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

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.

3.2 The distribution function

3.3 The survival function

3.3.1 The survival function: properties

3.3.2 The residual survival function

3.4 The hazard function

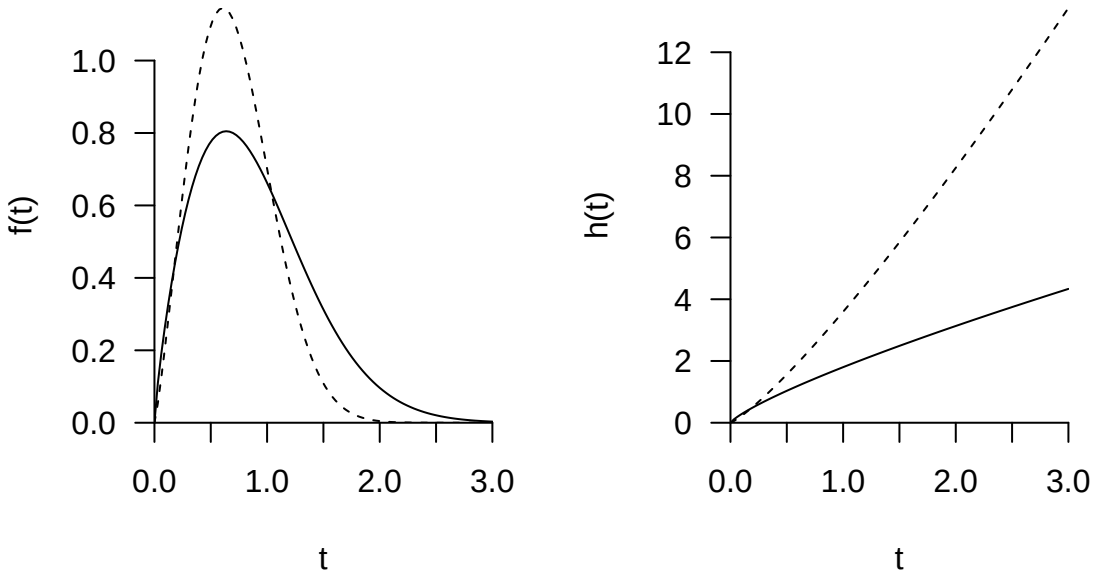
3.4.1 The hazard function: properties

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

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

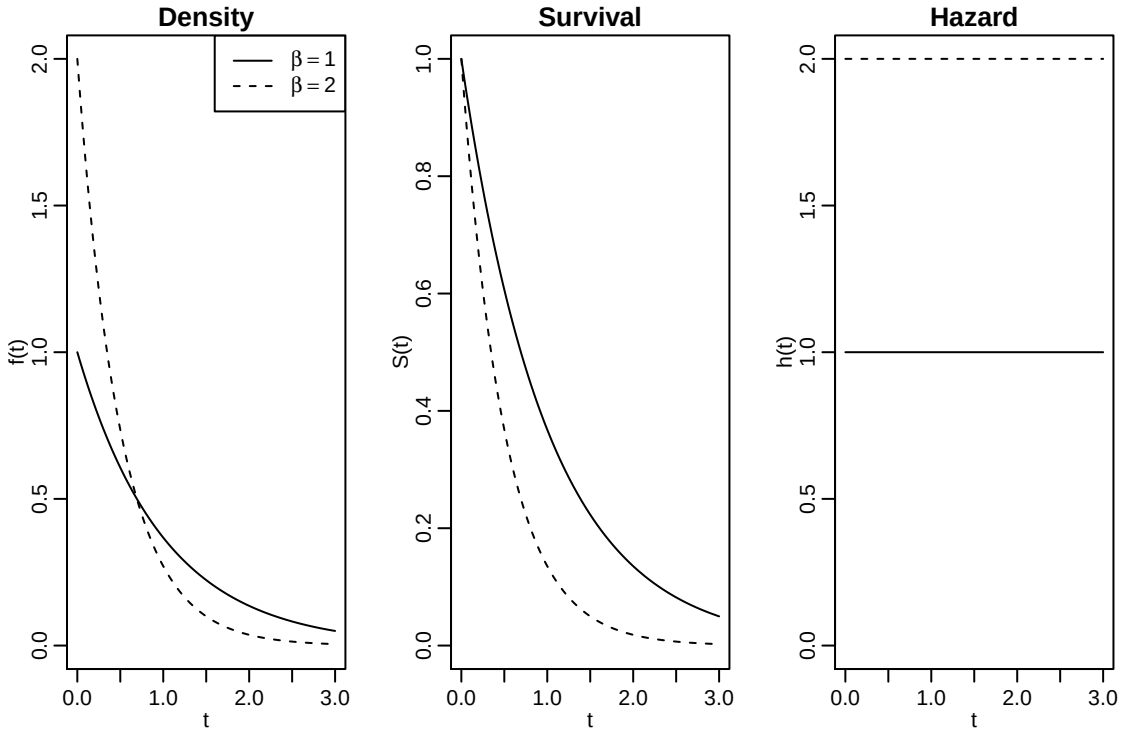
$$h_T(t) = \beta.$$

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 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

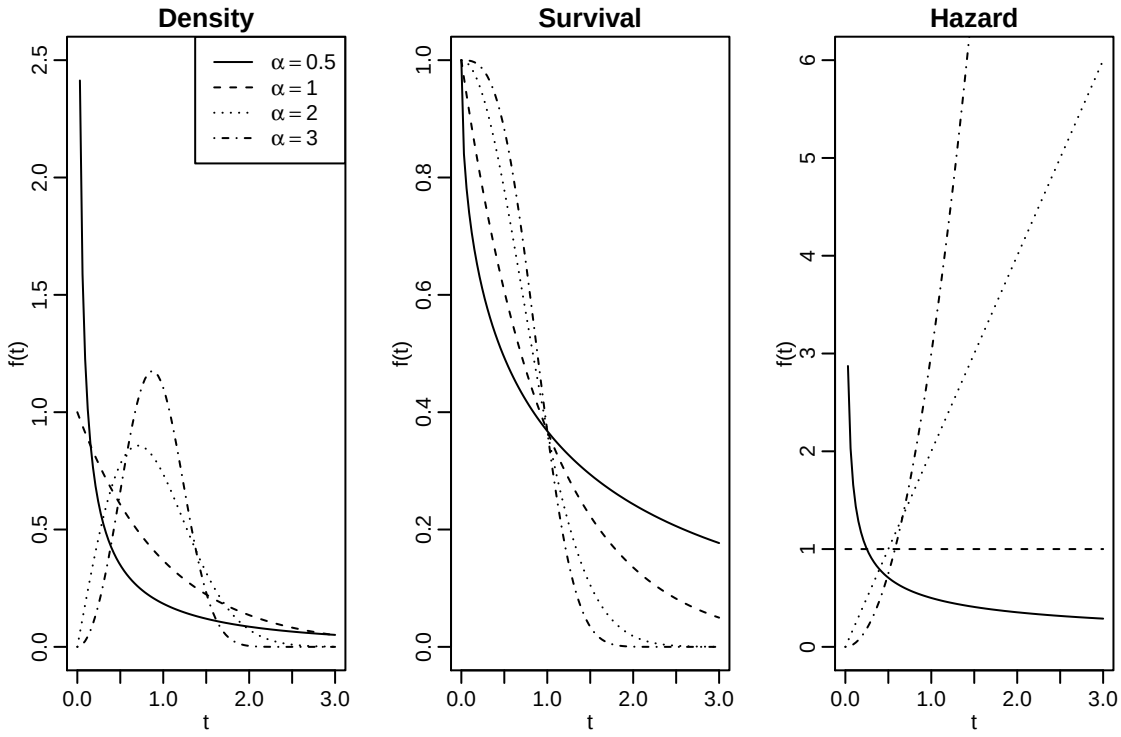
$$h_T(t) = \alpha\beta\,(\beta t)^{\alpha-1}.$$

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 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

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

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

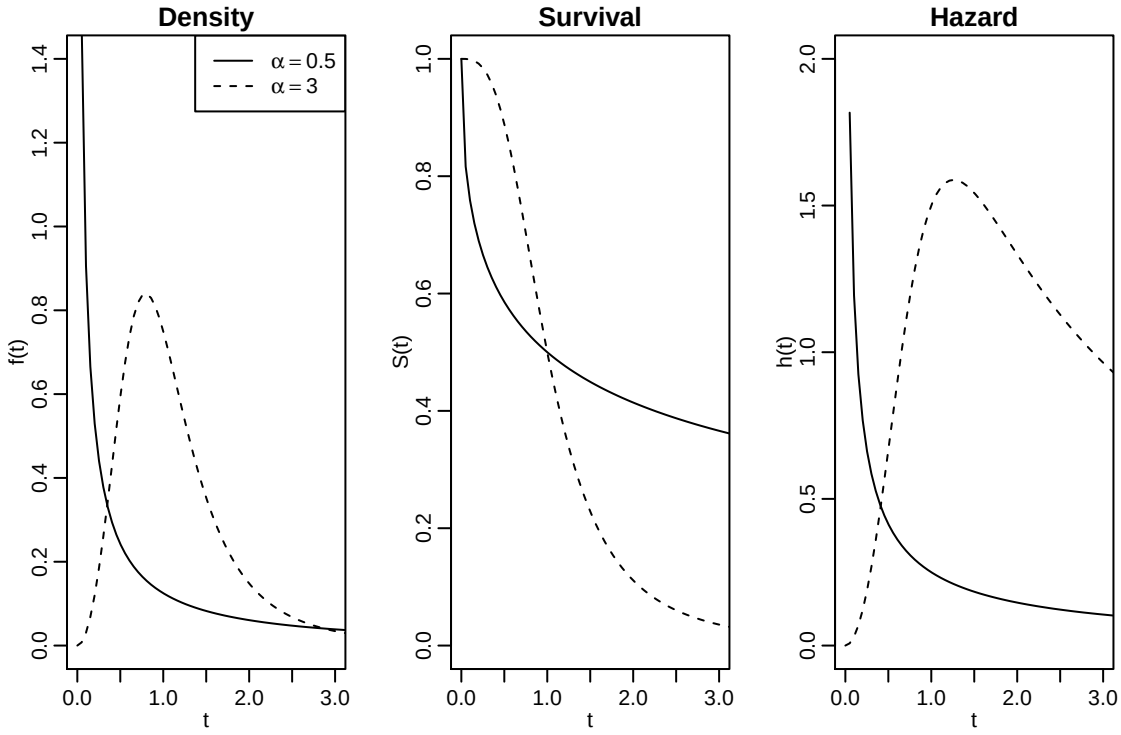
and hazard function

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

If $T \sim$ loglogistic(α, β) then

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

The 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

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

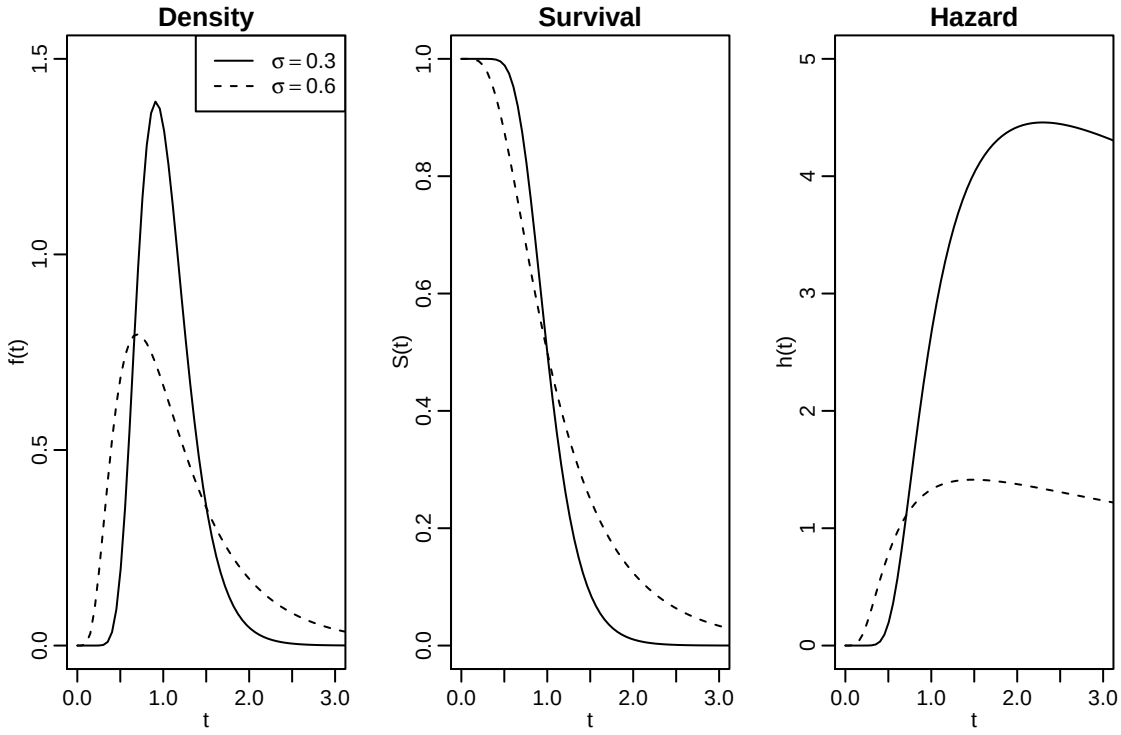
where ϕ and Φ are the standard normal density and distribution functions. The lognormal hazard function is $h_T(t) = f_T(t)/S_T(t)$.

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

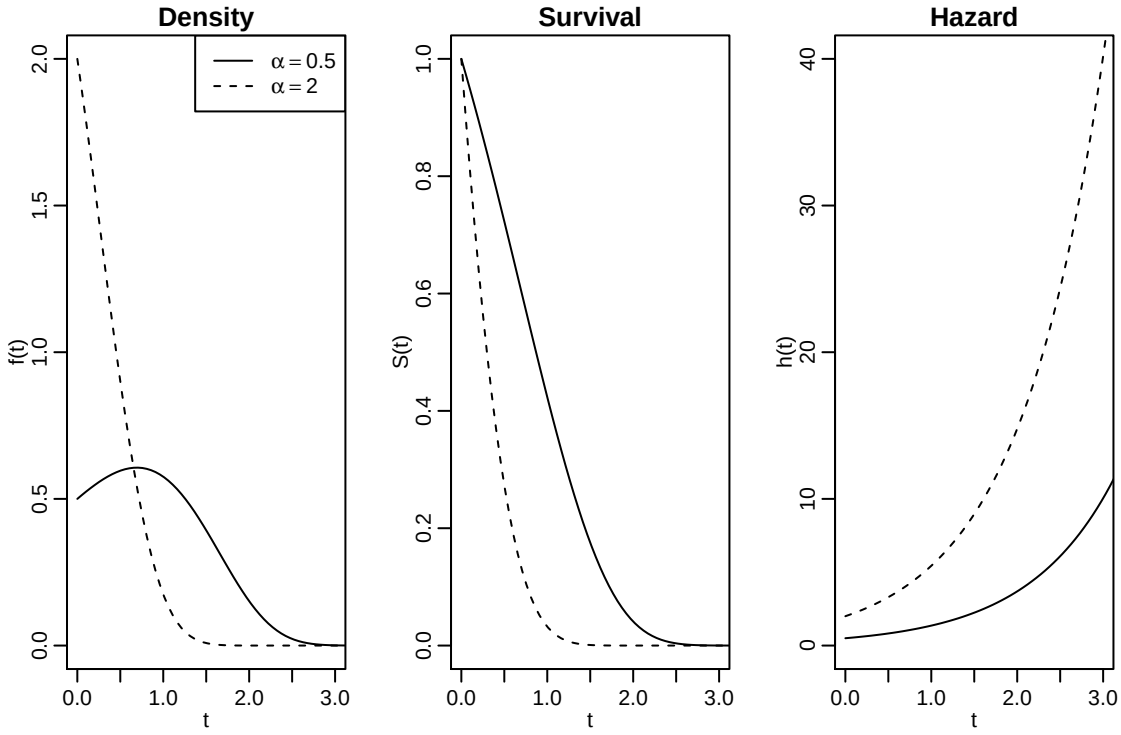
$$S_T(t) = \exp\left(-\frac{\alpha}{\beta}(e^{\beta t} - 1)\right),$$

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

5.4 Maximum likelihood properties

```
> 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

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.

5.6 Example: Weibull model likelihood

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

6.2 The Kaplan-Meier estimator

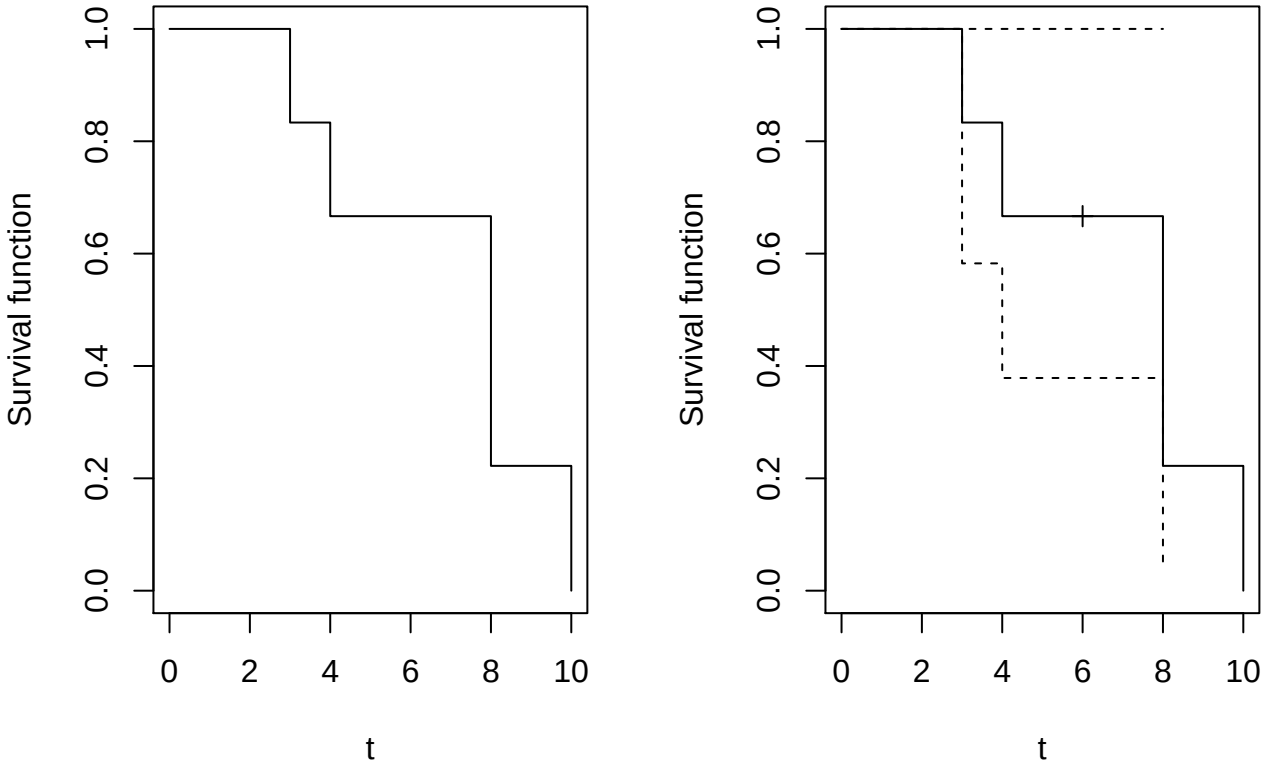
6.2.1 Notes on the Kaplan-Meier estimator

6.2.2 Example of calculating the Kaplan-Meier estimate by hand

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.

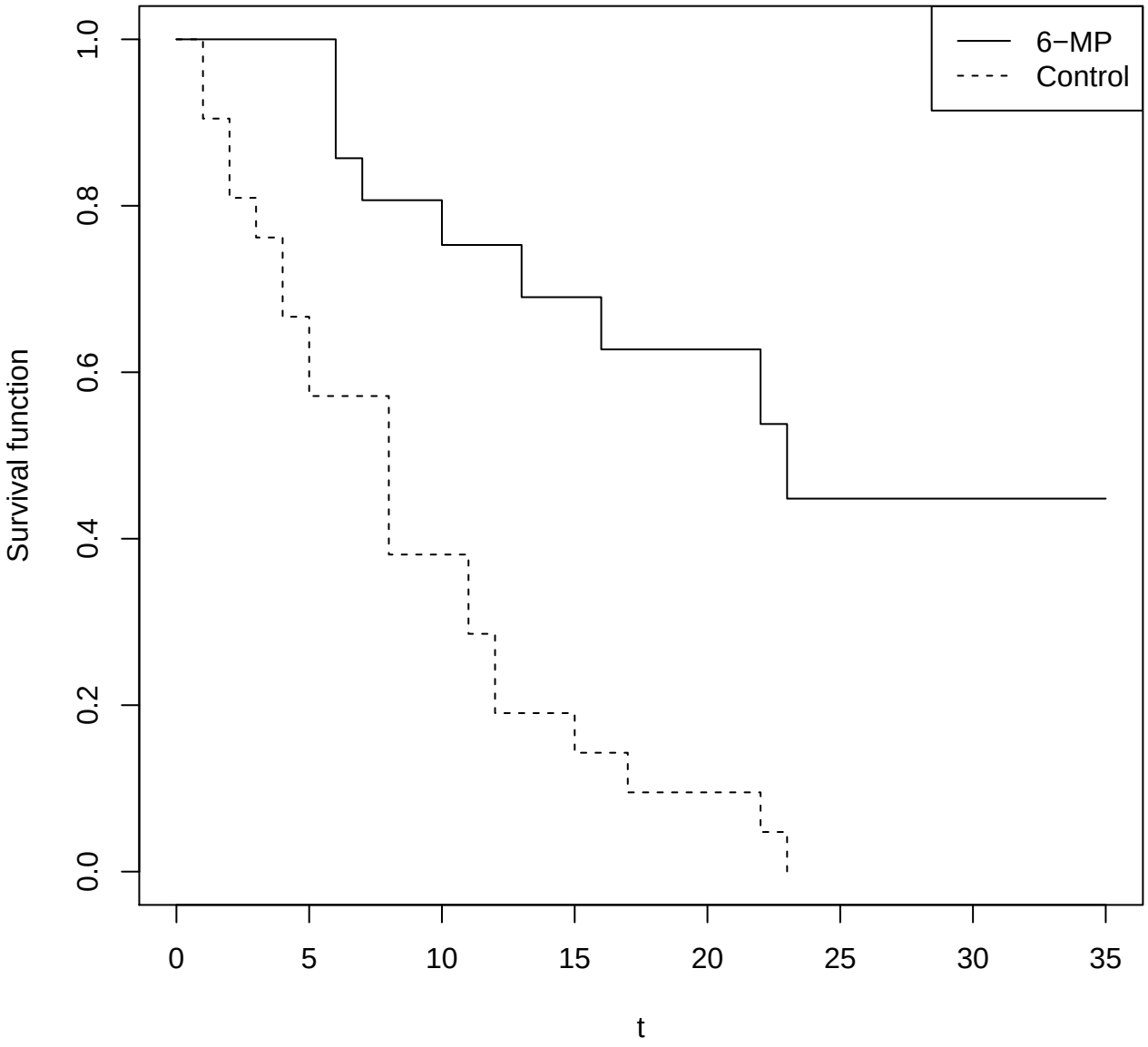
The below code plots survival curves for the two groups (control and 6-MP) in the `gehan` data.

```
> 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

6.3 The Nelson-Aalen estimator

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}.

7.3 Partial likelihood

7.4 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

7.7 Hypothesis testing

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

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

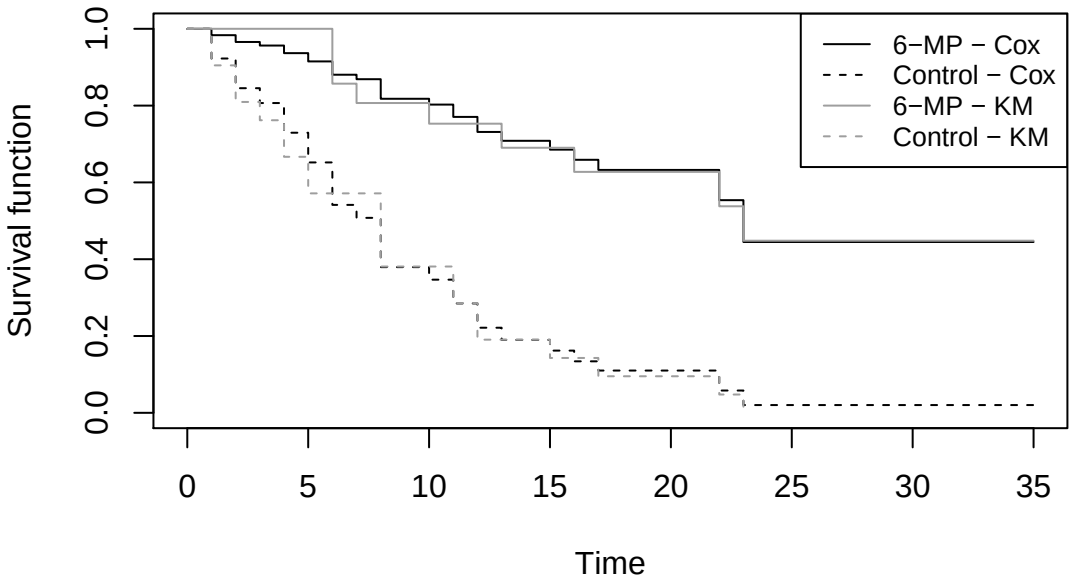
7.9 Fitting Cox models using R

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

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

```
> 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

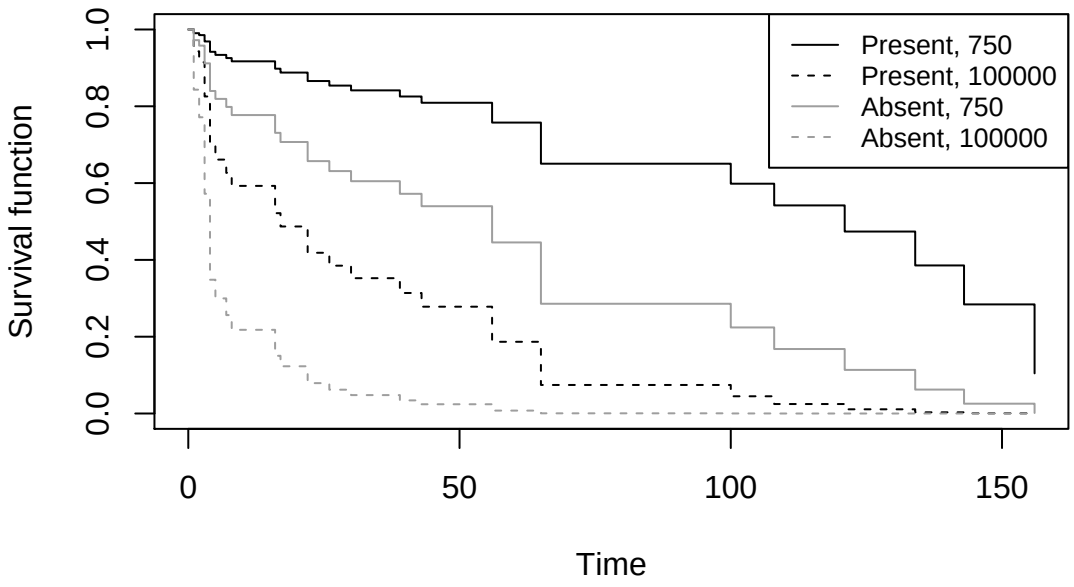
The `R` code below fits a Cox model with two explanatory variables: a dummy variable for *ag* and a variable given by the natural logarithm of *wbc*.

```
> 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) se(coef)      z Pr(>|z|)
agpresent -1.0691    0.3433   0.4293 -2.490  0.01276 *
log(wbc)   0.3677    1.4444   0.1360  2.703  0.00687 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef) exp(-coef) lower .95 upper .95
agpresent    0.3433     2.9126     0.148    0.7964
```

```
log(wbc)     1.4444     0.6923     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
```

```
> 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)
```



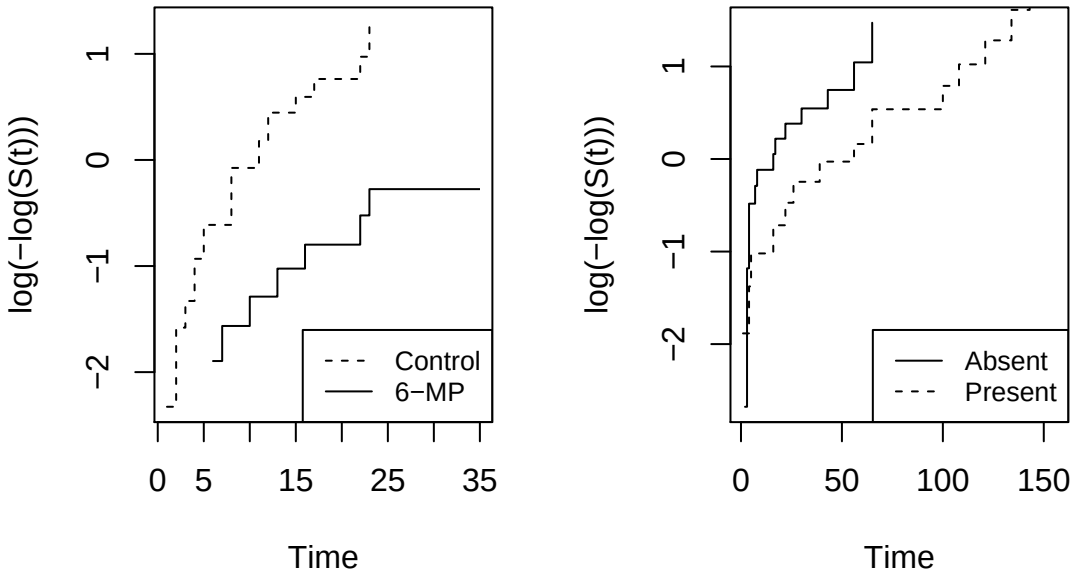
7.10 Checking proportional hazards

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

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.

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)

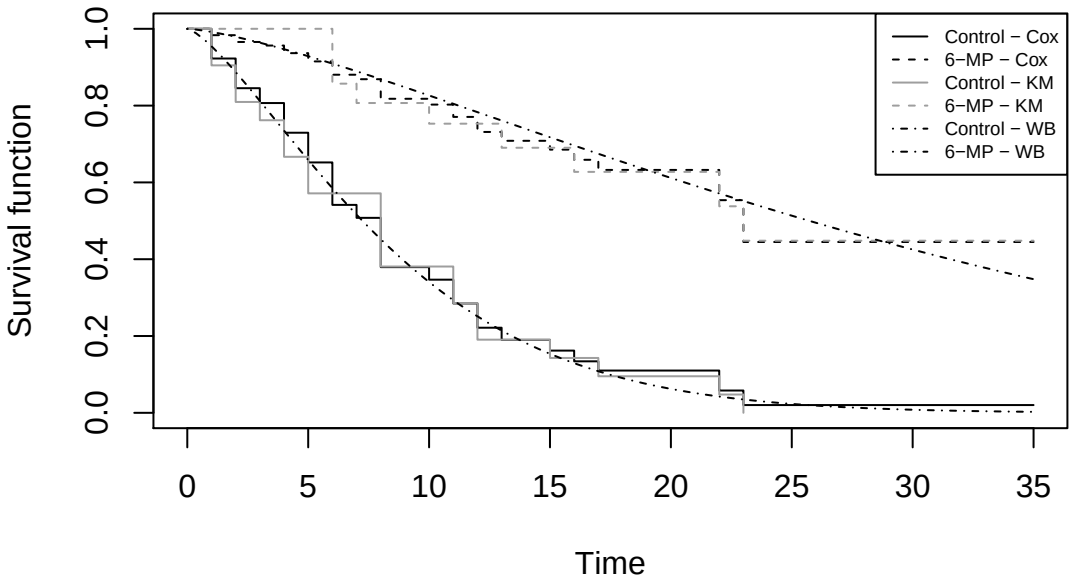
7.12.1 Example: Weibull ART model for the gehan data

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     z       p
(Intercept)   3.516      0.252 13.96 < 2e-16
treatcontrol -1.267      0.311 -4.08 4.5e-05
Log(scale)   -0.312      0.147 -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
```

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

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:

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

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 $p_{ij}(x,t)$.

8.6 The holding time distribution

8.7 Conditional transition probability

8.8 Examples

8.8.1 Two-state model (absorbing state)

8.8.2 Two-state model (no absorbing state)

8.8.3 Four-state process

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

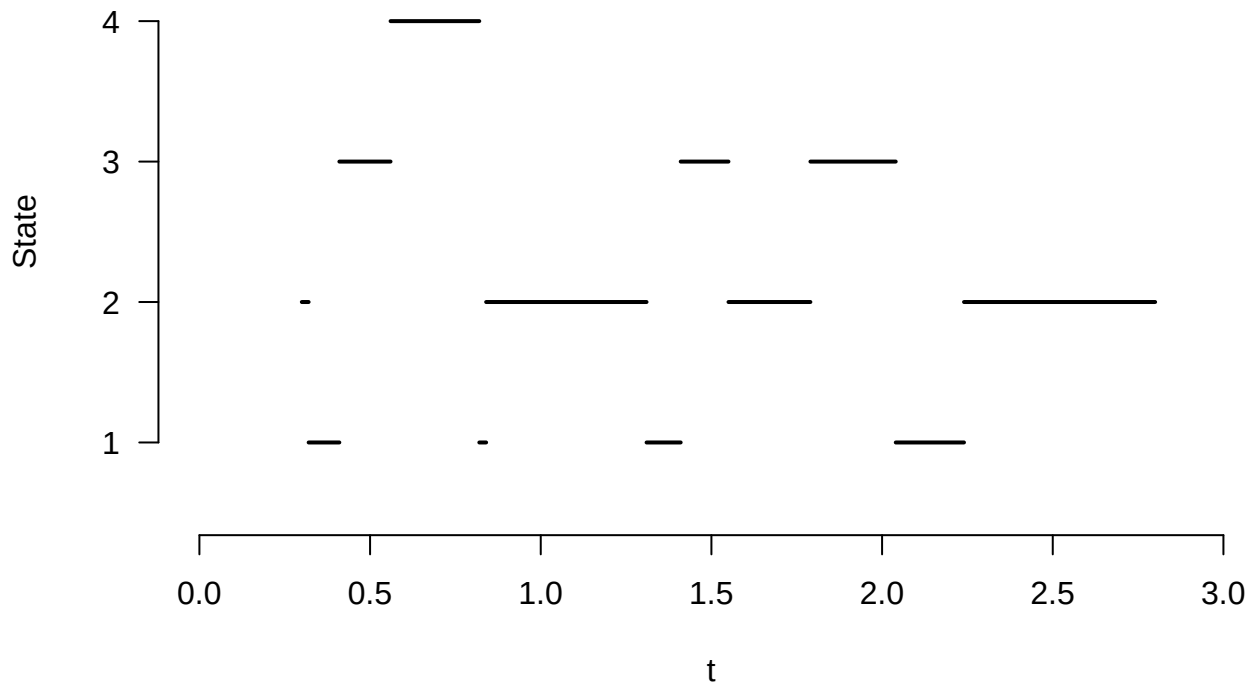
$$\mu_{k\ell}, \quad i, j = 1, \ldots, m, \quad k \neq \ell$$

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

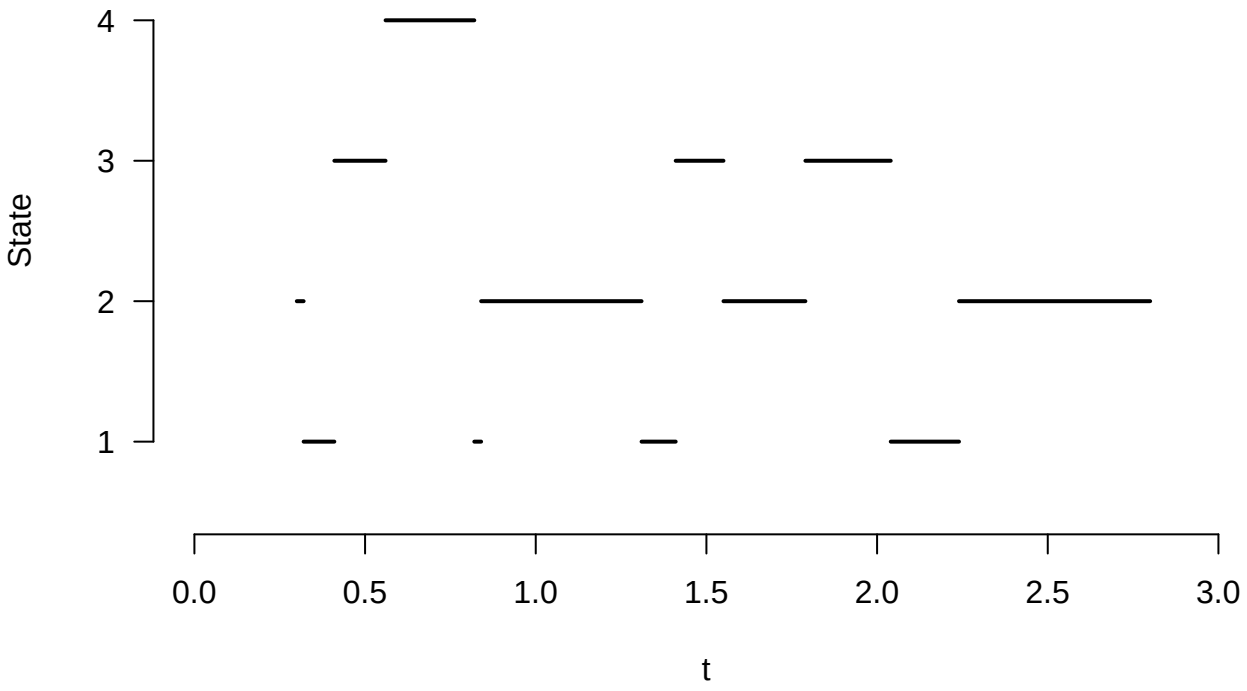
9.3 Likelihood

9.4 Maximum likelihood estimation for μ

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

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



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 $\{t_{ik}^{+}\}$.

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

10.4 Alternative notation

10.5 p_x and q_x notation

10.6 Force of mortality

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.

11.3 Life expectancy

11.3.1 Obtaining e_x from the life table

11.3.2 Obtaining $\overset{\circ}{e}_x$ from e_x

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

11.3.4 Expectation of life in an interval

11.3.5 Life expectancy using L_x and/or T_x

11.3.6 The trapezium rule approximation

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

12.2 Constant force of mortality

12.3 Balducci assumption

12.4 Shape of the survival function

12.5 Implied force of mortality

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

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.

13.1.1 The binomial likelihood

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 ${}_{b-a}q_{x+a}$

13.1.4 Approximate likelihood for q_x

13.1.5 Likelihood under UDD

13.1.6 Log-likelihood example

13.1.7 A simple estimator for q_x

13.1.8 Central exposed to risk

13.1.9 The actuarial estimator

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).

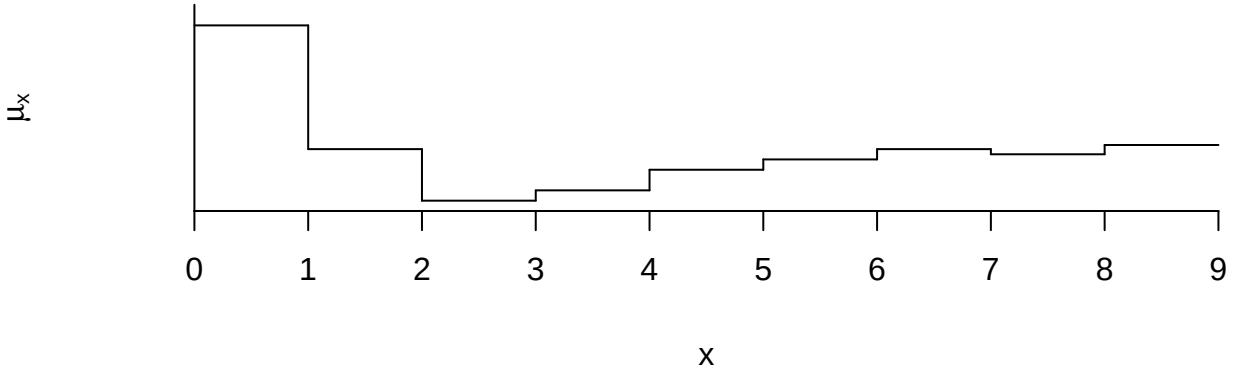
13.2.2 Estimating μ_x

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

$$\mu_{x+t} = \mu_{x+\frac{1}{2}}, \quad t \in [0, 1)$$

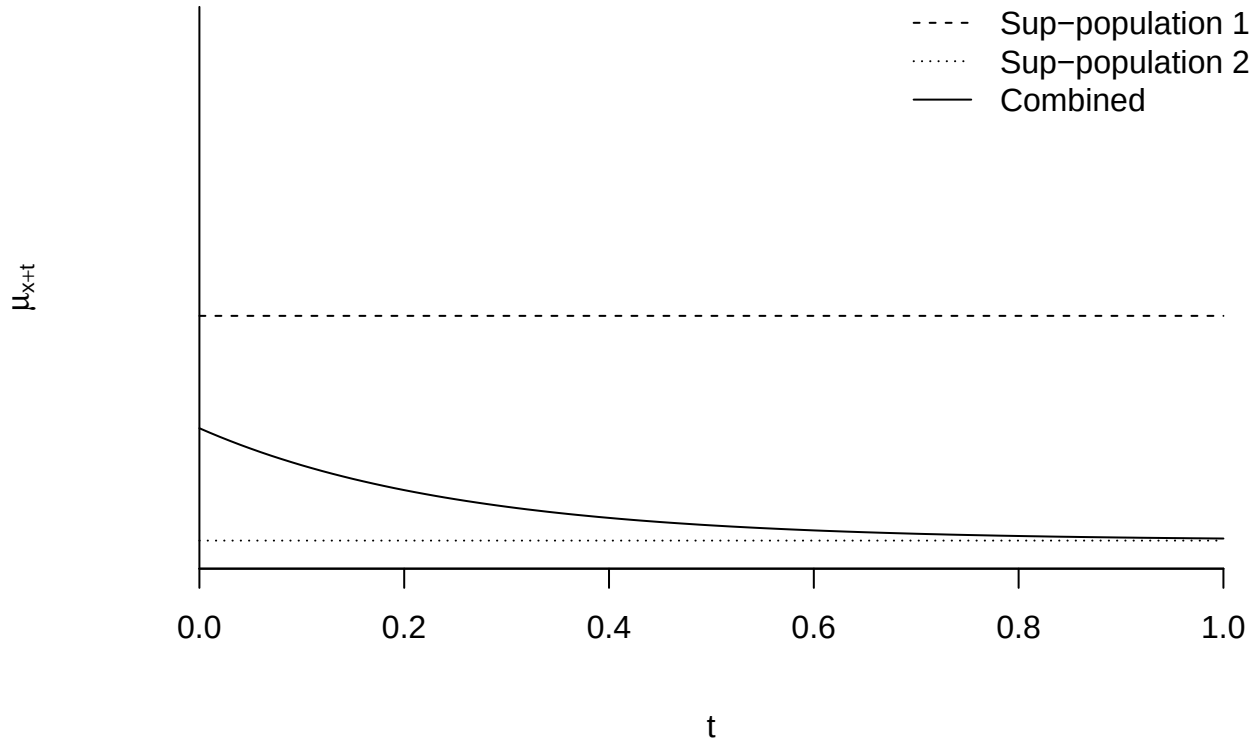
so the force of mortality is modelled as a step function:



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

13.2.5 Mortality rates

13.2.6 Central mortality rate

13.2.7 Estimating q_x from $\hat{m}_x$

13.2.8 Similarity of $\hat{q}_x$ and $\tilde{q}_x$

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.

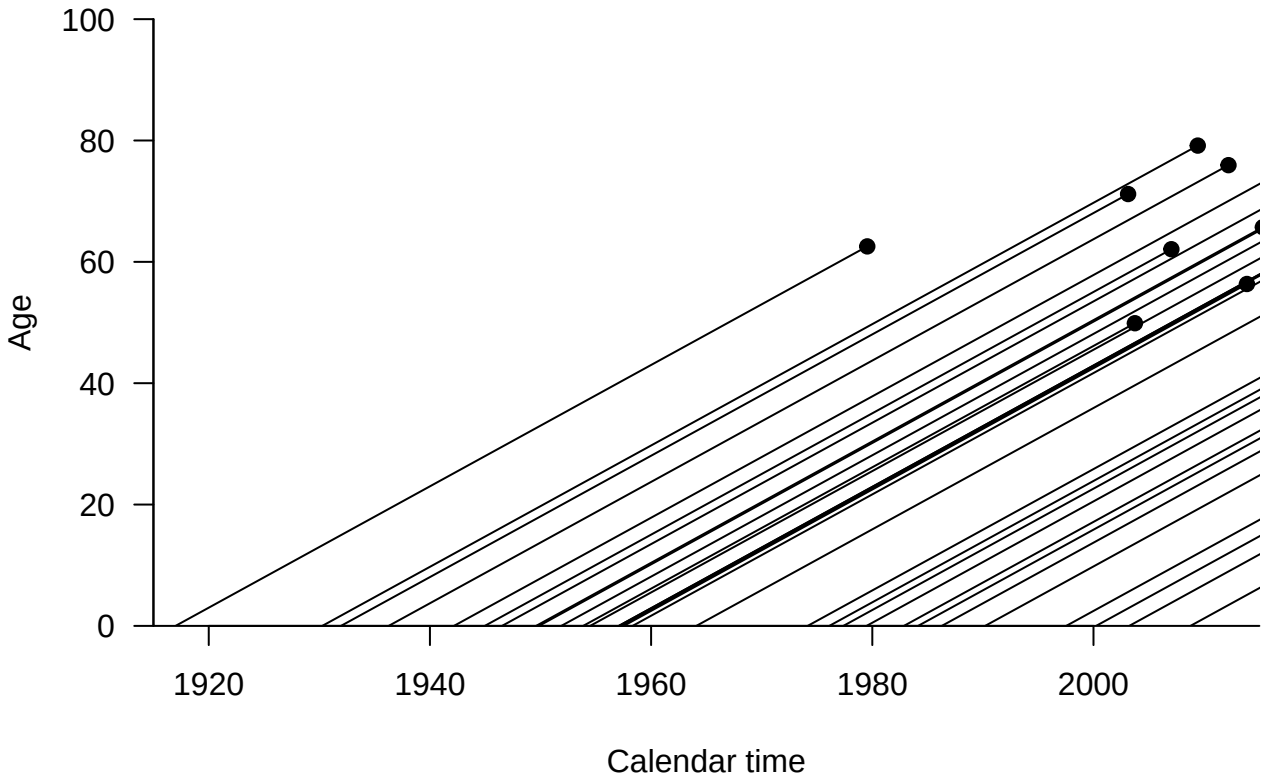
How can we calculate E_x^C using such information?

14.2 Lexis diagrams

The Lexis diagram is a useful way of visualising lifetime data.

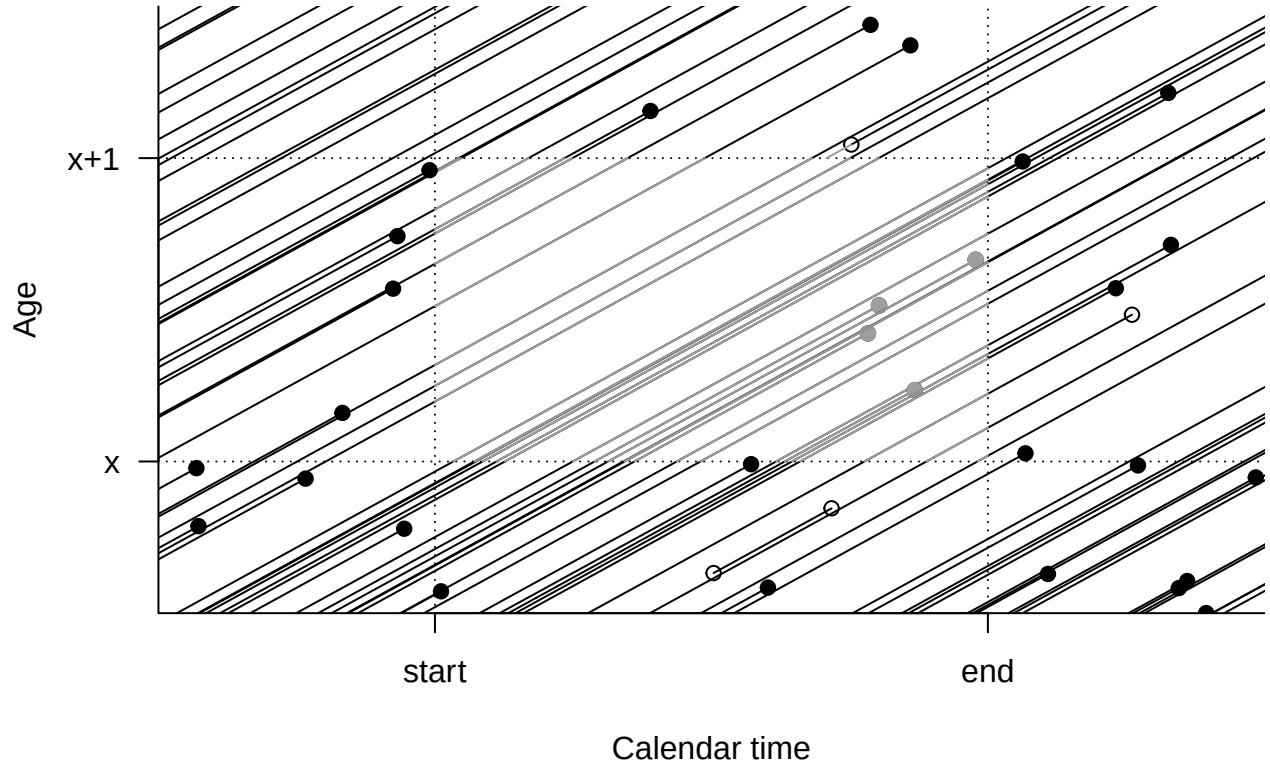
It plots an individual life trajectory as a line on a graph where the y-axis is age and the x-axis is calendar year.

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.

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

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)
```

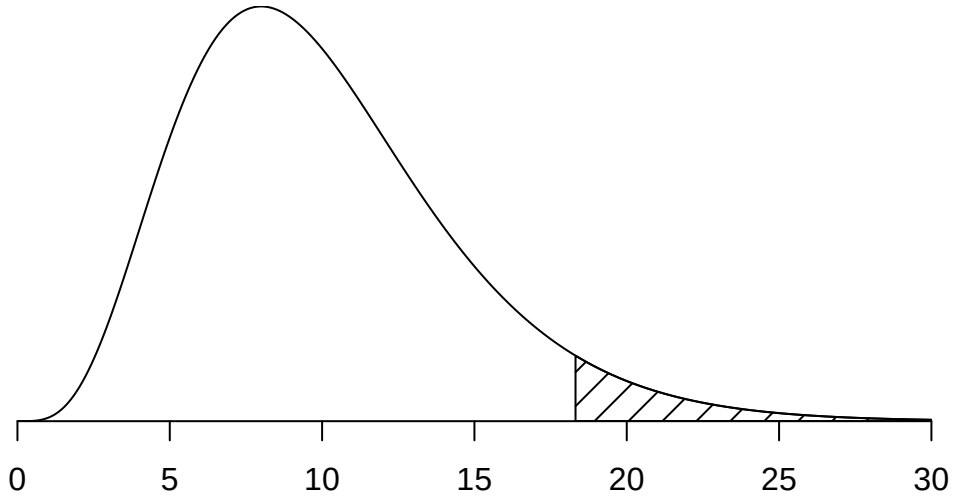


Figure 15.1: χ^2_{10} distribution with upper 5% tail indicated

```
[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.

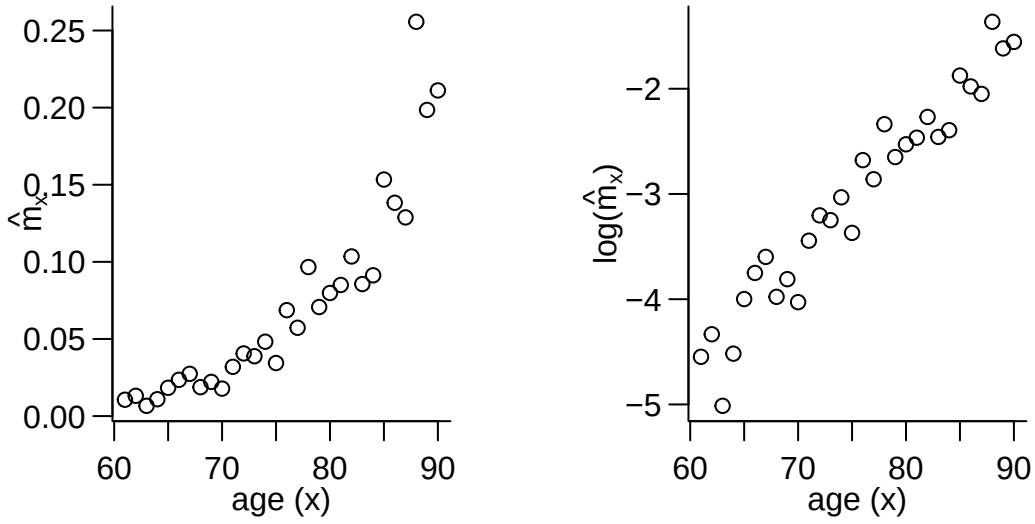


Figure 16.1: Plots of raw estimated mortality rates (original and log scales)

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

$$\log \tilde{m}_x = \hat{\beta}_0 + \hat{\beta}_1 \log m_x^S$$

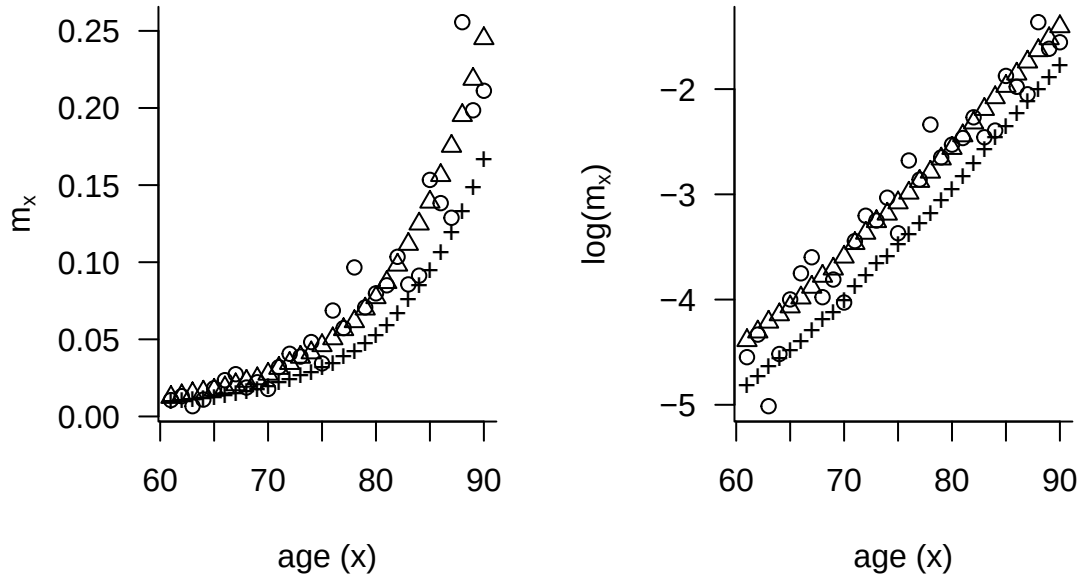
If the standard rates are smooth (as they typically will have been previously graduated) then the estimated rates will be smooth too.

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.

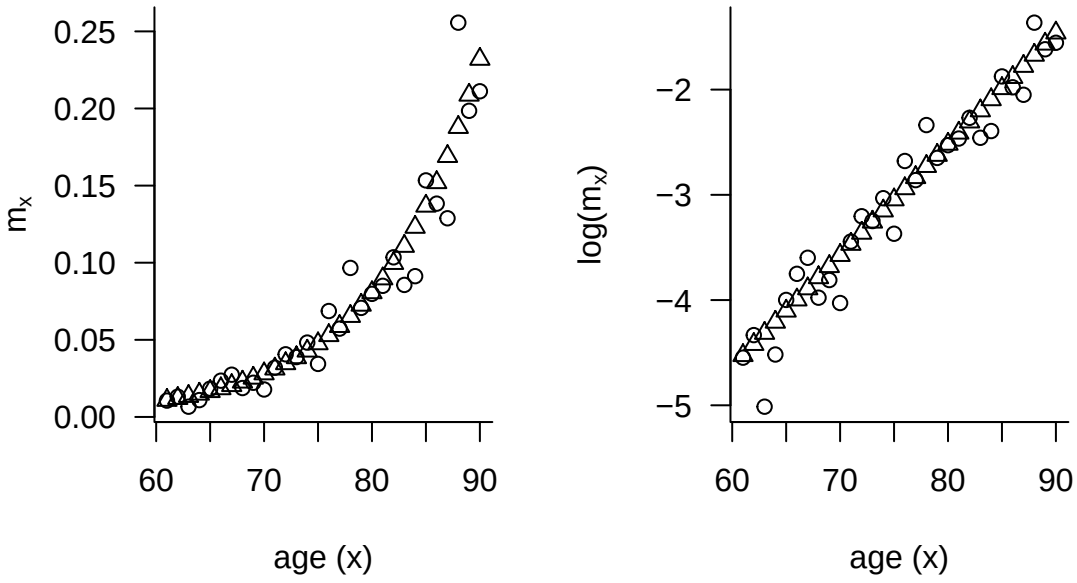
More complex Gompertz models replace the linear function $\beta_0 + \beta_1 x$ with a polynomial $\beta_0 + \beta_1 x + \beta_2 x^2 +$

$\cdots + \beta_p x^p$ 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).

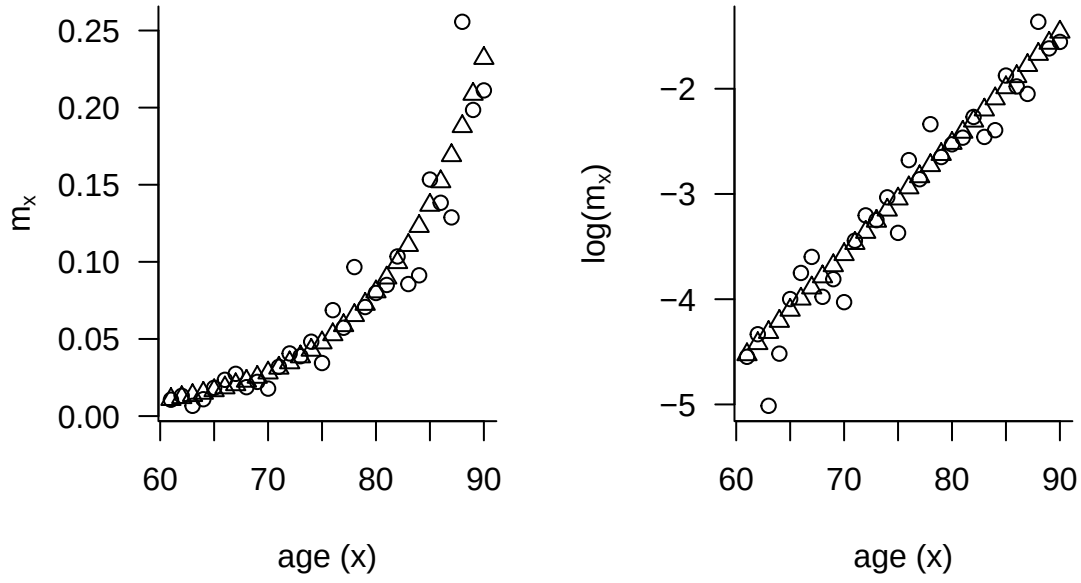
16.5.2 Return to example

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.