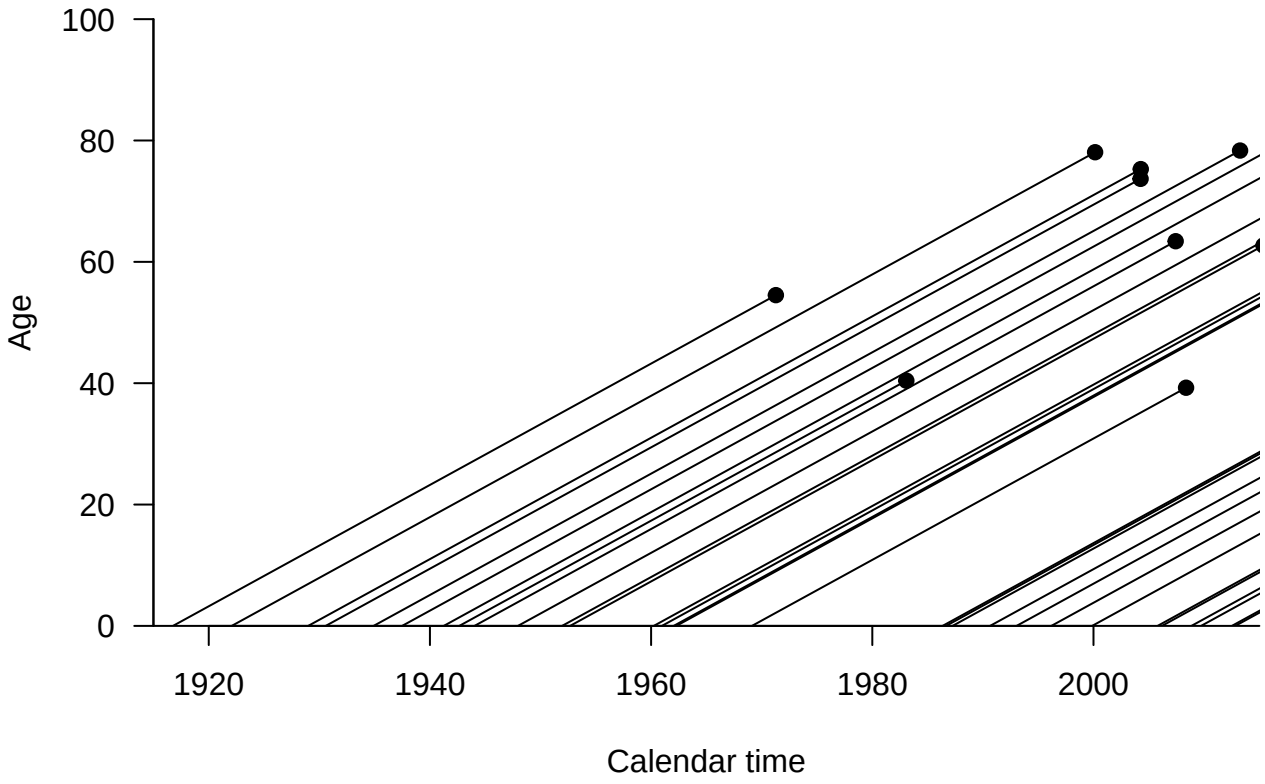
How can we calculate E_x^C using such information?

14.2 Lexis diagrams

The Lexis diagram is a useful way of visualising lifetime data.

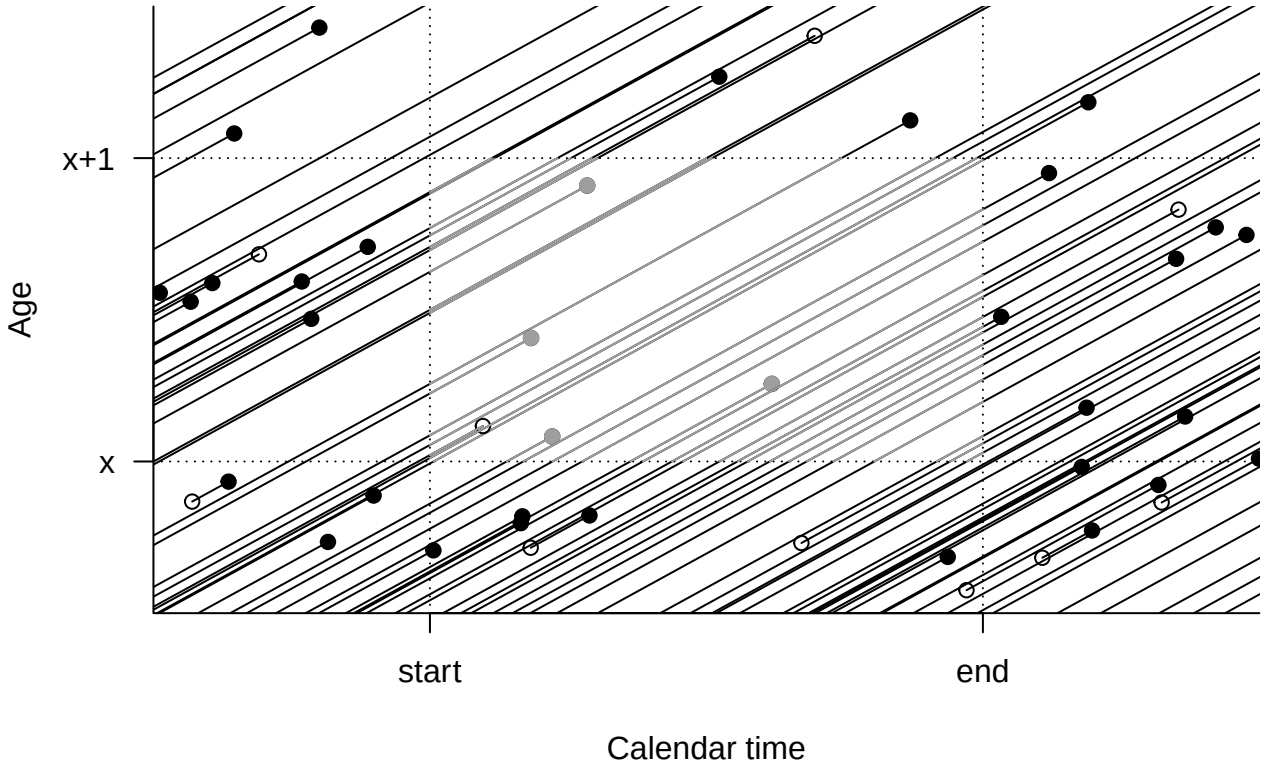
It plots an individual life trajectory as a line on a graph where the y-axis is age and the x-axis is calendar year.

The following Lexis diagram shows 30 randomly selected lives born between 1915 and 2015. A death (before end 2014) is marked by a solid circle.



14.2.1 Lexis diagram over a study period

The lifetimes which are relevant for the calculation of E_x^C can be identified in the Lexis diagram.



Life lengths included in the calculation of E_x^C (for the study period with start and end dates identified) are identified in grey. Deaths included in the calculation of $\hat{\mu}_{x+\frac{1}{2}}$ and/or $\hat{q}_x$ are identified in grey. Observation periods which end for reason other than birth and death are identified by unfilled circles.

14.3 Calculating E_x^C

The central exposed to risk at age x, E_x^C is the total time represented by all the grey lines in the picture in Section 14.2.1.

If we denote by $P_x(t)$ the number of individuals in our study with ages in the interval $[x, x+1)$ at exact time t (the number of red lines crossing the vertical line, calendar time $= t$), then

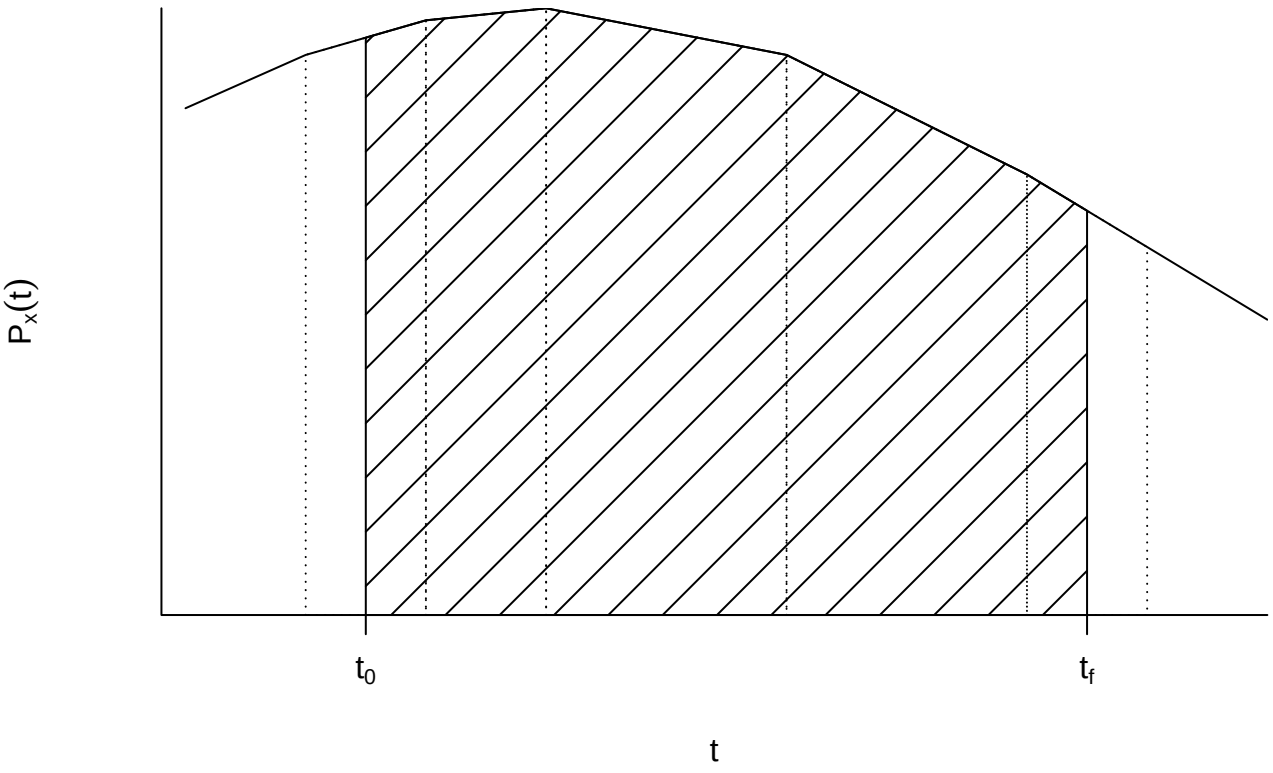
$$E_x^C \;=\; \int_{t_0}^{t_f} P_x(t)\mathrm{d}t$$

where t_0 and t_f are the start and end dates of the study, respectively.

14.4 Approximating E_x^C

In practice, we do not know $P_x(t)$ for every calendar time t, but we typically have census data which gives (or allows us to approximate) $P_x(t)$ for certain values of t.

Then, we approximate E_x^C using the trapezium rule, which approximates $P_x(t)$ by a linear function between the known values.



Trapezium rule approximation to E_x^C over the period $[t_0, t_f]$ is given by the shaded area. Census times (at which values of $P_x(t)$ are available) are indicated by the vertical dotted lines.

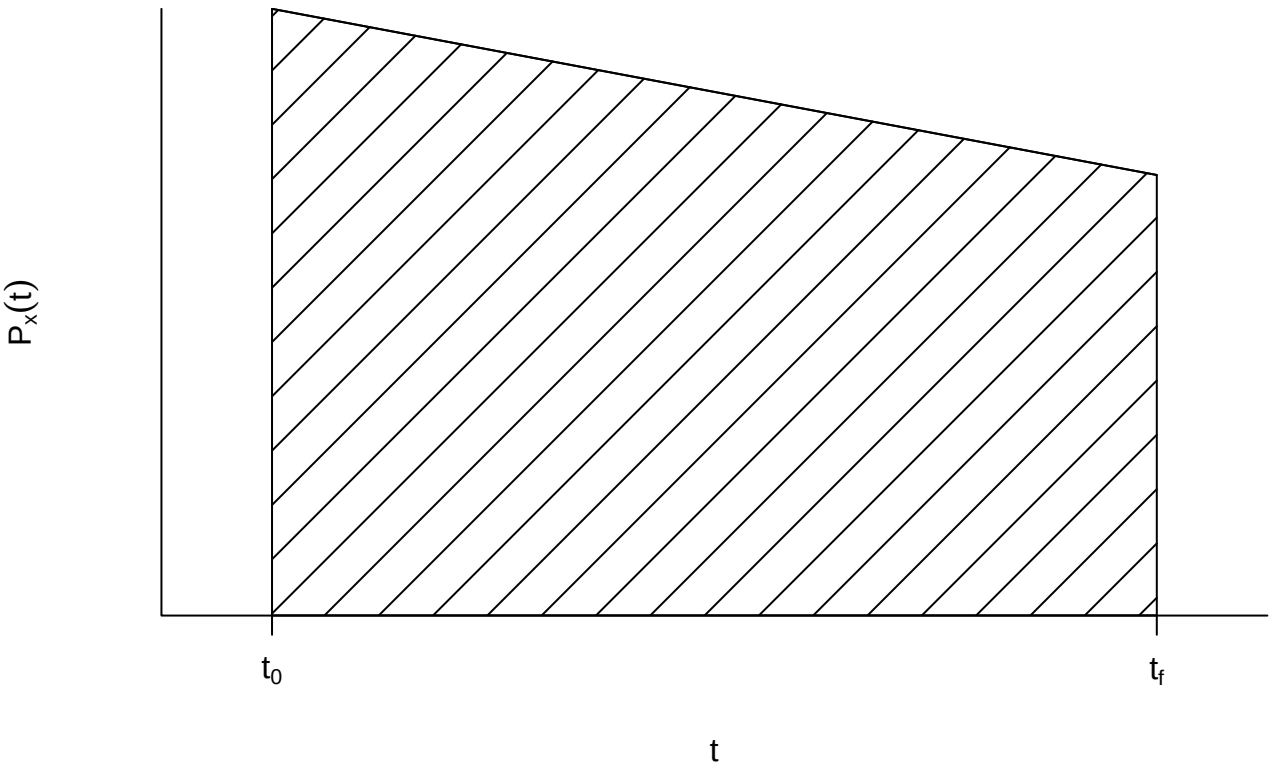
If $P_x(t)$ is available at $t_0, t_1, \ldots, t_{f-1}, t_f$ then the trapezium rule for E_x^C over the period $[t_0, t_f]$ is

$$E_x^C \;=\; \int_{t_0}^{t_f} P_x(t)\mathrm{d}t \;\approx\; \sum_{i=1}^{f} \frac{1}{2}[P_x(t_i) + P_x(t_{i-1})](t_i - t_{i-1})$$

This is known as the *census approximation* for the central exposed to risk.

When there is census information at t_0 and t_f but at no time in between, we have

$$E_x^C \;=\; \int_{t_0}^{t_f} P_x(t)\mathrm{d}t \;\approx\; \frac{1}{2}[P_x(t_0) + P_x(t_f)](t_f - t_0)$$



If $t_f - t_0 = 1$ year, then E_x^C is an estimate of the *mid-year* population at age x (population at time $t_0 + 1/2$).

14.5 The principle of correspondence

The principle of correspondence as it applies to the estimation of mortality rates states that:

A life alive at time t should be included in the exposure at age x at time t if and only if, were that life to die immediately, it would be counted in the deaths data at age x.

Thus far, we have been careful to respect this principle. For example, this is why lives entering into a study at age $x + t$ where $0 < t < 1$ can only be counted towards the exposure from age $x + t$ (even though we know they were alive at ages between x and $x + t$).

It is important to bear this principle in mind when data on deaths and census data use different age definitions.

14.6 Age definitions

Ages may be defined by

- Age at last birthday, so ages in $[x, x + 1)$ are recorded as x.
- Age at next birthday, so ages in $[x - 1, x)$ are recorded as x.
- Age at nearest birthday, so ages in $\left[x - \frac{1}{2}, x + \frac{1}{2}\right)$ are recorded as x.

If data on deaths and census data use different age definitions, then we transform (approximately) the census data to match the definition used in the death records, $\{D_x\}$.

This is generally straightforward.

For example if death data uses age at last birthday, and census data $P_x(t)$ records age at next birthday, then the census approximation

$$E_x^C \approx \sum_{i=1}^{f} \frac{1}{2}[P_x(t_i) + P_x(t_{i-1})](t_i - t_{i-1})$$

needs to be amended with P_x replaced by P_{x+1}.

14.7 Half year adjustments

Where census ages are defined by age at last birthday or age at next birthday and need to be adjusted to age at nearest birthday, or vice versa, then some further approximation is required.

Suppose that $P_x(t)$ records age at nearest birthday, but we require the number of lives aged x at last birthday (at time t). Then, at time t

- those lives with ages in $\left[x, x+\frac{1}{2}\right)$ are counted within $P_x(t)$
- those lives with ages in $\left[x+\frac{1}{2}, x+1\right)$ are counted within $P_{x+1}(t)$

If we make the additional assumption that, within each age x, the $P_x(t)$ birthdays are uniformly distributed through the calendar year, then the census count at time t under our required age definition (aged x at last birthday) is

$$\frac{1}{2}\left[P_x(t) + P_{x+1}(t)\right]$$

Other examples are similar - just use common sense!

14.8 Presenting the estimates

Care needs to be taken to be clear about which age range any estimated mortality rates $\hat{\mu}$ and estimated death probabilities $\hat{q}$ correspond to.

Recall that in the standard approach, where the age in the definition of deaths (and hence for the calculation of estimates) is age at last birthday, we use q_x to indicate a probability of death in $[x, x+1)$ given survival to x and $\mu_{x+\frac{1}{2}}$ to indicate a constant force of mortality over $[x, x+1)$.

If a different age definition is used for deaths, then we need to adjust these subscripts accordingly, so that the subscript on q is still the start of the interval, and the subscript on μ is the middle of the interval.

- For age at nearest birthday, our age range is $\left[x-\frac{1}{2}, x+\frac{1}{2}\right)$ and hence we are estimating $q_{x-\frac{1}{2}}$ and μ_x.
- For age at next birthday, our age range is $[x-1, x)$ and hence we are estimating q_{x-1} and $\mu_{x-\frac{1}{2}}$.

14.9 Rate interval

A mortality rate is defined with respect to a particular *rate interval*, the period over which a life has a particular age for the purposes of calculation of the mortality rate.

We have focussed on *life year rate intervals* where a life changes age at a date determined by birthday (either on the birthday or at the midpoint between birthdays).

Other possibilities are used when exact birthdays are unknown

- Policy year rate intervals, where a life changes age at a fixed date annually (usually the anniversary date of a particular insurance policy) – used when only age at last birthday, at inception of policy, is known.
- Calendar year rate intervals, where a life changes age annually on January 1 – used when year of birth is known.

The correct calculation in these cases can be derived by constructing the Lexis diagram where the y-axis is age as defined by the rate interval.

Chapter 15

Comparing Mortality Rates

Standard mortality rates are mortality rates for a (usually) large reference population. For example, national life tables provide standard mortality rates.

It is often of interest to compare a set of observed mortality rates, for a study population, with a relevant set of standard rates.

We denote the standard mortality rate at age x by m_x^S, and the corresponding observed death count and total exposure in the study population by D_x and E_x^C respectively (it is assumed that the observed and standard rates are comparable in terms of definition).

A *standardised mortality ratio* r_x at age x is the ratio of the number of observed deaths to the number of deaths expected in the standard population for the same exposure

$$r_x \;=\; \frac{D_x}{m_x^S E_x} \;=\; \frac{\hat{m}_x}{m_x^S}$$

15.1 Example

For example, suppose that we observe a regional population of 60-70 year old males over a short period, and wish to compare observed mortality rates with the latest interim life tables (England 2011-2013).

x	m_x^S	E_x^C	D_x	r_x
60	0.00794	762.0	8	1.32
61	0.00863	755.2	10	1.53
62	0.00949	761.8	5	0.69
63	0.01019	752.0	8	1.04
64	0.01100	733.0	13	1.61
65	0.01199	709.0	16	1.88
66	0.01330	680.8	18	1.99
67	0.01479	657.5	12	1.23
68	0.01586	640.7	14	1.38
69	0.01781	632.0	11	0.98

There seems to be some large discrepancies between mortality rates in our study population and the standard population, i.e. the values of r_x are not close to one.

15.2 Formal comparison by hypothesis test

A formal comparison involves testing the null hypothesis that our observed death pattern has been generated by the standard rates.

- Rejection of the null hypothesis leads to the conclusion that mortality rates in the study population are significantly different from the standard rates.
- Non-rejection of the null hypothesis leads to the conclusion that the standard rates could provide a reasonable model for the study population.

15.3 Large sample distributions

We know that, for the two-state or Poisson models, the central mortality rate

$$\hat{m}_x = \frac{D_x}{E_x^C}$$

is the maximum likelihood estimate for the central mortality rate m_x (equivalently an assumed constant force of mortality μ_x over $[x, x+1)$), and hence in large samples we can use the normal approximation

$$\begin{aligned}
\hat{m}_x \sim N\left(m_x, \frac{m_x}{E_x^C}\right) &\Rightarrow \frac{\hat{m}_x - m_x}{(m_x/E_x^C)^{1/2}} \sim N(0,1) \\
&\Rightarrow \frac{(\hat{m}_x - m_x)^2}{m_x/E_x^C} \sim \chi_1^2 \\
&\Rightarrow \frac{(E_x^C \hat{m}_x - E_x^C m_x)^2}{E_x^C m_x} \sim \chi_1^2 \\
&\Rightarrow \frac{(D_x - E_x^C m_x)^2}{E_x^C m_x} \sim \chi_1^2
\end{aligned}$$

15.4 The chi-squared test

As we assume that numbers of deaths in different age years are independent, we have that

$$\sum_x \frac{(D_x - E_x^C m_x)^2}{E_x^C m_x} \sim \chi_v^2$$

where n here is the number of age groups over which mortality rates are being compared.

To test the null hypothesis H_0: $m_x = m_x^S$, we have the test statistic

$$X^2 = \sum_x \frac{(D_x - E_x^C m_x^S)^2}{E_x^C m_x^S}$$

and reject H_0 if the observed value of X^2 is in the upper tail of the χ_v^2 distribution (typically, the threshold for rejection is the 95th percentile).

15.5 Return to the example

For the data in Section 15.1, we have

$$X^2 \;=\; \frac{(8-6.050)^2}{6.050} + \cdots + \frac{(11-11.256)^2}{11.256} \;=\; 23.66$$

In R:

```
> table <- read.table( file = "Chap15_cmr.txt", header = TRUE)
>
> head(table)
```

```
   x     mxs   ECx dx   rx
1 60 0.00794 762.0  8 1.32
2 61 0.00863 755.2 10 1.53
3 62 0.00949 761.8  5 0.69
4 63 0.01019 752.0  8 1.04
5 64 0.01100 733.0 13 1.61
6 65 0.01199 709.0 16 1.88
```

```
> X2 <- sum(((table$dx - table$ECx*table$mxs)^2)/(table$ECx*table$mxs))
> X2
```

```
[1] 23.65563
```

We compare this against a chi-squared distribution with 10 degrees of freedom, and we clearly reject at the 5% level of significance ($\chi^2_{10,0.95} = 18.31$) and even at the 1% level ($\chi^2_{10,0.99} = 23.21$).

In R, the quantiles of the χ^2 are

```
> qchisq(p = 0.95, df = 10)
```

```
[1] 18.30704
```

```
> qchisq(p = 0.99, df = 10)
```

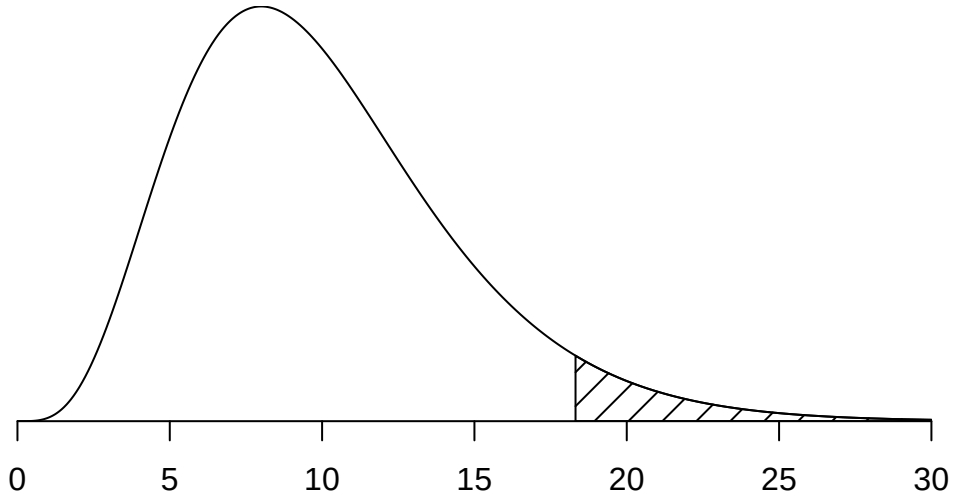


Figure 15.1: χ^2_{10} distribution with upper 5% tail indicated

```
[1] 23.20925
```

We can also calculate a p-value in R:

```
> 1 - pchisq(q = X2, df = 10)
```

```
[1] 0.008568838
```

15.6 Chi-squared test properties

A general rule of thumb is that conclusions from a chi-squared test can be considered reliable if all of the $E_x^C m_x^S$ values are greater than 1, and no more than 20% of them are less than 5.

The chi-squared test is usually a very sound *omnibus* test which can detect a wide range of discrepancies between a set of observed and standard rates.

However, other tests exist which can be more powerful to detect particular kinds of discrepancy, in particular,

- bias, where the observed rates are generally higher or lower then the standard rates.
- patterns of differences, where there are parts of the age spectrum with positive bias (observed rates higher than standard) and parts of the age spectrum with negative bias.
- extreme differences, where there are a small number age groups where the observed and standard rates are very different.

15.7 Multiple comparisons

Caution is required when performing multiple tests of (essentially) the same hypothesis (goodness-of-fit of standard mortality rates to a set of observed rates).

If we have a hypothesis which is true, and we perform 5 independent tests of the hypothesis then, with probability $1 - 0.95^5 = 0.226$, we will observe at least one significant result (at the 5% level).

So, sound advice is to just use a single test, usually an omnibus chi-squared test unless you have an a priori reason to expect a particular kind of discrepancy.

- If the null hypothesis is rejected, examine the patterns of observed and standard rates to identify why there is a significant difference.
- If the null hypothesis is not rejected, you can still examine the patterns of observed and standard rates to investigate whether there might be any cause for concern about a discrepancy not detected by your omnibus test, but be cautious about any further formal hypothesis testing.

15.8 Other tests

Bearing in mind the caution advised in using further tests, we will introduce some other possible tests to detect differences between observed and standard mortality rates.

- Bias
  - Cumulative deviations test
  - Sign test
- Patterns of differences
  - Difference of signs test

  - Serial correlation test
  - Grouping of signs test (not recommended) – better to use the Wald-Wolfowitz runs test

- Extreme differences
  - Examination of individual standardised differences

We consider tests for bias.

15.9 Tests for bias

15.9.1 Cumulative deviations test

We recall from Section 15.3 that we have the approximation

$$\hat{m}_x \sim N\left(m_x\,,\,\frac{m_x}{E_x^C}\right) \Rightarrow \sum_x E_x^C \hat{m}_x \sim N\left(\sum_x E_x^C m_x\,,\,\sum_x E_x^C m_x\right)$$
$$\Rightarrow \frac{\sum_x (D_x - E_x^C m_x)}{\left(\sum_x E_x^C m_x\right)^{1/2}} \sim N(0,1)$$

To test the null hypothesis H_0: $m_x = m_x^S$, we have the test statistic

$$Z = \frac{\sum_x (D_x - E_x^C m_x^S)}{\left(\sum_x E_x^C m_x^S\right)^{1/2}}$$

and reject H_0 if z, the observed value of Z, is in either tail of the standard normal distribution (typically, the threshold for rejection is $|z| > 1.96$, giving a significance level of 5%).

15.9.2 Sign test

The sign test is easy to apply, but generally has low power. The test statistic for the sign test is $S = \sum_x U_x$ where

$$U_x = \begin{cases} 1 & \text{if } \hat{m}_x \geq m_x^S \qquad (\text{so } \hat{m}_x - m_x^S \text{ has a positive sign}) \\ 0 & \text{if } \hat{m}_x < m_x^S \qquad (\text{so } \hat{m}_x - m_x^S \text{ has a negative sign}) \end{cases}$$

Under the null hypothesis H_0: $m_x = m_x^S$, we have that the U_x are independent with $P(U_x = 1) = 0.5$, and therefore that the distribution of S is binomial$(v, 0.5)$ where v is the number of age groups.

We reject H_0 if the observed value of S is in either tail of the binomial.

For all but very small values of v, a normal approximation to the binomial can be used, so we reject H_0 if $|z| > z_{1-\alpha/2}$, at the $\alpha\%$ level, where z is the observed value of

$$Z \;=\; \frac{S \pm \frac{1}{2} - \frac{v}{2}}{\left(\frac{v}{4}\right)^{1/2}}.$$

Note the continuity correction ($+\frac{1}{2}$ if $S < \frac{v}{2}$, $-\frac{1}{2}$ if $S > \frac{v}{2}$)

15.9.3 Return to the example again

Cumulative deviations test

For the data in Section 15.1, we have, for the cumulative deviations test, $\sum_x D_x = 115$ and $\sum_x E_x^C m_x^S = 84.220$, and therefore

$$z \;=\; \frac{115 - 84.220}{84.220^{1/2}} \;=\; 3.354$$

The 0.975th quantile of $N(0,1)$ is $z_{0.975} = 1.96$. So H_0 is rejected at 5% level.

In R:

```
> Z <- sum(table$dx - table$ECx*table$mxs)/sqrt(sum(table$ECx*table$mxs))
> Z
```

```
[1] 3.353931
```

```
> qnorm(0.975)
```

```
[1] 1.959964
```

Can also calculate a p-value:

```
> 2*(1 - pnorm(Z))
```

```
[1] 0.0007967234
```

Sign test

For the sign test, $u = (1, 1, 0, 1, 1, 1, 1, 1, 1, 0)$ and therefore $s = 8$ and

$$z \;=\; \frac{7.5 - 5}{2.5^{1/2}} \;=\; 1.581$$

This is not significant at the 5% level of significance since $z_{0.975} = 1.96$.

In R:

```
> v <- 10
> S <-sum( as.numeric(table$dx/table$ECx >= table$mxc))
> Z <- (S + 0.5 - 0.5*v)/sqrt(0.25*v)
> Z
```

```
[1] -2.84605
```

```
> 2*(1 - pnorm(q = Z))
```

```
[1] 1.995573
```

The actual binomial p-value (probability of observing $S = 8$ or more extreme) is 0.109 (very similar).

```
> sum(dbinom(x = c(0,1,2,8,9,10), size = v, prob = 0.5))
```

```
[1] 0.109375
```

This illustrates the lack of power of the sign test.

Chapter 16

Graduation

We expect the true underlying mortality rate in a population to vary smoothly with age.

An estimated mortality rate

$$\hat{m}_x = \frac{D_x}{E_x^C}$$

is an observation of a random variable, with distribution centred on the true underlying mortality rate, but subject to variation quantified by its (asymptotic) variance m_x/E_x^C which may be large if the exposure E_x^C is not large.

Graduation is the term actuaries and demographers use for the statistical *smoothing* of estimated mortality rates, to obtain a more realistic picture of how mortality changes with age, with random variation smoothed out.

16.1 Example

This is an expanded version of the example used in Chapter 15, where now we observe age groups $x = 60, \ldots, 89$.

16.2 Graduation methods

There are three main approaches to graduation, all of which involve fitting a statistical model

- Graduation against a set of standard mortality rates.

  This is used where we have good reason expect our study to share features with the mortality experience of the standard population.

- Graduation using a parametric model.

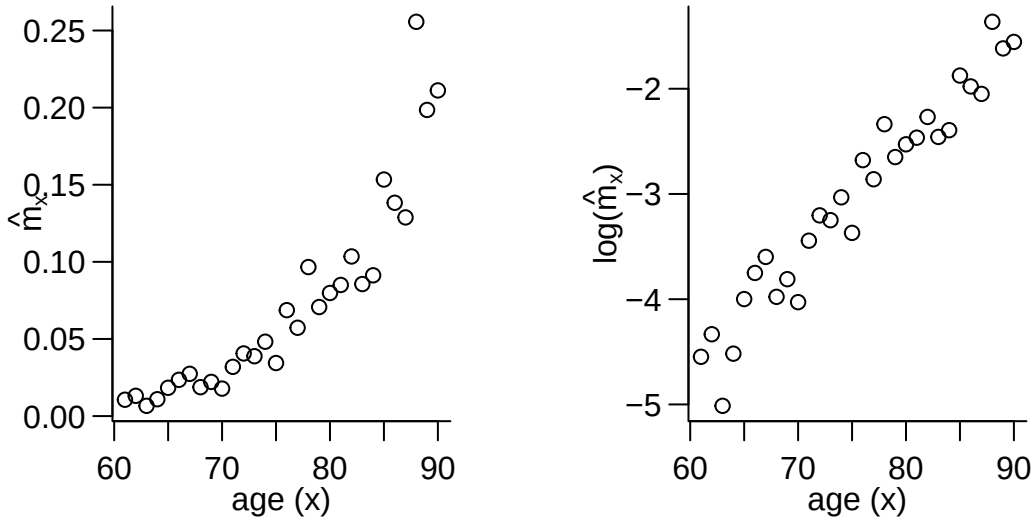


Figure 16.1: Plots of raw estimated mortality rates (original and log scales)

This is usually used where we have large amounts of data.

- Graduation using a semiparametric (smooth) regression model.

  This is usually used where we have small amounts of data and no suitable standard mortality rates.

16.3 Graduation against a set of standard rates

The simplest approach is to fit a regression-type model, such as

$$D_x \sim Poisson(E_x^C m_x) \quad \text{where} \quad \log m_x = \beta_0 + \beta_1 \log m_x^S$$

This is a generalised linear model (g.l.m.), which is easy to fit in software such as R. The graduated mortality rates are the maximum likelihood estimates, given by

$$\log \tilde{m}_x = \hat{\beta}_0 + \hat{\beta}_1 \log m_x^S$$

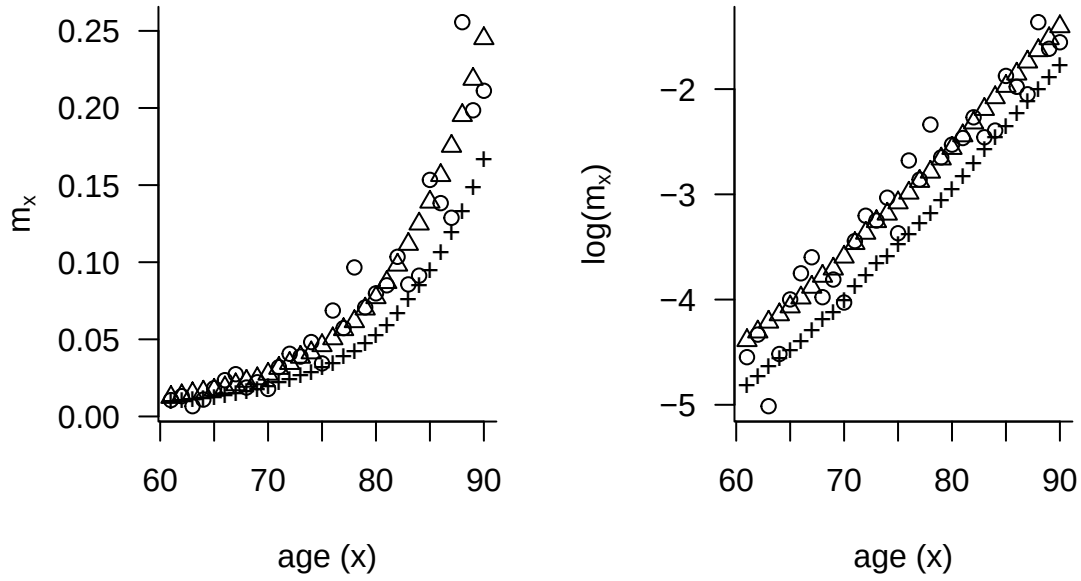
If the standard rates are smooth (as they typically will have been previously graduated) then the estimated rates will be smooth too.

Students taking MATH3012 will learn more about generalised linear models.

16.3.1 Return to example

Here we illustrate graduation against standard rates.

This shows the raw rates (circles) standard rates (crosses) and graduated rates (triangles).



16.4 Graduation using a parametric model

This approach assumes that there is a parametric formula describing the relationship between central mortality rate (force of mortality) and age.

The simplest model is the (log-linear) Gompertz model

$$D_x \sim Poisson(E_x^C m_x) \quad \text{where} \quad \log m_x = \beta_0 + \beta_1 x$$

This is a g.l.m., which is easy to fit in software such as R. The graduated mortality rates are the maximum likelihood estimates, given by

$$\log \tilde{m}_x = \hat{\beta}_0 + \hat{\beta}_1 x$$

The estimated rates are guaranteed to be smooth provided the model is not too complicated.

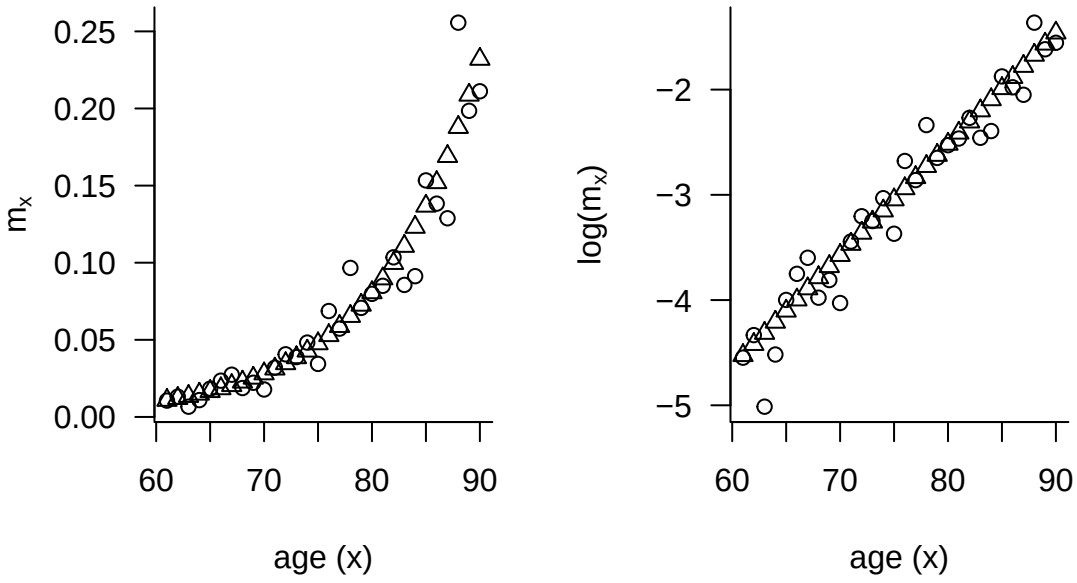
More complex Gompertz models replace the linear function $\beta_0 + \beta_1 x$ with a polynomial $\beta_0 + \beta_1 x + \beta_2 x^2 +$

$\cdots + \beta_p x^p$ and are also g.l.m.s

16.4.1 Return to example

Here we illustrate graduation using the log-linear Gompertz model.

This shows the raw rates (circles) and graduated rates (triangles).



16.5 Models for human mortality

A general Gompertz model

$$\log m_x = \beta_0 + \sum_{j=1}^{p} \beta_j x^j$$

is often found to fit mortality at higher ages well, for quite small values of p, such as $p = 1$ (linear model with two parameters) or $p = 2$ (quadratic model with three parameters).

If necessary, the Gompertz model can be extended to the Gompertz-Makeham family, which has

$$m_x = \alpha_0 + \sum_{j=1}^{q} \alpha_j x^j + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_p x^p\right)$$

This is no longer so easy to fit, as it is not a g.l.m.

16.5.1 Graduation using a semiparametric model

This approach estimates the underlying mortality rates by balancing the competing demands of fit to the data and smoothness of the underlying function. The aim is to find the closest fit without compromising smoothness.

The resulting model is sometimes called a *generalised additive model* (g.a.m.) and can be written as

$$D_x \sim Poisson(E_x^C m_x) \quad \text{where} \quad \log m_x = s(x)$$

where $s(x)$ denotes an arbitrary smooth function of x.

A g.a.m. is fitted by minimising an objective function which balances fit to the data (negative log-likelihood or similar) and smoothness (integrated squared second derivative or similar).

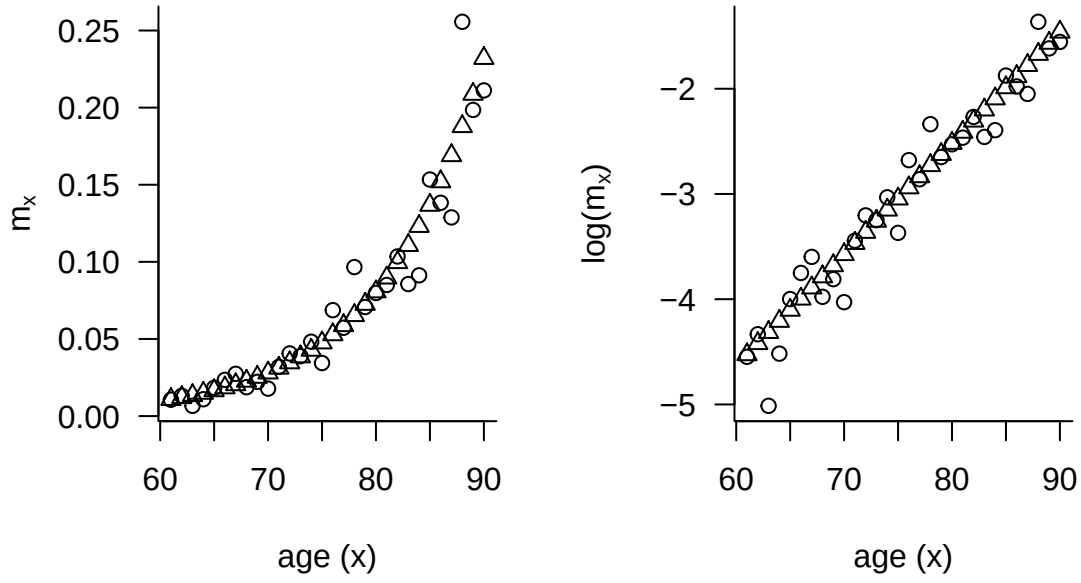
16.5.2 Return to example

Here we illustrate graduation using the semiparametric (smooth) model.

This shows the raw rates (circles) and graduated rates (triangles).

```
Error in library(gam): there is no package called 'gam'
```

```
Error in gam(d ~ s(x) + offset(log(e)), data = elt, family = poisson): could not find function
```



16.6 Testing the fit of a graduation

A set of graduated mortality rates should be compared with the original rates to check the fit of the graduation.

The chi-squared test, described in Chapter 15 can be used for this, with test statistic

$$X^2 = \sum_x \frac{(D_x - E_x^C \tilde{m}_x)^2}{E_x^C \tilde{m}_x} \tag{16.1}$$

where $\tilde{m}_x$ are the graduated rates.

A minor modification is required, because the components of the sum in (1) are no longer independent because the D_x have been used to obtain the graduated rates $\tilde{m}_x$.

This requires us to deduct, from v, one degree of freedom for each estimated parameter in the graduation process. So if we have estimated p parameters, the reference distribution for the test is χ^2_{v-p}. [Note there is no easy adjustment for the other tests in Chapter 15 which should be avoided for testing the fit of a graduation].

16.6.1 Return to example

For the graduations illustrated in Sections 16.3.1, 16.4.1 and 16.5.2 respectively, we have

- $X^2 = 31.74$ (graduation against standard rates)
- $X^2 = 28.20$ (graduation using Gompertz log-linear model)
- $X^2 = 25.84$ (graduation using semiparametric (smooth) model)

The number of age groups is $v = 30$ and the number of parameters is $p = 2$ for the first two models and $p = 3.0$ (an estimated function of the overall smoothness) for the semiparametric model.

So we compare our X^2 values with the 95% point of χ^2_{28} (41.34) and the 95% point of χ^2_{27} (40.11).

We see that there is no evidence to lead us to reject the fit of the graduated rates to the observed rates, so all three of the graduations seem satisfactory.