Survival Analysis

Course Manual 2023

Module organisers: Ruth Keogh, Aurélien Belot

Lecturers: Ruth Keogh, Manuela Quaresma, Aurélien Belot

Practical tutors: Aurélien Belot, Ruth Keogh, Leah Pirondini, Manuela Quaresma, Tim Russell, Wende Safari

Survival Analysis 2023: Module Overview

Dates and times

- All lectures will be pre-recorded.
- Computer practicals will take place on Fridays in weeks 1-4 (11.15-12.45 and 14.00-15.30) and Thursday in week 5 (11.15-12.45).
- Question and Answer sessions will take place on Fridays in weeks 2, 3, and 4 at 15.45-17.15.

Lecture Topics

1. Introduction to Survival Analysis (RK)
2. Non-parametric analysis of survival data (RK)
3. Parametric regression modelling (MQ)
4. The Cox proportional hazards model (RK)
5. More on the proportional hazards model and model checking (MQ))
6. Competing risks and multi-state models (AB)
7. Time dependent variables and frailty models (AB)
8. Alternative models for survival data (RK)

Software

Computer practicals are available in both Stata and R.

We recommend that you create a folder for this module and use this to store your data and your Stata do files and/or R script files.

Assignment

For MSc Medical Statistics students [Taking the module: Survival Analysis and Bayesian Statistics]: There will be a joint assignment between Survival Analysis and Bayesian Statistics. The assignment will be made available in week 3 and the date for submission will be Friday 24th March.

For MSc Health Data Science and MSc Epidemiology students [Taking the module: Analysis of Electronic Health Records Data]: There will be a single assignment for the Analysis of Electronic Health Records module. The date for submission will be Friday 24th March.

In week 3 one of the practical sessions is dedicated to the assignment. You will be able to use this time to work on your assignment and tutors will be available.

Some key references

Collett, D. 2003. Modelling Survival Data in Medical Research. Third Edition. Chapman and Hall. [Electronic access available to students of LSHTM from within the School network or via Remote Access.]
Covers most of the topics covered in this module, though some of the later topics that we consider are not discussed in depth.

Machin, D, Cheung, YB and Parmar, MKB. 2006. Survival Analysis. A Practical Approach. Second Edition. [Electronic access available to students of LSHTM from within the School network or via Remote Access.]
Covers the early parts of the module.

Cox, DR and Oakes, D. 1984. Analysis of survival data. Chapman and Hall.
A classic text on survival analysis.

Kalbfleisch, JD and Prentice, RL. 2002. The Statistical Analysis of Failure Time Data. Second edition. Wiley.
Mathematically technical, but comprehensive.

Texts covering more advanced topics:

Duchateau, L and Janssen, P. 2010. The Frailty Model. Springer.
For researchers interested in random-effect models (frailty models) for time-to-event outcome.

Geskus, R.B. 2015. Data Analysis with Competing Risks and Intermediate States. Chapman and Hall/CRC Biostatistics Series.
Covers in detail the situation when more than one event is of interest. R-code is given throughout the book.

Rizopoulos, D. 2012. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Chapman & Hall/CRC Biostatistics Series.
Comprehensive overview of methods to model jointly a continuous longitudinal process and a time-to-event outcome. Lots of R code (JM and JMBayes packages).

Van Houwelingen, H.C. and Putter, H. 2012. Dynamic Prediction in Clinical Survival Analysis. Chapman and Hall/CRC Press.
Covers matters relating to prognostic modelling in survival analysis, including extensions to dynamic prediction. The final part of the book focuses on genomic data. R code and example data used in the book are available online.

Table of Contents

Session 1

Introduction to Survival Analysis

1.1 Aims of this lecture and practical

At the end of this lecture and practical you will be able to:

- Describe the features of survival data, including censoring.
- Define the hazard function, survivor function, and distribution function when describing survival time distributions.
- Define the relationships between these functions.
- Explain the properties of the exponential distribution, the Weibull distribution and the log-logistic distribution for survival times.
- Formulate the likelihood for survival data under the three models listed above.
- Prepare data for survival analysis in Stata and R.
- Estimate the parameters of survival models in Stata and R and interpret the output.

1.2 What is survival analysis?

Survival analysis is the study of observations which are times at which some outcome or event of interest occurs. In this module we will use the term outcome. Examples of outcomes of interest in survival analysis are:

- Death (all causes)
- Death following a disease diagnosis or following a procedure
- Diagnosis with a disease
- A woman conceiving
- Return to work following sickness

The time at which the event occurs is referred to as a survival time. The terms 'failure time' and 'event time' are also used. Survival analysis is also referred to as time-to-event analysis.

Survival analysis may be used to answer different questions. Some investigations we may wish to make are:

- Studying the patterns of survival in a given population over a particular time scale. *Example:* For a person born in the UK in a particular year, what is the probability that the person lives to age 5, 40, 100?
- Comparing survival times for individuals in two groups, or more generally in several groups. *Example:* Following a disease diagnosis, do individuals given a new treatment have better survival prospects than individuals given a standard treatment?
- Studying the effects of several continuous and categorical variables on survival time, taking into account possible confounding. *Example:* How is adult body mass index associated with survival time after controlling for potential confounders, where the event of interest is a particular disease diagnosis?
- Predicting future survival based on features of an individual. *Example:* What is the probability that an individual with features x,y,z will survive 5 years following a particular disease diagnosis?

1.3 Sources of survival data

Survival data arise from a number of different sources and in a number of different types of study:

- In national registers of births and deaths.
- In randomized controlled trials, e.g. to study whether individuals with a particular medical condition who were randomized to a new treatment tended to live longer than individuals randomized to the standard treatment.
- In prospective observational studies (cohort studies), in which individuals are recruited to a cohort and followed-up for a range of different outcomes.
- Survival data also arise in non-medical settings, including in industrial studies, in which the lifetime of products is of interest, and in studies of occurrence of geological or weather-related events.

1.4 Features of survival data

Analysis of survival data requires careful definition of the following:

- The outcome of interest
- A time origin
- A time scale, measured in appropriate units
- The time at which the outcome occurs

The time origin and the time scale are closely related and it is important that they are carefully chosen and defined. The time origin is the time relative to which the event time is measured. The time scale may be age, or time since some particular originating

event (e.g. disease diagnosis, beginning of a treatment, beginning of the study) and we may measure it in hours, days, weeks, months, years depending on the situation and the data available to us.

We return to the four examples at the start of Section 1.2:

- Considering all causes of death in a general population, such as the UK population, the time origin is date of birth and the survival time may be measured by age in years.
- Death (or other adverse event) following a disease diagnosis or procedure (e.g. receiving an organ transplant): Here the appropriate time origin is likely to be the date of the diagnosis or procedure and the survival time may be measured in months to death from this point. Example: Time to Acute Graft versus Host Disease among bone marrow transplant recipients.
- Diagnosis with a disease (e.g. a particular cancer): Here the appropriate time origin may again be date of birth and the survival time measured in age in years. However, suppose the individuals we are studying are part of an occupational cohort of individuals being followed for a particular event, then the chosen time origin may be the time of recruitment to the cohort and the time scale may be time spent in the study. It may be necessary to account for age in some way as well in this setting. This is an example of when there may be more than one possible time scale to consider. We return to this later. Example: Study of risk factors for breast cancer in a large cohort.
- A woman conceiving: Here the appropriate time origin may be the date at which a woman starts trying to conceive, and the time scale may be the time in weeks up to conception. Example: A study of time taken to conceive in women of different ages.
- Return to work following sickness: The appropriate time origin may be the time of being signed off work by a doctor, with time to return measured in days or weeks. Example: a study of government interventions to reduce sickness time.

1.5 Censoring of survival times and left-truncation

A particular feature that nearly always arises in survival data from studies of human health is that not all individuals are observed to have the outcome of interest. We say that their event time is 'censored'. This can occur for different reasons:

- In both intervention studies and observational studies, some individuals may be lost to follow-up, meaning that the investigators lose contact with them.
- If individuals in a cohort study are followed up for the outcome 'death', then we nearly always wish to perform some analysis of the data before the point at which all members of the cohort have died. This is called administrative censoring.
- If we are interested in time to disease diagnosis, e.g. in a cohort study, then

some (in fact, usually most) individuals in our study population will never be diagnosed with the disease in question and will be observed to die of another cause. This is a situation of 'competing risks', which is covered in a later session.

In these situations, instead of observing the time of the outcome for each individual (the survival time), for some individuals we only observe a time up to which we know they have not had the outcome. This is referred to as 'right censoring'. Right censoring must be accounted for in our analyses of survival data. It is the major reason that we need special methods of analysis for time-to-event data.

An important assumption

It will be assumed throughout most of this module that censoring is uninformative about event times. This means that the time at which an individual is censored, or the fact that they are censored, does not give us any information about when that person may have the outcome that we are studying.

Other types of censoring

There are other types of censoring, notably 'left censoring' and 'interval censoring', but we do not cover these in this module.

Left truncation (delayed entry)

Left truncation or 'delayed entry' is another feature of survival data that we often encounter. It is different from censoring. Left truncation occurs when we do not observe all individuals from the time origin. It occurs especially when the time scale is age. When the time scale is age individuals may not be followed from birth (the time origin) but instead from, say, entry to a study. Analyses should take into account that individuals are not followed from the time origin, and this is done by conditioning on the fact that an individual did not have the event of interest prior to the start of their follow-up.

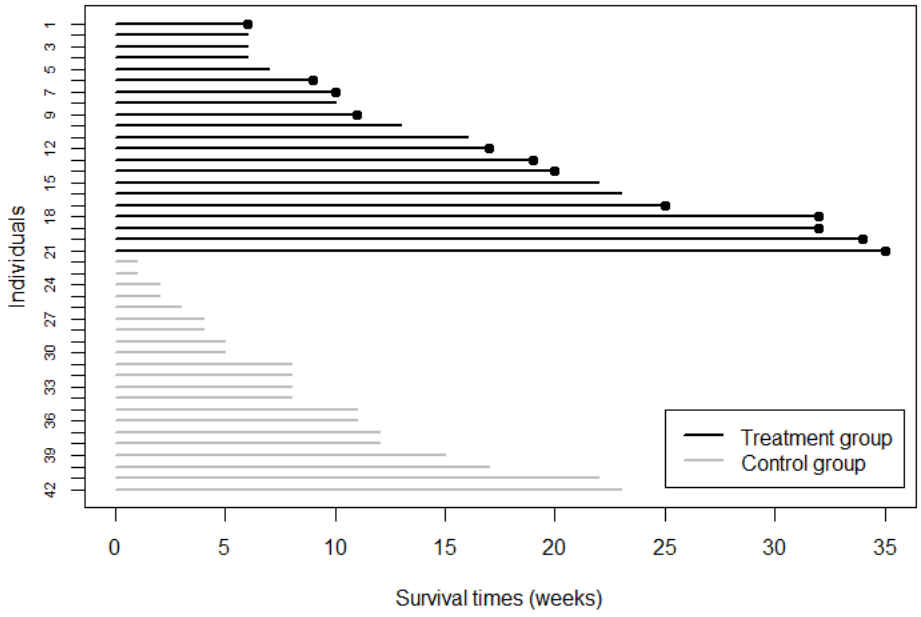
Example 1.1

Time to death for leukemia patients in a randomized trial.

Figure 1.1 shows the times to death among 42 leukemia patients from a randomized trial. The time origin is the time (i.e. a date) of leukemia diagnosis, at which point patients were randomized to treatment or control groups. The time scale is weeks since diagnosis. There are a large number of censored death times, all of which are in the treatment group. How can we formally compare the survival times in the treatment and control groups, taking into account the censoring?

[Example taken from Cox and Oakes 1984]

Figure 1.1: Times to death for 42 leukemia patients in a randomized trial, where a dot indicates a censoring time.



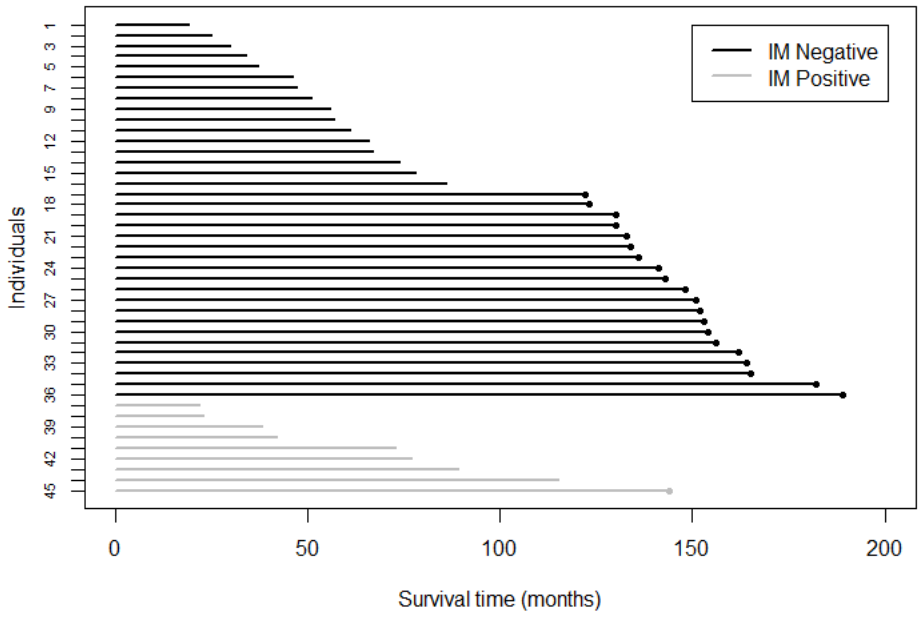
Example 1.2

Survival following breast cancer diagnosis

Figure 1.2 shows the data on 45 female breast-cancer patients from a cancer registry. The time origin is the date of breast cancer diagnosis. The time scale is months since diagnosis. There are some censored death times. It may be of interest to study the survival times in this group, or to investigate whether survival time depends on patient characteristics such as immuno-histochemical response (positve or negative).

[Example taken from Klein and Moeschberger 1984. Data available in the KMsurv package in R (the 'btrial' data set).]

Figure 1.2: Times to death for 45 breast cancer patients observed in a registry, showing individuals separately by immuno-histochemical responses, where a dot indicates a censoring time.



1.6 Analysis of survival data

We might think of studying patterns of observed survival times in a very simple way by looking at histograms of survival times and producing summary statistics, e.g. the mean or median survival. But it is not clear how to handle censoring using this very simple approach. Furthermore, we usually wish to describe survival times more formally. We come on to this in the next section.

Methods for analysis of survival data need to handle censoring and the fact that survival times are strictly non-negative. In order to investigate how survival times are associated with continuous exposures, or with multiple exposures, we want to use regression-type approaches that are analogous to those for continuous exposures (linear regression) and binary outcomes (logistic regression), for example.

Survival data are analysed using special methods. In this course we will learn about the following three approaches:

- Non-parametric methods: These are relatively simple methods and do not make assumptions about the distribution of survival times. They are an important starting point in many investigations of survival data, but do not extend to more complex problems.
- Fully parametric methods: In these we make assumptions (which can and should be carefully investigated) about the patterns of survival times by describing the survival times distribution using a parametric model. When we have explanatory variables, this leads to regression-type analyses for survival data, which are

analogous to regression approaches for other types of response variables.

- Semi-parametric methods: These parametrize how survival times are associated with exposures of interest, but they leave part of the full distribution of the survival times unspecified. This leads to another regression-type analysis for survival data, called Cox regression or proportional hazards regression.

1.7 Describing survival data

To analyse survival data we need to know how to describe it formally. In this section we will learn how to describe distributions of survival times mathematically.

We define a random variable T, which denotes the event time. There are a number of ways in which the distribution of this random variable can be described. We focus on three, which are all related:

- The survivor function.
- The hazard function and the cumulative hazard.
- The probability density function.

The survivor function

The survivor function at a time t (a realisation of the random variable T), which we denote by $S(t)$, is the probability that the survival time T exceeds a value t:

$$S(t) = \Pr(T > t) \tag{1.1}$$

The survivor function is simply the probability that an individual in our underlying population of interest survives beyond time t. For example, if T denotes age at death in the UK population (thus age is the time-scale), then the survivor function at age 80 ($t = 80$) is

$$S(80) = \Pr(T > 80)$$

In general, the survivor function is a smooth function of t. From previous modules you will be familiar with the cumulative distribution function for a random variable, often denoted F(t), which is defined as

$$F(t) = \Pr(T \leq t) \tag{1.2}$$

It is clear to see that the survivor function is 1 minus the cumulative distribution function:

$$S(t) = 1 - F(t) = 1 - \Pr(T \leq t) \tag{1.3}$$

The hazard function

Another way of describing the distribution of survival times is by the hazard function. We denote the hazard function at time t by $h(t)$. To define the hazard function we first

consider a discrete-time setting, where the event could be observed at times $1, 2, 3, \ldots$. The hazard at time t is

$$h(t) = \Pr(t \leq T < t + 1 | T \geq t). \tag{1.4}$$

The hazard is the instantaneous probability of having the event at a given time, conditional on survival up to that time. In a continuous time setting, the hazard function is defined as

$$h(t) = \lim_{\delta \to 0} \frac{1}{\delta} \Pr(t \leq T < t + \delta | T \geq t). \tag{1.5}$$

The hazard function is conceptually a bit more difficult to understand than the survivor function. It is the limit as δ gets very small of the probability that the outcome occurs between time t and time $t + \delta$ given that it has not yet occurred up to time t. Given a population all alive now, the hazard is the proportion of the population that will die in the next short unit of time, divided by the length of the short time unit, thus giving a rate. We can replace 'death' here by the occurrence of some other type of event, e.g. disease diagnosis, conception.

In future lectures we will see that it is particularly useful in the analysis of survival data to focus on the form of the hazard function.

The cumulative hazard is also often used. The cumulative hazard at time t is defined as

$$H(t) = \int_0^t h(u)du \tag{1.6}$$

The probability density function

The probability density function at time t is denoted $f(t)$ and is defined as

$$f(t) = \lim_{\delta \to 0} \frac{1}{\delta} \Pr(t \leq T < t + \delta). \tag{1.7}$$

Relationships between survivor, hazard and probability density functions

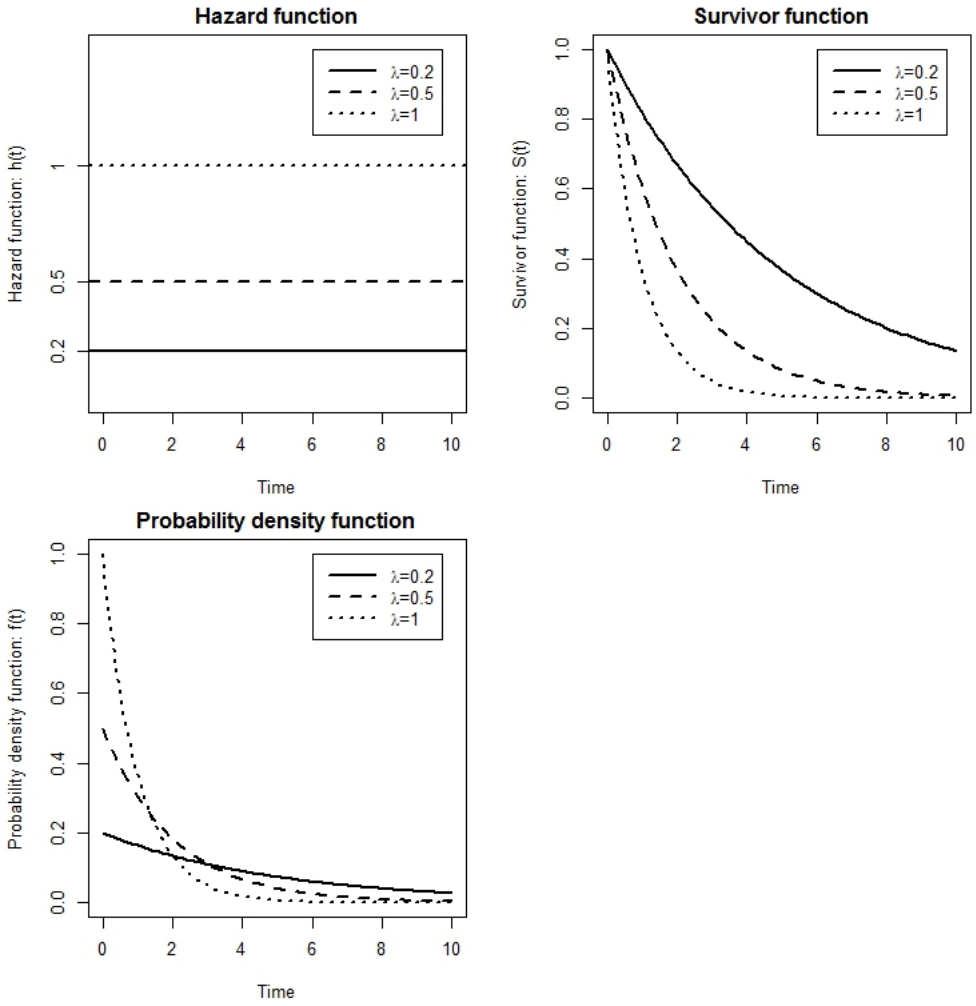
It is important and useful to understand the relationships between the three functions described above. Some useful relationships are as follows:

$$\begin{aligned} f(t) &= -\tfrac{d}{dt} S(t) \\ S(t) &= \textstyle\int_t^\infty f(u)du \\ h(t) &= \tfrac{f(t)}{S(t)} \\ h(t) &= -\tfrac{d}{dt} \log S(t) \end{aligned}$$

Exercise 1.1

Write the survivor function $S(t)$ in terms of the hazard $h(t)$, and then in terms of the cumulative hazard $H(t)$.

Figure 1.3: Graphs showing the hazard function, survivor function and probability density function under an exponential distribution for survival times.



1.8 Parametric distributions of survival times

In this section we learn about some commonly used distributions for survival times. That is, we learn about some functions that can be attached to the survivor, hazard and probability density functions.

The exponential distribution

The simplest distribution for survival times is the exponential distribution. Under the exponential distribution the hazard rate is constant, meaning that the rate of occurrence of the outcome of interest does not vary over time. The exponential distribution therefore depends on one parameter, λ, and we can write the hazard function, survivor function and probability density function as:

$$h(t) = \lambda, \quad S(t) = e^{-\lambda t}, \quad f(t) = \lambda e^{-\lambda t} \tag{1.8}$$

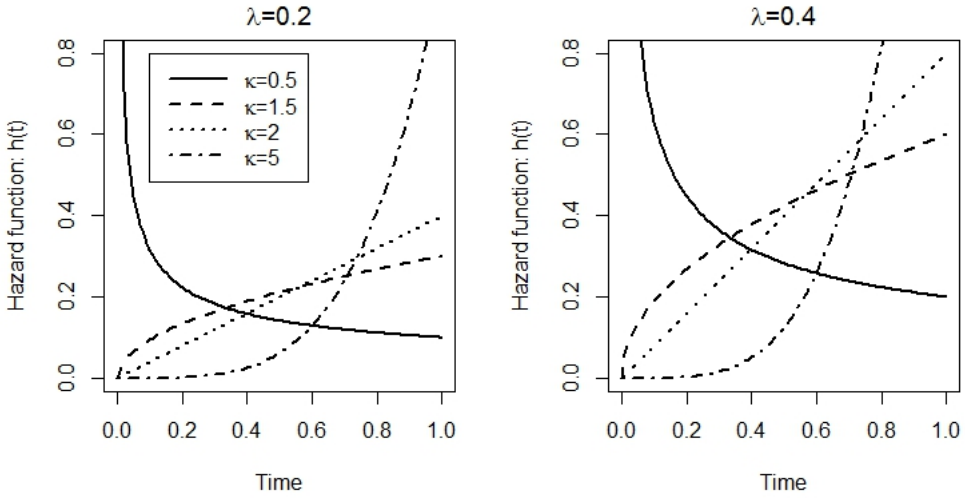
The Weibull distribution

The exponential distribution is often not a suitable way of describing survival data because it assumes a constant hazard rate. Another distribution that we will consider further is the Weibull distribution. This involves two parameters, with the hazard, survivor and probability density functions given by

$$h(t) = \kappa\lambda t^{\kappa-1}, \quad S(t) = e^{-\lambda t^{\kappa}}, \quad f(t) = \kappa\lambda t^{\kappa-1} e^{-\lambda t^{\kappa}} \tag{1.9}$$

The parameter κ is called the shape parameter and λ is called the scale parameter. A feature of the Weibull distribution is that the hazard is a monotonic function, i.e. it is either increasing or decreasing (or constant, in the special case where $\kappa = 1$). It does not change direction. Figures 1.4 shows the shape of the hazard under a Weibull model for different values of the two model parameters, κ and λ. Corresponding survival functions and density function are shown in Figure 1.5. We can see that the Weibull distribution allows for monotonically increasing or decreasing hazards with a wide range of shapes over time.

Figure 1.4: Illustrations of the hazard function under a Weibull distribution with different shape (κ) and scale (λ) parameters.



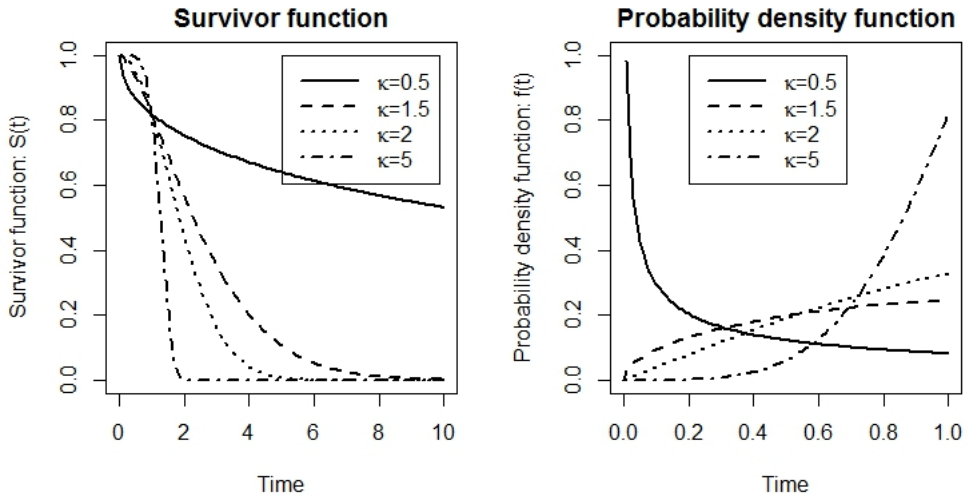
Exercise 1.2

What is the connection between the exponential distribution and the Weibull distribution?

Other distributions

There are many other possible distributions for survival times. In Practical 1 we will also learn about the log-logistic distribution.

Figure 1.5: Illustrations of the survivor function and probability density function under a Weibull distribution with scale parameter $\lambda = 0.2$ and different values for the shape parameter (κ).



1.9 Likelihoods

Suppose we have decided on a suitable distribution for our survival data, e.g. the exponential distribution. We can estimate the parameters of the survival distribution by maximum likelihood estimation. To do this, we introduce some new notation.

Consider a population of n individuals ($i = 1, \ldots, n$) followed up for some period of time from the time origin. Some individuals have the outcome of interest and a survival time t_{Ei} is observed. Some individuals are censored, and for them we observe the time of censoring, which we denote t_{Ci}. For a censored individual i, t_{Ci} is the time up to which we know they have survived without having the outcome of interest. We do not know what happens to them after time t_{Ci}.

We also introduce an indicator variable δ_i, which takes value 1 for individuals who have the event of interest (i.e. for whom t_{Ei} is observed) and which takes value 0 for individuals who are censored (i.e. for whom t_{Ci} is observed). We define t_i to be the observed time for each individual, which may be a survival time or a censoring time:

$$t_i = \begin{cases} t_{Ei} & \text{if } \delta_i = 1 \\ t_{Ci} & \text{if } \delta_i = 0 \end{cases} \tag{1.10}$$

The data for the n individuals can be displayed as in Table 1.1.

The contribution to the likelihood for an individual i who is observed to have the outcome of interest at time t_{Ei} is $f(t_{Ei})$. For an individual i who is censored at time t_{Ci}, their contribution to the likelihood is $S(t_{Ci})$. Using our notation we can therefore write the full likelihood (informally) as

$$L = \prod_{\text{events}} f(t_{Ei}) \prod_{\text{censorings}} S(t_{Ci}) \tag{1.11}$$

Table 1.1: Data on survival and censoring times for n individuals.

Individual	Event or censoring time	Event indicator
1	t_1	δ_1
2	t_2	δ_2
.	.	
.	.	
.	.	
n	t_n	δ_n

More formally we can write this as

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \tag{1.12}$$

Exercise 1.3

Using what we know about the relationships between the survivor, hazard and probability density function, write the likelihood above in terms of the hazard function and the survivor function.

The maximum likelihood estimates for parameters are found in the usual way using the following steps:

(i) Writing down the log likelihood:

$$l = \sum_{i=1}^{n} \delta_i \log f(t_i) + \sum_{i=1}^{n} (1 - \delta_i) \log S(t_i) \tag{1.13}$$

(ii) Differentiating the log likelihood with respect to the parameters of the survival model, which we denote collectively by θ, say, and equating the resulting equations to 0: $dl/d\theta = 0$

(iii) Solving the set of equations to obtain maximum likelihood estimates for θ

The variance-covariance matrix for the estimates θ is given by the inverse of the observed information matrix (which is the negative observed second derivatives of the log-likelihood).

Future lectures

In future lectures we will extend this discussion of the use of parametric models for survival data. In particular, we will learn about:

- Using fully parametric models to study associations between individual features and survival.
- Assessing the suitability different parametric survival models.

Exercise 1.4

(a) Write down the likelihood for survival data (including censoring) assumed to follow an exponential distribution.
(b) Derive a formula for the maximum likelihood estimate for λ.

References

Cox DR and Oakes D. 1984. Analysis of survival data. Chapman and Hall.

Klein JP, Moeschberger ML. 2003. Survival Analysis: Techniques for Censored and Truncated Data. Second Edition. Springer.

Practical 1

Datasets required: `pbcbase_2021`, `whitehall` and `surv_data_practical1`

R packages required for R users: survival, flexsurv.

Introduction

This first practical is in three parts.

A Looks at some fundamentals of setting up Stata and R for use with survival data, and explores the effect of changing the origin and entry dates.

B You will be asked to derive the general formula for the maximum likelihood estimator of λ for exponential survival data, and to use this formula to estimate $\hat{\lambda}$ from data.

C Investigates two other distributions for survival data: the Weibull distribution and the log-logistic distribution.

Aims

By the end of this practical you should:

- Understand how to prepare a dataset for survival analysis (in Stata and/or R)
- Be able to derive the formula for the maximum likelihood estimator of the hazard rate parameter λ from an exponential distribution
- Be able to calculate the maximum likelihood estimate, $\hat{\lambda}$, from data
- Be aware of the Weibull and log-logistic distributions, and recognise their parameters

Where code examples are given or explanations are given that are specific to Stata or R, text and code relating to Stata is shown in this colour and text and code relating to R is shown in this colour.

Part A: Using survival data in Stata and R

The first dataset we will use is called pbcbase_2021. This is data from a multicentre, double-blind, clinical trial for the treatment of Primary Biliary Cirrhosis (PBC). There are 184 individuals in the data and 17 variables; the variables of interest in this session are described below.

Variable	Description
`id`	Unique identifier for each participant
`datein`	Date person entered the study
`dateout`	Date of the end of follow-up due to either death or censoring
`d`	Event indicator at the end of follow-up: 0=alive (censored), 1=dead
`time`	Follow-up time in years

Open the pbcbase_2021 dataset and familiarise yourself with the key variables.

1. How many people died in the dataset, and how many were censored?
2. What was the earliest date of entry to the study? And the latest?
3. What was the earliest date of exit from the study? And the latest?

   Dates in Stata

   The `datein` and `dateout` variables are shown in Stata in the format (day,month,year). However, behind the scenes in Stata the dates are simply recorded as numbers. The number used by Stata to encode a particular date is as the number of days after 1st January 1960. So 1st January 1960 takes value 0, 2nd January 1960 takes value 1, etc. You can see the dates in this format by changing the format in which the dates are viewed:
   `format datein dateout %tg`

   To return to the more readable (day,month,year) format type
   `format datein dateout %td`

   You can find out more about how time is recorded and displayed in Stata using the help files for `datetime`, `format`, and `datetime_display_formats`. Dates can be quite tricky in Stata (and other software packages) so you should always carefully check dates when inputting data and running analyses.

   Dates in R

   The `datein` and `dateout` variables are stored as text in the csv file. Start by formatting these as dates in R as follows:

   ```
   pbc$datein=as.Date(pbc$datein,"%d%b%Y")
   pbc$dateout=as.Date(pbc$dateout,"%d%b%Y")
   ```

   Take a look at what has happened to `datein` and `dateout`. They are now stored in the default form: YYYY-MM-DD. Behind the scenes in R the dates are now recorded as numbers. The number used by R to encode a particular date is as the number of days after 1st January 1970. So 1st January 1970 takes value 0, 2nd January 1970 takes value 1, etc.

   The 'lubridate' package is also useful for handlng date, but we do not use it in this module. Here is a useful blog on dates and times in R:
   https://www.gormanalysis.com/blog/dates-and-times-in-r-without-losing-your-sanity/

4. The fact that Stata and R store dates as numbers makes calculating the time between two dates simple. To calculate the length of time each person was in the study we can type:

   ```
   gen days_in_study = dateout - datein
   ```

```
pbc$days_in_study = pbc$dateout - pbc$datein
```

Compare this new variable to the `datein`, `dateout` and `time` variables in the first few rows to make sure this has done what you expect.

Discuss: What do you think is the appropriate time origin in the PBC study? How is time measured relative to the time origin?

5. For Stata. Before we can perform analyses of survival data in Stata it is necessary to use the `stset` command. This informs Stata what the time origin and time scale are and how the event of interest is indicated.

   (a) One way to `stset` the data is using:
   ```
   stset time, failure(d)
   ```
   This method uses the follow-up time as the time scale (after the `stset`) and specifies the event indicator using the `failure()` option. As the `time` variable is in years, this will be the units in which time is used in the analysis.

   (b) Stata creates some new variables when you `stset` the data. Examine these new variables. What do they each represent? How are each of these new variables related to the variables you used in the `stset` command?

   (c) Another way to `stset` the data is using:
   ```
   stset dateout, origin(datein) failure(d)
   ```
   Here the time of the end of follow-up is provided as the date each person left the study, and the options tell Stata the time origin (the date of joining the study, in this case), and their event indicator. Because the entry and exit dates (`datein` and `dateout`) are stored in number of days, the unit of analysis will be days.

   Examine the `_t` variable, and its relationship with the `time` variable we created earlier.

   (d) We can change the time unit using the `scale()` option. For example, to use years, rather than days we would type:
   ```
   stset dateout, origin(datein) failure(d) scale(365.25)
   ```
   Check that this gives the same values for `_d`, `_t0`, and `_t` as you saw in part (a).

5. For R. In R the 'survival' package contains the key tools for analysis of survival data.

   (a) Install the survival package
   ```
   install.packages("survival")
   library(survival)
   ```

(b) The `Surv()` function is fundamental to survival analysis in R. It is used in functions that we use in survival analysis, e.g. in model formulas, to specify the timescale and to specify which individuals have the event and which are censored. In R `Surv()` is used within other functions, whereas in Stata `stset` is used upfront before any commands using survival analysis methods are used. Try the following:

```
Surv(time=pbc$time,event=pbc$d)
```

Take a look at the output. What do the + symbols represent? This method uses the follow-up time as the time scale and specifies the event indicator. As the `time` variable is in years, this will be the units in which time is used in any analysis in which the the above `Surv()` function is included.

(c) Another way of specifying the `Surv()` function is:

```
Surv(time=as.numeric(pbc$dateout),event=pbc$d,origin=as.numeric(pbc$datein))
```

Here the time of the end of follow-up is provided as the date each person left the study, and the options tell R the time origin (the date of joining the study, in this case), and their event indicator. Because the entry and exit dates (`datein` and `dateout`) are stored in number of days, the unit of analysis will be days.

Whitehall data

We will now switch to use the Whitehall Study data ('whitehall'): this is data from the Whitehall Study, which is a cohort study, started in 1967, of risk factors for mortality in British male civil servants employed in various government departments in London. We will focus on deaths due to coronary heart disease (CHD). The dataset contains 1677 individuals and 14 variables, with the variables used in this practical described below:

Variable	Description
`id`	Unique identifier for each participant
`timebth`	Date of birth
`timein`	Date person entered the study
`timeout`	Date of the end of follow-up due to either death or censoring
`chd`	Event indicator at the end of follow-up: 0=alive or died from other cause (censored), 1=death due to coronary heart disease

6. What is the earliest date of entry into the study? And the latest?

In R you will need to format the dates as before.

```
whl$timein=as.Date(whl$timein,"%d%b%Y")
whl$timeout=as.Date(whl$timeout,"%d%b%Y")
whl$timebth=as.Date(whl$timebth,"%d%b%Y")
```

7. What is the latest date of exit from the study? How many people left the study on this date? How do you explain this?

8. What do you think the appropriate time scale is in this study?

   In Stata, compare the following four ways of using stset for this data:

```
stset timeout, failure(chd) origin(timein)
stset timeout, failure(chd) origin(timebth)
stset timeout, failure(chd) origin(timebth) enter(timein)
stset timeout, failure(chd) origin(timebth) enter(timein) scale(365.25)
```

   What is the time scale being used in each case? Examine the variables created by Stata in each case.

   In R, compare the following four ways of using `Surv()` for this data. For this question it turns out to be convenient to have the date variables in numeric format.

```
whl$timein=as.numeric(whl$timein)
whl$timeout=as.numeric(whl$timeout)
whl$timebth=as.numeric(whl$timebth)

Surv(time=whl$timeout,event=whl$chd,origin=whl$timein)
Surv(time=whl$timeout,event=whl$chd,origin=whl$timebth)
Surv(time=whl$timein,time2=whl$timeout,event=whl$chd,origin=whl$timebth)
Surv(time=whl$timein/365.25,time2=whl$timeout/365.25,event=whl$chd,
        origin=whl$timebth/365.25)
```

   What is the time scale being used in each case? Examine the output from `Surv()`.

   The third case is an example of using `stset` or `Surv()` to allow for 'left-truncation' or 'delayed entry'. Left truncation occurs when individuals cannot be observed to experience the event of interest until something else has occurred – here, that something else is entry to the study.

   Discuss with your colleagues so that you are clear you understand how each `stset` command changes the way Stata treats each person's progress through the study, and/or how each `Surv()` command changes the way R treats each person's progress through the study.

   We'll look at how changing these dates of entry and origin affects calculated survival probabilities and other analyses in Practical 2.

Part B: Fitting models to survival data

The first two questions are pen & paper exercises.

1. Write down the likelihood for survival data (including censoring) assumed to follow an exponential distribution.

2. Derive a formula for the maximum likelihood estimate for λ.

   We will now use this estimate in a simulated dataset called `surv_data_practical1`. There are just two variables: `survtimes` contains the survival time for each participant, and `d` the outcome. In this dataset all participants experienced the event (d=1). The data were simulated for 100 individuals, generated from an exponential distribution with hazard $\lambda = 0.2$.

3. Use simple descriptive commands in Stata or R to help you calculate, by hand, the maximum likelihood estimate for λ for these data.

4. We will fit an exponential model to these data to confirm the maximum likelihood estimate for λ we calculated above. Compare the results with what you found by hand.

   In Stata use `stset` to declare the data to be survival data. An exponential model can be fitted using the streg command:

   ```
   streg, distribution(exponential)
   ```

   See what happens if you use the `nohr` option in `streg`.

   In R exponential model can be fitted using `survreg` command from the survival package:

   ```
   exp.model = survreg(Surv(survtimes)~1,dist="exponential",data=mydata)
   summary(exp.model)
   ```

   It is important to note that the `survreg` function with `dist="exponential"` estimates $-\log\lambda$ rather than λ.

   The flexsurv package provides another way of fitting an exponential model. Try the following and compare with the results you obtained above.

   ```
   install.packages("flexsurv")
   library(flexsurv)
   exp.model2 = flexsurvreg(Surv(survtimes)~1,dist="exponential",data=mydata)
   exp.model2
   ```

   Discuss the interpretation of your estimate $\hat{\lambda}$.

Part C: Two distributions of survival data

We will again use the simulated data set containing survival times generated from an exponential distribution (`surv_data_practical1`).

1. The form of the hazard function in a Weibull distribution is given in equation (1.9). Here we will investigate graphically how the values of the parameters λ, κ

affect the hazard function. Play around with the values of the two parameters to investigate the effect on the shape of the hazard.

In Stata we can plot the hazard function for example values of λ, κ as follows:

```
local kappa=2
local lambda=0.2
graph twoway function y=‘kappa’*‘lambda’*x^(‘kappa’-1)
```

Note that you must run all three lines from a Do file at the same time; Stata will forget the value of the local macros kappa and lambda as soon as the Do file has been run.

In R we can plot the hazard function for example values of λ, κ as follows:

```
wei.haz<-function(x,lambda,kappa){lambda*kappa*x^(kappa-1)}
curve(wei.haz(x,0.2,2),xlab="Time",ylab="Hazard function")
```

2. Another model for survival data is the log-logistic distribution. The log-logistic model has hazard function of the form

$$h(t) = \frac{e^{\theta}\kappa t^{\kappa-1}}{1 + e^{\theta}t^{\kappa-1}}$$

Investigate the shape of the hazard function for different values of the parameters θ and κ. What feature does the log-logistic distribution have that the Weibull distribution does not have?

Session 2

Non-parametric analysis of survival data

2.1 Aims of this lecture and practical

At the end of this lecture and practical you will be able to:

- Estimate survivor functions and cumulative hazard functions non-parametrically, allowing for censoring. The methods used are referred to as the Kaplan-Meier method and the life-table method.
- Estimate uncertainty in the non-parametric estimates using Greenwood's formula.
- Make plots of non-parametric estimates.
- Use non-parametric methods to compare survival in different groups of individuals, e.g. treatment and control groups.
- Perform a test of whether survival curves differ between two groups (the log rank test).
- Perform non-parametric analyses in Stata and R and interpret the output.

2.2 Why use non-parametric methods?

Non-parametric methods are a relatively simple starting point for most analyses of survival data. Some reasons for using non-parametric analyses are:

- Using non-parametric methods we can estimate survivor functions and cumulative hazards without having to make parametric assumptions.
- Non-parametric methods provide a nice way of graphically displaying survival data, including when there is censoring.
- Non-parametric methods provide a simple way of comparing patterns of survival in two (or more) groups of individuals.
- Non-parametric methods can be used to inform more complex modelling of survival data.

2.3 Estimating the survivor function: the Kaplan-Meier estimate

Situation without censoring

If everyone in our study had the outcome of interest (e.g. death) then we could observe a survival time on everybody. In this situation an obvious simple non-parametric estimate of the survivor function $S(t) = \Pr(T > t)$ at a given time t is

$$\widehat{S}(t) = \frac{\text{no. individuals with event time } > t}{\text{total no. of individuals}} \tag{2.1}$$

The value of the estimated survivor function can be found at each unique event time t_k $(k = 1, \ldots, K)$ and then plotted on a graph.

Example 2.1

Time to death among leukemia patients: estimated survivor function in the control group

We will illustrate the calculation of the above estimated survivor function using the data from Example 1.1. The times to death (in weeks) in the treatment and control groups are shown in table 2.1. We don't yet know how to handle censoring, so we focus first on the control group only, where there is no censoring.

Table 2.1: Times to death among leukemia patients in the control and treatment groups of a randomized trial. * indicates a censoring time.

Control group	1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23
Treatment group	6*,6,6,6,7,9*,10*,10,11*,13,16,17*,19*,20*,22,23,25*,32*,32*,34*,35*

For the control group, the data and the estimated survivor function at each survival time are shown in Table 2.2. The estimated survivor function can be plotted as in Figure 2.1. Note that it is a 'step function', because we assume under this non-parametric approach that the survivor function remains constant between observed survival times.

Table 2.2: Times to death and the estimated survival function in the control group.

Event time t_j	Number of events d_j	$\hat{S}(t_j)$
1	2	19/21=0.90
2	2	17/21=0.81
3	1	16/21=0.76
4	2	14/21=0.67
5	2	12/21=0.57
8	4	8/21=0.38
11	2	6/21=0.29
12	2	4/21=0.19
15	1	3/21=0.14
17	1	2/21=0.10
22	1	1/21=0.05
23	1	0

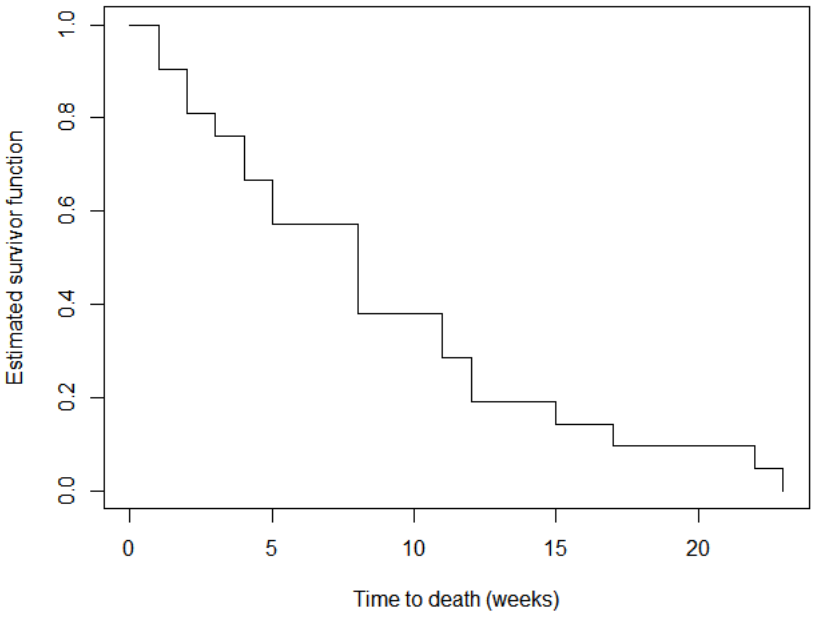


Figure 2.1: Leukemia example: Estimated survivor function corresponding to the data in Table 2.1 for the control group.

Incorporating censoring

The above non-parametric approach can be extended to incorporate censoring. Let the unique ordered event times be denoted $t_1 < t_2 < \cdots < t_K$ (these do not include the censoring times). The formula for the estimated survivor function in this situation is derived using the following steps:

1. First, we define h_j to be the hazard rate at time t_j. So there is a hazard associated with each event time: $h_1, h_2, \ldots, h_K$.
2. The probability that an individual in the population who is eligible to have the outcome at time t_1 (i.e. has not yet been censored) does not in fact have the outcome of interest at time t_1 is $1 - h_1$, i.e. 1 minus the probability of having the outcome instantaneously at time t_1. The probability that an individual in the population who is eligible to have the outcome at time t_1 and at time t_2 does not have the outcome at either time is $(1 - h_1)(1 - h_2)$. This is the probability that they do not have the outcome at time t_1 multiplied by the probability that, given they didn't have the outcome at time t_1, they do not have the event at time t_2.
3. The survival probability is the probability that an individual does not have the event at any time at which they are eligible to have the event. The survivor function at time t_j can therefore be estimated by

$$S(t_j) = \Pr(T > t_j) = \prod_{k=1}^{j}(1 - h_k) \tag{2.2}$$

where the product is over all observed survival times up to time t_j. Think about

this: if you survive beyond time t_j then you do not have the event at any of the times $t_1, t_2, \ldots, t_j$.

All that is left to do now is to estimate the hazard rate h_j at each survival time. An intuitive estimate of h_j is

$$\hat{h}_j = \frac{d_j}{n_j} \tag{2.3}$$

where d_j is the number of events at time t_j and n_j is the number of individuals 'at risk' at time t_j. Above we used the terminology 'eligible to have the event at time t': this refers to a person who has not had the event prior to time t and who has not been censored prior to time t. If a person has been censored we cannot say for sure whether or not they have had the event or not at any time after their censoring time. Another way of saying a person is 'eligible to have the event at time t' is to say that a person is 'at risk at time t'. The group of individuals who are at risk at time t is referred to as the 'risk set'. The above estimate of h_j can in fact be shown to be a maximum likelihood estimate. Using the above results, the estimated survivor function at an observed survival time t_j is

$$\hat{S}(t_j) = \prod_{k=1}^{j}(1 - \hat{h}_k) = \prod_{k=1}^{j}(1 - d_k/n_k) \tag{2.4}$$

More generally, we can write the estimated survivor function at *any* time t as

$$\hat{S}(t) = \prod_{j:t_j \leq t} (1 - d_j/n_j) \tag{2.5}$$

This called the Kaplan-Meier estimate of the survivor function.

Example 2.2

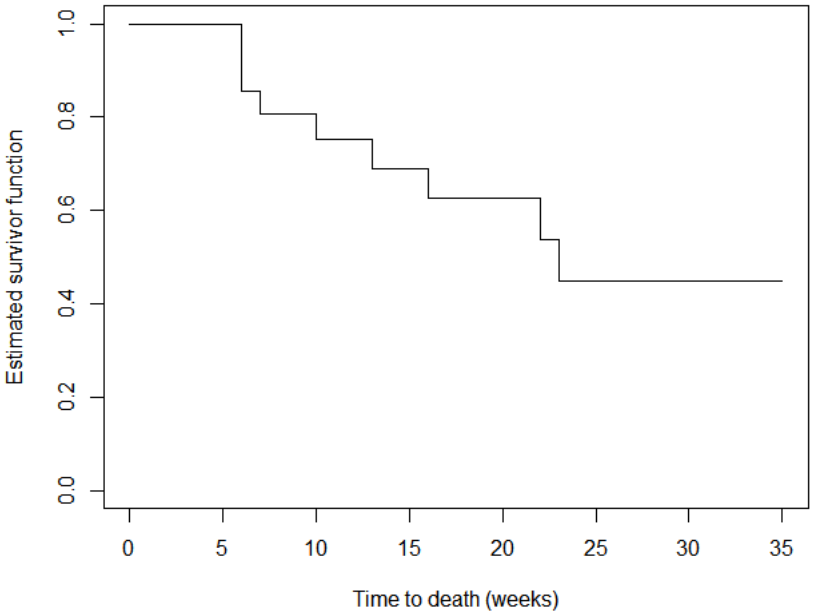
Time to death among leukemia patients in the treatment group: Kaplan-Meier estimate of the survivor function

The survival and censoring times are as given in Table 2.1. Table 2.3 shows the calculation of the Kaplan-Meier estimate of the survivor function and the resulting estimated survival curve is plotted in Figure 2.2. Note that if an event and censoring occur at the same time then the convention is to assume that the event occurred just before the censoring.

Table 2.3: Times to death for leukemia patients in the treatment group: Kaplan-Meier estimates of the survivor function.

Event and censoring times t_j	Number at risk	Number of events	Number of censorings	$\hat{S}(t_j)$
6	21	3	1	(1-3/21)=0.857
7	17	1	0	(1-3/21)(1-1/17)=0.807
9	16	0	1	-
10	15	1	1	0.753
11	13	0	1	-
13	12	1	0	0.690
16	11	1	0	0.627
17	10	0	1	-
19	9	0	1	-
20	8	0	1	-
22	7	1	0	0.538
23	6	1	0	0.448
25	5	0	1	-
32	4	0	2	-
34	2	0	1	-
35	1	0	1	-

Figure 2.2: Leukemia example: Estimated survivor function corresponding to the data in Table 2.1 for the treatment group.



Example 2.3

Survival following breast cancer diagnosis

Table 2.4 shows the death and censoring times of women in the breast cancer study introduced in Example 1.2.

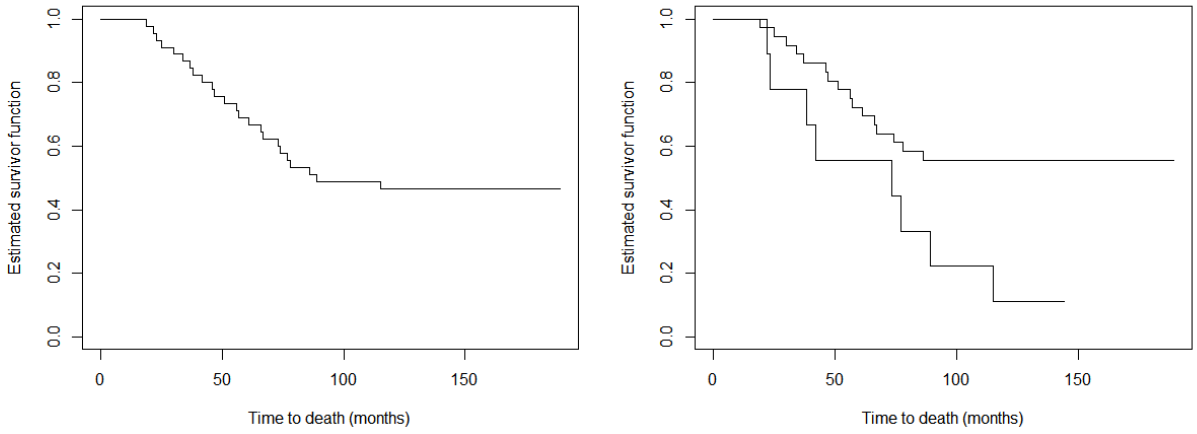
Table 2.4: Times to death (in months) among women diagnosed with breast cancer, by IM status. * indicates a censoring time.

IM negative	19,25,30,34,37,46,47,51,56,57,61,66,67,74,78,86,122*,123*,130*,130*, 133*,134*,136*,141*,143*,148*,151*,152*,153*,154*,156*,162*, 164*,165*,182*,189*
IM positive	22,23,38,42,73,77,89,115,144*

Exercise 2.1

Create a suitable table and obtain Kaplan-Meier estimates of the survivor function (a) in the whole group, (b) separately by IM status. The Kaplan-Meier plots are shown in Figure 2.3 - check that you can recreate these from your table.

Figure 2.3: Breast cancer example: Estimated survivor function corresponding to the data in Table 2.1. Left hand plot: the overall survivor function estimate. Right hand plot: separately by IM group.



In Stata:

```
stset time, failure(death)
sts graph
sts graph, by(im)
```

In R:

```
btrial.km <- survfit(Surv(time,death)~im,data=btrial)
ggsurvplot(btrial.km, data = btrial)
```

2.4 Estimating uncertainty in the Kaplan-Meier estimate: Greenwood's formula

Derivation of Greenwood's formula

As well as estimating the survivor function we want to be able to estimate the uncertainty in the estimates. That is, we want to find $\text{var}(\hat{S}(t))$. It helps to start by considering the variance of $\log \hat{S}(t)$. Using equation (2.5):

$$\begin{aligned} \text{var}\left(\log \hat{S}(t)\right) &= \text{var}\left(\log \prod_{j:t_j \leq t} (1-\hat{h}_j)\right) \\ &= \text{var}\left(\sum_{j:t_j \leq t} \log(1-\hat{h}_j)\right) \\ &= \sum_{j:t_j \leq t} \text{var}\left(\log(1-\hat{h}_j)\right) \end{aligned} \tag{2.6}$$

We now use the following linear approximation (first order Taylor series approximation):

$$\log(1-\hat{h}_j) \approx \log(1-h_j) - (\hat{h}_j - h_j)/(1-h_j) \tag{2.7}$$

which gives

$$\text{var}\left(\log(1-\hat{h}_j)\right) \approx \frac{\text{var}(\hat{h}_j)}{(1-h_j)^2} \tag{2.8}$$

Note that this is the delta method. Now we need to find the variance of $\hat{h}_j = d_j/n_j$. The number of events, d_j, at time t_j has a binomial distribution conditional on the number of people at risk, n_j:

$$d_j \sim \text{Binomial}(n_j, h_j) \tag{2.9}$$

It follows that

$$\text{var}(\hat{h}_j) = \frac{\text{var}(d_j)}{n_j^2} = \frac{n_j h_j (1-h_j)}{n_j^2} = \frac{h_j(1-h_j)}{n_j} \tag{2.10}$$

Putting all of this together gives

$$\text{var}\left(\log \hat{S}(t)\right) = \sum_{j:t_j \leq t} \frac{h_j}{n_j(1-h_j)} \tag{2.11}$$

Remember that what we really want is $\text{var}(\hat{S}(t))$. The final step is to use another linear approximation and to write

$$\log \hat{S}(t) \approx \log S(t) + (\hat{S}(t) - S(t))/S(t) \tag{2.12}$$

which gives us

$$\text{var}\left(\log \hat{S}(t)\right) = \frac{\text{var}(\hat{S}(t))}{S(t)^2} \tag{2.13}$$

It follows that the variance of the Kaplan-Meier estimate of the survival function can be estimated using

$$\text{var}(\hat{S}(t)) = \hat{S}(t)^2 \text{var}\left(\log \hat{S}(t)\right) = \hat{S}(t)^2 \sum_{j:t_j \leq t} \frac{h_j}{n_j(1-h_j)} \tag{2.14}$$

This is Greenwood's Formula.

95% confidence intervals

A 95% confidence interval for the estimated survivor function can be found using Greenwood's formula as follows:

$$\hat{S}(t) \pm 1.96\sqrt{\text{var}(\hat{S}(t))} \tag{2.15}$$

For obtaining this, we assumed the normality of the survival probability. This is not always a good way of finding confidence limits, because at extreme event times (the left and right of the survival plot), it can give values which are outside the range from 0 to 1. To find confidence intervals which do not have this problem, we focus on using a transformation of $\hat{S}(t)$ which is not bounded. One such transformation is $\log\left(-\log \hat{S}(t)\right)$. Using linear approximations like we did to derive Greenwood's formula we can find that

$$\text{var}\left\{\log\left(-\log \hat{S}(t)\right)\right\} \approx \frac{\text{var}\left(\log \hat{S}(t)\right)}{\left(\log \hat{S}(t)\right)^2} \tag{2.16}$$

If we denote the above variance estimate by $v^2(t)$, It can be found that a 95% confidence interval for S(t) is

$$\hat{S}(t)^{\exp\{\pm 1.96 v(t)\}} \tag{2.17}$$

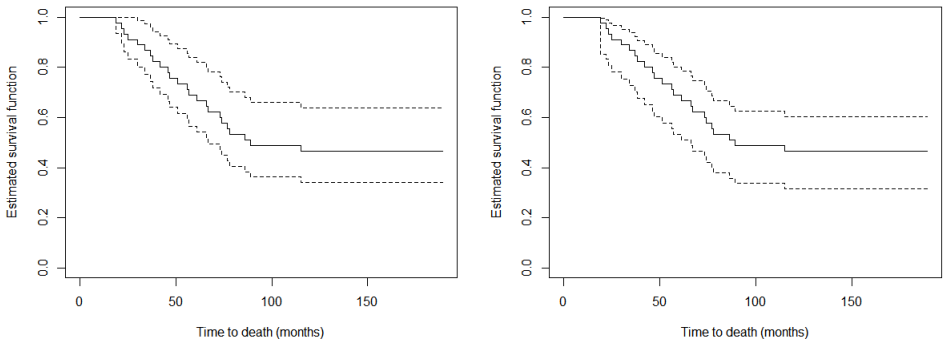
Figure 2.4 shows the overall estimated survivor function for the breast cancer example, showing the two different sets of confidence intervals. You can see some small differences in the two sets of confidence intervals.

2.5 The life table method

To obtain the Kaplan-Meier estimate of the survivor function we used data on individual event and censoring times. Sometimes we only have access to group-level data. For example we may have information on the number of deaths per year for several years in a particular study population, but not exact dates of death for individuals. The life-table method can be used to summarize data on survival of individuals when we only have group-level data within time-intervals.

We suppose that a period of follow-up time is divided into intervals $I_1, I_2, \ldots, I_K$, which are not necessarily of equal length. We denote the number of events in interval I_j by d_j. The number of individuals at risk at the start of interval I_j is denoted n_j. Finally,

Figure 2.4: Breast cancer example: Estimated survivor function corresponding to the data in Table 2.4 with 95% confidence intervals. Left hand plot: using Greenwood's Formula. Right hand plot: using the confidence intervals in 2.17.



the number of censorings within interval I_j, meaning the number of individuals who are censored before the start of the next interval, is denoted m_j. The probability of having the event of interest in interval I_j conditional on survival up to the start of that interval is estimated by

$$p_j = \frac{d_j}{n_j - m_j/2} \tag{2.18}$$

We use $m_j/2$ in the denominator because we're not exactly sure when the censorings occur, so we assume they occur evenly across the interval. Note that we are effectively assuming here that the hazard is constant within an interval. The life table estimate of the survivor function at time t is

$$\hat{S}(t) = \prod_{k=1}^{j}(1 - p_k) \qquad \text{for } t_j \leq t < t_{j+1} \tag{2.19}$$

You can see that the life table estimate is very similar to the Kaplan-Meier estimate. Usually the results are similar. Greenwood's formula can be used to find a variance estimate, by replacing n_j by $n_j - m_j/2$.

Example 2.4

Death among men with angina: the life table approach

Example taken from Van Belle et al. (2004). Biostatistics: A Methodology for the Health Sciences (2nd Ed)

In this example, 2418 men with angina were recruited and dates of death or censoring (loss-to-follow up) were recorded. The numbers of deaths and censorings in the first 10 years of follow-up are summarised by year in Table 2.3. The estimated survivor function is displayed graphically in Figure 2.4.

Table 2.5: Men with angina: Numbers of deaths (d_j), censorings (m_j), total numbers at risk (n_j), and the life-table estimate of the survivor function by year.

Year	n_j	d_j	m_j	p_j	$1-p_j$	$\hat{S}(t)$
0-1	2418	456	0			
1-2	1962	226	39			
2-3	1697	152	22			
3-4	1523	171	23			
4-5	1329	135	24			
5-6	1170	125	107			
6-7	938	83	133			
7-8	722	74	102			
8-9	546	51	68			
9-10	427	42	64			

Exercise 2.2

(a) Complete table 2.5 and compare your results with those shown graphically in Figure 2.5.
(b) What was the probability of surviving 5 years or more? What was the probability of surviving 2.5 years or more?

Figure 2.5: Men with angina data: Life table estimate of the survivor function.

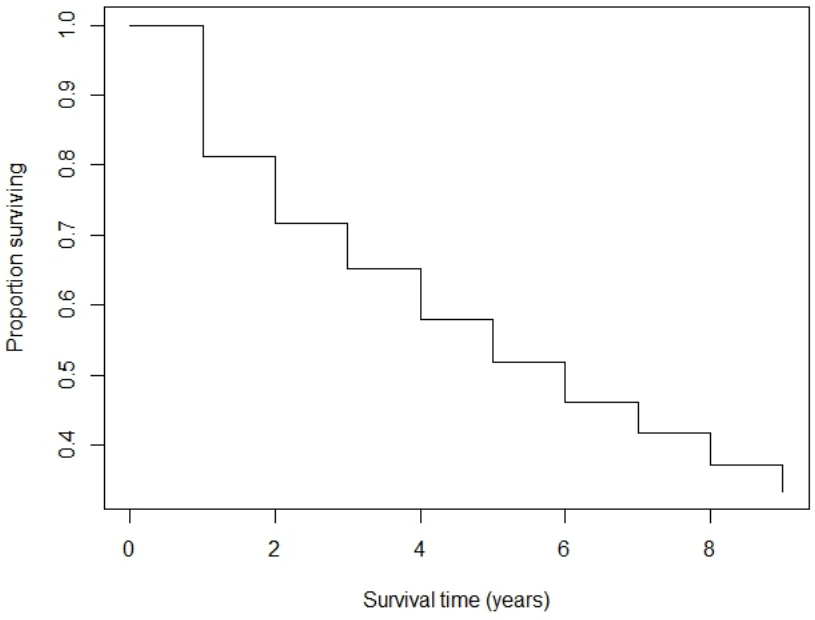


Table 2.6: Summary of numbers at risk and number of events at time t_j in two groups.

Group	Events at t_j	No. surviving beyond t_j	No. at risk at t_j
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

2.6 Comparing survival in two groups: the log rank test

The non-parametric methods introduced in this lecture can be used to compare survival in groups of individuals, i.e. individuals in the study population with different characteristics. Estimated survival curves can be created separately for individuals in different groups and plotted on the same graph, with corresponding confidence intervals, for visual comparison.

The survival curves for two (or more) groups can be compared formally using a statistical test. We denote the survivor curves in two groups $byS_1(t)$ and $S_2(t)$. A simple test of whether the survival probability differs between two groups at a particular time, say time u, can be performed by calculating the test statistic

$$\frac{\hat{S}_1(u) - \hat{S}_2(u)}{\sqrt{\text{var}\hat{S}_1(u) + \text{var}\hat{S}_2(u)}} \tag{2.20}$$

and by comparing this with the standard normal distribution.

However, we can do better than this single-time-point comparison, by comparing the whole survival time distributions between two (or more) groups. This can be done by using a test called the log rank test, or sometimes the Mantel-Haenszel test. The log rank test is test of whether the survivor curves for two (or more) groups are different. The null hypothesis is that the two survivor curves are the same. The alternative hypothesis is that the two survivor curves are different.

To derive the test, we first list the combined survival times in the two groups of individuals, groups 1 and 2 say. At a particular survival time t_j we denote the number of events at that time in the two groups by d_{1j} and d_{2j}, which occur within numbers at risk at that time of n_{1j} and n_{2j}. The data at survival time t_j can be summarized as in Table 2.4.

Under the null hypothesis of no association between group membership and the number of events, the number of events in group 1 has a hypergeometric distribution. Under this distribution the probability of d_{1j} events in a group of size n_{1j} given a total population of size n_j in which there are a total of d_j events is

$$\frac{\begin{pmatrix} d_j \\ d_{1j} \end{pmatrix} \begin{pmatrix} n_j - d_j \\ n_{1j} - d_{1j} \end{pmatrix}}{\begin{pmatrix} n_j \\ n_{1j} \end{pmatrix}} \tag{2.21}$$

Note that we could have written this in terms of d_{2j} and n_{2j} instead. At a given time t_j we can test whether the number of events in group 1 is much more (or much less) than would be expected under the null hypothesis that the number of events in the two groups at this time does not differ. Under this null hypothesis, the expected number of events at time t_j in group 1 is, according to the above distribution,

$$e_{1j} = \frac{n_{1j} d_{1j}}{n_j} \tag{2.22}$$

The difference between the observed and expected number of events at time t_j in group 1 is $d_{1j} - e_{1j}$. To use the information at all survival times, we sum the differences between observed and expected number of events in group 1 to give the difference between the total observed and total expected number of deaths over the whole follow-up period

$$\sum_j (d_{1j} - e_{1j}) \tag{2.23}$$

This statistic has expectation zero under the null hypothesis. We can also find its variance according to the hypergeometric distribution, using

$$v_{1j}^2 = \text{var}(d_{1j}) = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \tag{2.24}$$

This gives the test statistic

$$\frac{\left\{\sum_j (d_{1j} - e_{1j})\right\}^2}{\sum_j v_{1j}^2} \tag{2.25}$$

which, under the null hypothesis, has a chi-squared distribution with 1 degree of freedom, χ_1^2. This is the log rank test. The log rank test can be extended to enable comparison of survival curves in more than two groups. In that case, the degrees of freedom for the chi-squared test is the number of groups minus 1.

Example 2.5

Leukemia example: comparison of survival in the two treatment groups

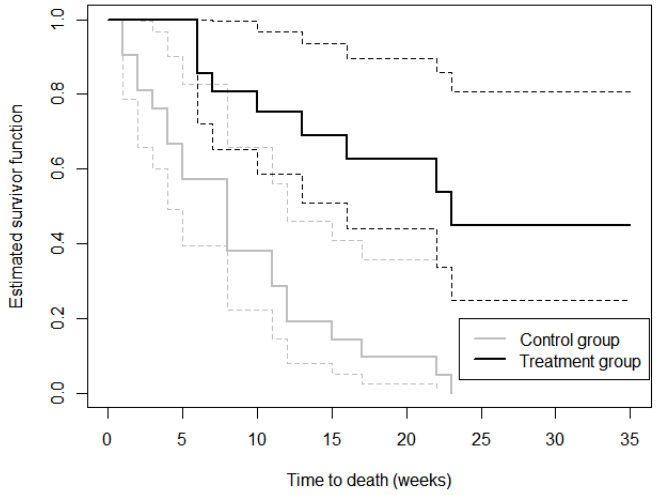
Figure 2.6 shows the Kaplan-Meier estimates of the survivor curves for the leukemia example separately by treatment group, including 95% confidence intervals.

The results from performing the log rank test in Stata are shown below:

```
. stset time, failure(death)
. sts test group, logrank

         failure _d:  death
   analysis time _t:  time
```

Figure 2.6: Leukemia example: Estimated survivor functions by treatment group, showing 95% confidence intervals.




```
Log-rank test for equality of survivor functions

          |   Events         Events
group     |  observed       expected
----------+-------------------------
Control   |        21          10.75
Treatment |         9          19.25
----------+-------------------------
Total     |        30          30.00

                chi2(1) =      16.79
                Pr>chi2 =     0.0000
```

The results from performing the log rank test in R are shown below:

```
> survdiff(Surv(time,d)~group,data=leuk)
Call:
survdiff(formula = Surv(time, d) ~ group, data = leuk)

          N Observed Expected (O-E)^2/E (O-E)^2/V
group=0 21       21     10.7      9.77      16.8
group=1 21        9     19.3      5.46      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

In this example we have $\sum_j (d_{1j} - e_{1j}) = 21 - 10.75 = 10.25$ and $\sum_j v_{1j}^2 = 6.56$, giving a test statistic of 16.79. This is compared to the χ_1^2 distribution. The p-value is < 0.0001. There is strong evidence against the null hypothesis that the survivor curves

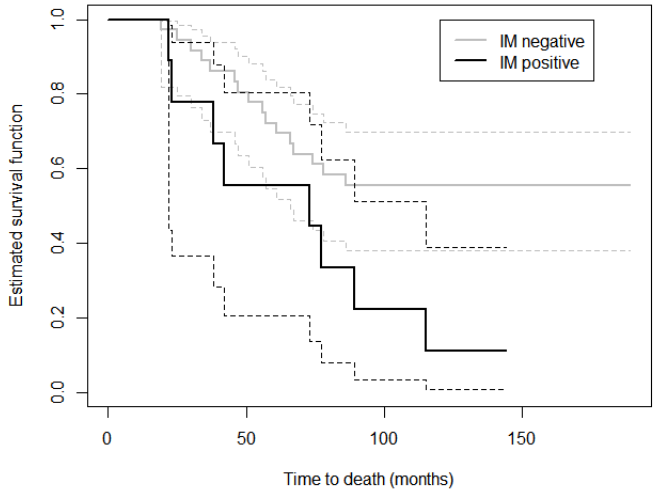
in the treatment and control groups are the same.

Example 2.6

Breast cancer example: comparison of survival by IM status

Figure 2.7 shows the Kaplan-Meier estimates of the survivor curves for the breast cancer data separately by IM status, including 95% confidence intervals.

Figure 2.7: Breast cancer example: Estimated survivor functions by IM status, showing 95% confidence intervals.



The results from performing the log rank test in Stata are shown below:

```
. sts test im, logrank

         failure _d:  death
   analysis time _t:  time

Log-rank test for equality of survivor functions

      |   Events         Events
im    |  observed       expected
------+-------------------------
1     |        16          20.19
2     |         8           3.81
------+-------------------------
Total |        24          24.00

            chi2(1) =       5.49
            Pr>chi2 =     0.0191
```

The results from performing the log rank test in R are shown below:

```
> survdiff(Surv(time,death)~im,data=btrial)
Call:
survdiff(formula = Surv(time, death) ~ im, data = btrial)

      N Observed Expected (O-E)^2/E (O-E)^2/V
im=1 36       16    20.19     0.869      5.49
im=2  9        8     3.81     4.599      5.49

 Chisq= 5.5  on 1 degrees of freedom, p= 0.02
```

Exercise 2.3

State the null hypothesis for the log rank test applied to the breast cancer example. Interpret the above results from performing the log rank test.

2.7 Estimating the cumulative hazard: the Nelson-Aalen estimator

We have focused in this lecture on non-parametric estimates of the survivor function. It can also be useful to use non-parametric estimates of the cumulative hazard. Recall from Lecture 1 that the cumulative hazard is defined as

$$H(t) = \int_0^t h(u)du \tag{2.26}$$

We have the following relationship between the survivor function and the cumulative hazard:

$$H(t) = -\log S(t) \tag{2.27}$$

A non-parametric estimate of $H(t)$ can be found using this relationship and by using the Kaplan-Meier estimate of the survivor function, $\hat{S}(t)$ giving

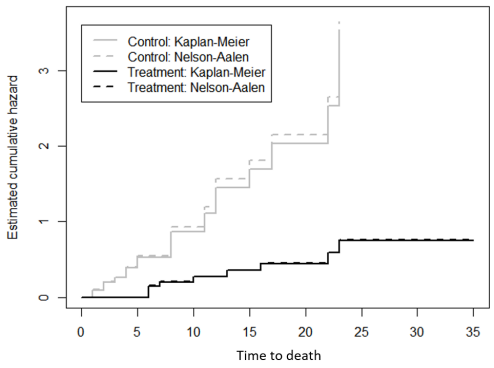
$$\hat{H}(t) = -\log \hat{S}(t) \tag{2.28}$$

However, a simpler estimate is given by summing estimates of the hazard h_j, giving the estimator

$$\hat{H}(t) = \sum_{j:t_j \leq t} \hat{h}_j = \sum_{j:t_j \leq t} d_j/n_j \tag{2.29}$$

This is called the Nelson-Aalen estimator of the cumulative hazard. Estimates obtained using the Kaplan-Meier formula or the Nelson-Aalen formula will usually be similar and are asymptotically equivalent. Figure 2.8 shows Kaplan-Meier and Nelson-Aalen estimates of the cumulative hazard for the leukemia example.

Figure 2.8: Leukemia example: Kaplan-Meier and Nelson-Aalen estimates of the cumulative hazard in the treatment and control groups



2.8 Further comments

We have focused on comparing survivor curves for two groups. Some extensions and further comments are as follows:

- We can compare survival in more than 2 groups by plotting several survivor curves on the same graph and by using a log rank test, based on a χ^2 distribution with suitable degrees of free (1 minus the number of groups being compared).
- It is often of interest to control for potential confounders in our analyses. If the confounders are categorical then we can investigate the impact of confounding by looking at survival curves for the main exposure (e.g. exposed/unexposed) within strata defined by the confounding variables. There exists a stratified version of the log rank test which can be used in this situation.

- This approach to investigating the impact of potential confounders becomes increasingly cumbersome as the number of confounders increases.
- The non-parametric methods covered in this lecture do not provide an easy way of *quantifying* differences in survival between groups.
- Non-parametric methods do not allow us to investigate the impact of continuous variables on survival (e.g. blood pressure). We could categorize the continuous variable and compare survivor curves within categories, but this is not usually a good idea.
- In the next lecture and beyond, we will learn about regression-based methods for investigating associations between explanatory variables and survival.

References

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958; 58: 457–481.

Greenwood M. The natural duration of cancer. Reports on Public Health and Medical Subjects (London: Her Majesty's Stationery Office) 1926; 33: 1–26.

Practical 2

Datasets required: `pbcbase_2021` and `whitehall`

R packages required for R users: survival, ggplot2, survminer.

Introduction

In this practical we will again use the Primary Biliary Cirrhosis (PBC) data and the Whitehall Study data, which are familiar to you from Practical 1. The practical is in two parts.

A You will use the PBC data and estimate survival probabilities, and compare survival curves between treatment groups, using the Kaplan-Meier method and log rank tests.

B You will use the Whitehall data to investigate the effect changing the origin and entry dates has on survival curves.

Aims

By the end of this practical you should:

- Understand survival tables, and be able to calculate the probabilities from data
- Be able to construct Kaplan-Meier plots in the software of your choice (Stata or R)
- Be able to interpret Kaplan-Meier plots
- Be able to perform and interpret a log rank test
- Understand the effect that changing the analysis timescale has on survival data

Where code examples are given or explanations are given that are specific to Stata or R, text and code relating to Stata is shown in this colour and text and code relating to R is shown in this colour.

Part A: Primary Biliary Cirrhosis data

Variable	Description
`id`	Unique identifier for each participant
`datein`	Date person entered the study
`dateout`	Date of the end of follow-up due to either death or censoring
`d`	Event indicator at the end of follow-up: 0=alive (censored), 1=dead
`time`	Follow-up time in years
`treat`	Treatment 1=placebo, 2=active

Open the PBC data and remind yourself of the variables. Today we will be using one variable we did not look at previously: `treat`.

1. How many events and censorings are there (overall and by treatment group), and what is the median event and censoring time (overall and by group)?

2. In this question we will create a Kaplan-Meier plot of overall survival.

   In Stata use `stset` to declare the data as survival data, using the date of entry into the study as the origin. There are two equivalent ways in which you could do this using the variables `d, time, datein` and `dateout`.

   Create a Kaplan-Meier plot of overall survival using
   `sts graph`

   Investigate what information is given when you type `sts list`.

   In R you can create Kaplan-Meier estimates of survival probabilities using `survfit`:

   ```
   pbc.km <- survfit(Surv(time,d)~1,data=pbc)
   ```

   You could alternatively use

   ```
   pbc.km <- survfit(Surv(as.numeric(dateout),d,origin=as.numeric(datein))~1,data=pbc)
   ```

   There are various ways of create a Kaplan-Meier plot using the results from the above `survfit` object. Here we will use the the 'ggplot2' and 'survminer' packages.

   ```
   ggsurvplot(pbc.km, data = pbc)
   ```

   Investigate what information is given when you run `summary(pbc.km)`.

   Discuss: What is the estimated probability of survival beyond 1 year? 5 years?

3. Below there is an incomplete table showing the calculations necessary to produce a Kaplan-Meier plot for the 26 active treatment patients who were suffering from cirrhosis at the start of the trial. By hand, complete the "Survival probability" column.

Time	At risk	Events	Censorings	Survival probability
0.104	26	1	0	0.9615
0.2628	25	1	0	0.9231
0.4572	24	1	0	0.8846
0.4846	23	1	0	0.8462
0.9172	22	0	1	0.8462
1.164	21	1	0	0.8059
1.369	20	1	0	0.7656
1.572	19	0	1	?
1.687	18	1	0	?
1.725	17	1	0	?
2.182	16	1	0	?
2.201	15	1	0	0.5954
2.634	14	0	1	?
2.667	13	1	0	?
3.047	12	0	1	0.5496
3.45	11	1	0	0.4997
.	.	.	.	
.	.	.	.	
8.89	2	0	1	0.1399
11.25	1	0	1	0.1399

4. Next we will describe the survival experience of the two treatment groups using the Kaplan-Meier method.

   In Stata, create a Kaplan-Meier plot of the estimated survival functions in the two treatment groups using the following command:

   ```
   sts graph, by(treat)
   ```

   Add confidence intervals to the plot using the `ci` option.

   You can see the underlying Kaplan-Meier table in Stata using `sts list, by(treat)`.

   In R, create a Kaplan-Meier plot of the estimated survival functions in the two treatment groups using:

   ```
   pbc.km <- survfit(Surv(time,d)~treat,data=pbc)
   ggsurvplot(pbc.km, data = pbc)
   ```

   Add confidence intervals to the plot using the `conf.int = T` option. You can see the underlying Kaplan-Meier table using `summary(pbc.km)`.

   Use your results to create the table from Question 3 and check your answers.

   Discuss: What do you conclude about the effect of the active treatment on survival?

5. We will use the log rank test to formally compare survival curves in the two treatment groups. What is the null hypothesis?

In Stata the log rank test is performed using:

```
sts test treat, logrank
```

In R the log rank test is performed using:

```
survdiff(Surv(time,d) treat,data=pbc)
```

Discuss: Interpret your results. What are your conclusions?

6. Lastly, we will look at the cumulative hazard functions in the two treatment groups. How does the cumulative hazard plot relate to the survival plot?

   In Stata the Nelson-Aalen estimates of the cumulative hazards are found using:

   ```
   sts graph, by(treat) cumhaz
   ```

   Whenever you produce survival graphs, it can be very useful to see how many people contribute to the curve at each point. To do this, add the option `risktable` to the above command.

   In R you can obtain the Nelson-Aalen estimate of the cumulative hazard as follows:

   ```
   pbc.km1 <- survfit(Surv(time,d)~1,data=subset(pbc,pbc$treat==1))
   pbc.km2 <- survfit(Surv(time,d)~1,data=subset(pbc,pbc$treat==2))
   cumhaz.1<-cumsum(pbc.km1$n.event/pbc.km1$n.risk)
   cumhaz.2<-cumsum(pbc.km2$n.event/pbc.km2$n.risk)
   plot(pbc.km1$time,cumhaz.1,type="s",col="red",xlab="Time",ylab="Cumulative hazard")
   lines(pbc.km2$time,cumhaz.2,type="s",col="black")
   ```

   Go back to the formula for the Nelson-Aalen estimate in the lecture notes and understand how this code follows from that. Note that the Kaplan-Meier estimate of the cumulative hazard is easy to obtain in R using

   ```
   plot(pbc.km,conf.int=F,col=c("red","black"),mark.time=F,
           xlab="Time", ylab="Survivor function",fun="cumhaz")
   ```

Part B: Whitehall Study

We now return to the Whitehall dataset. You should be able to use Stata code and R code from previous questions (and Practical 1) to answer these questions.

Variable	Description
`id`	Unique identifier for each participant
`timebth`	Date of birth
`timein`	Date person entered the study
`timeout`	Date of the end of follow-up due to either death or censoring
`chd`	Event indicator at the end of follow-up: 0=alive or died from other cause (censored), 1=death due to coronary heart disease

1. First we will investigate overall survival using time-in-study as the timescale. Examine the distribution of time to CHD mortality by looking at the Kaplan-Meier estimate of the overall survivor curve.

2. We will now investigate how changing the timescale changes the Kaplan-Meier plot. We will compare three different approaches:

   (a) Origin & start of period in which the participant is 'at risk' = Date a participant entered the study

   (b) Origin & start of period in which the participant is 'at risk' = Participant's date of birth

   (c) Origin = Date of birth; start of period in which the participant is 'at risk' = date a participant entered the study

   For all three approaches produce a Kaplan-Meier plot. Interpret the results. Use the 'risktable' option for `sts graph` in Stata to see the number of individuals at risk at selected time points. In R you can add the option `risk.table = T` into the `ggsurvplot` function.

   Discuss: when would the different time scales be appropriate?

3. Use the Kaplan-Meier approach to compare the survival experienced by civil servants who had different levels of SBP at entry into the study, using the variable `sbpgrp`. Use time-in-study as the timescale. Are the survival curves different?

4. Use a log rank test to compare survival across the blood pressure groups. What are the degrees of freedom for the test? Interpret the results.

   Discuss: Suppose your aim was to investigate the effect of systolic blood pressure on mortality. Do you think the above analysis provides an answer to this question?

Session 3

Parametric regression modelling

3.1 Aims of this lecture and practical

At the end of this lecture and practical you will be able to:

- Explain why regression modelling for survival data is useful.
- Define models using different parametric distributions for survival times with explanatory variables, including the exponential and Weibull distributions.
- Explain how these models differ.
- Formulate the likelihoods for survival data using parametric models, including the effects of explanatory variables, and allowing for censoring.
- Estimate the parameters of survival time distributions using maximum likelihood estimation for parametric models.
- Fit parametric models to survival data using Stata and R and interpret the output.

3.2 The purpose of parametric regression modelling for survival data

The methods for analysing survival data discussed in Lecture 2 are widely used and they provide a useful way of estimating survival probabilities and making comparisons between two or more groups of individuals. However, the methods of Lecture 2 also have some limitations. The major limitations are:

- The methods do not quantify the association between exposures and survival.
- The methods do not accommodate continuous exposures, unless we categorise the exposure.
- If we wish to adjust for potential confounders in our analyses, then we have to look separately at groups defined by the confounder and the methods quickly become cumbersome and the groups too small for meaningful analysis.
- In general, analyses involving several exposure variables, which may be binary continuous or categorical, are not possible using the methods of Lecture 2.

In this lecture we introduce a more advanced approach to analysing survival data using regression modelling. In regression modelling for survival data we assume a model for the survival times which includes how survival times depend on explanatory variables. Regression modelling for survival data is analogous to the use of linear regression to

study the dependence of continuous response variables on explanatory variables, and of logistic regression to study the dependence of a binary outcome on explanatory variables, and so on.

The models we use in this lecture are described as *parametric*. This is because our regression models assume a particular shape for the distribution of the survival times which depends on defined parameters. By contrast, the methods introduced in Lecture 2 are described as *non-parametric*.

3.3 Reminder of the likelihood for survival data

In this lecture we assume that a population of n individuals $i = 1, \ldots, n$ have been followed up over time for some outcome of event of interest, and their survival or censoring times recorded. The survival or censoring time for individual i is denoted t_i, and the indicator δ_i tells us whether this time indicates an event ($\delta_i = 1$) or a censoring time ($\delta_i = 0$). In Lecture 1 we found that the likelihood for survival data can be written as

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \tag{3.1}$$

where $f(t)$ is the probability density function and $S(t)$ is the survivor function at time t.

3.4 Incorporating explanatory variables

In the non-parametric setting we compared survivor curves in two groups of individuals, e.g. treatment and control groups. We can test for a difference between groups using the log rank test. However, this does not quantify the effect of the explanatory variable on survival.

For simplicity we start by considering a single binary explanatory variable X taking value 0 or 1. This is observed on each individual at the start of follow-up. Alongside a failure or censoring time for each individual we observe the explanatory variable, giving observed values $x_1, \ldots, x_n$.

Examples

- In a randomized trial setting X it may refer to treatment group.
- In an observational study of an occupational cohort, X may refer to occupational exposure to radiation. In a population-based cohort X may refer to smoking status [smoker or non-smoker].

In this lecture we focus on one particular way in which the effect of explanatory variables on survival can be expressed. For a binary exposure we suppose that the hazard in one group of individuals ($X = 1$, say) is a multiple of the hazard in the 'baseline' group ($X = 0$). We let $h_0(t)$ denote the hazard function in the $X = 0$ group and $h_1(t)$

denote the hazard function in the $X = 1$ group. Formally, we can write

$$h_1(t) = \psi h_0(t) \tag{3.2}$$

where ψ is a parameter to be estimated. Note that here the range of values that ψ could take must be restricted because the hazard cannot be negative. For this reason it is convenient instead to write

$$h_1(t) = e^{\beta} h_0(t) \tag{3.3}$$

Using this formulation the parameter β can take any value. A model of the form in (3.3) is referred to as a proportional hazards model, for fairly obvious reasons. The ratio of the hazards in the two groups is

$$\frac{h_1(t)}{h_0(t)} = e^{\beta} \tag{3.4}$$

and e^{β} is referred to as the *hazard ratio*. We will often work on the log scale, and β is referred to as the *log hazard ratio*. Note that this ratio does not depend on time (t). We are making an assumption here that the ratio of the hazards for the two groups is the same no matter how much time has passed, known as the *proportional hazards assumption*. This s a strong assumption and later we will describe methods for assessing whether it is valid in a particular data set. The proportional hazards assumption will not be reasonable (typically) if we believe that the effect of a treatment changes over time, for example.

For the rest of this lecture we focus on models for the survival times which are such that the ratio of the hazards in the two groups is constant over time – these are referred to as *proportional hazards models*. There are alternative assumptions one could make about how the explanatory variable is associated with the hazard, and we will discuss some of these in a later lecture.

3.5 The exponential model

We now apply the above in practice and we begin by considering the simplest model for survival times - the exponential model. This was introduced in Lecture 1. As a reminder, the exponential distribution has the following properties:

$$\begin{aligned}
\text{Hazard function: } h(t) &= \lim_{\delta \to 0} \frac{1}{\delta} \Pr(t \leq T < t + \delta | T \geq t) = \lambda \\
\text{Survivor function: } S(t) &= \Pr(T > t) = \exp(-\lambda t) \\
\text{Probability density function: } f(t) &= \lim_{\delta \to 0} \frac{1}{\delta} \Pr(t \leq T < t + \delta) = \lambda \exp(-\lambda t)
\end{aligned} \tag{3.5}$$

As noted in Lecture 1, under the exponential distribution the hazard function is constant over time. This means that the rate at which events occur is constant over the time scale.

To incorporate a binary explanatory variable X we write

$$\begin{aligned} h(t|X=0) &= \lambda \\ h(t|X=1) &= \lambda e^{\beta} \end{aligned} \tag{3.6}$$

It is even more convenient to write

$$h(t|x) = \lambda e^{\beta x} \tag{3.7}$$

It follows that the survivor function and probability density function are respectively

$$S(t|x) = \exp\left(-\lambda t e^{\beta x}\right) \tag{3.8}$$

$$f(t|x) = \lambda e^{\beta x} \exp\left(-\lambda t e^{\beta x}\right) \tag{3.9}$$

The likelihood for the data is therefore

$$L = \prod_{i=1}^{n} \left\{\lambda e^{\beta x_i} \exp\left(-\lambda t_i e^{\beta x_i}\right)\right\}^{\delta_i} \left\{\exp\left(-\lambda t_i e^{\beta x_i}\right)\right\}^{1-\delta_i} \tag{3.10}$$

where t_i is the survival or censoring time for individual i, δ_i is the indicator of whether the person had the outcome or was censored, and x_i is their exposure. Maximum likelihood estimates for λ and β can be found in the usual way by differentiating the log likelihood with respect to both parameters. We can estimate standard errors for the parameters λ and β using the inverse of the information matrix. These can be used to obtain 95% confidence intervals.

It is of interest to perform a test of the null hypothesis that the hazard ratio is 1, i.e. there is no difference between the hazard rates in the two groups. The null hypothesis is $e^{\beta} = 1$, or equivalently $\beta = 0$. The Wald test statistic, for example, is $\hat{\beta}/SE(\beta)$, which is normally distributed $N(0,1)$ under the null hypothesis. We could also perform a likelihood ratio test.

Exercise 3.1

Find the equations for the maximum likelihood estimates for λ and β using on the likelihood based on the exponential distribution in (3.10).

3.6 Fitting exponential models in Stata and R

In Stata parametric survival models can be fitted using the `streg` command, with the `distribution(exponential)` option. In R there are different functions for fitting parametric survival models, including the `phreg` or `weibreg` function in the 'eha' package (using options `dist="exponential",shape=1`), the `survreg` function in the 'survival' package, and the `flexsurvreg` function in the 'flexsurv' package, using the `dist="exponential"` option. The output shown from using different methods differs slightly. Also, the `survreg` function in R uses a different parameterization of the exponential model than we have used in the formulae for the exponential distribution in 3.7,

3.8, and 3.9. Table 3.1 summarises of output shown when fitting an exponential model in Stata (using `streg` with `nohr` option) and in R (using `phreg`, `flexsurvreg` or using `survreg`). We show some examples below. In R we will primarily use `weibreg`, as this gives output that is closest to that shown in Stata, and closest to the parameterization used in the lecture notes. Note that `phreg` and `weibreg` are the same for most purposes, but `weibreg` allows delayed entry/left truncation, whereas `phreg` does not. A drawback of `weibreg` and `phreg` is that they do not show confidence intervals. In the computer practical we provide a function that enables calculation of 95% confidence intervals.

Table 3.1: Summary of output shown when fitting an exponential model in Stata (using `streg` with `nohr` option) and in R (`weibreg` with `shape=1`, `flexsurvreg` or `survreg`). The parameters refer to a hazard model of the form $h(t|x) = \lambda e^{\beta x}$ (equation 3.7).

Parameter	Stata `streg`	R `weibreg`	R `flexsurvreg`	R `survreg`
$\log \lambda$	`_cons`	-	-	-
λ	-	-	`rate`	-
$-\log \lambda$	-	`log(scale)`	-	`(Intercept)`
β	name of variable x	name of variable x	name of variable x	-
$-\beta$	-	-	-	name of variable x

Example 3.1

Leukemia example: fitting an exponential model

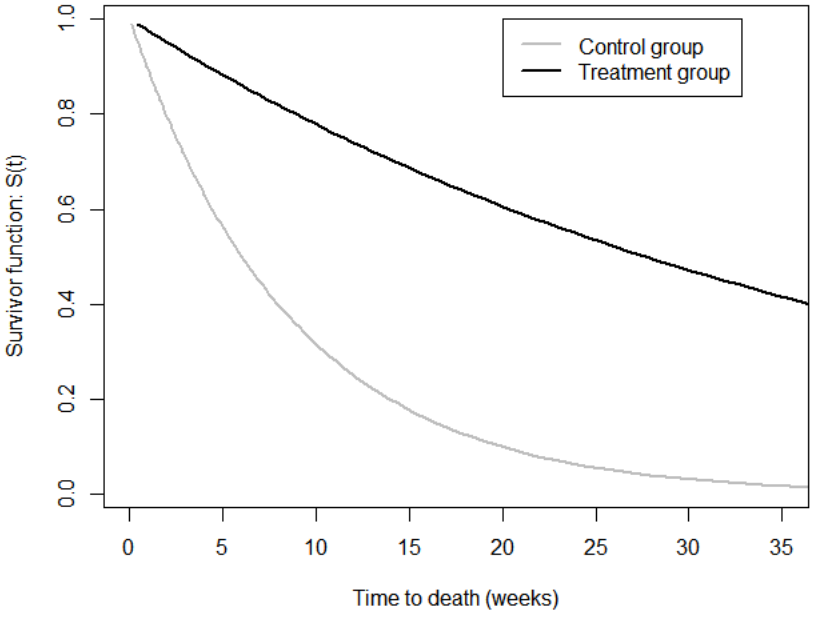
We fit an exponential model to the data on times to death in leukemia patients in two groups: control ($X = 0$) and treatment ($X = 1$). This example was introduced in the previous lectures. The results are shown below in Table 3.2.

Table 3.2: Results from fitting an exponential model to the leukemia data.

parameter	Estimate	Standard error	95% CI	p-value
λ	0.12	0.03	(0.08, 0.18)	< 0.001
β	-1.53	0.40	(-2.31, -0.75)	< 0.001
$\exp(\beta)$	0.22	0.09	(0.10,0.48)	< 0.001

The estimated hazard ratio is 0.22, meaning that the hazard rate for death is lower by 78% in the treatment group compared to the control group. This hazard ratio is highly statistically significantly different from 1 and the results provide strong evidence that the treatment has a beneficial effect on survival. We can also plot the estimated survivor curves in the two groups, which are shown in Figure 3.1. Later in this lecture we will see how to investigate whether the exponential model is a suitable model for this data.

Figure 3.1: Leukemia patient data: estimated survivor curves under an exponential model.



In Stata:

```
. streg group, distribution(exponential) nohr

        failure _d:  death
  analysis time _t:  time

Exponential PH regression

No. of subjects =           42                 Number of obs    =          42
No. of failures =           30
Time at risk    =          541
                                               LR chi2(1)       =       16.49
Log likelihood  =    -49.00866                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |  -1.526614   .3984095    -3.83   0.000    -2.307482   -.7457452
       _cons |  -2.159484   .2182179    -9.90   0.000    -2.587183   -1.731785
------------------------------------------------------------------------------

stcurve, survival at1(group=0) at2(group=1)
```

In R:

```
> leukaemia.exp<- weibreg(Surv(time=time,event=death)~as.factor(group),shape=1,
                                data=leukaemia)
> leukaemia.exp
Call:
weibreg(formula = Surv(time = time, event = death) ~ as.factor(group),
    data = leukaemia, shape = 1)

Covariate           Mean       Coef Exp(Coef)  se(Coef)    Wald p
as.factor(group)
               0    0.336     0         1           (reference)
               1    0.664    -1.527     0.217     0.398     0.000

log(scale)                    2.159     8.667     0.218     0.000

 Shape is fixed at  1

Events
Total time at risk           541
Max. log. likelihood      -108.52
LR test statistic         16.5
Degrees of freedom        1
Overall p-value           4.90309e-05
```

3.7 The Weibull distribution

In many applications it will not be reasonable to assume a constant hazard rate over time, as under the exponential distribution. An alternative distribution to consider is the Weibull distribution, which we first encountered in Lecture 1. As a reminder, the Weibull distribution has the following properties:

$$\begin{aligned} \text{Hazard function: } & h(t) = \kappa\lambda t^{\kappa-1} \\ \text{Survivor function: } & S(t) = \exp(-\lambda t^{\kappa}) \\ \text{Probability density function: } & f(t) = \kappa\lambda t^{\kappa-1}\exp(-\lambda t^{\kappa}) \end{aligned} \tag{3.11}$$

Recall that the exponential model is a special case of the Weibull with $\kappa = 1$.

To incorporate a binary explanatory variable X, assuming proportional hazards, we write

$$h(t|x) = \kappa\lambda t^{\kappa-1}e^{\beta x} \tag{3.12}$$

It follows that the survivor function is

$$S(t|x) = \exp\left(-\lambda t^{\kappa}e^{\beta x}\right) \tag{3.13}$$

The likelihood is therefore

$$L = \prod_{i=1}^{n}\left\{\kappa\lambda t^{\kappa-1}e^{\beta x_i}\exp\left(-\lambda t_i^{\kappa}e^{\beta x_i}\right)\right\}^{\delta_i}\left\{\exp\left(-\lambda t_i^{\kappa}e^{\beta x_i}\right)\right\}^{1-\delta_i} \tag{3.14}$$

3.8 Fitting Weibull models in Stata and R

Weibull models can also be fitted in Stata using the `streg` command with the `distribution(weibull)` option, and in R using `phreg` or `weibreg`, `survreg`, or `flexsurvreg`, all using the `dist="weibull"` option. There are several different ways of parameterizing the Weibull model, which you will see in different text books. The parameterization used for the Weibull model in Stata is that same as used in the lecture notes - equation 3.12 for the hazard. In R, the different functions (`phreg` and `weibreg`, `survreg`, `flexsurvreg`) use different parametrizations of the Weibull model. This means that the output shown from fitting Weibull models in Stata and R differs. Table 3.3 summarises of output shown when fitting a Weibull model in Stata (using `streg` with `nohr` option) and in R (`weibreg`, `flexsurvreg`, `survreg`), and we show some examples below.

Table 3.3: Summary of output shown when fitting a Weibull model in Stata (using `streg` with `nohr` option) and in R (using `weibreg`, `flexsurvreg` or `survreg`). The parameters refer to a hazard model of the form $h(t|x) = \kappa\lambda t^{\kappa-1}e^{\beta x}$ (equation 3.12).

Parameter	Stata `streg`	R `weibreg`	R `flexsurvreg`	R `survreg`
$\log\lambda$	`_cons`	-	-	-
κ	`p`	-	`shape`	-
$\log\kappa$	`ln_p`	`log(shape)`	-	-
$-\log\kappa$	-	-	-	`Log(scale)`
$\lambda^{-1/\kappa}$	-	-	`scale`	-
$\log(\lambda^{-1/\kappa})$	-	`log(scale)`	-	`(Intercept)`
β	name of variable x	name of variable x	-	-
$-\beta/\kappa$	-	-	name of variable x	name of variable x

Example 3.2

Leukemia example: fitting a Weibull model

The results from fitting a Weibull model to this data, rather than an exponential model are as in Table 3.4.

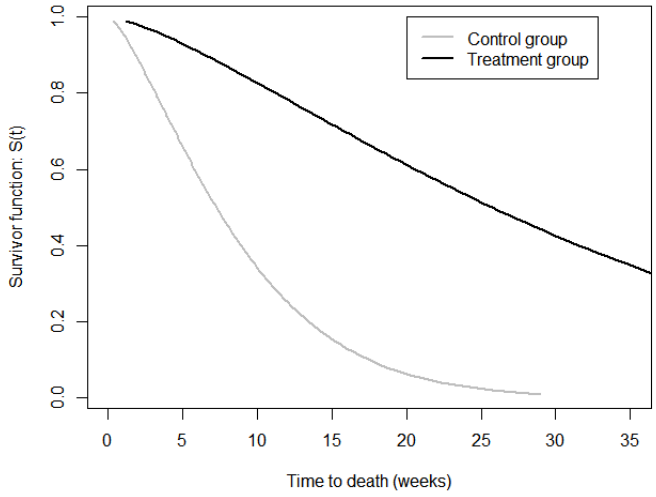
Table 3.4: Results from fitting a Weibull model to the leukemia data.

parameter	Estimate	Standard error	95% CI	p-value
$\log\lambda$	-3.071	0.558	(-4.165,-1.977)	< 0.001
$\log\kappa$	0.312	0.147	(0.023,0.600)	0.034
β	-1.731	0.413	(-2.54, -0.92)	< 0.001
$\exp(\beta)$	0.18	0.07	(0.08,0.40)	< 0.001

Fitting the Weibull model gives a hazard ratio estimate of 0.18. This compares with the estimated hazard ratio from the exponential model of 0.22. The estimates are

similar. Again we can plot the estimated survivor curves in the two patient groups - see Figure 3.2.

Figure 3.2: Leukemia patient data: estimated survivor curves under a Weibull model.



In Stata:

```
    . streg group, distribution(weibull) nohr

         failure _d:  death
   analysis time _t:  time

Weibull PH regression

No. of subjects =           42                 Number of obs    =          42
No. of failures =           30
Time at risk    =          541
                                               LR chi2(1)       =       19.65
Log likelihood  =   -47.064102                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |  -1.730872   .4130819    -4.19   0.000    -2.540497   -.9212463
       _cons |  -3.070704   .5580701    -5.50   0.000    -4.164501   -1.976907
-------------+----------------------------------------------------------------
       /ln_p |   .3117092   .1472919     2.12   0.034     .0230224     .600396
-------------+----------------------------------------------------------------
           p |   1.365757    .201165                       1.02329     1.82284
         1/p |   .7321944   .1078463                      .5485944    .9772406
------------------------------------------------------------------------------
```

In R:

```
> leukaemia.weib<- weibreg(Surv(time=time,event=death)~as.factor(group),
                           data=leukaemia)
> leukaemia.weib
Call:
weibreg(formula = Surv(time = time, event = death) ~ as.factor(group),
    data = leukaemia)

Covariate           Mean       Coef Exp(Coef)  se(Coef)    Wald p
as.factor(group)
              0    0.336     0         1           (reference)
              1    0.664    -1.731     0.177     0.413     0.000

log(scale)                   2.248     9.472     0.166     0.000
log(shape)                   0.312     1.366     0.147     0.034

Events
Total time at risk          541
```

```
Max. log. likelihood      -106.58
LR test statistic         19.6
Degrees of freedom        1
Overall p-value           9.29141e-06
```

Exercise 3.2

An exponential model and a Weibull model were fitted to the breast cancer example data. The results from fitting models in Stata are shown below, where `im` is the indicator of IM positive status (taking value 0 if IM negative and 1 if IM positive). Interpret the results.

In Stata:

```
. streg im, distribution(exponential) nohr

       failure _d:  death
analysis time _t:  time

Exponential PH regression

No. of subjects =           45                 Number of obs    =          45
No. of failures =           24
Time at risk    =         4425
                                               LR chi2(1)       =        5.67
Log likelihood  =   -52.901164                 Prob > chi2      =      0.0172

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          im |   1.115589   .4330127     2.58   0.010      .2668999    1.964278
       _cons |  -6.586283   .6123724   -10.76   0.000    -7.786511   -5.386055
------------------------------------------------------------------------------

. streg im, distribution(weibull) nohr

         failure _d:  death
   analysis time _t:  time

Weibull PH regression

No. of subjects =           45                 Number of obs    =          45
No. of failures =           24
Time at risk    =         4425
                                               LR chi2(1)       =        6.33
```

```
Log likelihood  =   -52.309099                     Prob > chi2      =      0.0119

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          im |   1.200892   .4405352     2.73   0.006      .3374589    2.064325
       _cons |  -7.757733   1.296752    -5.98   0.000     -10.29932   -5.216146
-------------+----------------------------------------------------------------
       /ln_p |   .2034871   .1792612     1.14   0.256    -.1478585    .5548326
-------------+----------------------------------------------------------------
           p |   1.225669    .219715                      .8625532    1.741649
         1/p |   .8158807   .1462558                      .5741683    1.159349
------------------------------------------------------------------------------
```

In R:

```
> breastcancer.exp<- weibreg(Surv(time=time,event=death)~as.factor(im),
                                  shape=1,data=breastcancer)
> breastcancer.exp
Call:
weibreg(formula = Surv(time = time, event = death) ~ as.factor(im),
    data = breastcancer, shape = 1)

Covariate           Mean       Coef Exp(Coef)  se(Coef)    Wald p
as.factor(im)
               1    0.859     0         1           (reference)
               2    0.141     1.116     3.051     0.433     0.010

log(scale)                    5.471   237.625     0.250     0.000

 Shape is fixed at  1

> breastcancer.weib<- weibreg(Surv(time=time,event=death)~as.factor(im),
                                  data=breastcancer)
> breastcancer.weib
Call:
weibreg(formula = Surv(time = time, event = death) ~ as.factor(im),
    data = breastcancer)

Covariate           Mean       Coef Exp(Coef)  se(Coef)    Wald p
as.factor(im)
               1    0.859     0         1           (reference)
               2    0.141     1.201     3.323     0.441     0.006

log(scale)                    5.350   210.524     0.224     0.000
log(shape)                    0.203     1.226     0.179     0.256
```

3.9 Comparing the fit of Weibull and exponential models

An important question is of how we can choose a good (i.e. well fitting) parametric model for our data. In this section we outline two ways of assessing whether an exponential or Weibull model is appropriate:

- Using plots
- Using statistical tests

Using plots

Plots obtained using non-parametric methods can be used to informally assess whether a particular parametric model may be appropriate for the data. We focus on a binary exposure X. Under a Weibull distribution the cumulative hazard is

$$H(t|x) = -\log S(t|x) = \lambda t^{\kappa} e^{\beta x} \tag{3.15}$$

Taking logs gives

$$\log H(t|x) = \log\{-\log S(t|x)\} = \log\lambda + \kappa \log t + \beta x \tag{3.16}$$

The log cumulative hazard is linear in $\log t$ and the difference in $\log H(t|x)$ between the two exposure groups is β. Therefore, if the Weibull model is valid then in a plot $\log H(t|x)$ against $\log t$ in the two exposure groups we should see two straight and parallel lines. We can do this by obtaining a non-parametric estimate of $\log H(t|x)$, e.g. a Kaplan-Meier estimate.

Exercise 3.3

What would expect a a plot of $\log H(t|x)$ against $\log t$ in the two exposure groups to look like if an exponential model is valid?

Using statistical tests

The exponential distribution is a special case of the Weibull with $\kappa = 1$. Having fitted a Weibull model, a test of the null hypothesis that the hazard is constant over time is a test of the null hypothesis that $\kappa = 1$, or equivalently $\log\kappa = 0$. This test can be performed using the estimate of $\log\kappa = 0$ and its standard error. The Wald test statistic is $\frac{\log\hat{\kappa}}{SE(\hat{\kappa})}$ which is normally distributed under the null hypothesis.

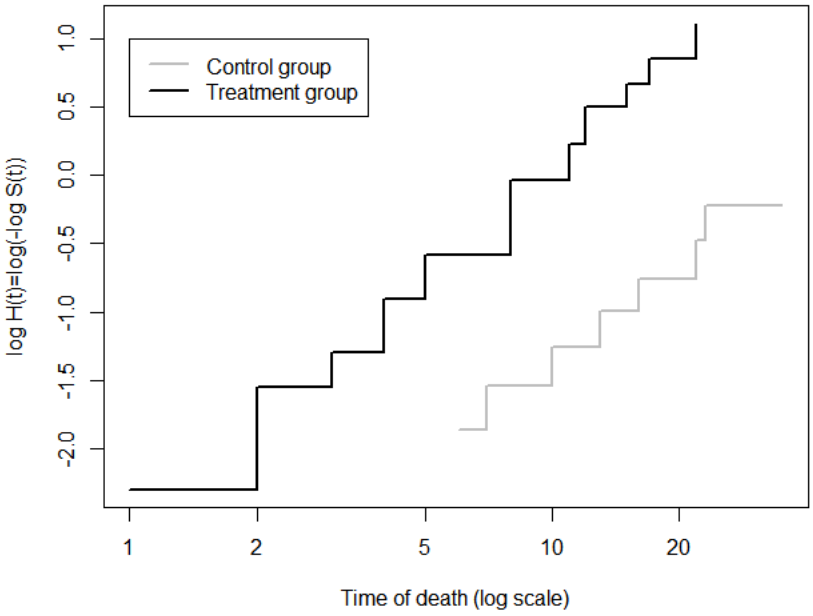
An alternative is to compare the likelihoods obtained from fitting the exponential model and the Weibull model. Because the exponential model is nested within the Weibull model, a likelihood ratio test is appropriate. For this the test statistic is minus twice the difference between the log likelihood for the exponential model and the likelihood for the weibull model. This test statistic has a χ^2_1 distribution under the null hypothesis.

Example 3.3

Investigating the exponential and Weibull models for the leukemia patient data.

We apply the methods of comparisons described in Section 3.9 to the Leukemia patient data. Figure 3.3 shows plots of Kaplan-Meier estimates of $\log H(t|x) = \log\{-\log S(t|x)\}$ against $\log t$. The lines are straight and they are parallel. This implies that the Weibull model was appropriate for these data.

Figure 3.3: Leukemia example: Plot of a Kaplan-Meier estimate of $\log H(t|x) = \log\{-\log S(t|x)\}$ against $\log t$ in treatment and controls groups.



Looking at the output from fitting the Weibull model to these data (given above), we see that the estimate of $\log\kappa$ is 0.312, with p-value 0.034. The log likelihoods for the exponential and weibull models are respectively -49.01 and -47.06, giving likelihood ratio test statistic $-2(-49.01 + 47.06) = 3.89$. Comparing this with χ^2_1, the p-value is 0.0486. There is some evidence against the null hypothesis that $\log\kappa = 0$, i.e. evidence that the Weibull model provides a better fit than the exponential model

In Stata:

```
. quietly: streg group, distribution(exponential) nohr
. est store a
. quietly: streg group, distribution(weibull) nohr
. est store b
. lrtest b a, force

Likelihood-ratio test                                 LR chi2(1)  =      3.89
(Assumption: a nested in b)                           Prob > chi2 =    0.0486
```

In R:

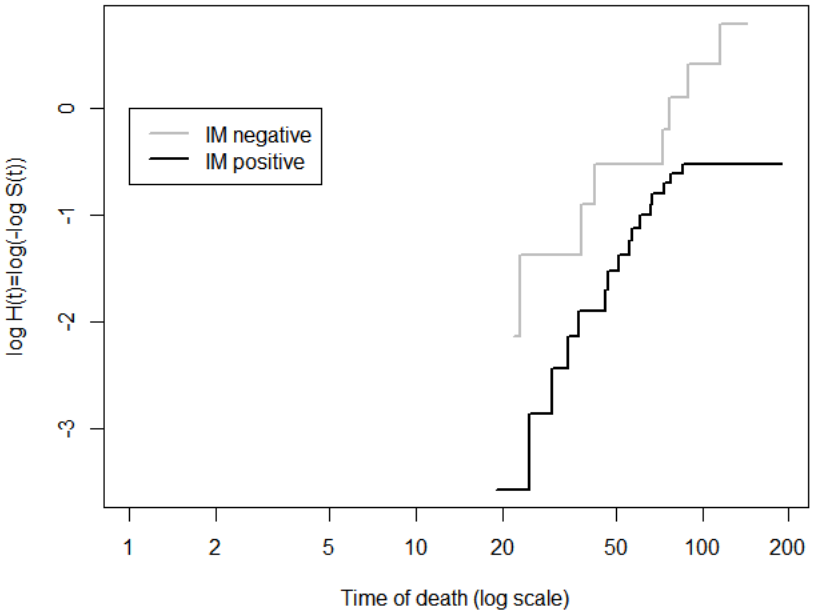
```
> leukaemia.exp<- weibreg(Surv(time=time,event=death)~as.factor(group),
                                  shape=1,data=leukaemia)
> leukaemia.weib<- weibreg(Surv(time=time,event=death)~as.factor(group),
                                  data=leukaemia)
> teststat=-2*(leukaemia.exp$loglik[2]-leukaemia.weib$loglik[2]) #test statistic
> teststat
[1] 3.889116
> 1-pchisq(teststat,df=1) #p-value from chi-squared test with 1-df
[1] 0.0486
```

Exercise 3.4

Breast cancer example

Figure 3.4 shows plots of Kaplan-Meier estimates of $\log H(t|x) = \log\{-\log S(t|x)\}$ against $\log t$. Use the plot and the output from fitting exponential and Weibull models to the breast cancer data (given above) to assess the fit of the Exponential and Weibull models.

Figure 3.4: Breast cancer example: Plot of a Kaplan-Meier estimate of $\log H(t|x) = \log\{-\log S(t|x)\}$ against $\log t$ in treatment and controls groups.



3.10 Extending beyond binary explanatory variables

We have so far focused on a single binary exposure variable of interest, $X = 0, 1$. The methods can be extended to continuous explanatory variables, categorical explanatory variables, and multiple explanatory variables, where X is a vector. For a vector of

explanatory variables $X = (X_1, X_2, \ldots, X_p)^\top$ the proportional hazards assumption is that

$$h(t|x) = h_0(t)e^{\beta x} \tag{3.17}$$

where $h_0(t)$ is the baseline hazard and $\beta = (\beta_1, \beta_2, \beta_p)^\top$ is a vector of parameters to be estimated. We outline the interpretation of β in different circumstances.

Continuous explanatory variable X

First consider a single continuous variable. The effect of an increase of 1 unit in the continuous variable X is to multiply the hazard by e^β. So the ratio of the hazards for a person with $X = 1$ and a person with $X = 0$ is

$$\frac{h(t|X=1)}{h(t|X=0)} = \frac{h_0(t)e^\beta}{h_0(t)} = e^\beta \tag{3.18}$$

As another example, the ratio of the hazards for a person with $X = 73$ and a person with $X = 72$ is

$$\frac{h(t|X=73)}{h(t|X=72)} = \frac{h_0(t)e^{73\beta}}{h_0(t)e^{72\beta}} = e^\beta \tag{3.19}$$

For a continuous variable X, β is the log hazard ratio associated with a 1 unit increase in X.

Categorical explanatory variable X with more than 2 categories

Next consider a single categorical variable. For a categorical variable with $K + 1$ categories, we define a series of indicator variables

$$X_k = \begin{cases} 1 & \text{if in category } k \\ 0 & \text{otherwise} \end{cases} \tag{3.20}$$

The hazard in category k is

$$h(t|X_k = 1) = h_0(t)e^{\beta_k}, \qquad k = 0, \ldots, K \tag{3.21}$$

The baseline hazard refers to an individual in the baseline category (assumed to be $X_0 = 1$) and it is assumed that $\beta_0 = 0$. So e^{β_k} is the hazard ratio which compares individuals in category k with individuals in category 0. This could be written equivalently as $h(t|x_1, \ldots, x_K) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K)$.

More than one explanatory variable

In general we let X denote a vector of explanatory variables, which may contain binary, categorical and continuous variables, and we let β denote a vector of corresponding log hazard ratio estimates. The interpretation of a particular element of β is as the log hazard ratio for a particular variable, holding all other elements of X fixed. That is, β_p, say, is the log hazard ratio for a unit increase in X_p conditional on all of the other variables in X.

3.11 Assessing different choices for the survival model when there are several explanatory variables

In Section 3.5 we outlined graphical ways of investigating whether an exponential or Weibull model may be suitable for our data. This was illustrated for the situation of a single binary exposure. The graphical procedure can be extended to more than one explanatory variable. Again we consider just two explanatory variables X_1 and X_2 and suppose now that these are both binary. There are therefore 4 combinations of values for X_1 and X_2. Plots such as those in Figure 3.3 could be made for all four combinations.

However, as the number of explanatory variables increases this graphical procedure becomes increasingly difficult: there will be many lines to assess and the groups of individuals may become too small to get a good non-parametric estimate of the survivor function. Furthermore, continuous variables are not accommodated, unless we categorize them, which we typically wish to avoid.

In Lecture 5 we will learn about special residuals which can be used to check model assumptions in survival analysis.

3.12 More on proportional hazards models

We continue to focus on a binary explanatory variable for now. At the start of this lecture we saw that the proportional hazards assumption can be written as $h_1(t) = h_0(t)e^{\beta}$ where the subgroups 0 and 1 refer to two groups of individuals in our data. Under the exponential distribution the hazards in the two groups can be written as

$$h_0(t) = \lambda, \qquad h_1(t) = \lambda e^{\beta} \tag{3.22}$$

We could rewrite this as

$$h_0(t) = \lambda_0, \qquad h_1(t) = \lambda_1 \tag{3.23}$$

We can see that the hazard function is constant over time in both exposure groups. What this means is that the survival distribution is the same in the two groups (i.e. exponential), but with a different parameter (to denote that we allow the constant hazard to be different in the two groups).

Let's do the same thing for the Weibull distribution. With a binary exposure under the proportional hazards assumption, the hazards in the two groups are

$$h_0(t) = \kappa\lambda t^{\kappa-1}, \qquad h_1(t) = \kappa\lambda t^{\kappa-1}e^{\beta} \tag{3.24}$$

We could rewrite this as

$$h_0(t) = \kappa\lambda_0 t^{\kappa-1}, \qquad h_1(t) = \kappa\lambda_1 t^{\kappa-1} \tag{3.25}$$

where $\lambda_0 = \lambda$ and $\lambda_1 = \lambda e^{\beta}$. The survival distribution is of the Weibull form in both groups.

So, under the exponential and Weibull models and when we assume that the effect of the exposure is to act proportionally on the hazard, the survival distribution is from the same model in the two exposure groups, only with different parameters. Because the exponential and Weibull distributions have the above property, they can be referred to as proportional hazards models, or as belonging to the proportional hazards family of models. Not all distributions for survival times have this property.

Recall from lecture 1 that another distribution for survival time is the log-logistic distribution, which has the following properties:

$$\begin{aligned} \text{Hazard function: } h(t) &= \frac{\kappa\lambda(t\lambda)^{\kappa-1}}{1+(t\lambda)^{\kappa}} \\ \text{Survivor function: } S(t) &= \frac{1}{1+(t\lambda)^{\kappa}} \\ \text{Probability density function: } f(t) &= \frac{\kappa\lambda(t\lambda)^{\kappa-1}}{(1+(t\lambda)^{\kappa})^2} \end{aligned} \tag{3.26}$$

Now suppose that we wish to assume proportional hazards in two exposure groups:

$$h_0(t) = \frac{\kappa\lambda(t\lambda)^{\kappa-1}}{1+(t\lambda)^{\kappa}}, \qquad h_1(t) = \frac{\kappa\lambda(t\lambda)^{\kappa-1}}{1+(t\lambda)^{\kappa}}e^{\beta} \tag{3.27}$$

It is clear that the distribution of the survival times cannot be written as being of log-logistic form in the two groups. The log-logistic model is therefore not a proportional hazards model. In fact, the log-logistic model is in a class of models referred to as accelerated failure time models. We return to models of this type in a later session.

Practical 3

Datasets required: `whitehall`

R packages required for R users: survival, eha, flexsurv, ggplot2, survminer.

Introduction

In this practical we will use the Whitehall data, which is familiar from the earlier practicals. We will investigate the association between job grade (`grade`) and risk of coronary heart disease (`chd`), with and without adjustment for age. We will use the time-in-study timescale except for in one question.

Variable	Description
`id`	Unique identifier for each participant
`timebth`	Date of birth
`timein`	Date person entered the study
`timeout`	Date of the end of follow-up due to either death or censoring
`chd`	Event indicator at the end of follow-up: 0=alive or died from other cause (censored), 1=death due to coronary heart disease
`grade`	Job grade at study entry. 1=admin & professional/executive; 2=clerical & other
`agein`	Age in years at study entry

Aims

By the end of this practical you should be able to

- Be able to fit exponential and Weibull distribution models to survival data (in Stata and/or R) and interpret the results
- Be able to check the constant hazard assumption of the exponential model
- Understand the effect of changing the analysis timescale on estimates from the exponential model

Where code examples are given or explanations are given that are specific to Stata or R, text and code relating to Stata is shown in this colour and text and code relating to R is shown in this colour.

Questions

1. Load the data and explore the grade variable. Summarize the numbers and timings of CHD deaths and censorings by job grade.
   In Stata use `stset` to reflect that we wish to use time-in-study as the time scale.
   In R format the dates as in previous practicals. How should `Surv()` be specified to use time-in-study as the time scale?
2. We begin by using simple methods to investigate the association between job grade and CHD.

(a) Use a Kaplan-Meier plot to compare survival in the two groups, including the 95% confidence intervals. Interpret the plots.

(b) How many individuals survived to 5, 10, 15 years of follow-up in each job grade category?
In Stata you may wish to consult the help file for `sts list`.
In R you may wish to consult the help file for `summary.survfit`.

(c) Use the log rank test to compare the estimated survivor curves in the two job grades.

3. We will now fit an exponential model to the Whitehall data using job grade as an explanatory variable.

(a) Write down the hazard and survivor functions and hence the likelihood.

(b) Fit the exponential model and interpret the parameter estimates. What is the association between job grade and survival?
As you saw briefly in Practical 1, in Stata parametric survival models can be fitted using the `streg` command.
In R try out `weibreg` with the `shape=1` option to fit the exponential model (in the 'eha' package) - see the lecture notes for some examples of this. Note that this does not automatically give confidence intervals - in the R script we have provided a function that allows for calculation of 95% confidence intervals. As you saw briefly in Practical 1, in R parametric survival models can also be fitted using `survreg` (in the 'survival' package) or `flexsurvreg` (in the 'flexsurv' package) - see the example R script.

(c) Change to the age time scale (accounting for delayed entry into the study) and refit the exponential model. Compare your results with those found when using time-in-study as the timescale.
In Stata you will need to use stset again to change the time scale. In R you will need to change the specification of `Surv()` to change the time scale

4. Revert to the time-in-study timescale for this question and all subsequent questions. By fitting an exponential distribution we are assuming that the hazard rate does not change over time. Because this may not be a reasonable assumption we investigate fitting a Weibull model. Fit the Weibull model. Interpret the parameters of the model. Compare your results with those from the exponential model.
In Stata Weibull models can be fitted using `streg` with the `dist(weibull)` option.
In R we will use `weibreg` to fit the Weibull model.

Tables 3.1 and 3.3 in the notes provide some information on what is shown in the output from fitting these models in Stata and R.

5. Create a suitable non-parametric plot to investigate whether you expect the Weibull model fitted above to be appropriate.

In Stata you can create these plots using `sts graph`. You may wish to use the `yscale(log)` and `xscale(log)` options to `sts graph`.
In R you can create the plots using

```
ggsurvplot(whl.km, data = whl,conf.int = T,fun="cloglog")
```

where `whl.km` is the `survfit` object used to obtain Kaplan-Meier curves.

Discuss: Does the Weibull model provide a good fit?

6. The age at which individuals entered the study may have an important part to play in the analysis. So we will add an age variable to the Weibull model.
   (a) Include `agein` as an additional explanatory variable in the Weibull model fitted for job grade in Question 4.
   (b) Interpret the hazard ratios for job grade and for age.

   Discuss: What effect does adjusting for age at entry to the study have on the hazard ratio for job grade? Can you explain why this might happen?

7. Create non-parametric plots to investigate whether you expect the Weibull model fitted in Question 6 to be appropriate for this data. Age is a continuous variable so we could categorize the age variable for use in making (approximate) assessments of whether the Weibull model is appropriate. We recommend using the age categories: 40-49, 50-54, 55-59,..., 65-69. Example code for creating these plots is provided in the example Stata do file and R script file.
8. Referring to the model fitted in question 6, perform a test of the null hypothesis that the hazard rate does not change over time. What do you conclude?
9. Using the Weibull model fitted in question 6, plot estimated survivor curves for individuals in job grade groups 1 and 2 aged 45, 55, 65.

   In Stata you can create these plots using the `stcurve` command.
   In R some code for producing these plots is provided in the example R script.

   Discuss: What do the plots show?

Extra exercises

1. We used the Weibull model above to allow the hazard to change over time. A different approach is to split the follow up time up into a few periods and fit a series of exponential models within each period. It can then be investigated whether the baseline hazard changes across the periods. To do this we need to create a record for each individual within each time period up to their event or censoring time.

This can be done in Stata using the `stsplit` command. Note that you will need to add the `id` option into your `stset` command before doing this. Try the following commands to see what happens:

```
list id _t0 _t _d if id==5001
list id _t0 _t _d if id==5350
stsplit period, at(0,5,10,15,20)
list id _t0 _t _d if id==5001
list id _t0 _t _d if id==5350
```

Fit a model using the exponential distribution to this newly split data including `period` as an additional explanatory variable. Write down the algebraic expression for the model being fitted here.

This can be done in R using the `survSplit` command. Try the following commands to see what happens:

```
whl[whl$id %in% c(5001,5350),c("id","timein","timeout","time","chd")]
whl.split<- survSplit(Surv(time=timeout,chd,origin=timein)~., dta=whl,
                        cut=c(0,5,10,15,20), episode="period")
whl.split[whl.split$id %in% c(5001,5350),c("id","tstart","timeout","chd","period")]
```

Fit a model using the exponential distribution to this newly split data including `period` as an additional categorical explanatory variable.

(a) Write down the algebraic expression for the model being fitted.

(b) Interpret the results and compare the results from this model with those from the Weibull and exponential models fitted earlier.

This approach of fitting exponential models within time bands is sometimes called 'Lexis expansion'.

2. We have used the exponential model to investigate the association between job grade and CHD. The exponential model is based on the assumption of a constant baseline hazard. An equivalent way of fitting this model is using Poisson regression, which should be familiar to you from earlier modules.

   Fit a poisson regression model to these data, with job grade as an explanatory variable. Check that you get the same results using Poisson regression and using an exponential model.

Session 4

The Cox proportional hazards model

4.1 Aims of this lecture and practical

At the end of this lecture and practical you will be able to:

- Write down algebraically the Cox proportional hazards model.
- Explain how the Cox proportional hazards model is fitted using a partial likelihood, and how this is used to estimate hazard ratios and corresponding standard errors, p-values and confidence intervals.
- Compare the Cox modelling approach with fully parametric and nonparametric methods.
- Interpret the results from Cox proportional hazards models with binary, categorical and continuous explanatory variables, and multiple explanatory variables of different types.
- Interpret assessments of whether the proportional hazards assumption is appropriate using plots.
- Fit Cox-proportional hazards models and perform assessments of these models using appropriate plots in Stata and R.

4.2 Introduction to the Cox proportional hazards model

Under the assumption that explanatory variables act proportionally on the hazard, the hazard function for an individual with vector of explanatory variables $X = x$ is

$$h(t|x) = h_0(t)e^{\beta^\top x} \tag{4.1}$$

where $h_0(t)$ is the hazard function for a baseline individual. Under the Weibull distribution, or the special case of the exponential distribution, the baseline hazard is parameterized. Cox (1972) suggested that the baseline hazard could be left unspecified, i.e. not written in terms of parameters to be estimated. The model in (4.1) is referred to as the Cox proportional hazards model. Under this model the baseline hazard is not parameterized, but the effect of the explanatory variables on the hazard is parameterized (using β). For this reason, this model is referred to as a semi-parametric model.

As in previous lectures we assume a study population of n individuals $i = 1, \ldots, n$. Individual i has explanatory variables x_i (which may be a vector, in general), survival

or censoring time t_i, and indicator δ_i which takes value 1 if t_i is a survival time and value 0 if t_i is a censoring time.

Under the Cox proportional hazards model the full likelihood of the data is

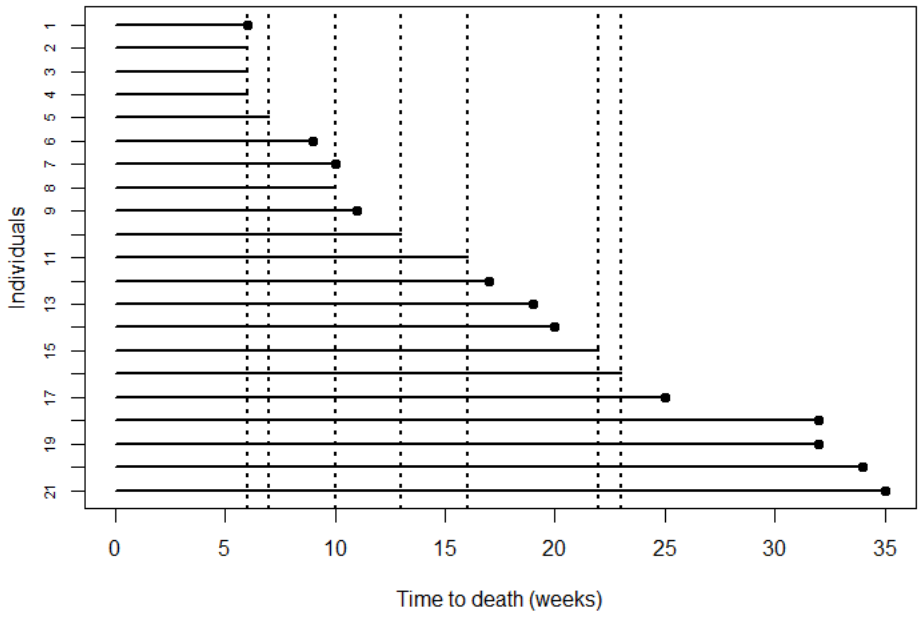
$$L = \prod_i^n \left\{ h_0(t_i) e^{\beta^\top x_i} \exp\left(-e^{\beta^\top x_i} \int_0^{t_i} h_0(u) du \right) \right\}^{\delta_i} \left\{ \exp\left(-e^{\beta^\top x_i} \int_0^{t_i} h_0(u) du \right) \right\}^{1-\delta_i} \tag{4.2}$$

How can we use this likelihood without choosing a particular form for the baseline hazard $h_0(t)$? The answer is that we can't (not using standard methods anyway) and so we use a special kind of analysis, which is outlined in the next section.

4.3 Partial likelihood

We begin by considering a particular time t_j at which an individual has the event of interest – we refer to this individual by the index (subscript) i_j and their explanatory variables are therefore x_{i_j}. For now we assume that there are no tied survival times (i.e. there are no individuals who share the same survival time). At time t_j there is a risk set – the risk set is the groups individuals who, up to just before time t_j, have not yet had the event of interest and have not been censored. The risk set is the group of individuals who might have been observed to have the event at time t_j. We denote this risk set at time t_j by R_j. Person i_j is in the risk set, along with all those who do not have the event at time t_j or before and who have not been censored.

Figure 4.1: Life lines showing times to death or censoring for 21 leukemia patients, with dotted lines showing the risk sets at each observed death time. Censoring times are indicated by a circle.



Now we ask the following question: given that the set of individuals R_j have survived

up to time t_j without having the event or being censored, what is the probability that it was individual i_j with explanatory variables x_{i_j} who had the event at time t_j when it might have been any one of the other individuals in the risk set R_j with their corresponding explanatory variables? This conditional probability is

$$\frac{h_0(t_j)\exp(\beta^\top x_{i_j})}{\sum_{k\in R_j} h_0(t_j)\exp(\beta^\top x_k)} = \frac{\exp(\beta^\top x_{i_j})}{\sum_{k\in R_j}\exp(\beta^\top x_k)} \tag{4.3}$$

This arises because the probability of having the event at time t_j for an individual k who is at risk just before time t_j is the hazard for that person at time t_j, i.e. $h_0 t_j \exp(\beta^\top x_k)$. The baseline hazard terms cancel out and we are left with a conditional probability which involves only the parameter vector β.

There is a conditional probability like that in (4.3) at each event time t_j. We multiply all these probabilities together to get what is called a partial likelihood, denoted by L_P:

$$L_P = \prod_j \frac{\exp(\beta^\top x_{i_j})}{\sum_{k\in R_j}\exp(\beta^\top x_k)} \tag{4.4}$$

where the product is over all event times. This is called a partial likelihood because it is not the likelihood for the full survival process, but for part of it. It has been found that the partial likelihood has the same asymptotic properties as a standard likelihood. This means we can use the partial likelihood to estimate β using maximum likelihood estimation, and that the variance of the MLEs $\hat{\beta}$ is given by the inverse of the information matrix. Estimating the parameters β in this way is referred to as *Cox regression*.

The log-partial likelihood, denoted by l_P, is

$$l_P = \sum_j \beta^\top x_{i_j} - \sum_j \log\left(\sum_{k\in R_j}\exp(\beta^\top x_k)\right) \tag{4.5}$$

Assuming univariable X for a moment, the maximum likelihood estimates for β are therefore found by solving

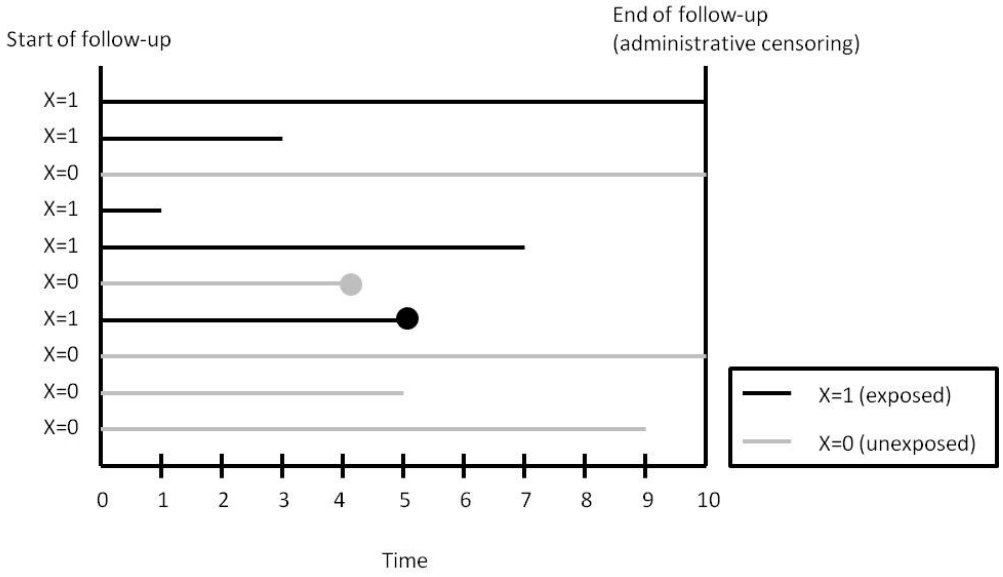
$$\frac{dl_P}{d\beta} = \sum_j x_{i_j} - \sum_j \frac{\sum_{k\in R_j} x_k \exp(\beta^\top x_k)}{\sum_{k\in R_j}\exp(\beta^\top x_k)} = 0 \tag{4.6}$$

If β is a vector of parameters then there will be a vector of terms like this, giving a set of equations to be solved simultaneously. In general, iterative methods are required to find the maximum likelihood estimates.

Exercise 4.1

Formulate the partial likelihood for the data displayed in Figure 4.2, by working out the contributions to the partial likelihood at each event time.

Figure 4.2: Example data on 10 individuals. Censorings are indicated by the circles (apart from those three individuals who are censored due to the end of follow-up at time 10).



Example 4.1

Cox model fitted to the leukemia example data

We fitted the Cox proportional hazards model to the leukemia patient data used in previous examples. Recall that these data are from a randomized controlled trial of 42 leukemia patients, of whom 21 were taking a treatment and 21 were in the control group. The event of interest is death and time is measured in weeks from diagnosis. The exposure is treatment group, with the controls being the baseline group. The hazard is $h_0(t)$ in the control group and $h_0(t)e^{\beta}$ in the treatment group.

Performing the Cox regression analysis gives the log hazard ratio estimate $\hat{\beta} = -1.51$ with standard error $0,410$ and 95% CI (-2.31,-0.71). The hazard ratio estimate is therefore $e^{\hat{\beta}} = 0.22$, with 95% confidence interval (0.10,0.49). The hazard for death is reduced by 78% in the treatment group relative to the control group. The 85% CI for the hazard ratio excludes 1. There is strong evidence against the null hypothesis of no association between treatment and the hazard for death.

In Stata:

```
. stcox group

        failure _d:  death
  analysis time _t:  time
```

```
Cox regression -- Breslow method for ties

No. of subjects =           42                  Number of obs    =          42
No. of failures =           30
Time at risk    =          541
                                                LR chi2(1)       =       15.21
Log likelihood  =   -86.379622                  Prob > chi2      =      0.0001

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |   .2210887   .0905501    -3.68   0.000     .0990706    .4933877
------------------------------------------------------------------------------
```

In R:

```
> leukaemia.cox<- coxph(Surv(time=time,event=death)~as.factor(group),
                                data=leukaemia,ties="breslow")
> summary(leukaemia.cox)
Call:
coxph(formula = Surv(time = time, event = death) ~ as.factor(group),
    data = leukaemia, ties = "breslow")

  n= 42, number of events= 30

                     coef exp(coef) se(coef)      z Pr(>|z|)
as.factor(group)1 -1.5092    0.2211   0.4096 -3.685 0.000229 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
as.factor(group)1    0.2211      4.523   0.09907    0.4934
```

Exercise 4.2

The Cox regression model was fitted to the breast cancer example data using Stata and R and the output is shown below. Interpret the results.

```
. stcox im

         failure _d:  death
   analysis time _t:  time

Cox regression -- no ties

No. of subjects =           45                  Number of obs    =          45
```

```
No. of failures =           24
Time at risk    =         4425
                                                LR chi2(1)       =        4.45
Log likelihood  =   -81.520649                  Prob > chi2      =      0.0350

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          im |   2.664988   1.158975     2.25   0.024     1.136362    6.249912
------------------------------------------------------------------------------
```

```
> breastcancer.cox<- coxph(Surv(time=time,event=death)~as.factor(im),
                                data=breastcancer,ties="breslow")
> summary(breastcancer.cox)
Call:
coxph(formula = Surv(time = time, event = death) ~ as.factor(im),
    data = breastcancer, ties = "breslow")

  n= 45, number of events= 24

                 coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(im)2 0.9802    2.6650   0.4349 2.254   0.0242 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

               exp(coef) exp(-coef) lower .95 upper .95
as.factor(im)2     2.665     0.3752     1.136      6.25
```

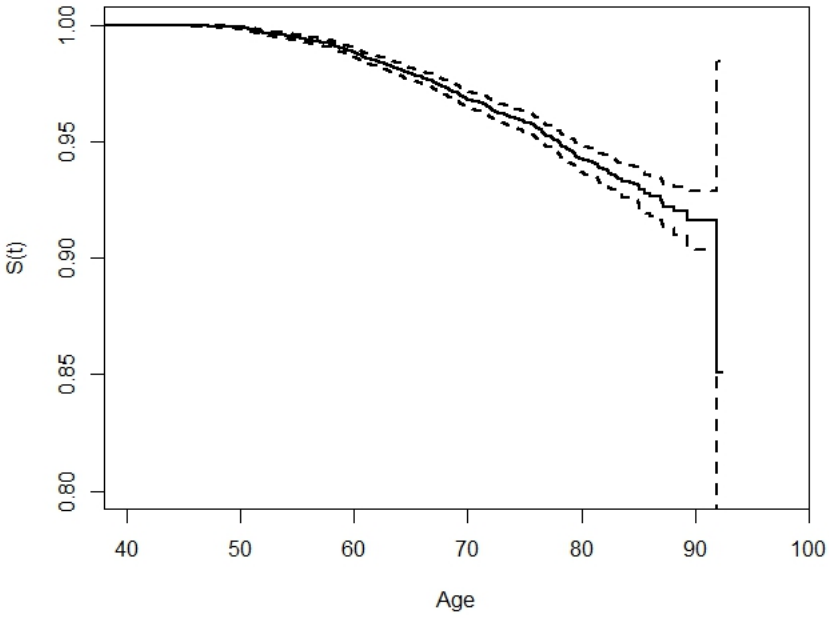
Example 4.2

A more complex example: Risk factors for breast cancer in the EPIC-Norfolk cohort

EPIC-Norfolk is a cohort of 25,639 individuals who were recruited in the 1990s aged about 40 and older. This cohort has been followed up for disease diagnoses and death. A large amount of data on lifestyle exposures, including detailed data on dietary intake, was obtained on participants. In this example we look at times to breast cancer diagnosis (as a first cancer) among 12,576 women in this cohort. This excludes women who had any cancer prior to joining the cohort and women were censored if they had another cancer prior to a breast cancer. Individuals were also censored due to death from other causes and due to loss to follow-up. The time scale is age. Figure 4.3 shows the Kaplan-Meier estimate of the survival curve.

Suppose we are interested in studying the association between alcohol intake and breast cancer risk. To investigate this we would want to adjust for any potential confounders

Figure 4.3: Kaplan-Meier estimate of the survival curve for women in the EPIC-Norfolk cohort (solid line), showing the 95% confidence limits (dotted lines).



of the association. For illustration, we consider adjustment for family history of breast cancer and smoking status as potential confounders. In reality, there are a large number of variables that would be considered as confounders. Alcohol intake was measured using a questionnaire and is measured in this example in units of 10 grams per day. Family history is a binary variable (yes/no). Smoking status is a categorical variable with three categories: never smoker (the baseline group), former smoker, current smoker. All information was obtained at recruitment to the cohort. The proportional hazards model is

$$h(t|X) = h_0(t) \exp\left(\beta_1 X_{\text{alc}} + \beta_2 X_{\text{FH}} + \beta_3 X_{\text{former-smoker}} + \beta_4 X_{\text{current-smoker}}\right) \qquad (4.7)$$

Where X_{alc} is a continuous variable indicating alcohol intake, X_{FH} is a binary variable indicating family history and taking value 1 for Yes and 0 for No, $X_{\text{former-smoker}}$ is an indicator taking value 1 for former smokers and 0 otherwise, and $X_{\text{current-smoker}}$ is an indicator taking value 1 for current smokers and 0 otherwise. The results from performing the Cox proportional hazards regression analysis are shown in Table 4.1.

Table 4.1: Results from fitting a Cox proportional hazards regression model to the EPIC-Norfolk data on risk factors for breast cancer.

Variable	Hazard ratio	95% CI	p-value
X_{alc}	1.140	(1.041,1.249)	0.005)
X_{FH}	1.766	(1.337,2.334)	< 0.001
$X_{\text{former-smoker}}$	0.875	(0.714,1.047)	0.197
$X_{\text{current-smoker}}$	1.001	(0.749,1.337)	0.995

Exercise 4.3

Interpret the results in Table 4.1.

4.4 Handling tied survival times

In some studies, there will be tied survival times - that is, some individuals will have the event recorded at the same time, e.g. on the same day. This can be incorporated into the partial likelihood analysis used for the Cox proportional hazards model. Suppose that at a particular time t_j there are m_j individuals who have the event. We now consider the question: given that the set of individuals R_j have survived up to time t_j without having the event of being censored, what is the probability that it was the set of individuals $\{i_{1j}, i_{2j}, \ldots, i_{m_j j}\}$ with vectors of explanatory variables $\{x_{i_{1j}}, x_{i_{2j}}, \ldots, x_{i_{m_j j}}\}$ who had the event at time t_j when it might have been any other set of size m_j formed from the individuals in the risk set R_j? This conditional probability is

$$\frac{h_0(t)e^{\beta^\top x_{i_{1j}}} \times h_0(t)e^{\beta^\top x_{i_{2j}}} \times \cdots, \times h_0(t)e^{\beta^\top x_{i_{m_j j}}}}{\sum_{L \in R_{m_j j}} \prod_{l \in L} h_0(t_j)e^{\beta^\top x_l}} = \frac{\exp\left(\beta^\top x_{i_{1j}} + \beta^\top x_{i_{2j}} + \cdots + \beta^\top x_{i_{m_j j}}\right)}{\sum_{L \in R_{m_j j}} \prod_{l \in L} e^{\beta^\top x_l}} \tag{4.8}$$

where here $R_{m_j j}$ denotes the set of all possible sets of size m_j from the risk set R_j. So the denominator is the sum of the product of the hazards for all possible sets of size m_j from the risk set R_j. The full partial likelihood is the product of these expressions over all event times t_j. The above expression becomes difficult to use from a computational point of view if the number of tied event times is quite large – this is because this results in a very large number of terms in the denominator. Therefore an approximation is usually used. This approximation takes the form

$$L_P^* = \prod_j \frac{\exp\left(\beta^\top x_{i_{1j}} + \beta^\top x_{i_{2j}} + \cdots + \beta^\top x_{i_{m_j j}}\right)}{\left(\sum_{k \in R_j} e^{\beta^\top x_k}\right)^{m_j}} \tag{4.9}$$

This is called 'Breslow's approximation' and is available in the software packages that we are using.

In Stata the default method for handling tied event times is Breslow's method. R uses a different default, which is why we specified `ties="breslow"` in `coxph` in the examples from R shown above.

4.5 Assumptions of the Cox model

Like any model applied to data, the Cox proportional hazards model relies on some assumptions. Two primary assumptions are

1. The proportional hazards assumption: that the explanatory variables act on survival in such a way that the hazard ratio is constant over time. In other words, the assumption that the model $h(t|x) = h_0(t)e^{\beta^\top x}$ is correct.
2. The assumption that we have correctly specified the form for how the explanatory variables act on the hazard. For example, for a continuous variable X is it appropriate to have $h(t|x) = h_0(t)e^{\beta x}$ or $h(t|x) = h_0(t)e^{\beta_1 x + \beta_2 x^2}$?

These two assumptions are bound up together. That is, the proportional hazards assumption (1) may hold for one specified form for the explanatory variables but not another (2).

We also have the assumptions as discussed previously: that the censoring is uninformative about the event of interest; that individuals are independent.

Later in this session and in session 5, we will cover some methods for assessing the proportional hazards assumption. Methods for assessing assumption 2 above will also be considered in session 5.

A primary reason for using the Cox proportional hazards model, and the special analysis by partial likelihood, is so that we do not have to assume a particular parametric form for the baseline hazard $h_0(t)$. Thus this semi-parametric approach is making fewer assumptions than the alternative fully parametric approach (eg. using a Weibull model). Cox regression is a convenient analysis which works well and is extremely widely used and understood.

Some points to consider when choosing a survival model are:

- If an exponential or Weibull model is suitable for our survival data then a Cox proportional hazards model is also suitable.
- If an exponential or Weibull model is suitable for our survival data then a Cox proportional hazards modelling approach and the fully parametric modelling approach will give very similar results. The fully-parametric approach may give slightly more precise estimates of the hazard ratio parameters (i.e. with smaller standard errors). However, this gain in precision will usually be small.
- The possible small gain in precision (or efficiency) that we may gain from a fully-parametric analysis has to be traded off against the concern that the parametric form for the baseline hazard may have been mis-specified, i.e. the model we have chosen to use is not the true model. By using the Cox proportional hazards model we avoid this issue completely.
- A 2002 paper Royston and Parmar (Statistics in Medicine 2002; 21: 2175-2197) introduced a new class of parametric survival models, referred to as 'flexible parametric survival models', in which the log cumulative baseline hazard is modelled smoothly using cubic splines. These combine some advantages of the Cox model and some of parametric models and have started to become popular. We do not cover flexible parametric survival models in this course but it is a good idea to be aware of them for the future.

4.6 Estimating survivor curves

Often when we perform an analysis of survival data, the thing we are most interested in is the association between explanatory variables and survival. Inference about this association comes from the hazard ratio estimates $e^{\hat{\beta}}$ and the corresponding standard errors and confidence intervals. As we have discovered in this lecture, we can perform such investigations without specifying a form for the baseline hazard.

Sometimes, however, we are interested in describing the survival of people in our study, for example we may wish to present estimated survival curves. It is also sometimes of interest to make prediction of survival for an individual with particular explanatory variables. For example, in the leukemia patient data we may wish to estimate the probability of survival to 20 weeks for individuals in the treatment and control groups. This requires information about the baseline hazard as well as the hazard ratio.

In fully parametric models, such as those considered in Lecture 3, the baseline hazard is parameterized and so we can easily estimate the survivor curve using the estimated parameters. It is desirable to be able to do the same when we have used a proportional hazards model to estimate the hazard ratios.

We let a 'baseline' person have a particular set of values for the set of explanatory variables, and we denote these values by $x^* = (x_1^*, x_2^*, \ldots, x_p^*)^\top$. Often all of these values are zero. Alternatively, the values may be the mean value in the study population (e.g. because a height of 0 doesn't make sense). An estimate of the baseline cumulative hazard at a time t in the range $t_k \leq t < t_{k+1}$ is

$$\widehat{H}_0(t) = \sum_{j=1}^{k} \frac{d_j}{\sum_{l \in R_j} \exp\left(\hat{\beta}^\top (x_l - x^*)\right)} \tag{4.10}$$

where d_j denotes the number of events at time t_j. This formula is referred to as Breslow's estimate of the cumulative hazard. It follows that an estimate of the survivor function is

$$\hat{S}(t|x) = \exp\left(-\widehat{H}_0(t) e^{\hat{\beta}^\top x}\right) \tag{4.11}$$

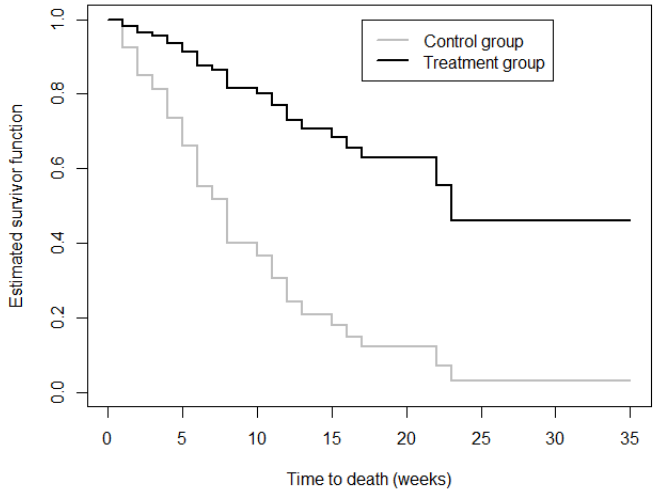
The estimated survivor function can be plotted for different values of the covariates X.

Example 4.3

Leukemia example: Estimating survivor curves

Figure 4.4 shows the estimated survivor curves in the treatment and control group obtained from fitting a Cox regression model and then using the formula for estimating the survival probabilities given in (4.11). The survivor curves have a 'step-wise' appearance. However, they are not to be confused with the Kaplan-Meier estimates of the survivor curves. The Kaplan-Meier estimates are non-parametric, whereas the survivor curves in Figure 4.4 were obtained from the semi-parametric Cox model in which we made an assumption of proportional hazards.

Figure 4.4: Leukemia example: Estimated survivor curves using the Cox proportional hazards model.



In Stata:

```
stcox group
stcurve, survival at1(group=0) at2(group=1)
```

In R (two examples of how to make the survival curves plot):

```
plot(leukaemia.survfit,mark.time=F,col=c("black","grey"),
                xlab="Time",ylab="Estimated survivor function")
legend(25,1,c("Placebo","Active"),col=c("black","grey"),lty=1,cex=0.8)

ggadjustedcurves(leukaemia.cox, data = leukaemia,variable="group")
```

Example 4.4

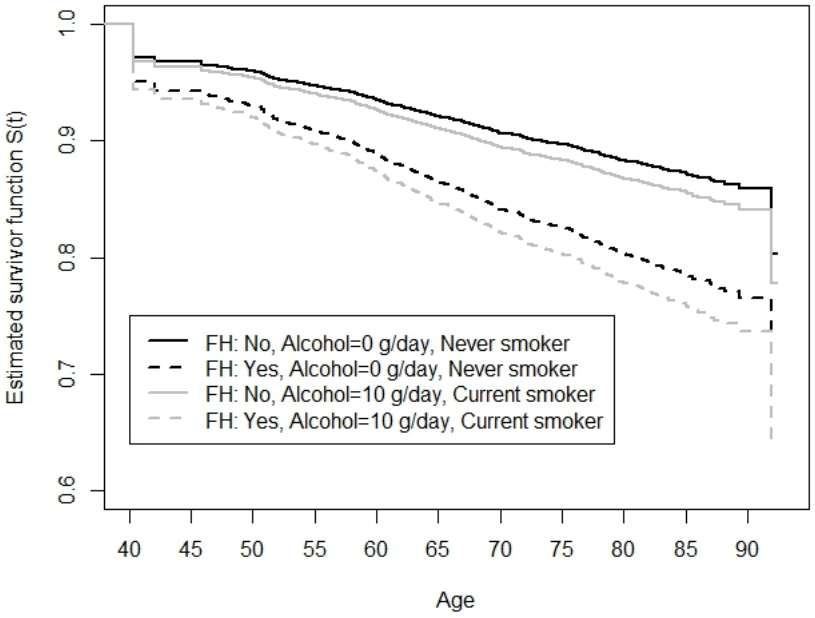
Risk factors for breast cancer in the EPIC-Norfolk cohort: Estimating survivor curves

In this example there are three explanatory variables (alcohol intake (continuous), family history of breast cancer (binary), smoking status (categorical)). When there are several explanatory variables it can be useful to look at the estimated survivor curves for individuals with given combinations of values for the explanatory variables. Figure 4.5 shows the estimated survivor curves under the Cox proportional hazards model for individuals with four different sets of values for the three explanatory variables.

4.7 Beyond the hazard ratio

As stated earlier, the Cox regression model is extremely popular. It has become the default method of analysis in both randomized and observational studies with time-

Figure 4.5: EPIC-Norfolk cohort example: Estimated survivor curves using the Cox proportional hazards model. [FH: family history].



to-event outcomes. The results are typically presented in terms of the hazard ratio. However, over the past 10 years or so, a series of papers have been written which explain that a hazard ratio does not have a straightforward interpretation in terms of a 'causal effect'. For example see the papers of Hernan (2010), Stensrud et al (2019). We do not go into the details of the issues here, but the basic idea is as follows. In a randomized trial the treatment and control groups are 'balanced' in terms of their characteristics at time 0 due to the randomization. However, at any time t^* after time 0 the characteristics of the two groups are no longer balanced, because if the treatment is effective then there will be proportionally more individuals in the treatment group who are 'frail' (i.e. sicker and more likely to have the event) compared with the control group. This is because people in the treatment group who are 'frail' at time t^* might have had the event before t^* if they had been in the control group (and therefore not be in the risk set at time t^*), but by being in the treatment group they have been saved a little (i.e their expected time to the event is longer). This is a form of selection bias. The hazard ratio involves comparing individuals in the treatment and control groups at every time at which an event occurs, and assuming the ratio does not change over time. The above issue means that the individuals in the treatment and control groups are not comparable (balanced) at any time after time 0. This doesn't mean that we can't use the Cox model though.

A solution is to quantify the effects of covariates not in terms of a hazard ratio, but in terms of another quantity which does not suffer the above issues. Two such quantities are the risk difference and the risk ratio. For simplicity, consider a randomized controlled trial, where X denotes treatment group. The risk difference at time t^* is

defined as

$$\Pr(T \leq t^*|X=1) - \Pr(T \leq t^*|X=0) = (1 - S(t^*|X=1)) - (1 - S(t^*|X=0)) \quad (4.12)$$

and the risk ratio is defined as

$$\frac{\Pr(T \leq t^*|X=1)}{\Pr(T \leq t^*|X=0)} = \frac{1 - S(t^*|X=1)}{1 - S(t^*|X=0)} \quad (4.13)$$

The risk difference and risk ratio can be estimated following a Cox regression by making use of the estimated survival probabilities using the result in equation (4.11).

4.8 Introduction to assessing the proportional hazards assumption

The proportional hazards assumption is that the ratio of hazards between two values of a covariate is constant over time. The hazard model could be written in the general form

$$h(t|x) = h_0(t) \exp\{\beta x \times g(t)\} \quad (4.14)$$

The proportional hazards model is the case where $g(t) = 1$. If $g(t)$ depends on t (time) then the hazard ratio depends on t, i.e. it is not constant over time. In some simple situations (e.g. a binary or categorical explanatory variable) we can use plots to study the proportional hazards assumption.

Under the proportional hazards model we can write the survival function as

$$S(t|x) = \exp\left\{-\int_0^t h_0(u)e^{\beta x}du\right\} \quad (4.15)$$

By taking the log and moving the minus sign we get

$$-\log S(t|x) = \int_0^t h_0(u)e^{\beta x}du = H_0(t)e^{\beta x} \quad (4.16)$$

where $H_0(t)$ is the baseline cumulative hazard. Taking logs again gives

$$\log\{-\log S(t|x)\} = \log H_0(t) + \beta x \quad (4.17)$$

When X is binary we have $\log\{-\log S(t|0)\} = \log H_0(t)$ and $\log\{-\log S(t|1)\} = \log H_0(t) + \beta$. We can see from this that if we plot $\log\{-\log S(t|0)\}$ and $\log\{-\log S(t|1)\}$ against $\log H_0(t)$ then the two curves should be separated by a constant (β) over the whole time range if the proportional hazards assumption is valid, i.e. the two curves are parallel. In practice we plot Kaplan-Meier estimates of $\log\{-\log S(t|0)\}$ and $\log\{-\log S(t|1)\}$ against $\log t$, because to plot against $\log H_0(t)$ would require some parametric assumption in order to estimate $H_0(t)$.

It can also be shown that non-proportional hazards tend to result in survivor curves that cross or that converge or diverge over time. Observing Kaplan-Meier plots of the survivor curves in two (or more) groups is therefore another way of assessing the proportional hazards assumption visually.

Another visual assessment of the proportional hazards assumption is to compare the estimated survivor curves from the proportional hazards model in the two (or more) groups, as found using the formula in (4.11), with the corresponding Kaplan-Meier survivor curves.

Extensions

The same procedure can be performed for a categorical explanatory variable. In this case there would be an estimated survivor curve for each category, and these should all be approximately parallel under the proportional hazards model.

Things are more tricky when we have continuous variables or several explanatory variables to consider simultaneously, which is often the case. For multiple categorical covariates we would have to look at the survivor curves for every combination of the different variables. For example, if we have two binary explanatory variables there would be four curves to plot. Continuous variables have to be categorized for this. It is clear that this approach quickly becomes cumbersome and unrealistic once there are more than two or three variables.

Example 4.5

Leukemia example: Assessing the proportional hazards assumption

Figure 4.6 shows plots of $\log\{-\log S(t|x)\}$ against $\log t$. If the proportional hazards model were true the estimated survival curves in the two groups would be parallel. There is no visual evidence against the proportional hazards assumption here. Note that this plot is that same as one that we saw in Lecture 3.

Figure 4.7 shows the estimated survivor curves found using the Cox proportional hazards model (and Breslow's estimate in (4.11) and using Kaplan-Meier. The two sets of estimates are similar. This provides further evidence that the proportional hazards assumption appears to have been appropriate.

In Stata:

```
stphplot, by(group)
quietly: stcox group
stcoxkm, by(group)
```

In R:

The code is more complicated and we will see an example in the practical.

Exercise 4.4

Acute graft versus host disease (AGVHD) example: assessing the proportional hazards assumption

Figure 4.6: Leukemia example: Plot of a Kaplan-Meier estimate of $\log\{-\log S(t|x)\}$ against $\log t$ in treatment and controls groups.

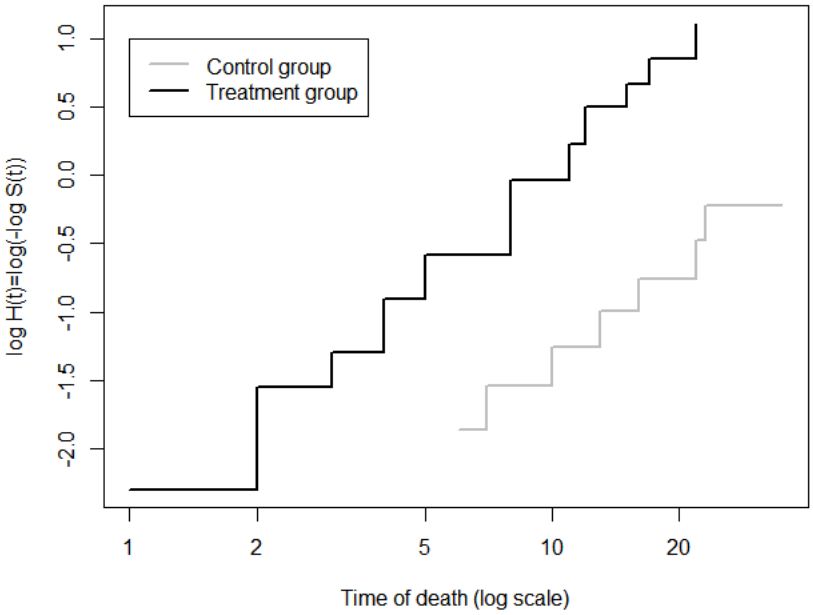
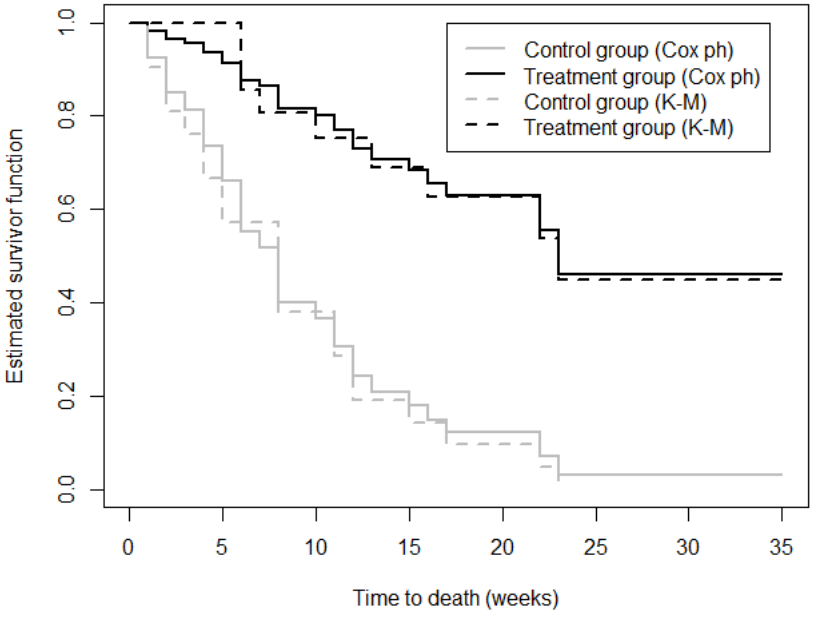


Figure 4.7: Leukemia example: Plot of survivor curves in the treatment and control groups estimated using Cox regression (and Breslow's estimate) and using Kaplan-Meier.

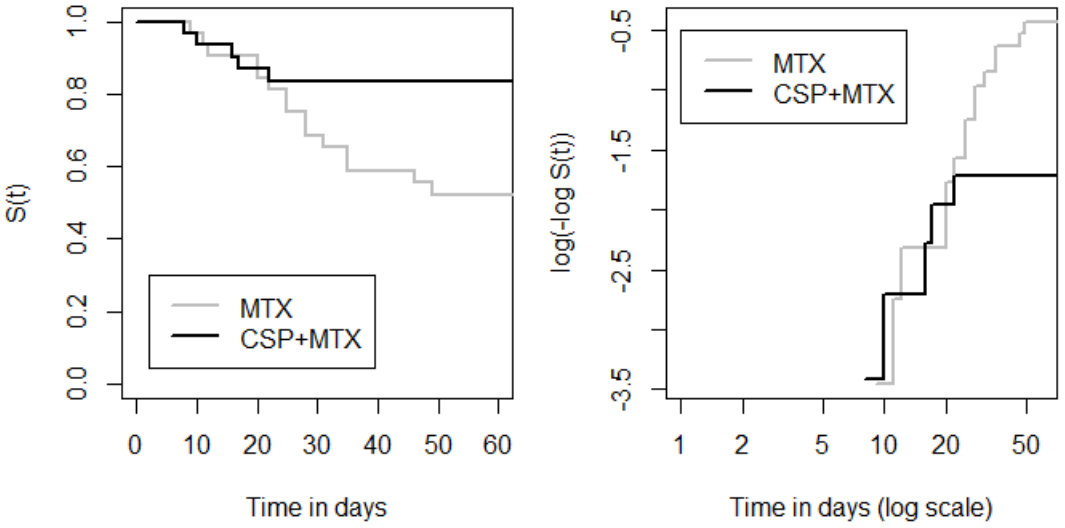


For this example we use data from a randomized trial involving 64 patients with severe aplastic anaemia who had received bone marrow donated by a family member. In this trial there were two treatment groups; one group received methotrexate alone (MTX) and the other group received methotrexate plus cyclosporine (CSP+MTX). The data

are given in the book by Kalbfleisch and Prentice (2nd edition 2003, Table 1.2, pp. 3). The event of interest was diagnosis of a life-threatening stage of acute graft versus host disease (AGVHD). A plot of the Kaplan-Meier estimates of the survivor functions on the two treatment groups is shown in Figure 4.8 (left-hand panel). There is a suggestion from this plot that those in the CSP+MTX group do better in the long term compared with those in the MTX group.

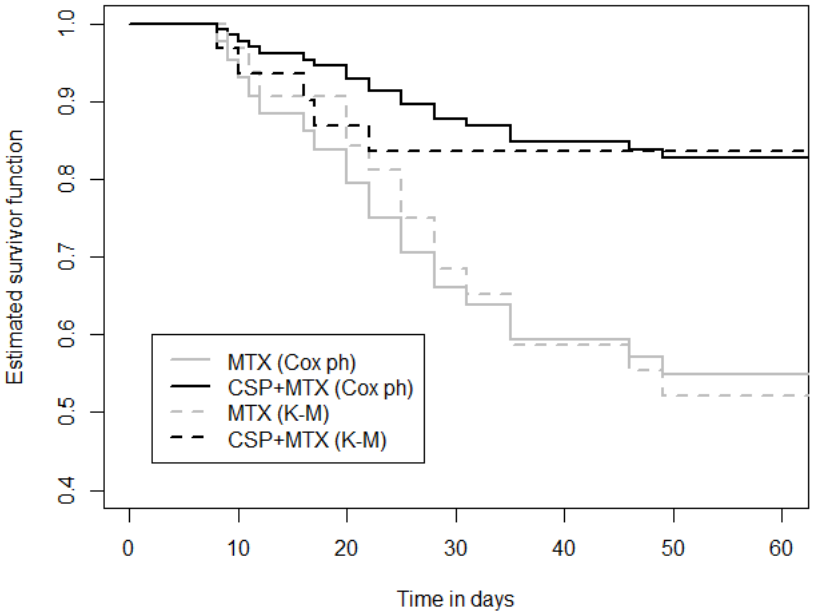
The right-hand panel of Figure 4.8 shows the plots of $\log\{-\log S(t|x)\}$ against $\log t$ in the two treatment groups. Figure 4.9 shows the comparison between the survivor curves predicted under the Cox proportional hazards model and the Kaplan-Meier curves.

Figure 4.8: Acute graft versus host disease (AGVHD) example: Plot of a Kaplan-Meier estimate of $\log\{-\log S(t|x)\}$ against $\log t$ in the two treatment groups.



Do you think a Cox proportional hazards model would be appropriate for this data?

Figure 4.9: Acute graft versus host disease (AGVHD) example: Plot of survivor curves in the two treatment groups estimated using Cox regression (and Breslow's estimate) and using Kaplan-Meier.



References

Cox DR. Regression models and life tables. Journal of the Royal Statistical Society (Series B) 1972; 34: 187-220.

Cox DR. Partial likelihood. Biometrika 1975; 62: 269-276.

Hernan MA. The hazards of hazard ratios. Epidemiology 2010; 21; 13-15.

Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine 2002; 21: 2175-2197.

Stensrud MJ, Aalen JM, Aalen OO, Valberg M. Limitations of hazard ratios in clinical trials. European Heart Journal, 2019; 40: 1378–1383.

Practical 4

Datasets required: `pbcbase_2021` & `alloauto`

R packages required for R users: survival, ggplot2, survminer.

This session introduces Cox regression. This practical is in two parts.

A We will use the PBC data to fit a Cox model, check the proportional hazards assumption, and estimate survival curves for different values of the covariates

B We will use a new dataset, called alloauto, to investigate the proportional hazards assumption of the Cox model

Aims After completing this practical you should be able to:

- Fit a Cox regression model (in Stata or R) and interpret the results
- Obtain estimates of hazard ratios to compare groups of individuals with different values of the covariates
- Obtain estimates of survival curves from a Cox regression model for particular values of the covariates
- Check the proportional hazards assumption of the Cox model

Where code examples are given or explanations are given that are specific to Stata or R, text and code relating to Stata is shown in this colour and text and code relating to R is shown in this colour.

Part A: PBC data

In this session we will focus on estimating the association between treatment group (`treat`) and the hazard, using Cox regression. later in the practical we will also use the `bil0` variable. Load the PBC data and re-familiarise yourself with the key variables. We will analyse the data on the time-in-study timescale.

Variable	Description
`id`	Unique identifier for each participant
`datein`	Date person entered the study
`dateout`	Date of the end of follow-up due to either death or censoring
`d`	Event indicator at the end of follow-up: 0=alive (censored), 1=dead
`time`	Follow-up time in years
`treat`	Treatment 1=placebo, 2=active
`bil0`	serum bilirunbin (mg/dl), measured at the start of the trial

1. Write down:
   (a) the form of the hazard assuming a Cox proportional hazards model
   (b) the partial likelihood for this model

2. In Cox regression, there is a contribution to the partial likelihood from each event time. In this question we will derive the contribution to the partial likelihood at the second time at which an event occurred in the PBC data, which is time $t = 0.052$.

   (a) What is the value of `treat` for the individual who has the event at that time?

   (b) How many individuals are at risk at that time?

   (c) What values of `treat` do these individuals have?

   (d) Using the information from (b) and (c) find the contribution to the partial likelihood at time $t = 0.052$ in the model including `treat`.

   In Stata you may find it helpful to use `sts list, by(treat)` to obtain some of the above information

   In R you may find it helpful to use `survfit` to obtain Kaplan-Meier estimates of the survival function by treatment group, as you did in Practical 2 (pbc.km, say) and then look at the output from `summary(pbc.km)`.

3. Fit the Cox model and interpret the results.

   In Stata (you must use stset first):

```
stcox i.treat
```

   In R:

```
pbc.cox<-coxph(Surv(time,d)~as.factor(treat),data=pbc)
summary(pbc.cox)
```

4. Obtain the estimated survivor curves in the two treatment groups based on the Cox model. What is the probability of survival beyond time 5 in the two treatment groups?

   In Stata:

```
stcurve, survival at1(treat=1) at2(treat=2)
```

   In R one way of creating the estimated survival curves is:

```
pbc.survfit=survfit(pbc.cox,newdata=data.frame(treat=c(1,2)))

plot(pbc.survfit,mark.time=F,col=c("black","grey"),xlab="Time",
                          ylab="Estimated survivor function")
legend(8,1,c("Placebo","Active"),col=c("black","grey"),lty=1,cex=0.5)
```

   Discuss: Why do the estimated survivor curves have 'steps'? How do these survivor curves differ from the Kaplan-Meier estimates?

5. Assess the proportional hazards assumption graphically.

   In Stata use `stphplot` and `stcoxkm`. What is being shown in each case?

In R try the code given below. What is being shown in each case?

```
plot(survfit(pbc.cox,newdata=data.frame(treat=c(1,2))),
     col=c("blue","red"),xlab="time",ylab="S(t)")
lines(pbc.km,mark.time=F,col=c("blue","red"),lty=2,add=T)
legend(8,1,c("Placebo, Cox","Active, Cox",
             "Placebo, Kaplan-Meier","Active, Kaplan-Meier"),
       col=c("blue","red","blue","red"),lty=c(1,1,2,2),cex=0.5)

plot(pbc.km,fun="cloglog",xlab="time (log scale)",ylab="log(-log S(t))",
     col=c("blue","red"),xlim=c(0.02,12))
legend(0.02,0,c("Placebo","Active"),col=c("blue","red"),lty=1,cex=0.5)

ggsurvplot(pbc.km, data = pbc,conf.int = T,fun="cloglog",censor=F,
           legend.title="",legend.labs = c("Placebo","Active"))
```

Discuss: What do you conclude about the proportional hazards assumption in this model?

6. The researchers are also interested in how the level of bilirubin measured at the start of the trial (`bil0`) is associated with the outcome. Write down the form of a hazard model including both treatment group and baseline bilirubin (you do not need to include an interaction term).
7. Fit the above model and interpret the results.
8. We will now compare the hazards in different types of individual.
   (a) What is the hazard ratio comparing: (i) a person in the active treatment group with bil0=75, (ii) a person in the active treatment group with bil0=30.
   (b) What is the hazard ratio comparing (i) a person in the placebo group with bil0=75, (ii) a person in the placebo group with bil0=30.
   (c) What is the hazard ratio comparing (i) a person in the active treatment group with bil0=75, (ii) a person in the placebo group with bil0=30.
9. Obtain the estimated survivor curves for individuals in the two treatment groups with baseline bilirubin value equal to 15, 30 and 75 (these are approximately the 25th, 50th and 75th percentiles). You can do this by extending the code used in question 4.
10. Fit a Weibull model containing treatment and bilirubin levels as explanatory variables.

Discuss: Compare your results from the Cox model with those from a Weibull model. Which model you prefer for these data?

Part B: Bone marrow transplant data

The `alloauto` dataset is from a study of 101 individuals with advanced acute myelogenous leukemia. 51 of the patients received treatment using their own bone marrow (an autologous bone marrow transplant) and 50 patients received bone marrow from a sibling (an allogenic bone marrow transplant). The event of interest was a composite of death or relapse. There are just three variables in the dataset.

Variable	Description
`time`	Time in months to event or censoring
`delta`	Event indicator: 0=Censored, 1=Death or relapse
`type`	Treatment type: 1=allogenic, 2=autologous

1. Load the data. Obtain Kaplan-Meier estimates of the survivor curves in the two treatment groups and perform a log rank test. Interpret your results.
2. Use graphical methods to investigate whether the proportional hazards assumption is appropriate for these data. If you are satisfied that the proportional hazards assumption is met, fit the Cox model and interpret the results.

Session 5

More on the proportional hazards model: stratified Cox model and model checking

5.1 Aims of this lecture and practical

At the end of this lecture and practical you will be able to:

- Describe the stratified Cox proportional hazards model and compare its assumptions with those of the standard Cox proportional hazards model.
- Perform and interpret assessments of the proportional hazards assumption using a test of whether exposure effects change over time and using Schoenfeld residuals.
- Use plots of Martingale residuals and deviance residuals to comment on model fit.
- Use Martingale residuals to investigate the appropriate functional form for an explanatory variable in a proportional hazards model.
- Calculate residuals and make appropriate plots of them in Stata and R.
- Perform likelihood ratio tests to compare Cox proportional hazards models in Stata and R.

5.2 Stratified Cox proportional hazards model

The Cox proportional hazards model is by far the most commonly used model in survival analysis. Sometimes, however, we may find that this model does not provide a good fit to our data and wish to consider some alternatives. In this lecture, we start by focusing on a direct extension to the Cox proportional hazards model; the *stratified Cox proportional hazards model*.

Under the Cox proportional hazards model the hazard function for an individual with a vector of explanatory features $x = (x_1, \ldots, x_p)^T$

$$h(t; x) = h_0(t)e^{\beta^T x} \tag{5.1}$$

The vector x may be made up of variables of various types; binary, categorical, continuous. Under this model the effect of each explanatory variable on the hazard is assumed to be such that the ratio of hazards is constant across the time scale (the proportional hazards assumption). In applications with several explanatory variables, the effect of some of these variables may not be proportional. When the aim of the

analysis is not focused on these particular variables, for example if they are just being used as adjustment variables and are not the main exposures of interest, then the proportionality assumption can be relaxed just for those variables by fitting a *stratified Cox proportional hazards model*.

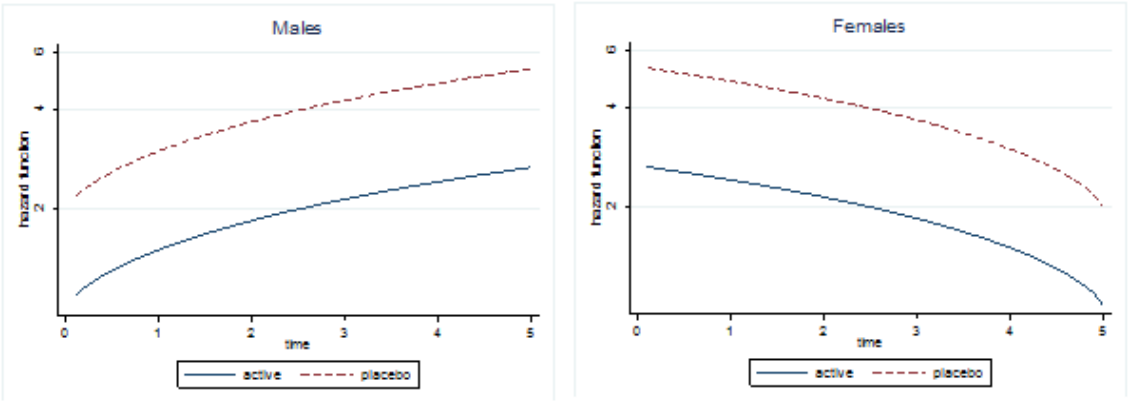
In the stratified Cox proportional hazards model, instead of assuming that the proportional hazards model holds overall, we assume that the proportional hazards model holds *within groups* (or *strata*) of individuals. This method, therefore, no longer provides an estimate of the effect of the factors defining the groups on the hazard, so is not a suitable approach where the factor exhibiting non-proportionality of primary interest.

Example 5.1

A stratified Cox proportional hazards model with the strata being men and women

Consider the scenario depicted in Figure 5.1. The hazard for men is initially very low but then increases steadily, while the hazard for women is initially high but then declines steadily. Note that in both cases the effect of the exposure variable, treatment, is constant over the follow-up time (the hazards are plotted on a log-scale).

Figure 5.1: Example of non-proportional hazards between males and females but proportional between treatment groups



Consider the standard Cox proportional hazards model:

$$h(t;SEX;TRT) = h_0(t)e^{\beta_{sex}SEX+\beta_{trt}TRT} \tag{5.2}$$

This implies the proportional hazards between men and women as well as between the two treatment groups. We clearly do not have proportional hazards between men and women. However, *within men* we have proportional hazards for the two treatment groups and *within women* we have proportional hazards for the two treatment groups. The following model would therefore be appropriate:

$$h(t;SEX;TRT) = \begin{cases} h_{0M}(t)e^{\beta_{trt}TRT} & \text{if Male} \\ h_{0F}(t)e^{\beta_{trt}TRT} & \text{if Female} \end{cases} \tag{5.3}$$

Note that the parameter β_{trt} is assumed to be the same for men and women - a single treatment effect is being estimated even though there are two strata (men and women) with non-proportional hazards.

Example of fitting a stratified Cox proportional hazards model in Stata for this (artificial) example, after appropriately `stset`-ing the data:

```
stcox treatment, strata(sex)
```

And the equivalent in R:

```
cox.strat=coxph(Surv(time,d)~treatment+strata(sex),data=dataset)
```

More generally, suppose that individuals can be separated into S groups (or strata) on the basis of a particular feature or features, $s = 1, \ldots, S$. Examples are sex (as in the above example) or age group. Other features of individuals that we are interested in are still denoted by the vector x. The proportional hazards model is now assumed to hold within strata, and can be written as

$$h(t; x, s) = h_{0s}(t)e^{\beta^T x} \tag{5.4}$$

Each stratum, s, has a separate baseline hazard $h_{0s}(t)$. However, the other explanatory variables x are assumed to act in the same way on the baseline hazard in each stratum, i.e. the β are the same across strata.

The stratified Cox proportional hazards model is simple to fit using a direct extension to the familiar partial likelihood analysis. In the partial likelihood analysis for the stratified Cox proportional hazards model, each case is compared at its time of failure with all individuals at risk at that time and in the same stratum as the case.

$$L_P = \prod_j \frac{e^{\beta^T x_{i_j}}}{\sum_{k \in R_{sj}} e^{\beta^T x_k}} \tag{5.5}$$

where the sum in the denominator is over the risk set at time t_j for individuals in stratum s, denoted R_{sj}.

Further comments to make on the stratified Cox proportional hazards model are:

- We cannot estimate the effect on survival of the variables represented by the strata. This is because their effect on survival is represented only through the baseline hazards, and the baseline hazards are eliminated in the partial likelihood analysis.
- It is possible, however, to estimate interactions between other explanatory variables and the strata. For example, if one of the explanatory variables is a binary 'treatment' variable, we can allow the effect of treatment on survival to differ across strata.
- We have to present estimated survivor curves from the stratified Cox proportional hazards model separately within strata, by estimating the survivor function within strata. See Lecture 4 for estimating the survivor function from a Cox proportional hazards model.

- Using a stratified model can sometimes solve issues of model fit.

We have focused on stratified cox proportional hazards models in this section, but the same idea can be extended to fully parametric proportional hazards models, e.g. the exponential and Weibull models that we are familiar with.

Exercise 5.1: How could we investigate graphically whether a stratified proportional hazards model is appropriate? Consider the simple case of one categorical main exposure in addition to the stratifying variable(s).

5.3 Introduction to model checking

Before reporting results we should check, as far as possible, that the fitted model is correctly specified. If not, our inferences could be invalid, and we may draw incorrect conclusions regarding our research question of interest.

Model checking is more complex in survival analysis than in, say, linear regression. This is due to the more complicated forms of the models which are involved in analysis of survival data and due to the presence of censoring. In Lecture 4 we learned how to use simple plots to check whether the proportional hazards assumption is appropriate. In the second part of this lecture, we will expand on this by looking at two other methods for checking the proportional hazards assumption. We will also consider some other aspects of model checking:

- How good is the overall fit of the model?
- Are there some individuals for whom the model does not provide a good fit?
- Is the functional form for the explanatory variables correct?

To address some of these questions we will learn about some new types of residuals for survival data: Martingale residuals, Schoenfeld residuals and deviance residuals.

5.4 Investigating the proportional hazards assumption

There are three main ways of assessing whether the proportional hazards assumption is reasonable:

- Using non-parametric plots.
- By performing a test of whether the effect of explanatory variables on the hazard depends on time.
- Using plots of special residuals.

Method 1 was outlined in Lecture 4. Here we focus on methods 2 and 3, which are closely related to each other.

5.4.1 Performing a test for proportional hazards

Under the proportional hazards model the ratio of hazards at different levels of an exposure is constant over time. In other words, the effect of explanatory variables on the hazard is the same over time. This excludes a situation such as when an explanatory variable has a strong effect at the start of the time scale but a weaker effect later on. This would be a violation of the proportional hazards assumption.

We can perform a test of whether the effect of an explanatory variable on the hazard changes over time. For an explanatory variable X this is done by including an interaction between time and the explanatory variable in the model:

$$h(t;x) = h_0(t)e^{\beta x + \gamma(x*t)} \tag{5.6}$$

If the effect of the explanatory variable X on the hazard changes over time then the parameter γ will be non-zero. We can perform a test of the null hypothesis that $\gamma = 0$ using a likelihood ratio test, or using the Wald test

$$\frac{\hat{\gamma}}{SE(\hat{\gamma})} \sim \mathcal{N}(0,1) \tag{5.7}$$

In order to fit the model in 5.6 we cannot simply calculate $x * t_i$ for each individual i (where t_i denotes their own survival or censoring time). This is because the fitting of the partial likelihood requires the value of $x * t$ at each event time t at which an individual is in the risk set. Luckily, the software easily handles this for us.

If there are several explanatory variables then one can include interactions with time for each explanatory variable, in which case γ will be a vector of parameters. Sometimes it may be reasonable to use $log(t)$, or some other function of t, in place of t in the exponential term in (5.6). This can be applied for any type of explanatory variable.

5.4.2 Using Schoenfeld residuals

Another way of assessing the proportional hazards assumption is by using **Schoenfeld residuals**, which we introduce in this section. First we note that the log partial likelihood under the proportional hazards model is

$$l_P = \sum_j \beta^T x_{i_j} - \sum_j log\left(\sum_{k \in R_j} e^{\beta^T x_k}\right) \tag{5.8}$$

where x_k is a vector of explanatory variables for individual k, $x_k = (x_{1k}, x_{2k}, \ldots, x_{pk})^T$, and β is a corresponding vector of log hazard ratios to be estimated, $\beta = (\beta_1, \ldots, \beta_p)^T$. The sums are over all event times t_j, thus excluding censoring times. The first derivate of l_P with respect to β_1, say, is

$$\frac{\partial l_P}{\partial \beta_1} = \sum_j x_{1i_j} - \sum_j \frac{\sum_{k \in R_j} x_{1k} e^{\beta^T x_k}}{\sum_{k \in R_j} e^{\beta^T x_k}} \tag{5.9}$$

The maximum likelihood estimate for β_1 is given by the value $\hat{\beta}_1$ for which the above derivative is equal to zero. The Schoenfeld residual at time t_j for the first explanatory variable, which we denote by $r_{S_{1j}}$, is given by the contribution of t_j to the above derivative, evaluated at $\hat{\beta}$:

$$r_{S_{1j}} = x_{1i_j} - \frac{\sum_{k \in R_j} x_{1k} e^{\widehat{\beta}^T x_k}}{\sum_{k \in R_j} e^{\widehat{\beta}^T x_k}} \tag{5.10}$$

Given that the maximum likelihood estimation of β_1 is such that the above derivative (5.9) is equal to zero, the Schoenfeld residuals must sum to zero. The Schoenfeld residual compares the observed values of the explanatory variable for the case at a given event time with the weighted average of the explanatory variable in the risk set. The residuals should not show any dependence on time - this would indicate that the proportional hazards assumption is not met.

It is actually more convenient usually to use the 'scaled Schoenfeld residuals'. The scaled Schoenfeld residual at event time t_j for the first explanatory variable, which we denote by $r_{SS_{1j}}$, is

$$r_{SS_{1j}} = d * r_{S_{1j}} var(\hat{\beta}_1) \tag{5.11}$$

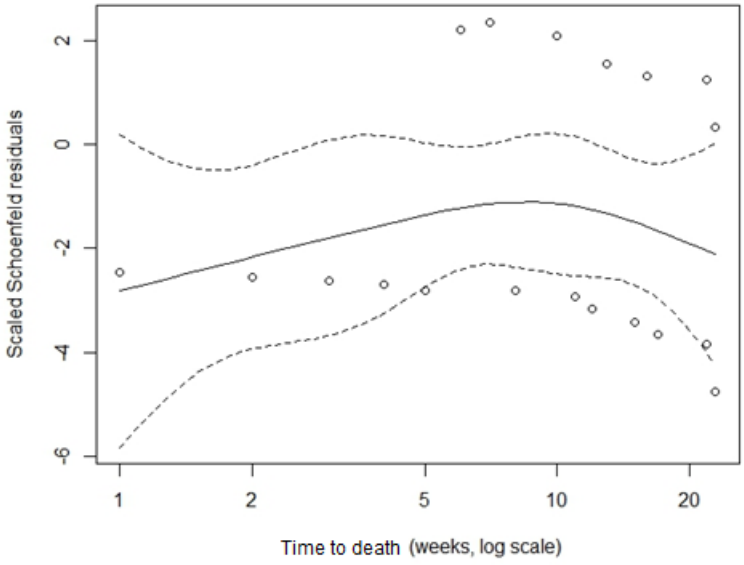
where $var(\hat{\beta}_1)$ is found using the inverse of the information matrix and d is the number of events observed at event time t_j (e.g. deaths). The scaled Schoenfeld residuals have a mean which is the true log hazard ratio under the proportional hazards assumption, and the average values of the scaled Schoenfeld residuals over time can be interpreted as the *time-varying log hazard ratio*. A plot of the scaled Schoenfeld residuals over time is therefore directly informative about how the log hazard ratio changes over time. It is useful to show a smoothed average curve on these plots.

Example 5.2

Leukemia patient data: assessing the proportional hazards assumption

The estimate of the parameter associated with the interaction between treatment group and time is $e^{\hat{\gamma}}$=1.008, which has confidence interval (0.894,1.137) and p-value 0.894. There is no suggestion here that the proportional hazards assumption was not appropriate. The scaled Schoenfeld residuals are shown in Figure 5.2.

Figure 5.2: Plot of the scaled Schoenfeld residuals for the leukemia patient data



The smoothed average curve is not a straight line, however the confidence intervals are wide and a straight line is compatible with the confidence limits, i.e. we could fit a straight line easily within the limits.

Example 5.3

Acute graft versus host disease (AGVHD) data: assessing the proportional hazards assumption

In this example the estimate of the parameter associated with the interaction between treatment group and time is $e^{\hat{\gamma}}$=0.854, which has confidence interval (0.726,1.003) and p-value 0.055. There is some evidence here, therefore, that the effect of the explanatory variable changes over time, which is consistent with what we saw in the plot of the transformed survivor curves in Lecture 4 - see Figure 5.3 for a reminder of these.

The scaled Schoenfeld residuals are shown in Figure 5.4. Here the smoothed curve is quite far from being a straight line and is curving downwards quite substantially as time increases. We could just about fit a straight line within the confidence bounds. However there does appear to be some evidence here that the proportional hazards assumption is not appropriate.

Figure 5.3: AGVHD data: plots of the estimated survivor curve in the two treatment groups

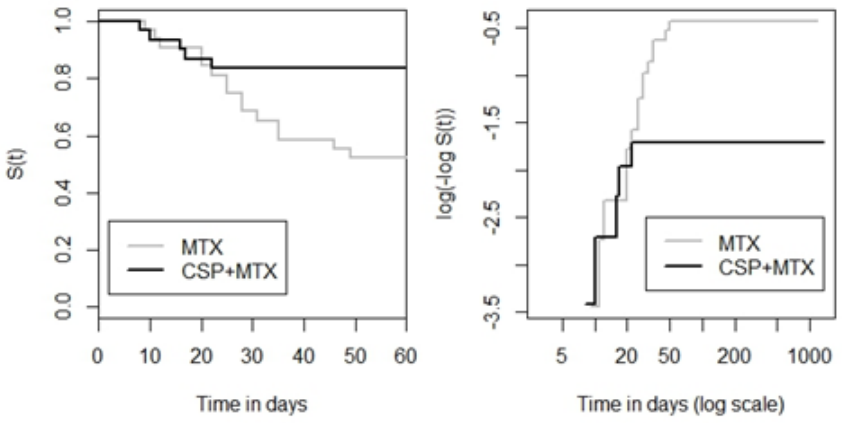
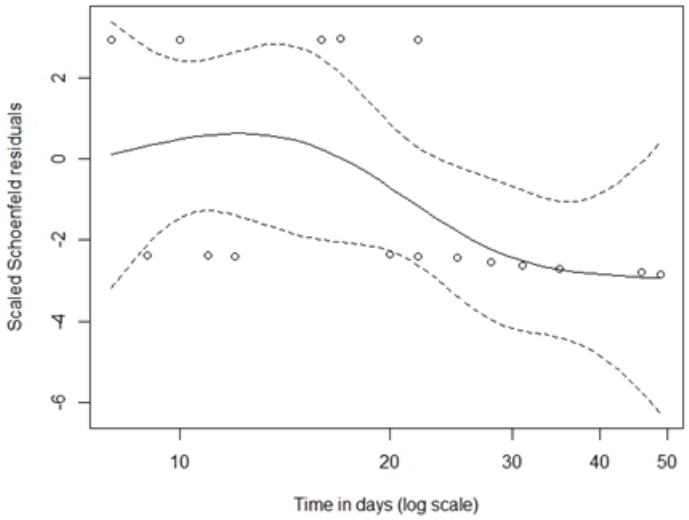


Figure 5.4: Plot of the scaled Schoenfeld residuals for the AGVHD data



Example 5.4

EPIC-Norfolk breast cancer data: assessing the proportional hazards assumption

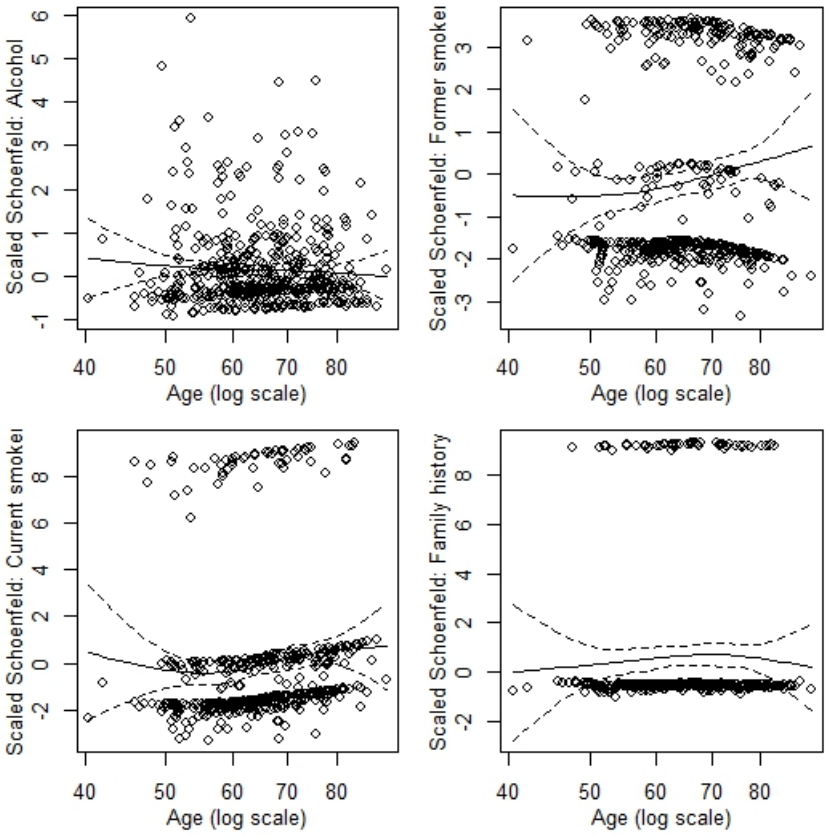
In this example there are three explanatory variables, one of which is a categorical variable with 3 levels. We can investigate whether the effects of the covariates vary over time by fitting the model
$h(t;X) = h_0(t)exp(\beta_1 X_{alc} + \beta_2 X_{FH} + \beta_3 X_{former-smoker} + \beta_4 X_{current-smoker} + \gamma_1 X_{alc}t + \gamma_2 X_{FH}t + \gamma_3 X_{former-smoker}t + \gamma_4 X_{current-smoker}t)$

The estimates of the $exp(\gamma)$ parameters are in the following table:

Parameter	Estimate	%95CI	p-value
Alcohol ($exp\gamma_1$)	0.995	(0.995,1.005)	0.306
Family history ($exp\gamma_2$)	1.007	(0.978,1.036)	0.656
Former smoker ($exp\gamma_3$)	1.029	(1.008,1.051)	0.008
Current smoker ($exp\gamma_4$)	1.038	(1.005,1.071)	0.022

There is a separate set of Scaled Schoenfeld residuals for each explanatory variable.

Figure 5.5: Plots of the scaled Schoenfeld residuals for the breast cancer data.



Exercise 5.2: What do you conclude for this example?

5.5 Assessing other aspects of model fit using residuals

5.5.1 Martingale residuals: assessing the functional form of continuous variables

Martingale residuals can be used as a way of investigating the appropriate functional form for continuous variables in the proportional hazards model. For example, consider a study with a continuous explanatory variable of interest, X. It may be appropriate to use a transformed version of X in the proportional hazards model. For example, we may wish to compare whether it is better to use X untransformed or log transformed

in the proportional hazards model:

$$h(t;x) = h_0(t)e^{\beta_1 x} \text{ or } h(t;x) = h_0(t)e^{\beta_1 logx} \tag{5.12}$$

A 'Martingale' is a residual for an event process - it is the difference between what happened to a person (whether they had the event or not) and what is predicted to happen to a person under the model that has been fitted. The Martingale residual for individual i is

$$r_{Mi} = \delta_i - \widehat{H}_0(t_i)e^{\hat{\beta}x_i} \tag{5.13}$$

where δ_i is the indicator of whether individual i had the event (1) or was censored (0), t_i is the event or censoring time, x_i denotes the explanatory variable (or more generally a vector of explanatory variables), and $\widehat{H}_0(t_i)$ is the estimated baseline cumulative hazard at time t_i. If the model is correct then the Martingale residuals should sum to 0.

It can be shown that a plot of the martingale residuals r_{Mi} from a null model (i.e. a model without any explanatory variables) against a continuous variable that we are interested in entering into the model can be used to indicate the appropriate functional form for the continuous variable when it is entered in the model. If there are binary and categorical variable we know that we want to include in the model, then we could include these in the model used to obtain the r_{Mi}, instead of using the null model.

Example code for plotting the Martingale residuals in Stata:

```
stset time, failure(event)
*Martingale residuals
stcox ag
predict mgale_res, mgale
lowess mgale_res wbc
lowess mgale_res log_wbc
stcox ag log_wbc
predict mgale_res2, mgale
lowess mgale_res2 log_wbc
```

And the equivalent in R:

```
leukaemia.cox=coxph(Surv(time,d)~as.factor(ag),data=leukaemia)
mgale_res1<-resid(leukaemia.cox,type="martingale")
plot(leukaemia$wbc,mgale_res1)
lines(lowess(leukaemia$wbc,mgale_res1))
plot(leukaemia$log_wbc,mgale_res1)
lines(lowess(leukaemia$log_wbc,mgale_res1))
leukaemia.cox2=coxph(Surv(time,d)~as.factor(ag)+log_wbc,data=leukaemia)
mgale_res2<-resid(leukaemia.cox2,type="martingale")
plot(leukaemia$log_wbc,mgale_res2)
lines(lowess(leukaemia$log_wbc,mgale_res2))
```

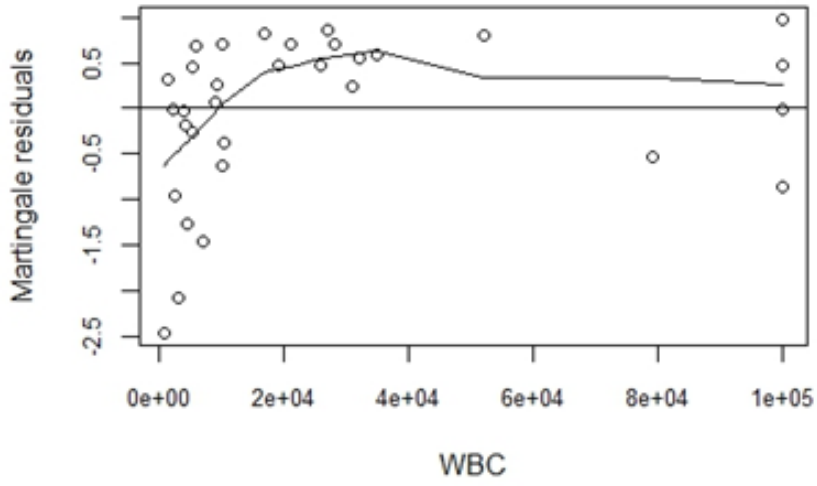
Example 5.5

Observational leukemia data (following on from Example 5.1): How should WBC be entered into the model?

In this example we illustrate the use of Martingale residuals for investigating the appropriate functional form for continuous variables in a proportional hazards model. There are two explanatory variables of interest: AG (binary) and WBC (continuous). The question is how should WBC be entered into the proportional hazards model.

We fitted a Cox model with only AG as an explanatory variable, i.e. excluding WBC. The Martingale residuals were obtained based on this model and were plotted against WBC - see Figure 5.6.

Figure 5.6: Plot of Martingale residuals (from a model with AG only) against WBC.



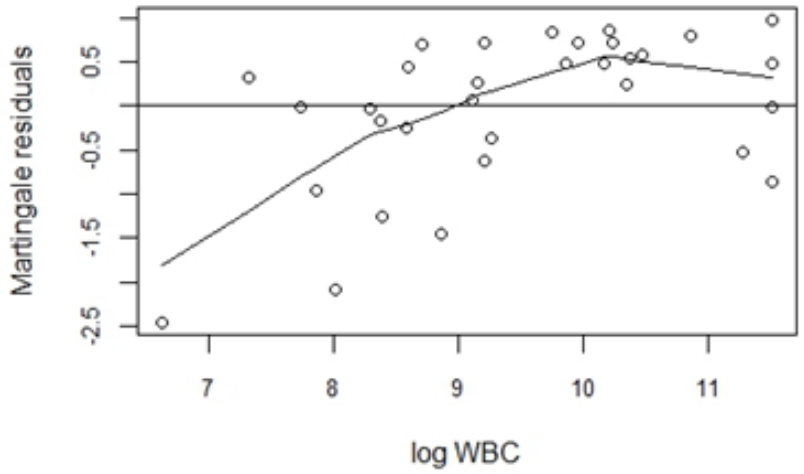
What are we looking for? If the smoothed curve appears linear then this suggests that entering WBC in its untransformed form would be appropriate, i.e.

$$h(t|AG, WBC) = h_0(t)e^{\beta_1 AG + \beta_2 WBC}$$

The smoothed curve is non-linear. This implies that the above form for WBC is not appropriate and therefore we should consider a transformation. In figure 5.7 we see a plot of the residuals against log WBC. The plot looks a little more linear, but is still not perfect. Although the log transformation doesn't look perfect, further investigations didn't reveal a better transformation. So let's fit the model with log WBC:

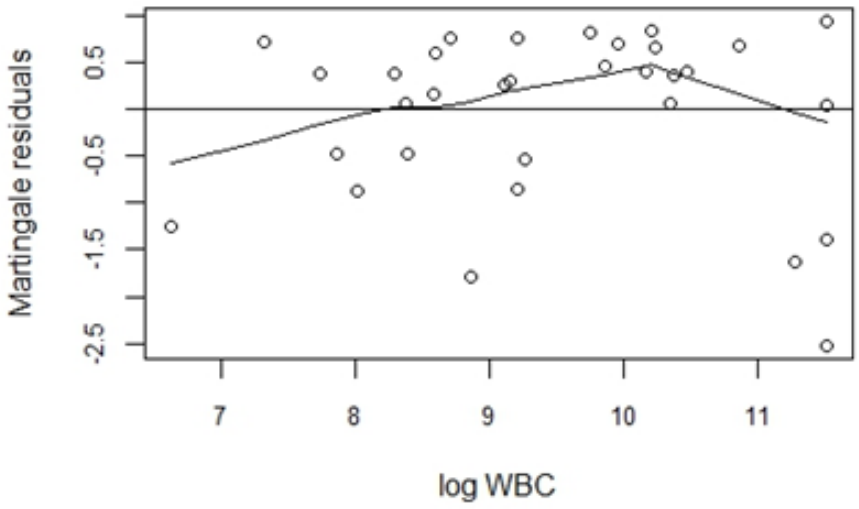
$$h(t|AG, WBC) = h_0(t)e^{\beta_1 AG + \beta_2 log(WBC)}$$

Figure 5.7: Plot of Martingale residuals (from a model with AG only) against log WBC.



After fitting the above model we can obtain the residuals and plot against log(WBC), as shown in Figure 5.8. If the functional form for WBC in model were appropriate then this plot would show a flat line. The line is not completely flat at 0, but depending on the circumstances we might be happy to live with this.

Figure 5.8: Plot of Martingale residuals (from a model with AG and log(WBC) against log WBC.

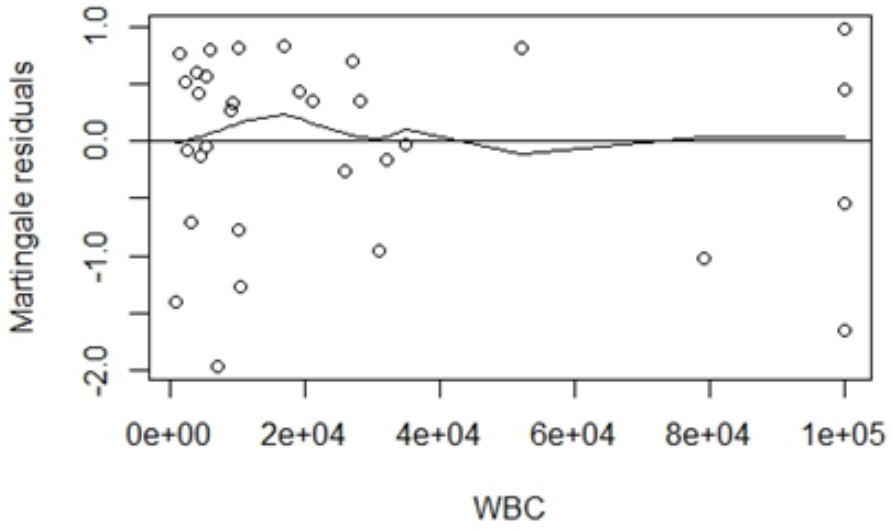


It could be possible to improve on the log transformation, for example by including a flexible form for WBC using a spline. More simply, one might consider categorizing WBC. We investigated categorizing WBC in the model. It was divided into 5 categories and we fit the following model, where $WBC_2, \ldots, WBC_5$ are dummy variables:

$$h(t|AG, WBC) = h_0(t)e^{\beta_1 AG + \beta_2 WBC_2 + \beta_3 WBC_3 + \beta_5 WBC_5}$$

Figure 5.9 shows a plot of the Martingale residuals from this model including categories of WBC against WBC. The smoothed line appears approximately flat at 0, suggesting that the form of WBC in the model is appropriate. In general, however, it is preferable not to categorize continuous variables.

Figure 5.9: Plot of Martingale residuals (from a model with AG and categorized WBC) against WBC.



5.5.2 Deviance residuals: identifying individuals for whom the model does not provide a good fit

The deviance residuals are a transformation of the Martingale residuals and are defined as follows:

$$r_{Di} = sign(r_{Mi})[-2(r_{Mi} + \delta_i log(\delta_i - r_{Mi}))]^{1/2} \tag{5.14}$$

The deviance residuals transform the Martingale residuals so that they are symmetrically distributed about the line at zero. We do not give the derivation of these residuals here but focus instead on their use in assessing model fit.

A large deviance residual indicates the model does not fit well for that individual. More specifically, a large positive deviance residual suggests that the individual has the event sooner than predicted by the model, and a large negative deviance residual indicates that the individual had the event later than predicted by the model.

A plot of the deviance residuals against the risk score (which is $\beta^T x$) can help to identify whether, say, individuals with a high risk score tend to have higher deviance residuals and therefore the model provides a worse fit for this type of individual. See the book by Collett, pg 126, for an example.

After assessing the residuals we must consider what to do about people for whom the model does not fit well. We can investigate their features – do they have particular features? Are there additional explanatory variables which could be included in the

model to give a better fit for these individuals? Or would a different functional form for some explanatory variables provide a better fit?

5.5.3 Other residual plots

There are various other plots of residuals that can be useful in assessing model fit, but which we will not go into details about:

- Another type of residual is called a **delta-beta**. The proportional hazards model provides an estimate of a particular log hazard ratio parameter of interest, $\hat{\beta}_1$ say. Suppose now that we remove individual i from the data and fit the model again, finding an estimate which we denote $\hat{\beta}_{1(i)}$. We can do this for each individual $i = 1, \ldots, n$. The delta-betas are the differences between the estimate using the full data and the estimate with individual i removed: $\hat{\beta}_1 - \hat{\beta}_{1(i)}$. A large delta-beta occurs if individual i has a large influence on the estimate from the full data. Plots of the delta-betas for all individuals can identify individuals who are having perhaps an unduly large influence on the overall estimate.
- Plots of **Cox-Snell** residuals can be used as a general assessment of model fit. The Cox-Snell residuals are defined as $r_{CSi} = \widehat{H}_0(t_i)e^{\hat{\beta}^T x_i}$. If we plot $-logS(r_{CSi})$ against r_{CSi} then, if the model is correct, we should see a straight line relationship. The straight line will have intercept 0 and slope 1. The Cox-Snell residuals can sometimes appear OK even though the model fit is poor, and so they are not much used now.

5.5.4 Residuals for fully parametric proportional hazards models

Residuals can also be obtained for parametric proportional hazards models, e.g. the exponential and Weibull models that we focused on in Lecture 3. In fact, the definitions of the Martingale residuals and deviance residuals only require a small tweak to be applied for parametric models – this is that the non-parametric estimate of the baseline hazard is replaced by its parametric estimate which is found when fitting the fully parametric model. We do not give the full details here. The interpretations of the residuals are the same.

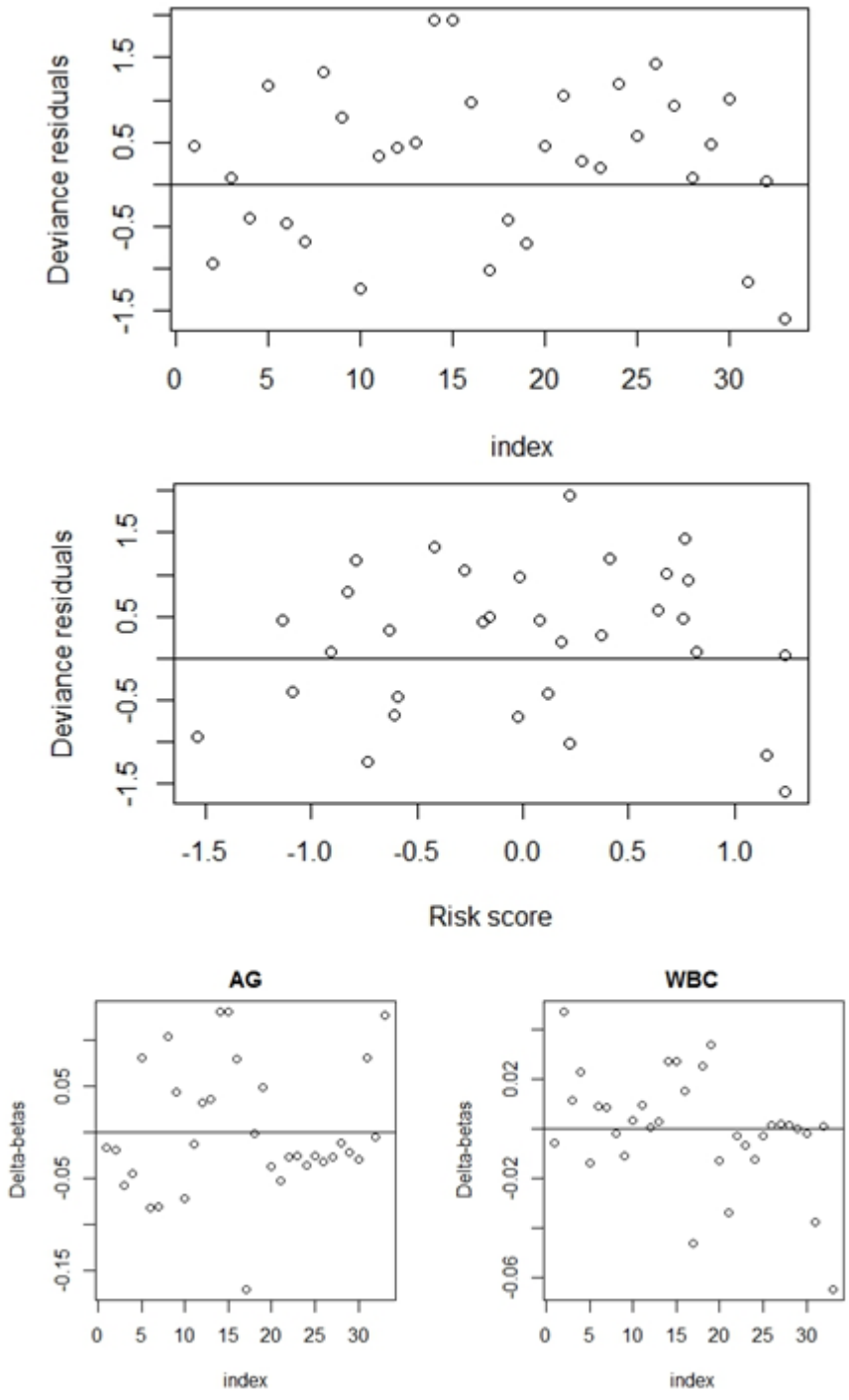
Example 5.6

Observational leukemia data: model checking using residuals

A Cox model was fitted with AG and log(WBC) as explanatory variables. Figure 5.10 shows the deviance and the delta-beta residual plots. The deviance plots do not suggest there are any outliers, i.e. any individuals for whom the model does not fit well. The second plot does not suggest that model fit is better or worse at different values for the risk score. Notice that there are two delta-beta plots – one for each parameter in

the model. The delta-beta plots do not suggest that there are any individuals having a large influence on the estimates.

Figure 5.10: Plots of residuals from A Cox model with explanatory variables AG and log(WBC)



Example code for plotting the deviance residuals and the Delta-Beta residuals in Stata:

```
*Cox model of interest
stcox ag log_wbc
*deviance residuals, plotted by individual
predict dev, deviance
scatter dev id
*deviance residuals plotted against risk score
predict riskscore, xb
scatter dev riskscore
*delta-betas
predict deltabeta_ag deltabeta_wbc, dfbeta
scatter deltabeta_ag id
scatter deltabeta_wbc id
```

And the equivalent in R:

```
leukaemia.cox3=coxph(Surv(time,d)~as.factor(ag)+log_wbc,data=leukaemia)
devres<-resid(leukaemia.cox3,type="deviance")
plot(devres,xlab="index", ylab="Deviance residuals")
abline(h=0)
#delta betas
delta.betas<-resid(leukaemia.cox3,type="dfbeta")
head(delta.betas)
plot(delta.betas[,1],xlab="index",ylab="Delta-betas",main="AG")
abline(h=0)
plot(delta.betas[,2],xlab="index",ylab="Delta-betas",main="wbc")
abline(h=0)
```

Practical 5

Datasets required: `alloauto` and `pbcbase_2021`

R packages required for R users: 'survival', 'ggplot2', 'survminer'.

Introduction

In this practical we will introduce ways to check the assumptions which underpin the Cox proportional hazards model. We will also offer some suggestions of what to do if any of the assumptions are not met. This practical is in two parts.

A Investigates the proportional hazards assumption using Schoenfeld residuals, and methods for fitting a Cox model when the proportional hazards assumption is not met for a particular variable

B Investigates how to ascertain the correct form to model a continuous variable, using Martingale residuals. We also investigate using deviance residuals and delta-beta residuals.

Aims

After completing this practical you should be able to

- Investigate the Proportional Hazards assumption of a Cox model using Schoenfeld residuals
- Fit a Cox regression model with an interaction between time and an explanatory variable in two different ways
- Interpret the results of a Cox model which includes such an interaction

Where code examples are given or explanations are given that are specific to Stata or R, text and code relating to Stata is shown in this colour and text and code relating to R is shown in this colour.

Part A: Alloauto data

In this first part we use the `alloauto` data set which was introduced in Practical 4. This data set contains information on 101 individuals with advanced acute myelogenous leukemia.

Variable	Description
`time`	Time in months to event or censoring
`delta`	Event indicator: 0=Censored, 1=Death or relapse
`type`	Treatment type: 1=allogenic, 2=autologous

1. Read the data into Stata or R and identify the outcome variables (event/censoring time and event indicator).

   In Stata, use `stset` using a time-in-study timescale.

Note what the correct form is for Surv() in R, for use in later questions.

2. We will initially repeat the visual checks of the proportional hazards assumption we performed in Practical 4.

   (a) Obtain a Kaplan-Meier plot of the survivor function in the two treatment groups

   (b) Produce a plot of $\log\{-\log S(t|x)\}$ against $\log t$ for $x = 0, 1$

   What do you think about the proportional hazards assumption for these data?

3. (a) Fit a Cox model including treatment type as the only explanatory variable.

   (b) Produce a plot of the Scaled Schoenfeld residuals. Note that the null hypothesis for this test is that there is an association between the residuals and time.

   In Stata, after fitting the cox model: `estat phtest, plot(2.type)`

   In R, after fitting the Cox model (called allo.cox):

```
sch.resid=cox.zph(allo.cox, transform = 'identity')
plot(sch.resid)
```

   (c) Perform a Schoenfeld test of the proportional hazards assumption

```
estat phtest
```

```
sch.resid
```

   Discuss: Interpret the results from the plot and the test. What do you conclude about the proportional hazards assumption for treatment type?

4. One way to deal with non-proportional hazards for a key variable is to allow the hazard ratio to change over time. We will demonstrate two ways to do this: first by allowing the HR to change in a continuous way over time, and second by estimating separate hazards in different timeperiods, for example we will consider estimating one HR for the early part of the study follow-up (up to 18 months), and one for the later part of the follow-up (after 18 months).

   (a) To allow the HR to change continuously over time we fit a Cox model including an interaction between treatment group and time.
   i) Write down the form of this model
   ii) Fit the model

```
 stcox i.type, tvc(i.type) texp(_t)
```

```
allo.mod.t=coxph(Surv(time,delta)~as.factor(type)+tt(type),
                    data=allo,tt=function(x,t,...){x*t})
```

(b) To estimate two HR's instead, we will use 18 months as the cutoff. This is approximately when the Schoenfeld residuals levelled off.
i) Write down the form of this model
ii) Fit the model

```
stcox i.type, tvc(i.type) texp(_t>18)
```

```
allo.cox.t2=coxph(Surv(time,delta)~as.factor(type)+tt(type),
                  data=allo,tt=function(x,t,...){x*(t>18)})
```

Discuss: Interpret the results from both models. What conclusions can you draw about the effect of the treatment type based on your analysis so far? How you would present these results to a clinician involved in the study?

5. EXTRA EXERCISE IF YOU HAVE TIME (please go on to Part B first).

An alternative way of fitting the models in question 4(b) is to split the follow-up time for each individual into two time periods, and then fit the Cox model including the interaction between treatment type and the binary time variable (before / after the split).

(a) Split the follow-up time for each individual into two time periods at 18 months using the code below. Take a look at the new form of the data.

Notice that to do this in Stata we need to generate a unique ID number for each individual in the dataset, and then include this in the `id()` option of the `stset` command. The `stsplit` command is used to split the follow-up time for each individual into two time periods.

```
gen id=_n
stset time, failure(delta) id(id)
stsplit time_period, at(18)
```

In R the `survSplit` function is used to split the follow-up time for each individual into two time periods.

```
allo.split=survSplit(Surv(time,delta)~., data=allo, cut=18, end="time",
                event="delta", start="time0", episode="time_period")
```

(b) Fit a Cox model including an interaction between treatment type and time period, without using the `tvc` and `texp` options (in Stata) or `tt` option in R. Compare the results with what you got in 4(b).

(c) Revert to the original format for the data (by reading the data in again). Next we show another way of fitting the model in 4(a). Use the following commands:

```
gen id=_n
stset time, failure(delta) id(id)
stsplit, at(failures)
```

```
event.times=alloauto$time[alloauto$delta==1]
alloauto.split=survSplit(Surv(time,delta)~., data=alloauto, cut=event.times,
                    end="time", event="delta", start="time0")
```

What has happened to the data? How many rows of data are there now?

With the data in this form, fit the model fitted in 4(a), but without using the `tvc` and `texp` options (in Stata) or `tt` option in R.

Discuss: Compare the results from this question to those from question 4.

Part B: PBC data

In this part we will use the familiar pbcbase data set. We will consider a survival model including treatment group and baseline bilirubin measurement (`bil0`) as explanatory variables, as in Practical 4. The aim of the analysis in this section is to conduct an an exploratory investigation into how different variables (measured at diagnosis) are associated with the hazard of death.

1. Open the PBC data and use stset in Stata.
2. Bilirubin (`bil0`) is a continuous variable, measured at baseline. We will use Martingale residuals to investigate the appropriate functional form for this variable in a Cox model.

   (a) This can be done by first fitting a Cox model including only the treatment, and then plotting the Martingale residuals from this model against `bil0`:

   ```
   stcox i.treat
   predict mgale_res1, mgale
   lowess mgale_res1 bil0
   ```

   ```
   pbc.cox=coxph(Surv(time,d)~as.factor(treat),data=pbc)
   summary(pbc.cox)


   mgale_res1<-resid(pbc.cox,type="martingale")
   plot(pbc$bil0,mgale_res1)
   ```

   (b) Create a variable which is the log-transformed bilirubin and plot this new variable against the Martingale residuals (from the model including the untransformed bilirubin). What do you conclude from these plots?

   (c) Fit a Cox model including treatment group and baseline bilirubin in your preferred form. Obtain the Martingale residuals based on this model and plot them against the bilirubin variable.

Discuss: Interpret these plots. What is the association between the Martingale residuals and the bilirubin variables?

3. Using the model that includes treatment and log-bilirubin, we will now assess the proportional hazards assumption for the two explanatory variables:

   (a) First, use plots of the scaled Schoenfeld residuals and the corresponding test. In Stata, use the 'detail' option for `estat phtest`.

   (b) Second, use interactions between each variable and time. Do this in two separate models (one for each explanatory variable), writing down the models being fitted each time.

   Discuss: What are your conclusions regarding the proportional hazards assumption for the two variables?

4. The following variables measured at baseline are also believed to be associated with the outcome: age (a continuous variable) and presence of cirrhosis (`cir0`: a binary variable).

   (a) Use appropriate residuals to investigate how age should be entered in the model.

   (b) Assess the proportional hazards assumption for all covariates in the model.

5. Imagine that one of your colleagues proposes stratifying by cirrhosis status, but not adjusting for age.

   (a) Write down the model they are suggesting (which includes treatment, log bilirubin, and stratified by cirrhosis).

   (b) Fit the stratified Cox model and compare the results with the model which includes cirrhosis as a covariate.

   Discuss: What are the advantages and disadvantages of using the stratified model?

6. Assess the model fit of your model in question 4 by looking at the deviance residuals and the delta-betas. This model includes the following covariates: treatment, log bilirubin, age, age-squared, cirrhosis.

```
stcox i.treat log_bil0 age cir0

* Deviance residuals
    predict dev, dev
    gen num=_n
    scatter dev num

* Delta-betas
    predict double dfb*, dfbeta
    scatter dfb* num

#deviance residuals
```

```
devres<-resid(pbc.cox3,type="deviance")
plot(devres,xlab="index", ylab="Deviance residuals")
abline(h=0)

#delta betas

delta.betas<-resid(pbc.cox3,type="dfbeta")
head(delta.betas)

plot(delta.betas[,1],xlab="index",ylab="Delta-betas",main="treat")
abline(h=0)
plot(delta.betas[,2],xlab="index",ylab="Delta-betas",main="logbil0")
abline(h=0)
plot(delta.betas[,3],xlab="index",ylab="Delta-betas",main="age")
abline(h=0)
plot(delta.betas[,4],xlab="index",ylab="Delta-betas",main="cir0")
abline(h=0)
```

Discuss: Interpret the plots. What do you conclude about the fit of this model?

Session 6

Competing risks and multi-state models

6.1 Aims of this lecture and practical

At the end of this lecture and practical you will be able to:

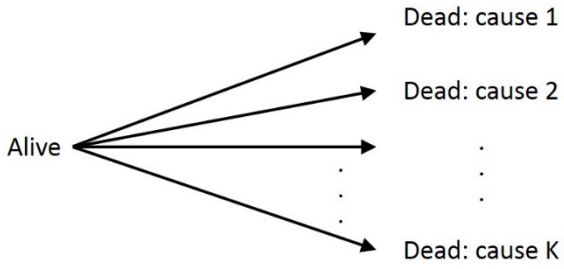
- Define what is meant by competing risks and explain why different methods are needed to analyse competing risks data.
- Define, plot and interpret the cumulative incidence function.
- Define and interpret results from cause-specific hazard models.
- Define and interpret results from subdistribution hazard models.
- Explain the difference between the subdistribution hazard and the cause-specific hazard approaches.
- Fit competing risks models in Stata and R.
- Explain key terms used in multi-state modelling, including the Markov property.
- Interpret results from multi-state models.

6.2 The censoring assumption

Recall that in handling censoring in Kaplan-Meier estimator or in Cox regression, it is assumed that the remaining uncensored individuals are representative of the survival experience in the censored individuals; i.e. those who are censored have the same future hazard of the event of interest as those who are not censored. This assumption is called non-informative censoring and is generally a reasonable assumption to make when censoring occurs due to the end of the study. However, if the censoring occurs due to some other event taking place, the assumption of non-informative censoring may not be so reasonable. When the outcome of interest is a particular cause of failure, which occurs alongside other possible causes of failure, this is called competing risks. These other competing risks events may prevent the event of interest from taking place. The obvious example of this is mortality: when a particular cause of mortality is of interest in the study, all other causes of death are then competing risks preventing the event of interest from taking place, see 6.1.

Example 6.1

Figure 6.1: Competing risks diagram showing K causes of death.

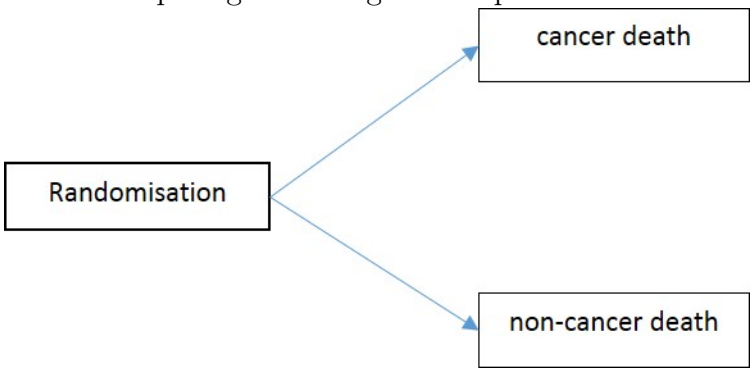


Treatment for prostate cancer

The example we will be using throughout this section on competing risks is a trial of a new treatment for prostate cancer involving 506 patients randomised either to the control or treatment arm (for the purposes of the exercise 6.1, 16 patients who exit the study in < 1 month are excluded, leaving 490 patients for analysis). Patients were followed-up for mortality outcomes for up to 76 months. The primary outcome of interest is cancer death, but other deaths are also recorded.

There are therefore two competing events of interest: death from cancer or death from another cause.

Figure 6.2: Competing risks diagram for prostate cancer example.



6.3 Kaplan-Meier

If we use the classical non-parametric Kaplan-Meier (KM) approach to estimate the cancer survival probability, thus censoring the other causes of death, our survival estimate will be biased. Moreover, the quantity given by 1 minus this survival probability will overestimate the probability of death due to cancer. Indeed, by using the KM estimator and censoring the other causes of death, we assume that those censored due to (say) a heart disease-related death have the same future marginal hazard of cancer death as those who have not yet had any event. This clearly is not true, since those who have already died of heart disease can never experience death from cancer. The Kaplan-Meier approach therefore over-estimates the probability of failure (by assuming

those already dead from other causes are still at risk of cancer death, and that they can be represented by those not yet experiencing any event) and under-estimates the probability of surviving to time t. It follows that scenarios with larger proportions of competing risks events produce larger bias in results.

Alternatively, the Kaplan-Meier results can be considered a valid interpretation of the hypothetical world where

- other causes of death have all been removed (i.e. we treat these censored individuals as though they were still at risk of the event of interest, as would be the case if the competing risks were removed), and
- we assume that removing one cause-specific hazard does not change or affect the other cause-specific hazard.

6.4 Cause-specific hazard

Although using a cause-specific outcome creates problems for Kaplan-Meier estimates of survival, the hazard function associated with this cause-specific outcome remains useful, as we will now discuss in detail. Let us consider a general competing risks set-up where K types of failure are possible. For each individual, the observed data are a time of (failure or censoring) $T = min(T, C)$ and the corresponding cause of failure, D (set $D = 0$ if censored). From these observed data, we can define the cause k-specific hazard function defined as:

$$h_k(t) = \lim_{\delta t \to 0} \frac{Pr(t \leq T < t + \delta t, D = k \mid T \geq t)}{\delta t} \tag{6.1}$$

This can be interpreted as the instantaneous rate of dying from cause k given the individual is alive at time t (i.e. they have not died of any cause before time t).

We can also define the cumulative cause-specific hazard by:

$$H_k(t) = \int_0^t h_k(u)du \tag{6.2}$$

And the "survivor function"

$$S_k(t) = \exp(-H_k(t)) \tag{6.3}$$

However, the interpretation of this as the survival function only holds in the hypothetical world and under the assumption detailed above. With these assumptions, it represents the marginal survival in the hypothetical situation where competing events have been removed – the individuals dying from other causes would have had the same future risk of the event of interest as those who are alive at time t. There is no way of testing this (it is an untestable assumption).

From formula 6.1, one can see that the overall hazard is the sum of all cause-specific hazards, that is $h(t) = \sum_{k=1}^{K} h_k(t)$

$$h(t) = \sum_{k=1}^{K} \lim_{\delta t \to 0} \frac{Pr(t \leq T < t + \delta t, D = k \mid T \geq t)}{\delta t} = \lim_{\delta t \to 0} \frac{Pr(t \leq T < t + \delta t \mid T \geq t)}{\delta t} \tag{6.4}$$

It follows that the overall survival can be written as useful application of this cause-specific "survivor function". By calculating the product over all causes of these cause-specific "survivor functions", we obtain the probability of not having failed from any cause at time t:

$$S(t) = \exp(-\sum_{k=1}^{K} H_k(t)) = \prod_{k=1}^{K} \exp(-H_k(t)) \tag{6.5}$$

6.5 Cumulative incidence function

In the competing risks situation, the way to describe the data is to estimate the cumulative incidence function $I_k(t)$, which corresponds to the probability of failing from cause k before time t. For continuous distributions, we have:

$$I_k(t) = P(T \leq t, D = k) = \int_0^t h_k(u) S(u) du \tag{6.6}$$

For discrete distributions, the relation is:

$$I_k(t) = P(T \leq t, D = k) = \sum_{t_l \leq t} h_k(t_l) S(t_l -) du \tag{6.7}$$

From this expression, one can see that this probability involves all cause-specific hazards through the overall survival in 6.6. Intuitively, this quantity can be understand as follows: the probability that an event of type k occurs at some time s is the product of (i) surviving up to time s (which is the survival probability $S(s)$) and then die from cause k at time s (which is the cause-k-specific hazard $h_k(s)$). The cumulative sum of these probabilities over times between 0 and t gives the cumulative incidence function $I_k(t)$. Notice that the Cumulative Incidence Function (CIF) has been sometimes also called the Absolute cause-specific Risk or the Crude Probability of event in the medical statistics literature. Why the CIF leads to different estimates than the "naïve" Kaplan Meier approach in this context? It differs from the naïve Kaplan-Meier estimate of the probability of failing from cause k before or at time t (given by $1 - S_k(t)$) in its use of $S(s)$; i.e. it incorporates the probability of surviving from all causes up to time t, rather than just the probability of surviving from cause k. In doing this, the CIF

accounts for the fact that those dead from competing risks causes are no longer at risk of event k.

6.5.1 Non-parametric estimation of the CIF

Using the prostate cancer example, we can calculate a non-parametric estimate for the CIF, $\hat{I}_k(t)$, for cancer death and other-cause death. Non-parametric estimation of the constituent parts is analogous to standard calculations for $h_j(t)$ and $S(t)$ (see Equations 2.3 and 2.5).

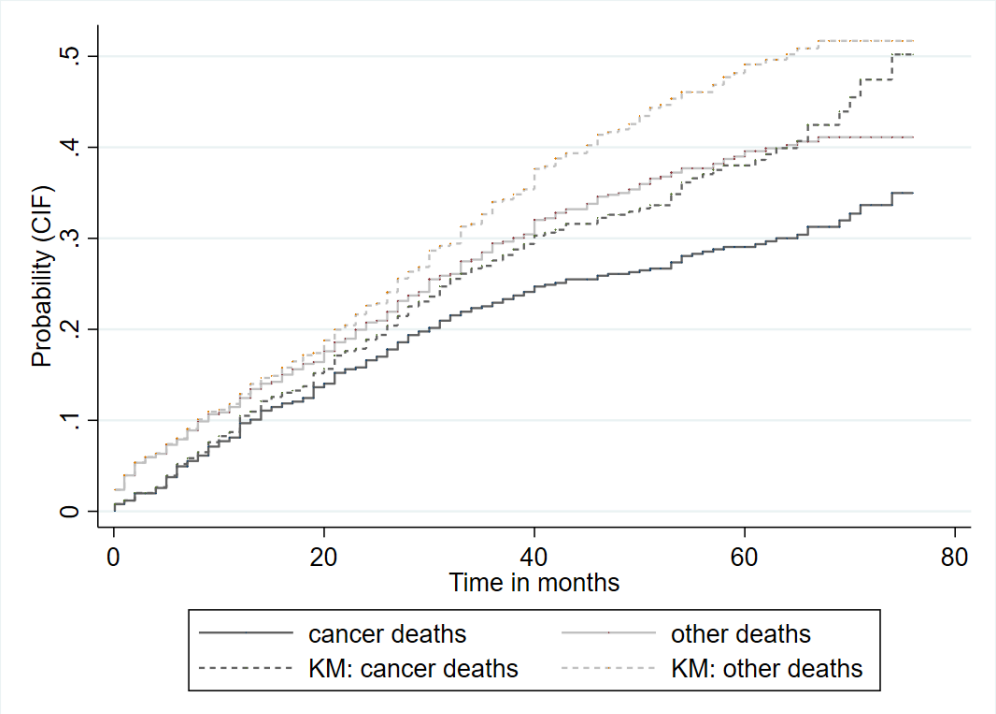
Exercise 6.1 Complete the table below

Time in months (t_j)	Cancer death (d_{1j})	Other death (d_{2j})	Cens.	n_j	d_j	$\hat{h}_1(t_j)$	$\hat{h}_2(t_j)$	$\hat{S}(t_j)$	$\hat{I}_1(t_j)$	$\hat{I}_2(t_j)$
1	2	8	0	490						
2	4	7	0							
3	0	3	0							
4	3	2	4							
5	6	5	0							

Table 6.1: Non parametric estimation of the CIF

We can then plot these alongside the crude KM-based estimates for each cause treating the other as censored:

Figure 6.3: Complement of Kaplan-Meier and non-parametric cumulative-incidence plots for prostate cancer example.



In Stata:

```
* Recoding CVD deaths and other deaths both as Other
. replace status=2 if status==3

. stset time, f(status==1) id(id)

. stcompet cuminc = ci hilim = hi lowlim = lo, compet(2)

. sort time
. gen cuminc1=cuminc if status==1
. gen cuminc2=cuminc if status==2

* Data management step to make the plot nicer:
* CIF starts at 0, and ends at the last observed time if people still at risk
. replace cuminc1=0 if cuminc1==. & _n==1
. replace cuminc2=0 if cuminc2==. & _n==1
. replace cuminc1=cuminc1[_n-1] if cuminc1==.
. replace cuminc2=cuminc2[_n-1] if cuminc2==.

. twoway (scatter cuminc1 time, sort c(J) msym(p) lcol(gs6) ///
           legend(lab(1 "cancer deaths" ))) ///
       (scatter cuminc2 time, sort c(J) msym(p) lcol(gs12) ///
       legend(lab(2 "other deaths" )))

* Comparison with 1-KaplanMeier
. sts gen surv1=s
. gen invKM1=1-surv1

. stset time, f(status==2) id(id)
. sts gen surv2=s
. gen invKM2=1-surv2

. twoway (scatter cuminc1 time, sort c(J) msym(p) lcol(gs6) ///
           legend(lab(1 "cancer deaths" ))) ///
       (scatter cuminc2 time, sort c(J) msym(p) lcol(gs12) ///
           legend(lab(2 "other deaths" )))  ///
   (scatter invKM1 time, sort c(J) msym(p) clp(shortdash) lcol(gs6) ///
        legend(lab(3 "KM: cancer deaths"))) ///
   (scatter invKM2 time, sort c(J) msym(p) clp(shortdash) lcol(gs12) ///
        legend(lab(4 "KM: other deaths")) ///
   xtit("Time in months") ytit("Probability (CIF)") graphr(fcolor(white)))
```

In R:

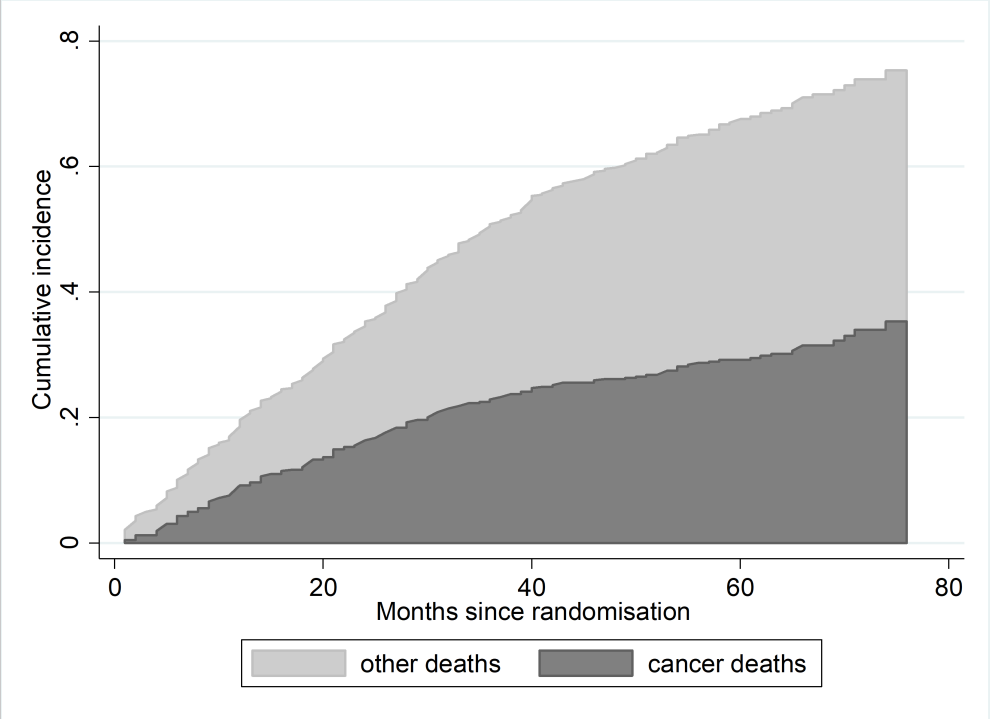
```
> library(cmprsk)
```

```
* Recoding CVD deaths and other deaths both as Other
> prostatecancer$status <- ifelse(prostatecancer$status==3, 2,
                                  prostatecancer$status)

# Non-parametric estimation on the CIF
> CIF <- cuminc(ftime=prostatecancer$time,fstatus=prostatecancer$status)
> plot(CIF, ylim=c(0,0.5), lwd=2)
```

It is also sometimes useful to stack cumulative incidence curves for different causes, particularly when there are more than two causes of interest. The distance between these curves represents the probability of each event, and the sum is equal to 1 minus the overall survival probability.

Figure 6.4: Stacked non-parametric cumulative incidence for prostate cancer example.



In Stata:

```
* stacking non-parametric CIFs
* ie cuminc holds CIF for cancer deaths with other deaths
* as competing risk for status=1 & holds CIF for other deaths
* with cancer deaths as competing risk for status=2
. stset time, fail(status==1) id(id)
. stcompet cuminc=ci, compet(2)
. reshape wide cuminc, i(id) j(status)
. replace cuminc1=cuminc1[_n-1] if cuminc1==.
. replace cuminc2=cuminc2[_n-1] if cuminc2==.
```

```
. recode cuminc1 .=0
. recode cuminc2 .=0
. gen ci_all=cuminc1+cuminc2
. twoway area ci_all time, lcolor(gs12) fcolor(gs12) c(J) ||
area cuminc1 time, lcolor(gs6) fcolor(gs6) c(J)
legend(lab(1 "other deaths") lab(2 "cancer deaths"))
xtit("Months since randomisation")
ytit("Cumulative incidence") tit("") graphr(fcolor(white))
```

In R:

```
> CIF <- cuminc(ftime=prostatecancer$time,fstatus=prostatecancer$status)

> timep <- sort(unique(c(CIF$`1 1`$time, CIF$`1 2`$time)))
> CIFpred <- timepoints(CIF, times=timep)
> CIFAll <- CIFpred$est[1,]+CIFpred$est[2,]
> plot(timep, CIFpred$est[1,], ylim=c(0,1), lwd=2, type="s",
       panel.first=abline(h=seq(0,1,0.1), lty=8, col="grey"),
       xlab="Time in Month", ylab="Cumulative incidence",
       main="CIF, stacked format")
> lines(timep, CIFAll, lwd=2, type="s", lty=8)
```

6.6 Explanatory variable effects

6.6.1 Log-rank test for comparison of CIFs

The most common log-rank test equivalent for cumulative incidence functions is a test devised by Gray (1988), although it has (to date) not been implemented in Stata (note: it is available in R through the `cuminc` function). An alternative test for testing equality of CIFs between groups was proposed by Pepe and Mori (1993). This is available in Stata using the downloadable function `stpepemori`. Applying this test to the prostate cancer data gives p=0.01 for equality of the CIFs for cancer death by treatment group. We do not give the details of these tests here. However, the main aim of explanatory variable analysis lies in quantifying the effects of these variables, which requires a regression approach.

6.6.2 Regression analysis: the problem with CIF

Since the cumulative incidence function depends on all the cause-specific hazard functions (see equations 6.5 and 6.6), covariables may therefore have a different association with this CIF function than with the cause-specific hazards. In other words, the standard approach of modelling of the hazard (as in Cox models) in order to make inferences about the relationship between covariables and the cumulative incidence is no longer possible because the 1-1 relationship between hazard and failure probability (risk) is lost.

6.6.3 Subdistribution hazard

One way to get around the problem of different associations of variables with cause-specific hazards and with cumulative incidence is to model directly the relationship of covariables with the CIF. This model was originally proposed by Fine and Gray (1999), and is sometimes referred to by this name. This approach uses an alternative definition of the hazard, called the subdistribution hazard, which represents the instantaneous rate of dying from cause k in the small interval $[t, t + \delta t]$ given that an individual has not already died from cause k, that is

$$h_k^s(t) = \lim_{\delta t \to 0} \frac{Pr(t \leq T < t + \delta t, D = k \mid T \geq t \text{ or } (T \leq t, D \neq k))}{\delta t} \tag{6.8}$$

This differs from the cause-specific hazard in its risk set; here individuals are not removed from the risk set if they die from another competing cause of death than cause k. So the risk set for the subdistribution hazard includes patients that are alive but also those who died from non-k causes before time t (this differs from the risk set for the cause-specific hazard function, which only includes those who are currently event free).

The subdistribution hazard is related to the CIF via the following one-to-one relationship:

$$h_k^s(t) = -\frac{d}{dt} log(1 - I_k(t)) \tag{6.9}$$

6.6.4 Semi-parametric estimation of covariable effects through subdistribution hazards

The effects of explanatory variables on the CIF can be modelled using the following parametrisation of the subdistribution hazards:

$$h_k^s(t; x) = h_{0,k}^s(t) \exp(\beta_k^\top x) \tag{6.10}$$

where x is a vector of explanatory variables and β_k is a vector of parameters to be estimated – the ratios of the subdistribution hazards.

Returning to the prostate cancer example, we can fit a subdistribution hazard model for the effect of treatment; this estimates the subdistribution hazard ratio (SHR) as 0.64 (95 % CI 0.46 to 0.89, p=0.007). The interpretation of this is that treatment decreases the instantaneous cancer death rate by 36%, amongst those who have not yet died from cancer (i.e. including those who have died from other causes). This is a problematic interpretation as it requires the assumption that those dying from other causes remain in the risk set for the event of interest indefinitely (since they will never actually have this event). However, recall that the purpose of setting up this model is to make inferences about the CIF.

Substituting equation 6.10 into equation 6.9, the CIF is estimated as:

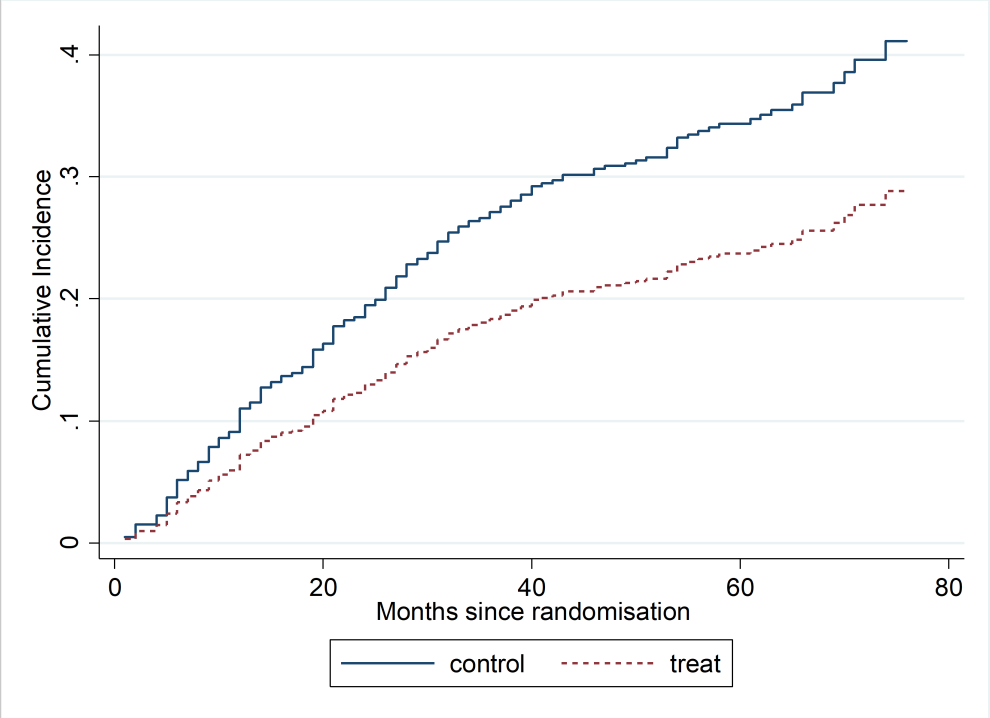
$$\hat{I}_k(t;x) = 1 - \exp(-\int_0^t \hat{h}^s_{0,k}(u)\exp(\hat{\beta}_k^\top x)du)) \tag{6.11}$$

Rearranging $h^s_{0,k}(t) = -ln[1 - I_k(t;0)]$ and $h^s_{0,k}(t)\exp(\beta_k^\top) = -ln[1 - I_k(t;1)]$ the relationship between the CIFs in the two treatment groups is given by:

$$1 - I_k(t;1) = [1 - I_k(t;0)]^{\exp(\beta_k)} \tag{6.12}$$

Semi-parametric estimates for the CIF can then be calculated and plotted in the same way as survivor curves following Cox regression (Figure 6.5). These plots can also be compared to non-parametric estimates as a basic indicator of goodness-of-fit.

Figure 6.5: Parametric cumulative incidence estimates by treatment group, for prostate cancer example.



In Stata:

```
* semi-parametric estimation of CIF
. stcrreg treat, compet(status==2)
. stcurve, cif at(treat=0) at(treat=1) graphr(fcolor(white)) tit("")
clp(solid shortdash) leg(lab(1 "control") lab(2 "treat"))
xtit("Months since randomisation")
```

In R:

```
> subhaz <- crr(ftime=prostatecancer$time,fstatus=prostatecancer$status,
              cov1=prostatecancer$treatment, failcode = 1)
> predCIF <- predict(subhaz, cov1=c(0,1))
> plot(predCIF, lty=c(1,8), color=2:3,
      xlab="Months since randomisation", ylab="Cumulative Incidence",
      panel.first=abline(h=seq(0,1,0.1), lty=8, col="grey"))
> legend("bottomright", c("control", "treatment"), lty=c(1,8),
        lwd=2, col=2:3)
```

6.6.5 Semi-parametric estimation of covariable effects through cause-specific hazards

It is important to note that estimation of effects on cause-specific hazards is a valid approach with interpretable results. The problem only lies in translating these results onto the cumulative incidence scale. It is therefore generally recommended that the subdistribution hazard approach is the best choice for modelling with view to predicting competing-event outcomes, whereas the cause-specific hazard approach is the appropriate choice for modelling with view to estimating effects of explanatory variables in an etiological sense. Fitting a cause-specific hazards model to the prostate cancer example, we get a cause-specific hazard ratio of 0.65 (95% CI 0.47 to 0.90, p=0.009). This can be interpreted as the treatment reduces the cancer death rate by 35% amongst those who are eligible for this event, i.e. those who have not yet died of any cause. The key difference in interpretation here compared to the interpretation of the subdistribution hazard ratio is in the risk set. This can be seen in figures 6.6 and 6.7, where triangles represent the event of interest (cancer death), circles represent competing risk events (other deaths) and crosses represent end-of-follow-up censoring:

Exercise 6.2 Complete the table, calculating the cause-specific hazard (CSH) and the subdistribution hazard (SH) for cancer death at each time-point shown in figures 6.6 and 6.7:

Time	1	2	3	4	5	6	7	8
CSH								
SH								

Table 6.2:

Figure 6.6: Risk sets for cause-specific hazards for a subset of 6 patients in the prostate cancer example.

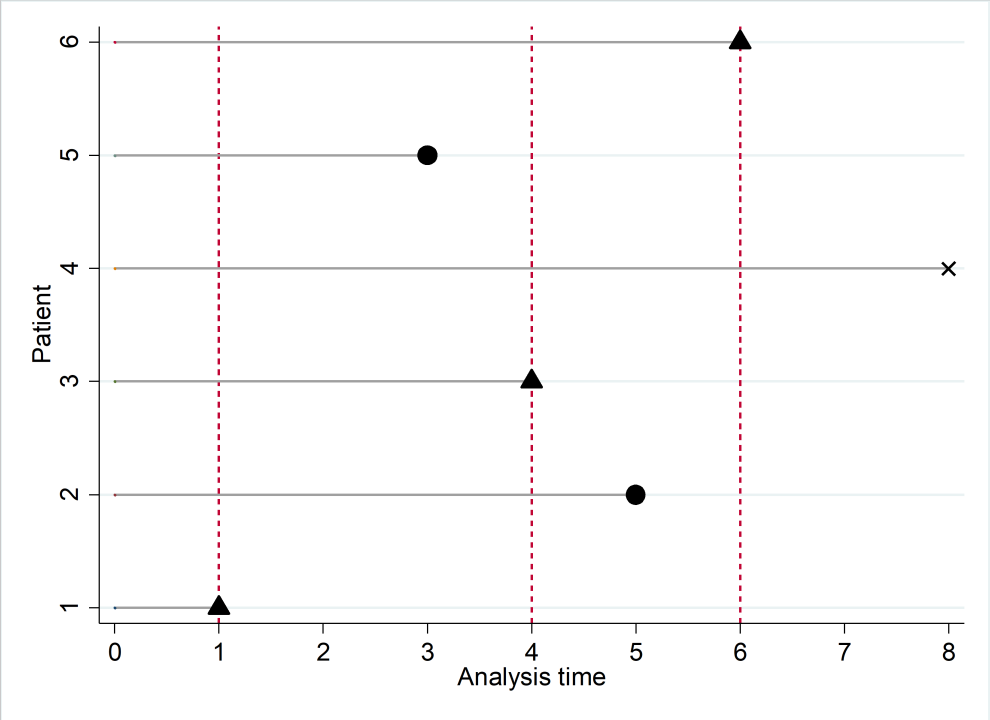
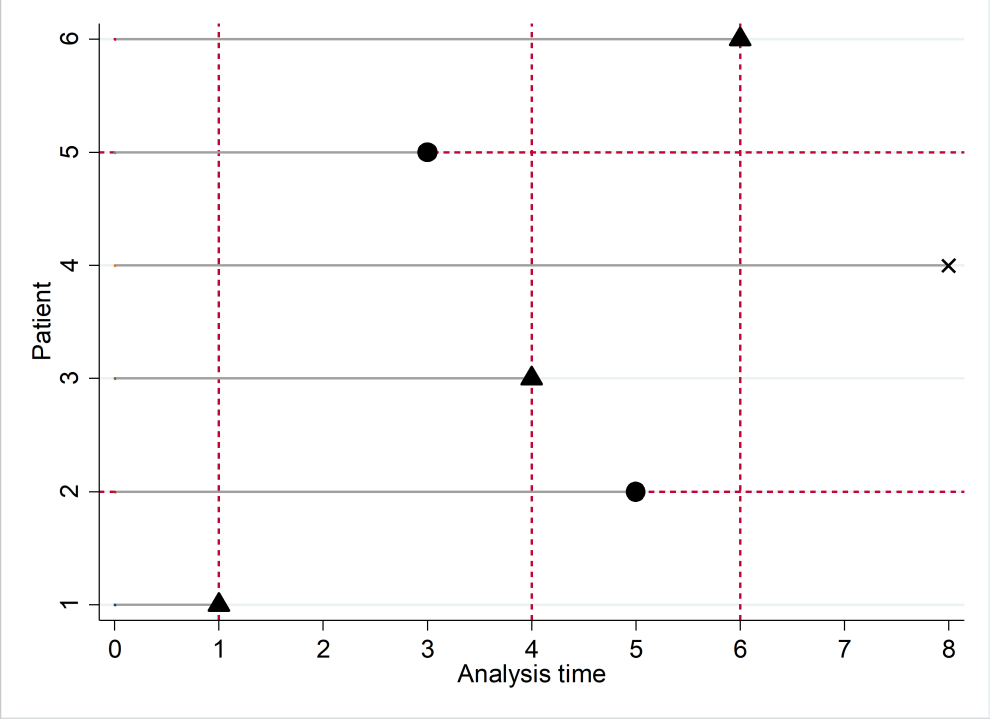


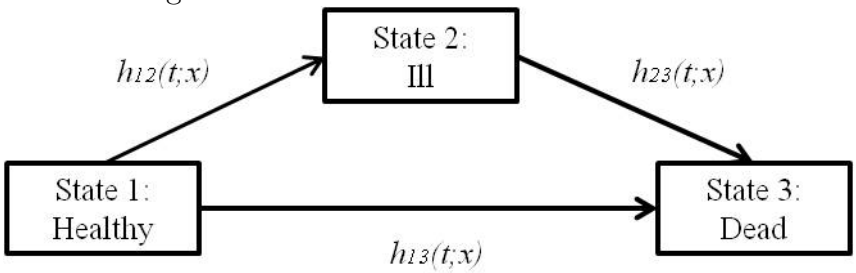
Figure 6.7: Risk sets for subdistribution hazards for a subset of 6 patients in the prostate cancer example.



6.7 Multi-state models

So far we have only considered first events happening, therefore considering them as if they were 'absorbing events' – competing risks events that are final (usually this means a death) and thus prevent the event of interest from subsequently taking place. In other examples, intermediate event may be of interest, and may occur before the absorbing event; they are called non-absorbing event. Taking account of these intermediate events is of particular importance where their occurrence substantially changes the hazard of the event of interest. The simplest example of a multi-state model is the so-called illness-death model, which consists of just three states (see Figure 6.8). More complex models can also be fitted, which allows more detailed modelling of specific illnesses (see the example below on breast cancer relapse).

Figure 6.8: Illness-death multi-state model.



The hazard rates defining movement from one state to another are referred to as transition intensities – the instantaneous rate of moving from state i to state j at time t; these are indicated on the figure as $h_{ij}(t;x)$. These transition intensities are equivalent to the cause-specific hazards described previously, the competing risks situation being a particular case of multi-state models. This example only permits transitions in one direction and is known as a uni-directional model. However, it is also possible to model transitions in both directions (for example, modelling recovery from the illness state) – this is then called a bi-directional model.

Notation

All states are numbered, $1, \ldots, S$. Define a random process $X(t); t \geq 0$, taking values 1 to S. The history of the process until s is given by: $\mathfrak{F}_s = X(u); 0 \leq u \leq s$. We can then define the transition rate from state i to state j, which for some transitions may be 0 for all t, where T denotes the time of reaching the state j from state i:

$$h_{ij}(t) = \lim_{\delta t \to 0} \frac{Pr(t \leq T < t + \delta t \mid T \geq t)}{\delta t} \tag{6.13}$$

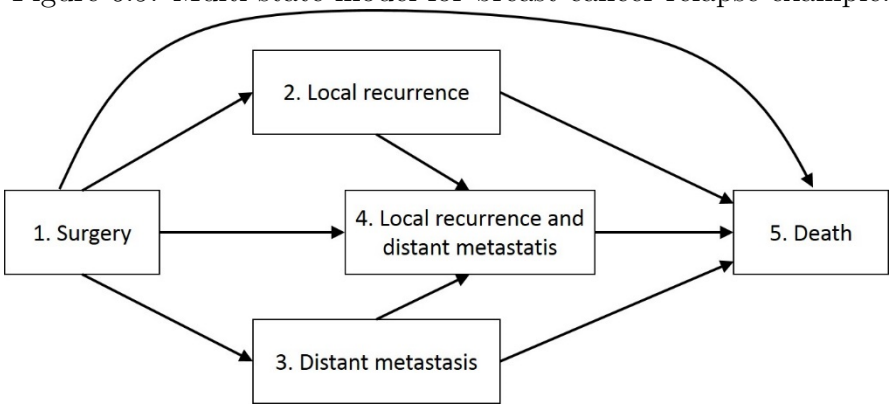
The time-scale may be either counted from the time of entry into the starting state ("clock forward") or from the time of entry into the current state ("clock reset"). The choice between these is largely determined by clinical context, although the difference in outputs is often small.

Example 6.2

Breast cancer relapse

The example we will be using throughout this section is of relapse in breast cancer. In this example, people who had received treatment for breast cancer were followed up from the time of initial tumour removal for 60 months. There were two events of interest: local relapse at an old tumour site and distant relapse at a new site. Deaths were also recorded, and may occur with or without a prior relapse. In this example we focus just on the control group, which comprises 248 individuals. The competing risks of interest in this study are not absorbing events – that is, a person could have a local recurrence and then go on to have a distant recurrence later, or vice versa. Only the first of these events was recorded in this study, therefore what we are studying is the time to the first recurrence event, which could be either a local or distant recurrence. The relapse events are also intermediate events that would impact on the future hazard of death. These types of data can be modelled using multi-state models. Using the example described above, a multi-state model for this study could be set up as follows, see 6.9 (Figure from Putter, 2007):

Figure 6.9: Multi-state model for breast cancer relapse example.



It is common for multi-state models to have a single starting state (as shown in this example), defined by entry criteria into the study. Here, individuals undergoing surgery are entered in to the starting state for the study. The state labelled "surgery" here is actually a state representing individuals who are alive post-surgery and without relapse.

6.8 Markov models

6.8.1 The Markov assumption

A common assumption for multi-state models is that upon entering a particular state i, individuals are subject to common transition rates for movement to state j, irrespective of their history, $\mathfrak{F}$. In other words, we assume that the transition rate does not differ

according to the previous states an individual has been in. This is called the Markov assumption, and is often quite a strong assumption to make. For example, if we extend the breast cancer study to permit recovery from relapse to the healthy "post-surgery" state, the Markov assumption would mean that we assume the future hazard of relapse is the same for those who have never had a relapse as those who have had a past relapse:

$$P(X(t) = j \mid X(s) = i, \mathfrak{F}_{s^-}) = P(X(t) = j \mid X(s) = i) \tag{6.14}$$

6.8.2 Extended Markov models

In practice, it is often not very reasonable to make the Markov assumption. In our breast cancer example, we may think that individuals who have spent longer in the local relapse state have an increased hazard of developing a distant relapse. We may also think that for those with both local and distant relapse, that the order that these occur influences the hazard of death. One way around the problem of making unrealistic Markov assumptions is to duplicate a state to reflect different histories. This means that the Markov assumption will hold for each individual version of this state. For example, we could create two different states for relapse – one for primary relapse and one for secondary relapse – these states could potentially have different transition rates leading to death. Such models are referred to as extended Markov models.

6.8.3 Other assumptions

Multi-state models may either be time-homogeneous or non-homogeneous. In the former, the transition rate does not depend on t, i.e. $h_{ij}(t) = h_{ij}$ for all t. Conversely, for non-homogeneous models, transition rates may change over time. Time non-homogeneity may be modelled with parametric distributions, such as Weibull or using Cox modelling (including adjustment for covariables). Semi-Markov models are those that allow the transition rate to depend on the amount of time spent in the state (known as the sojourn time).

6.9 Transition probabilities

A key output from multi-state modelling is the estimation of transition probabilities for moving from state i at time s to state j at time t (equivalent to survivor function estimation for multi-state models). For the simple (Markovian) illness-death model described in Figure 6.8, the transition probabilities have simple explicit expressions:

$$P_{00}(s,t) = \exp[-\int_s^t (h_{02}(u) + h_{01}(u))du] \tag{6.15}$$

$$P_{01}(s,t) = \int_s^t P_{00}(s,u^-)h_{01}(u)P_{11}(u,t)du] \tag{6.16}$$

$$P_{11}(s,t) = \exp[-\int_s^t h_{12}(u)du] \tag{6.17}$$

Non-parametric estimators of these are available using a generalisation of the Nelson-Aalen estimator (called the Aalen-Johansen estimator), although we do not give details here.

6.10 Incorporating explanatory variables

We can use Cox regression modelling for each transition separately by using appropriate subsets of the data. This will provide estimates for the effect of covariables on the transitions. We could also combine the data into one Cox model and stratify on transition, thus estimating a common effect for the covariables, although this may not always be a reasonable assumption. Note that as with competing risks, the relationship of covariables with transition intensities are not directly related to the effects of these covariables on the transition probabilities. The transition intensity for transition i to j is given by:

$$h_{ij}(t;x) = h_{ij,0}(t)\exp(\beta_{ij}^\top x) \tag{6.18}$$

So that in a stratified model for a 3-state uni-directional model, we have:

$$h_{12}(t;x) = h_{12,0}(t)\exp(\beta_{12}^\top x) \tag{6.19}$$

$$h_{13}(t;x) = h_{13,0}(t)\exp(\beta_{13}^\top x) \tag{6.20}$$

$$h_{23}(t;x) = h_{23,0}(t)\exp(\beta_{23}^\top x) \tag{6.21}$$

Example 6.3

Bone marrow transplantation

This study comprises 2204 patients in the European Blood and Marrow Transplant Registry who had received a bone marrow transplant between 1995 and 1998. Three states were considered:

- State 1: Transplanted
- State 2: Platelet recovery (meaning that blood platelet levels returned to normal after transplant)
- State 3: Relapse or death (this may occur from either State 1 or State 2)

Using these data, we can set up Cox model to investigate the effect of some explanatory variables on these transition rates. Entry is delayed into the relapse states if we use a "clock forward" approach (this is achieved in Stata using the `enter()` option in stset, and in R using the `start` and `stop` notations). Results are shown below for just two

explanatory variables; age at transplant and whether the donor was the same sex as the transplant recipient.

A model stratified on transition was fitted first:

		Parameter estimate	95% Conf Int
Age at transplant	≤ 20	Reference	-
	20-40	-0.050	[-0.17;0.071]
	>40	0.145	[0.02;0.27]
Donor recipient	No sex mismatch	Reference	-
	Sex mismatch	0.042	[-0.058;0.142]

Table 6.3: Results from a Cox model stratified by transition

Further models were then fitted for each transition (see Putter *et al.* (2007) for the full results):

		Parameter estimate	95% Conf Int
Transition from State 1 to 2			
Age at transplant	≤ 20	Reference	-
	20-40	-0.183	[-0.333;-0.034]
	>40	-0.139	[-0.298;0.021]
Donor recipient	No sex mismatch	Reference	-
	Sex mismatch	0.039	[-0.09;0.17]
Transition from State 1 to 3			
Age at transplant	≤ 20	Reference	-
	20-40	0.221	[-0.064;0.507]
	>40	0.476	[0.183;0.768]
Donor recipient	No sex mismatch	Reference	-
	Sex mismatch	-0.074	[-0.290;0.142]
Transition from State 2 to 3			
Age at transplant	≤ 20	Reference	-
	20-40	0.119	[-0.173;0.412]
	>40	0.677	[0.388;0.967]
Donor recipient	No sex mismatch	Reference	-
	Sex mismatch	0.192	[-0.032;0.416]

Table 6.4: Results from different Cox models used for each transition

Write down the models fitted, any assumptions made and interpret the results

In Stata:

```
* data manipulation
* use ebmt3, clear
. gen trans=12
. append using ebmt3
. recode trans .=13
. append using ebmt3
. recode trans .=23
. gen timein=0
. gen timeenter=0 if trans!=23
. replace timeenter=prtime if trans==23
. gen timeout=prtime if trans==12
. replace timeout=rfstime if trans==13 & prstat==0 | trans==23
. replace timeout=prtime if trans==13 & prstat==1
. gen fail=1 if prstat==1 & trans==12
. replace fail=1 if rfsstat==1 & trans==13 & prstat==0 | rfsstat==1 &
  trans==23 & prstat==1
. recode fail .=0

* analysis
* Model stratified on transition
stset timeout, fail(fail) origin(timein) enter(timeenter)
stcox i.age i.drm, strata(trans) nohr

* Model for each transition
stcox i.age i.drm if trans==12, nohr
stcox i.age i.drm if trans==13, nohr
stcox i.age i.drm if trans==23, nohr
```

In R:

```
# Notice that we re-used the dataset prepared in stata

# Model stratified on transition
> coxph(Surv(timeenter, timeout, fail) ~ strata(trans) + as.factor(age) +
      as.factor(drmatch),data=ebmt3, ties="breslow")

# Model for each transition
> coxph(Surv(timeenter, timeout, fail)~ as.factor(age)+as.factor(drmatch),
      data=ebmt3, ties="breslow", subset=trans==12)
> coxph(Surv(timeenter, timeout, fail)~ as.factor(age)+as.factor(drmatch),
      data=ebmt3, ties="breslow", subset=trans==13)
> coxph(Surv(timeenter, timeout, fail)~ as.factor(age)+as.factor(drmatch),
      data=ebmt3, ties="breslow", subset=trans==23)
```

References

Thorough but accessible introduction to a wide range of competing risks topics, including multi-state modelling:

- Putter et al. Tutorial in biostatistics: competing risks and multi-state models. Stat Med 2007.

Intuitive, less technical introduction papers to the concepts underpinning competing risks analysis:

- Noordzij et al. When do we need competing risks methods for survival analysis in nephrology? Nephrol Dial Transplant 2013.
- Andersen et al. Competing risks in epidemiology: possibilities and pitfalls. Int J Epi 2012.

Introduction papers to issues in multi-state modelling, including technical details:

- Keiding. Event history analysis. Ann Rev Stat Appl. 2014
- Andersen et al. Interpretability and importance of functionals in competing risks and multistate models. Stat Med 2012.

Classic competing risks publications:

- Fine and Gray. A proportional hazards model for the subdistribution of a competing risk. JASA 1999.
- Gray. A class of K-sample tests for comparing the cumulative incidence of a competing risk. Ann Stat 1988
- Pepe and Mori. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? Stat Med 1993.

Practical 6

Dataset required: `aaatrial_2016`

Stata command for Stata users: stcompet (ssc install stcompet).

R packages required for R users: survival, Epi, cmprsk.

Introduction

This session looks at the issue of competing risks in survival analysis. You will use Stata/R to estimate cause-specific hazards and cumulative incidence functions. There is also an optional section on multi-state modelling.

We will use a dataset from a randomised trial of screening for abdominal aortic aneurysm (AAA), where older men were either invited to be screened (invited group) or not contacted (control group). The outcome of interest was deaths relating to AAA (including deaths following repair operations), but men in the study also died from a range of other causes.

If an AAA was found at screening, a repair operation was carried out, essentially removing the risk of future death from AAA. However, the operation itself carries around a 5% risk of mortality. Undetected AAAs may rupture at any time, resulting in either an emergency repair operation (which carries around a 40% risk of mortality) or death. Men were followed up for 8-10 years, until the end of 2007.

The key variables are described below.

Variable	Description
`id`	Unique identifier for each participant
`group`	Randomisation group: 0=Control, 1=Invited
`dateran`	Date of randomisation
`aaadeath`	AAA-death indicator: 0=Censored, 1=AAA-death
`alldeath`	All-cause mortality: 0=Censored, 1=Died
`deathtype`	0=Censored; 1=non-AAA death; 2=AAA-death
`timeout`	Date of study exit

Aims

By the end of this session you should be able to:

- Estimate, plot and interpret the cumulative incidence functions
- Fit and interpret results from cause-specific hazard models
- Fit and interpret results from subdistribution hazard models

Where code examples are given or explanations are given that are specific to Stata or R, text and code relating to Stata is shown in this colour and text and code relating to R is shown in this colour.

Questions

1. Summarise the data using some basic descriptive analyses to familiarise yourself with the dataset. How many people died from any cause, and how many had an AAA-death and non-AAA-death? How many people were in the two randomisation arms?

2. We will begin by focusing on death from any cause. The time scale for the analysis is time-in study (measured in years).

   In Stata `stset` the data using time-in-the-study as the time scale and `alldeath` as the failure indicator.

   In R calculate the time to death from any cause:
   `aaatrial$futime <- as.numeric(aaatrial$timeout-aaatrial$dateran)/365.25`

3. (a) Plot the Kaplan-Meier survival curves for all-cause mortality by randomisation arm.

   (b) Perform a log rank test of whether the survival curves differ by randomisation arm. Recall that you can do this in Stata using `sts list` and in R using `survdiff`.

   (c) Fit a Cox model for all-cause mortality with randomisation arm as the only explanatory variable and interpret the results.

   Discuss: What is the evidence for any benefit to inviting these men to screening?

4. We will now move on to focusing on AAA-deaths. From your investigations of the data in question 1, you should have seen that there are also a large number of non-AAA deaths in this study, which is a competing risk for AAA-death.

   (a) We first explore how to use stset in Stata and `Surv` in R when there are competing risks.

   In Stata re-stset your data using the `deathtype` failure variable, indicating AAA-death as failure. Study the output from `stset` and how it differs from that in question 2.

   In R use `survfit(Surv(time=futime,event=aaadeath) 1, data=aaatrial)` and investigate the output. What information is given for (i) people who have an AAA-death, (ii) people who have a non-AAA-death, (iii) people who are censored?

   (b) Estimate the non-parametric cumulative incidence functions for each randomisation arm separately.

   In Stata cumulative incidence functions can be obtained using a user-contributed program called `stcompet`. Note that you will have to use `stcompet` for each arm separately, defining the value of your failure variable for each competing risk, e.g.: `stcompet mycuminc1=ci if group==0, compet1(1)`

```
stcompet mycuminc2=ci if group==1, compet1(1)
```
Investigate the dataset and the columns `mycuminc1` and `mycuminc2`

In R the `cuminc` function from the R-package `cmprsk` can be used to estimate the cumulative incidence functions. Look at the help file from `cuminc` and make sure you understand its first 3 arguments. Run
```
cumincfit1 <- cuminc(ftime=aaatrial$futime, fstatus=aaatrial$deathtype,
group=aaatrial$group)
```
Inspect the cumincfit1 object and make sure you understand the output. (You might want to cross-tabulate `deathtype` and `group` for help in interpreting correctly the results `table(aaatrial$deathtype, aaatrial$group)`

(c) Plot the cumulative incidence functions for AAA deaths. Interpret the plot. What is the probability of an AAA-death within 5 years in the two randomisation arms?
In Stata :
```
twoway line cuminc1 _t if deathtype==2, connect(step)
sort lcolor(black) || line cuminc2 _t if deathtype==2, connect(step)
sort lcolor(black) lpattern(dash)
legend(lab(1 "cuminc, controls") lab(2 "cuminc, invited"))
```

In R:
```
plot(cumincfit1, lwd=2, col=1:2, lty=1:4, ylim=c(0,0.4),
curvlab = c(paste0(levels(aaatrial$group), " Non-AAA death "),
paste0(levels(aaatrial$group), " AAA Death")),
panel.first=abline(h=seq(0,1,0.1), col="grey", lty=2))
```

5. Next we consider a cause-specific Cox regression analysis for AAA-death. Fit the cause-specific Cox model with randomisation arm as the explanatory variable. What is the interpretation of the cause-specific hazard ratio in the presence of the competing event of non-AAA death?

6. We will now carry out a competing risks analysis based on the subdistribution hazard, still focusing on AAA-death.

   (a) We begin by fitting the subdistribution hazard model.

   In Stata the subdistribution hazard model can be fitted using `stcrreg`. Fit the model with randomisation arm as the explanatory variable.

   In R the subdistribution hazard model can be fitted using the function `crr` from the R-package `cmprsk`:
   ```
   crrfit1 <- crr(aaatrial$futime, aaatrial$deathtype, cov1=aaatrial$group,
   failcode=2)
   ```

   Interpret your results.

   (b) How does the interpretation of the results differ from your interpretations based on the cause-specific hazards model in the previous question.

(c) Lastly, we will use the subdistribution hazard model to obtain an estimate of the cumulative incidence curves in each randomisation arm.

In Stata you can use the `predict` command to estimate the baseline cumulative incidence function in the control arm, then create another variable with the semi-parametric estimate of the CIF in the screened arm (Hint: use the subdistribution Hazard ratio and formula 6.12 in the lecture notes). To find the estimated baseline cumulative incidence:

```
predict newvar, basecif
```

In R we can use of the `predict` function associated with the `crr` function to obtain the cumulative incidence functions for AAA-death in both randomisation arms:

```
mypredCIF <- predict(crrfit1, cov1=c(0,1))
```

As an additional exercise, predict the CIF for the reference group, and use the subdistribution hazard ratio and formula 6.12 in the lecture notes to derive the CIF for the screened group

(d) Check your calculations by comparing your plots with the functions produced from `stcurve` or `plot(mypredCIF)`.

(e) Compare your estimated cumulative incidence curves from the subdistribution hazard modelling with those obtained using the non-parametric analysis. What assumption did you make in the subdistribution hazard analysis that you did not make in the non-parametric analysis? How would you investigate this assumption?

Discuss: What do you conclude about the effect of inviting men to screening on the death from AAA?

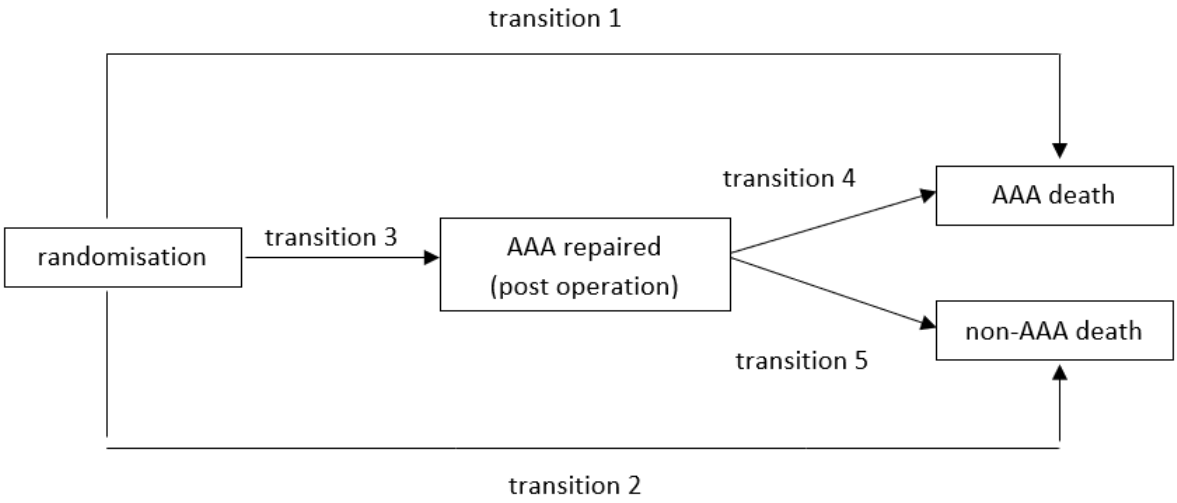
7. Repeat the investigations for non-AAA-death as the cause of interest, with AAA-death treated as a competing risk. How does randomisation arm affect non-AAA-deaths.

Discuss: How would you summarise the results from this trial? Write a short paragraph suitable for the abstract of a journal paper, to summarise the statistical methods and results from this study.

Optional section on multi-state modelling

We recommend using the code provided for this section.

1. We have so far modelled competing risks into the absorbing states for different causes of death. However, some men in the trial had an operation to repair their AAA, which substantially influences their future survival prospects. We would like to include this information in the investigation. Here is a basic multi-state model showing the progress of individuals post-randomisation:



Split your data so that individuals having an operation have separate records pre- and post-operation.
In Stata, you will need to use `stsplit postop=opdate, at(0)` and then do a recoding of the new postop variable to prepare the data for the subsequent sections:

```
replace postop=postop*-1
recode postop 0=.  if _t0!=0
```

In R, you will need to calculate the time to operation and then use `Lexis` and `cutLexis` functions from the `Epi` package.

```
 > idx <- as.numeric(aaatrial$timeout-aaatrial$opdate)
> aaatrial$timeop <- as.numeric(aaatrial$opdate-aaatrial$dateran)/365.25
> aaatrial[idx==0&!is.na(idx),]$futime <- aaatrial[idx==0&!is.na(idx),]$futime
+ 0.5/365.25
> mydat <- Lexis(entry.status="Rand", id=aaatrial$id,
exit = list(tft = futime),
exit.status = factor(alldeath, labels = c("Censor", "Dead")),
data = aaatrial)
> mydattr <- cutLexis(mydat, cut=mydat$timeop,
precursor.states="Rand",
new.state="Oper")
```

Make sure you understand the new form of the data.

2. Investigate the effect of randomisation arm on transition 3 (the rate of progression to an operation after randomisation), after accounting for competing risks.

   In Stata: Note that you will need to re-stset your data, carefully defining the origin and failure indicator to account for competing risks corresponding to the other possible transitions after randomisation.

   In R: Note that you will need to or to subset your data to analyse only patients "at

risk" for operation, and use the `Surv(Start, Stop)` format, carefully defining the origin and failure indicator to account for competing risks corresponding to the other possible transitions after randomisation.

(a) Which group has a higher hazard for receiving an operation? Why might this be?

(b) Investigate whether the effect of group on this transition changes over time. What is the problem with modelling any time dependency in the relationship?

Session 7

Time dependent variables and frailty models

7.1 Aims of this lecture and practical

At the end of this lecture and practical you will be able to:

- Give examples of time-dependent explanatory variables.
- Distinguish between external and internal time-dependent variables
- Write down equations showing how time-dependent explanatory variables are incorporated into the Cox model
- Describe the form of the survival data when there are time-dependent explanatory variables.
- Interpret the results from fitting a survival model with time-dependent explanatory variables.
- Fit survival models with time-dependent explanatory variables in Stata and R.
- Distinguish between unshared and shared frailty models
- Write down the form of survival models that include frailty terms, to include both unshared frailty models and shared frailty models, and outline in broad terms how such models are fitted.
- Interpret the results of both unshared and shared frailty models.
- Fit frailty models in Stata and R.

7.2 Introduction to time-dependent variables

It is quite common for us to be interested in explanatory variables which change over time in an individual. We refer to these as time-dependent variables. They are also sometimes referred to as time-updated variables and time-varying variables. Examples are:

- Consider a study of factors associated with survival in an observational study of individuals with cystic fibrosis using data from a patient registry. Over the course of follow-up of the cohort, some individuals have a lung transplant and this has an important impact on survival. Here the time dependent variable is 'transplant status' (a binary variable) and individuals can only move in one direction – from untransplanted to transplanted.

- Consider a study of the association between blood pressure and heart disease in an observational cohort. Blood pressure may be observed at several study visits during the course of follow-up, giving several measurements over time. Here the time dependent variable (blood pressure) is continuous.

Two main types of time-dependent (TD) explanatory variables should be distinguished: internal (also called endogeneous) TD variables and external (also called exogeneous) TD variables. As we will see later, those two types of variables will require specific attention, and more particularly the internal TD variables.

External time-dependent variables are those that do not require contact with the patient to be known, and it does not require that the patient is alive to exist (and so be measured). A simple example is age. Another example of an external time-dependent variable occurs in a trial setting where the dosage of a drug is to be altered over time in a pre-determined way. Another example, relevant for an observational setting, is air pollution level and its association to asthma attacks.

Internal TD variables can only be measured when an individual is alive and still in the study and they cannot be determined without contact with the patient. Examples of internal TD variables are biological measurements which are made over the course of the study (such as the level of a biomarker, blood pressure), or whether a patient has received an organ transplant. The level of the biomarker could take any continuous value and increase or decrease over time, while the transplant status could only change from 0 (not transplanted) to 1 (transplanted). In the next two sections when describing the parametrisation of the Cox model with TD variable and defining the partial likelihood, we do not distinguish between external or internal TD variables. However, we come back on interpretation of the results, and the implications of those 2 types of TD later on.

In previous lectures we have denoted explanatory variables by x which may be a vector. We now extend this and let $x(t)$ denote an explanatory variable, or a vector of explanatory variables, at time t. If $x(t)$ a vector then it may contain variables which do not change over time, i.e. take the same value no matter what the value of t.

7.3 Analysis using time-dependent variables

The Cox proportional hazards model has been extended to accommodate time-dependent explanatory variables, and in this situation takes the form

$$h(t; x(t)) = h_0(t) \exp(\beta^\top x(t)) \tag{7.1}$$

In this formulation it is assumed that we are interested in the explanatory variable at the time of the event of interest. In other words, it means that only the current value of the covariates (i.e. at time t) affects the hazard (we'll mention some other possibilities later). Here, the baseline hazard function is interpreted as the hazard function for an individual for whom all the variables are zero (from the time origin and during all the follow-up).

What is the interpretation of the log hazard ratios in this situation? Consider a single time-dependent explanatory variable $x_1(t)$. The hazard ratio for individuals r and s is

$$\frac{h(t; x_{1r}(t))}{h(t; x_{1s}(t))} = \frac{h_0(t)\exp(\beta^\top x_{1r}(t))}{h_0(t)\exp(\beta^\top x_{1s}(t))} = \exp(\beta(x_{1r}(t) - x_{1s}(t))) \tag{7.2}$$

Therefore β_1 is the log hazard ratio for two individuals whose explanatory variable at (any) time t differs by 1 unit. Notice here that the effect of 1-unit change of TD variable is assumed to be the same over time. However, the quantity $x_1(t) - x_2(t)$ varies with time, and therefore model (8.2) is no longer a proportional hazard model. We refer to it hereafter as extended Cox model.

Exercise 7.1 Let's consider a binary TD variable $x_1(t)$ e.g. transplant status at time t (yes/no). Suppose that person r is transplanted at time 5 and person s is never transplanted. What is the ratio of hazards for these two individuals: (a) at time 4, (b) at time 6 (assuming both individuals are at risk at times 4 and 6.

Under the extended Cox model with time-dependent explanatory variables the partial likelihood is

$$\prod_j \frac{\exp(\beta^\top x_{i_j}(t_j))}{\sum_{k\in R_j}\exp(\beta^\top x_k(t_j))} \tag{7.3}$$

In this partial likelihood we compare the hazard for the actual case at time t_j with explanatory variables $x_{i_j}(t_j)$ the hazards for individuals in the risk set at time t_j ($k \in R_j$) with explanatory variables $x_k(t_j)$. One important think to notice here is that, by using the partial likelihood 7.3 we need to know the values of the explanatory variables for each individual at every event-time t_j at which they feature in the risk set, including the TD variable. For binary TD variable, as transplanted yes/no, this is not a problem. However, it may be more problematic for continuous TD variables like blood pressure, which will be measured only at periodical check-ups. By default, the partial likelihood is fitted using the last known value of the time-dependent covariate (assuming that it stays constant from the last time it has been observed).

7.4 Structure of data with time-dependent variables

Consider a binary time dependent explanatory variable $x(t)$, and suppose we have three individuals as follows (all individuals entering the data at time 0):

- Individual 1 has the same value $x(t) = 0$ over the full course of follow-up and is censored at time 10.
- Individual 2 has $x(t) = 0$ up until time 5, then their exposure changes to $x(t) = 1$ and they are censored at time 20.

- Individual 3 has $x(t) = 0$ up until time 15, then their exposure changes to $x(t) = 1$ and they have the event at time 25.

The data for these individuals can be displayed as in table 7.1. When the explanatory variable changes for a given individual a new row is added to the data, containing the updated value of TD variable. So in the dataset used for the analysis, each individual may contribute multiple rows to the data.

Individual	Time origin	Start time	Stop time	$x(t)$	Status
1	0	0	10	0	0
2	0	0	5	0	0
2	0	5	20	1	0
3	0	0	15	0	0
3	0	15	25	1	1

Table 7.1: Example data for individuals with time-dependent explanatory variable $x(t)$. Each individual has a line of data for each time period over which the explanatory variables takes a different value. The status refers to whether the individual has the event (1) or not (0) at the end of the interval.

The time intervals are continuous on the left and closed on the right, meaning that individual 2 has exposure 0 in the interval (0,5] and exposure 1 in the interval (5,20], for example.

Example 7.1

The Stanford Heart Transplant Data The Cox model for time-dependent exposures will be illustrated using a classic data set based on the Stanford Heart Transplant programme. Subjects needing a heart transplant were accepted onto a programme and then had to wait for a suitable heart to become available. They were followed up from the date of their acceptance onto the programme. Not all individuals received a transplant. The question of interest is: "Does transplant increase survival?". The data contains the following variables (amongst others):

- datein: date of entry into the programme
- datetr: date of transplant (missing if none)
- dateout: date of exit from the study (because of death or censoring)
- dead: status at end of study (dead (1) or censored (0))

Here is an example of what the data look like:

```
+-----------------------------------------------+
| id datein     datetr      dateout     dead    |
|-----------------------------------------------|
| 25 28apr1969  22may1969   01apr1974   0       |
```

```
    | 26 01may1969  .           01mar1973   0       |
    | 27 04may1969  .           21jan1970   1       |
    | 28 07jun1969  16aug1969   17aug1969   1       |
    | 29 14jul1969  .           17aug1969   1       |
    |-----------------------------------------------|
    | 30 19aug1969  03sep1969   18dec1971   1       |
    +-----------------------------------------------+
```

To use the data in a Cox proportional hazards regression we need to put it in the correct form for analysis. This can be done using the `stsplit` command in Stata or using the `tmerge` function in R (see below). Individuals who have a transplant have one row of data pre-transplant and one row of data post-transplant. Individuals who do not have a transplant just have one row of data.

Stata code

```
. use stanford.dta, clear

* Correcting the data
. replace dateout=mdy(01,21,1968) if id==3
. replace id=100 if dob==td("31jan1939")
. replace id=101 if dob==td("25aug1924")
. replace id=102 if dob==td("30oct1933")
. replace id=103 if dob==td("20may1928")
. replace dateout=mdy(09,28,1968) if id==15

. replace datetr=mdy(01,01,2001) if datetr==.
. stset dateout, id(id) origin(datein) scale(365.25) f(dead)
. stsplit post=datetr, at(0)
. replace post=post+1
. stcox post
```

R code

```
> library(haven)
> library(survival)
# Some already prepared data :  data(heart)
> stanford <- read_dta("stanford.dta")
> stanford <-zap_formats(stanford)
> stanford <-zap_labels(stanford)

# Correcting the data
> stanford$id[stanford$dob=="1939-01-31"]=100
> stanford$id[stanford$dob=="1924-08-25"]=101
> stanford$id[stanford$dob=="1933-10-30"]=102
> stanford$id[stanford$dob=="1928-05-20"]=103
```

```
# calculate age at acceptance
stanford$ageaccept <- (stanford$datein-stanford$dob)/365.25

# myfup contains futime=Total follow-up time
#                txtime=Transplant time
> myfup <- with(stanford, data.frame(id = id,
                                futime= as.numeric(dateout-datein),
                                txtime= as.numeric(datetr-datein),
                                dead = dead))
# 1 mistake in the data (id=3 ==> fu.date should be 1968-01-21)
> myfup$futime <- ifelse(myfup$futime<0, 15, myfup$futime)
# 1 patient died on the same day than entering the study==>0.5 day
> myfup$futime <- ifelse(myfup$futime==0, 0.5, myfup$futime)

> s2data <- tmerge(stanford, myfup, tstop=futime, id=id,
                death = event(futime, dead),
                trt   =  tdc(txtime))
# transforming f-up time: days to years
s2data$tstart <- s2data$tstart/365.25
s2data$tstop <- s2data$tstop/365.25

coxph(Surv(tstart, tstop, death)~ trt+ageaccept, data=s2data, ties = "breslow")
```

The data above now look like this in Stata:

```
+-------------------------------------------------------------------------------------------------+
| id datein     datetr      dateout     dead    _st _d  _origin _t          _t0         post      |
|-------------------------------------------------------------------------------------------------|
| 25 28apr1969  22may1969   22may1969   .       1   0   3405    .06570842   0           0         |
| 25 28apr1969  22may1969   01apr1974   0       1   0   3405    4.9253936   .06570842   1         |
| 26 01may1969  01jan2001   01mar1973   0       1   0   3408    3.8329911   0           0         |
| 27 04may1969  01jan2001   21jan1970   1       1   1   3411    .71731691   0           0         |
| 28 07jun1969  16aug1969   16aug1969   .       1   0   3445    .19164956   0           0         |
|-------------------------------------------------------------------------------------------------|
| 28 07jun1969  16aug1969   17aug1969   1       1   1   3445    .19438741   .19164956   1         |
| 29 14jul1969  01jan2001   17aug1969   1       1   1   3482    .09308693   0           0         |
| 30 19aug1969  03sep1969   03sep1969   .       1   0   3518    .04106776   0           0         |
| 30 19aug1969  03sep1969   18dec1971   1       1   1   3518    2.329911    .04106776   1         |
+-------------------------------------------------------------------------------------------------+
```

And these data now look like this in R:

```
id     datein     datetr    dateout     tstart      tstop trt death
25 1969-04-28 1969-05-22 1974-04-01 0.00000000 0.06570842   0     0
25 1969-04-28 1969-05-22 1974-04-01 0.06570842 4.92539357   1     0
26 1969-05-01       <NA> 1973-03-01 0.00000000 3.83299110   0     0
27 1969-05-04       <NA> 1970-01-21 0.00000000 0.71731691   0     1
28 1969-06-07 1969-08-16 1969-08-17 0.00000000 0.19164956   0     0
28 1969-06-07 1969-08-16 1969-08-17 0.19164956 0.19438741   1     1
```

```
29 1969-07-14       <NA> 1969-08-17 0.00000000 0.09308693   0     1
30 1969-08-19 1969-09-03 1971-12-18 0.00000000 0.04106776   0     0
30 1969-08-19 1969-09-03 1971-12-18 0.04106776 2.32991102   1     1
```

Exercise 7.2 For Stata users, explain what is contained within the columns labelled *_t*0, *_t*, *post* and *_d*, and for R users, explain what is contained within the columns labelled `tstart`, `tstop`, `trt` and `death`?

Exercise 7.3 A Cox proportional hazards model can be fitted on the data in this format either in Stata or in R as shown at the end of the example above

```
* Stata
. stcox post ageaccept
# R
> coxph(Surv(tstart, tstop, death) ~ trt + ageaccept
```

.

For the binary time-dependent variable `post/trt`, the estimated hazard ratio is 0.91, 95% CI [0.49;1.67]. What is the interpretation of this hazard ratio?

7.5 Refinements of the extended Cox model

The ability to accommodate time-dependent explanatory variables through the partial likelihood is an important feature of the Cox model. Different kinds of refinements of the extended Cox model are possible, depending on the research question.

- For example, in the example of transplantation, it may be of interest to investigate if the effect of transplantation depends on the number of days from admission to transplantation. To deal with this question, one way would be to create multiple TD variables, one for each pre-defined intervals indicating whether the transplantation as occurred. So we create new variables between defined as interaction between transplantation and the time-interval where transplantation occurred. For example, we could create 2 TD variables, the first one being defined as "transplantation during the first 2 days after admission" (denoted I_1), and another defined as "transplantation after the third day onward" (denoted I_2). It's worth to notice that the results of this model should be interpreted carefully as the 2 TD variables are linked (they cannot be 1 simultaneously). So the subjects who have a value of 0 need to be carefully considered (See Hosmer & Lemeshow's book, pages 224-226 for a full discussion on this topic).

- Another refinement would be to assess if the effect of being transplanted is constant or varies over time after the transplantation. Let's define z_i as the time of transplantation for patient i. In such situation, a model of the form $h(t; x_i(t)) = h_0(t) \exp(\beta(t - z_i)^\top x_i(t))$ could be useful, assuming a linear-in-time change of the effect. A more flexible formulation of this model could be also used, defining $h(t; x_i(t)) =$

$h_0(t)\exp(f(t-z_i)^\top x_i(t))$ and where in this case $f()$ is a flexible function of time such as a polynomial or a spline verifying $f(u) = 0$ when $u < 0$ (Heinzl, Stat in Med 1996). A binary variable has been used for this example, but this extension could also be done for a continuous time-dependent variable. However, the model may be quite complicated to interpret.

- In the model 7.1 it is assumed that this is the current exposure which is relevant for the hazard; that is the hazard at time t depends on the exposures at time t. Sometimes, past exposure may have an effect on survival independently of current exposure. Some information on past values of explanatory variables could be incorporated into the proportional hazards model, e.g. we could include 'exposure 1 year ago' as well as 'exposure now'. This leads to more complicated models. Another possibility is that it is the pattern of an explanatory variable over time (e.g. an increase or decrease over time) that is important for survival, or the cumulative exposure over time.

Time-dependent explanatory variables in other survival models

- The use of TD explanatory variables is a natural extension in a proportional hazards model, because of the form of the model whereby covariates act on the hazard. It is therefore a straightforward extension to allow the covariate value at a given time to act on the hazard at that time. So TD explanatory variables can also be used in extensions to parametric proportional hazards models, e.g Weibull, exponential.
- The extension of TD explanatory variables to accelerated failure time models is not straightforward, because in the AFT model, covariates act on the survival time itself rather than on the hazard. We do not give any further details here about the use of time-dependent explanatory variables in AFT models. More about this topic can be find in the book of Kalbfleisch and Prentice, second edition, chapter 7.4.4.

7.6 Cautionary notes

It is easy to write down a Cox model with time-dependent covariates, but harder to fit (computationally) and harder to interpret the results. For interpretation, we need to go back to the distinction between internal and external TD variables. A major difference between these two types of TD variables lies in the relationship between the conditional hazard and the conditional survival.

For external TD variables, the classical relationship holds:

$$S(t \mid x(u), u \leq t) = \exp(-\int_0^t h_0(u)\exp(\beta^\top x(u))du) \tag{7.4}$$

This quantity may be complicated to calculate because it involves in the integral not only the baseline hazard function $h_0(t)$ but also the values of the TD variables over the interval from 0 to t. So it seems odd to estimate the probability of survival beyond time t for a particular individual, because it depends on TD variables that vary in the

future. However, because external TD variables could be defined at any time, we can imagine predicting survival up to time t given a TD variable's path up to time t.

On the other hand, for internal TD variables the classical relationship 7.4 between the survival probability and the hazard does not hold anymore. We can still calculate the quantity on the right hand side of formula 7.4, but it does not represent the survival probability anymore. This is because the survival probability does not make any sense since the TD variable was measured when the individual was alive at time t. In other words, the probability to be alive at (i.e. having survived up to) the time t is 1. This is the main specificity of internal TD variable: it requires the survival of the subject for their existence. On the other hand, those remarks do not invalidate the estimation and interpretation of association between TD variables and the mortality hazard.

Another important point to bear is mind concerns the causal interpretation of the results, as for example a treatment effect. Indeed, models with TD variables carry a great risk of controlling for variables in the causal pathway. In the PBC trial, a model could be fitted with the updated values of bilirubin, for example, as well as treatment. However the active treatment was designed to have a direct effect on bilirubin and therefore updated values of bilirubin might mediate the treatment effect, if the treatment is successful.

7.7 Frailty Models

Introduction

Frailty models are random effects models for time-to-event data. Two reasons for use of frailty models in a survival analysis setting are:

- Situation 1 (Individual frailty): When there are believed to be intrinsic features of an individual which impact on survival, and which cannot be or have not been observed. This leads to heterogeneity between individuals which is not explained by observed covariates. In other words, we do not believe the usual assumption that each individual has the same underlying hazard of having the event, after conditioning on the variables used in the analysis. An extreme example is where there are some individuals in a population who are virtually guaranteed not to have the event of interest and others who are more highly susceptible. One example of where this might occur is in a study of time to HIV infection in the general population. Those models are a kind of overdispersion/heterogeneity models.
- Situation 2 (Shared frailty): When survival times are observed from individuals who are intrinsically related in some way. For example, if the study involves the survival times of groups of individuals from different hospitals then individuals from the same hospital may have survival times which are more likely to be similar than individuals from a different hospital, e.g. due to treatment practices and expertise within a hospital. This is an example of clustered (or hierarchical) data and the corresponding model which accounts for the clustering can be seen

as a random effects model.

We can deal with these situations in survival model by introducing a random effect. In the context of survival models this random effect is called 'frailty' for historical reason, so the models we describe below are called frailty models. The term 'frailty' simply refers to the fact that these models allow for the fact that some individuals are intrinsically more 'frail' (assuming a medical context) than others.

Situation 1: Individual frailty model
In the survival models considered so far we have described a hazard function for an individual i with vector of explanatory variables x_i; $h_i(t; x_i)$. It was assumed that all individuals with a given vector of explanatory variables were subject to the same hazard. This can be extended to include a term representing additional heterogeneity between individuals (due for example to an unmeasured important explanatory variable) by instead considering the hazard function

$$h(t; x_i) = \alpha_i h(t; x_i) = \alpha_i h_0(t) \exp(\beta^\top x_i) \tag{7.5}$$

where α_i is the frailty term for individual i. The frailty term represents the effect on the hazard of being individual i, aside from those effects encompassed by the explanatory variables x_i. The frailty term α_i is a random effect, that is, it is not observed. It is assumed to follow a particular distribution with a positive support, such that the α_i are positive. Individuals with a frailty term $\alpha_i > 1$ have an increased hazard whilst individuals with $\alpha_i < 1$ have a decreased hazard. We will show below how to fit this survival model by assuming a particular distribution for the frailty term.

Associated with this conditional hazard, we could define a conditional survival using the classical relationship between hazard and survival

$$S(t; x_i) = S_0(t)^{\alpha_i \exp(\beta^\top x_i)} \tag{7.6}$$

To derive the full likelihood, we first need to define the conditional likelihood (conditional on the random effect) and then integrate it out to finally obtain the full likelihood.

The likelihood conditional on the random effects, denoted L^c, is therefore

$$L^c = \prod_{i=1}^{n} [h(t_i; x_i) S(t_i; x_i)]^{\delta_i} S(t_i; x_i)^{1-\delta_i} \tag{7.7}$$

To obtain the full likelihood, we need to integrate out the frailty distribution, under the assumption that the frailties come from a particular distribution $f(\alpha_i)$. The full likelihood is then

$$L = \prod_{i=1}^{n} \int_0^\infty [h(t_i; x_i) S(t_i; x_i)]^{\delta_i} S(t_i; x_i)^{1-\delta_i} f(\alpha_i) d\alpha_i \tag{7.8}$$

This definition of the full likelihood contains an integral that complicates the calculation. However, in some cases, it can be re-written to get an analytical form; this

is the case when we assume (i) a gamma distribution for the frailty variable, and (ii) a parametric distribution for the hazard model (such as Weibull). However, in most cases, there's no closed form of this likelihood, and it needs to be approximated by specific techniques like numerical integration.

It is important to notice that the quantities defined in 7.5 and 7.6 are conditional hazard and conditional survival. The unconditional quantities (i.e. population hazard and survival) can be obtained by integrating out the unobserved random effect. We are not going to give the details here of the mathematical expression for those quantities, but those interested could read the chapter 3.5 of Duchateau and Jansen's book.

Application to a Weibull model

Frailty terms can be incorporated as extensions to models that we are already familiar with. Consider the Weibull model. The form of the hazard function in a Weibull model is

$$h(t; x_i) = \kappa\lambda t^{\kappa-1} \exp(\beta^\top x_i) \tag{7.9}$$

The extended hazard function which includes frailty terms is

$$h(t; x_i, \alpha_i) = \alpha_i\kappa\lambda t^{\kappa-1} \exp(\beta^\top x_i) \tag{7.10}$$

The likelihood conditional to the random effect, L^c is therefore

$$L^c = \prod_{i=1}^{n} [\alpha_i\kappa\lambda t_i^{\kappa-1} \exp(\beta^\top x_i) \exp(-\alpha_i\lambda t_i^\kappa \exp(\beta^\top x_i))]^{\delta_i} [\exp(-\alpha_i\lambda t_i^\kappa \exp(\beta^\top x_i))]^{1-\delta_i} \tag{7.11}$$

To obtain the full likelihood, assuming that the frailties come from a particular distribution $f(\alpha_i)$, we need to integrate out the frailty distribution. The full likelihood is then

$$L = \prod_{i=1}^{n} \int_0^\infty [\alpha_i\kappa\lambda t_i^{\kappa-1} \exp(\beta^\top x_i) \exp(-\alpha_i\lambda t_i^\kappa \exp(\beta^\top x_i))]^{\delta_i} [\exp(-\alpha_i\lambda t_i^\kappa \exp(\beta^\top x_i))]^{1-\delta_i} f(\alpha_i) d\alpha_i \tag{7.12}$$

It is common to assume a gamma distribution with parameters (a, b) for the frailty. This assumption is mainly due to mathematical tractability as we can obtain an analytical expression of the likelihood with such assumption.

Statistical software packages now allow us to fit frailty models for parametric survival models without difficulty. See below for an example of this.

Frailty terms can also be included in other types of model, including accelerated life models and the Cox proportional hazards model. The methods for fitting the Cox proportional hazards model with frailty terms are more complex than described above and we do not give the details here. To my knowledge, Stata does not currently allow fitting of Cox proportional hazards models with individual frailty terms.

Example 7.2

Evidence of frailty in lung cancer mortality (Taken from Manton et al. 1986 and Aalen 1994)

Manton *et al.* found that mortality from lung cancer tailed off at older ages 7.1. This may be interpreted as a frailty effect, because for example the people who are most susceptible to the negative effects of smoking have died at younger ages. Those who survive to older ages are those whose lifestyle, occupation, environment and intrinsic features such as genetics are such that they have survived this long; these characteristics may also mean that these individuals are less susceptible to lung cancer mortality. Manton *et al.* used frailty models for the lung cancer mortality hazard, assuming a Weibull distribution and a Gamma frailty.

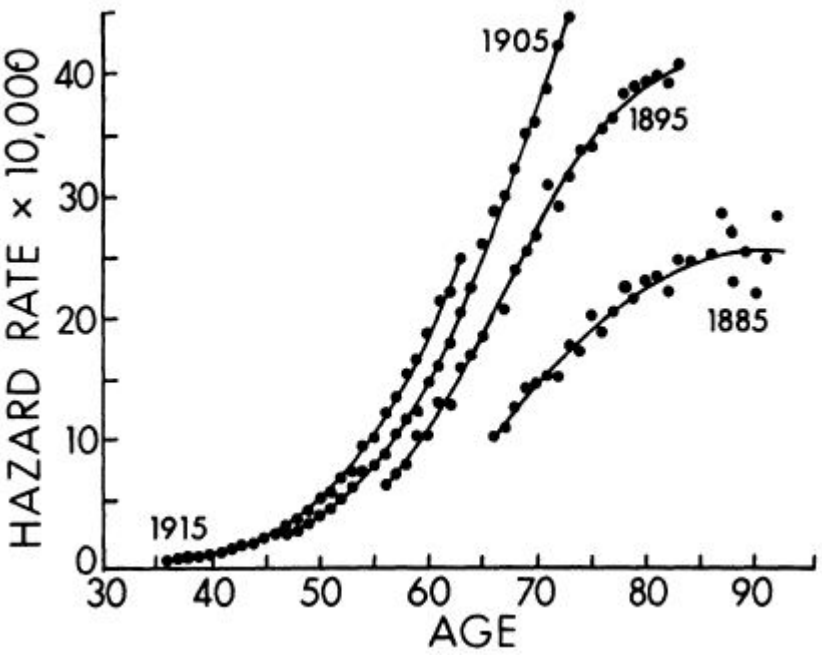


Figure 7.1: Hazard rates for lung cancer mortality in 4 male birth cohorts, showing fitted curves from frailty models.

Situation 2: Shared frailty model

In Situation 1 there were unobserved features of individuals, leading to heterogeneity between individuals, which were represented by individual frailty terms. Next, we consider the more common situation encountered in practice, where the heterogeneity is between groups of individuals, due to clustering or hierarchical structure in the data (e.g. individuals in hospitals). Here we suppose that the study population is formed from G groups, which we index $j = 1, ..., G$. Within group j there are n_j individuals,

indexed by $i = 1, \ldots, n_j$. The vector of exposures for the i^{th} individual in group j is denoted $x_i j$.

The hazard function, assuming a proportional hazards model just for example, is of the form

$$h(t; x_{ij}, \alpha_j) = \alpha_j h_0(t) \exp(\beta^{top} x_{ij}) \tag{7.13}$$

where α_j is the frailty term. This term is shared by all individuals in group j and it represents the effect on the hazard of being in group j. For this reason the model in 7.13 is referred to as a shared frailty model. Groups with a frailty term $\alpha_j > 1$ have an increased hazard and groups with $\alpha_j < 1$ have a decreased hazard. As before, the α_j are usually assumed to follow a specific distribution (e.g. gamma) with additional constraint for identifiability reasons, as mean 1 and variance σ^2. The variance σ^2 is a measure of the variability between clusters (and hence this model reduces to the standard Cox model when $\sigma^2 = 0$).

We can write the model in 7.13 as a classical mixed-effect model, with the random effect being included in the linear predictor:

$$h(t; x_{ij}, \alpha_j) = h_0(t) \exp(\beta^{top} x_{ij} + w_j) \tag{7.14}$$

The definition of the likelihood follows the same principle detailed before for the individual frailty model (see equations 7.11 and 7.12). Briefly, it consists of the following steps (i) defining the contribution to the likelihood of one individual in a specific cluster, then (ii) the product of these individual contributions to obtain the cluster specific (or conditional) likelihood, then (iii) defining the conditional likelihood as the product of all cluster-specific likelihood and finally, (iv) the full likelihood by integrating out the random effect distribution. As mentioned previously, depending on the assumptions made, we can work with an analytical form of the likelihood, or we need to use specific numerical integration techniques to work out the likelihood that needs to be optimised. Many different possibilities are available to the user for both the distribution of the hazard and the distribution of the frailty. From a practical point of view, different possibilities are available in different software.

Stata allows fitting of shared frailty models using both parametric survival models and the Cox proportional hazards model (this is the only type of frailty model that can be fitted for the Cox proportional hazards model in Stata).

Shared frailty models are also appropriate for use in studies of recurrent events, for example time between migraines. In this situation the data may comprise multiple observations from each individual, the individual being the cluster in this case. An example of this type of analysis is given below.

Example 7.3

Using frailty models to study recurrence of infection in kidney dialysis patients

This example is given by Gutierrez (2002), who in turn took the data from McGilchrist and Aisbett (1991). The data contain times to recurrence of infection due to catheter insertion for kidney patients using portable dialysis equipment. When infection occurs, the catheter is removed and the infection is treated, and then, after a pre-determined period of time, the catheter is reinserted. When the catheter is removed for reasons other than infection, the time to infection is treated as censored. Data were obtained from 38 patients, some of whom had more than one infection and therefore provide more than one recurrence time. The data are available in Stata by typing: `use http://www.stata-press.com/data/r8/catheter, clear`. In R the data are directly available from the `survival` package by typing `data(kidney)`.

The data for the first 5 patients look like this, where time gives time in days to catheter removal:

patient	time	infect	age	female
1	16	1	28	0
1	8	1	28	0
2	13	0	48	1
2	23	1	48	1
3	22	1	32	0
3	28	1	32	0
4	318	1	31.5	1
4	447	1	31.5	1
5	30	1	10	0
5	12	1	10	0

Stata code

We first use the stset command, which gives

```
    . stset time infect

     failure event:  infect != 0 & infect < .
obs. time interval:  (0, time]
 exit on or before:  failure

------------------------------------------------------------------------------
       76  total observations
        0  exclusions
------------------------------------------------------------------------------
       76  observations remaining, representing
       58  failures in single-record/single-failure data
     7424  total analysis time at risk and under observation
                                             at risk from t =         0
                                  earliest observed entry t =         0
                                       last observed exit t =       562
```

We start by fitting a standard Cox proportional hazards model, with age and female as explanatory variables.

```
    . stcox age female

[some output omitted]
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.002245   .0091153     0.25   0.805     .9845377    1.020271
      female |   .4499194   .1340786    -2.68   0.007     .2508832    .8068592
------------------------------------------------------------------------------
```

Next, we fit a Cox model with shared frailty. This model recognises that some observations come from the same individual.

```
    . stcox age female, shared(patient)

         failure _d:  infect
   analysis time _t:  time

Fitting comparison Cox model:

Estimating frailty variance:

Iteration 0:   log profile likelihood = -182.06713
```

```
Iteration 1:   log profile likelihood =  -181.9791
Iteration 2:   log profile likelihood = -181.97453
Iteration 3:   log profile likelihood = -181.97453

Fitting final Cox model:

Iteration 0:   log likelihood = -199.05599
Iteration 1:   log likelihood = -183.72296
Iteration 2:   log likelihood = -181.99509
Iteration 3:   log likelihood = -181.97455
Iteration 4:   log likelihood = -181.97453
Refining estimates:
Iteration 0:   log likelihood = -181.97453

Cox regression --
         Breslow method for ties                 Number of obs      =        76
         Gamma shared frailty                    Number of groups   =        38
Group variable: patient

No. of subjects =           76                  Obs per group: min =         2
No. of failures =           58                                 avg =         2
Time at risk    =         7424                                 max =         2

                                                Wald chi2(2)       =     11.66
Log likelihood  =   -181.97453                  Prob > chi2        =    0.0029

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.006202   .0120965     0.51   0.607     .9827701    1.030192
      female |   .2068678    .095708    -3.41   0.001     .0835376    .5122756
-------------+----------------------------------------------------------------
       theta |   .4754497   .2673108
------------------------------------------------------------------------------
Likelihood-ratio test of theta=0: chibar2(01) =     6.27 Prob>=chibar2 = 0.006

Note: standard errors of hazard ratios are conditional on theta.
```

R code

We start by fitting a standard Cox proportional hazards model, with age and female as explanatory variables.

```
> kfit <- coxph(Surv(time, status) ~ age + sex, data=kidney, ties = "breslow")
> summary(kfit)
```

```
    Call:
    coxph(formula = Surv(time, status) ~ age + sex, data = kidney,
        ties = "breslow")

      n= 76, number of events= 58

            coef exp(coef)  se(coef)      z Pr(>|z|)
    age  0.002182  1.002184  0.009225  0.236  0.81305
    sex -0.820995  0.439994  0.298720 -2.748  0.00599 **
    ---
    Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
    age     1.002     0.9978    0.9842    1.0205
    sex     0.440     2.2728    0.2450    0.7902

    Concordance= 0.662  (se = 0.045 )
    Likelihood ratio test= 7  on 2 df,   p=0.03
    Wald test            = 7.87  on 2 df,   p=0.02
    Score (logrank) test = 8.28  on 2 df,   p=0.02
> kfitshared <- coxph(Surv(time, status)~ age + sex +  frailty(id, dist='gamma'),
                      data=kidney, ties = "breslow")
> summary(kfitshared)
    Call:
    coxph(formula = Surv(time, status) ~ age + sex + frailty(id,
        dist = "gamma"), data = kidney, ties = "breslow")

      n= 76, number of events= 58

                              coef      se(coef) se2      Chisq DF    p
    age                        0.005466 0.01178  0.008785  0.22  1.00 0.64000
    sex                       -1.556750 0.45563  0.349881 11.67  1.00 0.00063
    frailty(id, dist = "gamma                               22.33 12.71 0.04500

        exp(coef) exp(-coef) lower .95 upper .95
    age    1.0055     0.9945   0.98253    1.0290
    sex    0.2108     4.7434   0.08631    0.5149

    Iterations: 6 outer, 46 Newton-Raphson
         Variance of random effect= 0.3976791   I-likelihood = -182.1
    Degrees of freedom for terms=  0.6  0.6 12.7
    Concordance= 0.813  (se = 0.033 )
    Likelihood ratio test= 45.58  on 13.86 df,   p=3e-05
```

Exercise 7.4 What is your interpretation of the results from the two models (ie the models with and without the shared frailty term)? What are your conclusions?

References

- Fisher DL, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. Annual review of public health 1999; 20: 145-57
- Heinzl H, Kaider A, Zlabinger G. Assessing interactions of binary time-dependent covariates with time in cox proportional hazards regression models using cubic spline functions. Statistics In Medicine 1996; 15: 2589-601.
- Hosmer DW, Lemeshow SJr, May S. Applied Survival Analysis: Regression Modeling of Time to Event Data, 2nd Edition 2008; Wiley Series in Probability and Statistics. Chapter 7.
- Aalen O. Effects of frailty in survival analysis. Statistical Methods in Medical Research 1994; 3: 227-243. Gutierrez RG. Parametric frailty and shared frailty survival models. The Stata Journal 2002; 2: 22-44.
- Manton KG, Tallard ER, Vaupel JW. Alternative Models for the Heterogeneity of Mortality Risks Among the Aged . Journal of the American Statistical Association 1986; 81: 635-644.
- McGilchrist CA, Aisbett CW. Regression with frailty in survival analysis. Biometrics 1991; 47, 461-466. June 1991.

Practical 7

Datasets required: `hip4` for part A. For part B: `catheter` (direct from Stata) or load `kidney_frailty.csv` for R users)

R packages required for R users: survival, Epi, ggplot2, survminer

Introduction

This practical is in two parts.

- Part A is on **time-dependent variables** and uses data from a trial aiming to protect elderly women from hip fractures.
- Part B explores **frailty models** using data from a study which measured time to recurrence of infection after catheter insertion in kidney patients.

Aims

By the end of this session you should be able to:

- Write down equations showing how time-dependent explanatory variables are incorporated into the Cox model
- Fit survival models with time-dependent explanatory variables (in Stata or R)
- Interpret the results from fitting a survival model with time-dependent explanatory variables
- Fit and interpret frailty models (in Stata or R)

Part A: Time-dependent variables

We will use a study of 48 women over the age of 60. The aim of the study was to quantify the benefit of a new inflatable device (randomly given to 28 of the 48 women) to protect the elderly from hip fractures resulting from falls. Each woman's blood calcium level was measured every five months.

Variable	Description
`id`	Patient ID
`time0`	Begin of span
`time1`	End of span
`fracture`	0= No fracture, 1=Fracture
`protect`	Random assignment: 0=no device, 1=protective device
`age`	Age at enrollment
`calcium`	Blood calcium level (mg/dL)
`gap`	0=No gap, 1=gap
`init_drug_level`	Drug level at start of period (mg)

Where code examples are given or explanations are given that are specific to Stata or R, text and code relating to Stata is shown in this colour and text and code relating to R is shown in this colour.

1. We will begin by investigating what is in the data and how the data are arranged. Read/load the dataset and explore the variables. We will look closely at the data from two particular women to understand how the data were collected.

   (a) How many records (rows) are there in the dataset? And how many individual women?

   (b) How many fractures were observed in the study?

   (c) Examine the records for woman 11 and woman 18.

   Discuss: Why are the multiple rows for each of these women? Is the reason the same for both women? What does the `gap` variable indicate? Why are the `fracture` and `calcium` variables different on each row?

2. When we have data which allows participants to have a period of time when they are not observed we must account for that.

   In Stata, use the `time0` option of stset:

   ```
   stset time1,failure(fracture) time0(time0) id(id)
   ```

   Look up the help file for `stset` for more details on why we need the time0 option.

   In R, you should use the 'counting process' notation for the `Surv` object.

   ```
   Surv(hip$time0, hip$time1, hip$fracture, type="counting")
   ```

   (a) For Stata users: Check that the output from stset corresponds to the number of women and number of fractures you counted earlier.
   For R users: Study the output from using the `Surv()` function as given above. How many of the rows of `Surv()` end with the `+` symbol?

   (b) What are the calcium level [mg/dL] and the initial drug level [mg] for woman 16?

3. Next, we will investigate how wearing a protective device (variable `protect`) is associated with time to fracture event.

   (a) Describe graphically (using a non-parametric plot) how wearing a protective device is associated with time to fracture event.

   (b) Use a Cox regression model to quantify the association between use of the protective device and risk of a fracture. Interpret the results.

4. It is also of interest to study how calcium level and age at enrolment to the study are associated with the hazard of a fracture event.

   (a) Write down the form of a Cox proportional hazards model which includes three explanatory variables; protect, age and calcium.

(b) Fit the model.

Discuss: Interpret the results from this Cox model.

5. The investigators wish to investigate the impact of a new bone-fortifying drug. The inital dose of this drug is given by `init_drug_level`.

   (a) First change the units of the variable `init_drug_level` by dividing the values by 50 (call the new variable `init_drug_level_50`). Why is this a good idea?

   (b) Include `init_drug_level_50` in the model fitted in question 4 and interpret the results.

   What is this model assuming about the association between the initial drug level in the patient's bloodstream and its association with the hazard for hip fractures?

6. It may be more reasonable to assume that the hazard changes with the *current* level of the drug rather than the *initial* level. Suppose that the drug level in the patient's bloodstream remains at its initial level for the first 5 months after which it suddenly reduces to zero. This is an example of an external time-dependent variable. We are going to refit the model including `init_drug_level` under this new scenario.

   (a) Write down the form of the hazard under the Cox proportional hazards model for the new scenario.

   (b) Fit the model
   In Stata you will use the `tvc` and `texp` options of `stcox` (you used these in Practical 5). (Hint: You want to multiply `init_drug_level` by 1 for the first 5 months and by 0 thereafter).

   In R, you will need to use the option `tt` in `coxph` (you used this in Practical 5). (Hint: You want to multiply `init_drug_level` by 1 for the first 5 months and by 0 thereafter).

   (c) Now suppose that rather than changing suddenly, the level of the drug decays in a non-linear fashion in the patient's bloodstream over time. We assume the decay happens at an exponential rate `init_drug_level_50 * exp(-0.35t)`. Fit a Cox model that accommodates this assumption. Does this make any difference to the results?

   In Stata, use the `tvc` and `texp` options to `stcox`:

```
stcox protect age calcium, tvc(init_drug_level_50) texp(exp(-0.35*_t))
```

   In R, use the option `tt` in `coxph`

```
hip.cox <- coxph(Surv(time0,time1,fracture) ~ protect + age + calcium +
            tt(init_drug_level_50), tt=function(x,t,...)(x*exp(-0.35*t)),
            data=hip, ties="breslow")
```

7. The analyses in Question 6 can alternatively be performed by splitting the data and then using a Cox model on these data. Remember that under the Cox regression approach there is a contribution to the partial likelihood at each event time. The contribution to the partial likelihood at a given event time involves information from every individual who is at risk at that time. Therefore information on time dependent variables is required at every time at which a given individual is at risk. The data for an analysis in which variables are changing potentially continuously in time should therefore show time-dependent variables at each event time. Hence it is appropriate to arrange the data by using `stsplit` in Stata or `survSplit` in R so that the records will be split at all observed failure times, i.e. times of fracture.

   (a) Split the data as described above by using the following command:

   ```
   stsplit, at(failures)
   ```

   ```
   failure.times <- unique(hip$time1[hip$fracture==1])
   hip.split <- survSplit(Surv(time0,time1,fracture)~.,
                          data=hip, cut=failure.times)
   ```

   Inspect the data. How many rows of data are there now?

   (b) Generate a new variable that expresses the current level of drug in the bloodstream according to the scenario in Question 6(c) and repeat the analysis on the split data. Did you get the same result?

   In Stata:

   ```
   gen current_drug = init_drug_level_10*exp(-0.35*_t)
   stcox protect age calcium current_drug
   ```

   In R

   ```
   with(hip.split, current_drug<-(init_drug_level/10)*exp(-0.35*time1))
   hip.split.cox <- coxph(Surv(time0,time1,fracture)~ protect+age+calcium
                    + current_drug, data=hip.split, ties="breslow")
   summary(hip.split.cox)
   ```

Part B: Frailty models (optional)

Go through Example 7.3 using the kidney data. The data can be loaded in to Stata by typing:

```
use "http://www.stata-press.com/data/r15/catheter", clear
```

This data set is discussed in the Stata Journal article by R Gutierrez, which can be found here:
http://www.stata-journal.com/sjpdf.html?articlenum=st0006

Consult the article for further details about the data and about frailty models.

A similar analysis can be performed in R by using the kidney data from the survival R-package

```
# Loading the data in R
data(kidney)
coxph(Surv(time, infect)~age+female+frailty(patient),data=kidney)
```

Session 8

Alternative models for survival data

8.1 Aims of this lecture and practical

At the end of this lecture and practical you will be able to:

- Describe the accelerated failure time model and how it differs from a proportional hazards model.
- Interpret the acceleration factor in an accelerated failure time model and contrast this with a hazard ratio.
- Show how the Weibull model can be parameterized as a proportional hazards model or an accelerated life model and compare the parameters under each formulation.
- Describe the features of another parametric form for an accelerated life model, the log-logistic distribution
- Describe the additive hazards model and its advantages and disadvantages, and explain how it can be fitted.
- Fit accelerated failure time models and additive hazards models in Stata and/or R and interpret the results appropriately.

8.2 Beyond the proportional hazards model

The Cox proportional hazards model is by far the most commonly used model in survival analysis. Sometimes, however, we may find that this model does not provide a good fit to our data. In session 5 we covered two options for modelling survival data when the proportional hazards assumption does not hold:

- Using a stratified proportional hazards model, in which the baseline hazard is allowed to differ across strata defined by one or more binary or categorical variables.
- Including a covariate-by-time interaction in an extended Cox model.

Using a stratified proportional hazards model is not always an option, including when the proportional hazards assumption does not hold for the main exposure and when the proportional hazards assumption does not hold for a continuous variable. Including a covariate-by-time interaction in an extended Cox model can be a good option, though this requires a functional form to be specified for the time interaction. This can be done flexibly using e.g. a spline.

Sometimes it is desirable to consider another type of model for survival data altogether. In this session we discuss two different models for survival data:

- the accelerated failure time model
- Aalen's additive hazards model

Chapter 6 of the book by Collett (2014) discusses the accelerated failure time model in detail. References relating to the Aalen's additive hazards model are provided in subsequent sections.

8.3 The accelerated failure time (AFT) model: Introduction

We begin with a reminder of the proportional hazards assumption. For a binary explanatory variable (e.g. denoting treatment or control group) we let $h_0(t)$ denote the hazard function at time t in the baseline group (controls), and $h_1(t)$ denote the hazard function at time t in the comparison group (treatment). Under a proportional hazards model, the effect of explanatory variables on survival is such that explanatory variables act multiplicatively on the hazard, so we can write

$$h_1(t) = \psi_{PH} h_0(t) \tag{8.1}$$

where ψ_{PH} is the multiplying factor to be estimated, and we write this usually as $\psi^{PH} = e^{\beta_{PH}}$. The baseline hazard may be parameterized. For example, in an exponential distribution for the survival times we have $h_0(t) = \lambda$, and in a Weibull distribution we have $h_0(t) = \kappa\lambda t^{\kappa-1}$. In the Cox proportional hazards model we do not specify the baseline hazard and use the special form of analysis, the partial likelihood, which avoids having to estimate $h_0(t)$.

Sometimes it may be more appropriate to think of a completely different way in which explanatory variables affect survival. One way is to think of the explanatory variables affecting the actual survival time (rather than the hazard) in a multiplicative way. Again considering a binary explanatory variable (treatment or control group), let T_0 denote the random variable representing survival time in the baseline group (e.g. controls), and T_1 denote the random variable representing survival time in the comparison group (e.g. treatment group). If the explanatory variable acts multiplicatively on survival time then we can write the survival time is the treatment group as

$$T_1 = \psi_{AFT} T_0 \tag{8.2}$$

where ψ_{AFT} is the multiplying factor to be estimated. We call this an *accelerated failure time (AFT) model*, or sometimes an *accelerated life model*. Suppose the multiplying factor is $\psi_{AFT} = 2$. This means that the survival times in the treatment group tend to be two times the survival times in the control group. In other words the event tends to happen later for the individuals in the treatment group. Under the AFT model, the survival times for individuals in the treatment group are 'speeded up' or 'slowed down' compared to individuals in the control group. The parameter ψ_{AFT} determines how much faster (sooner) or slower (later) an individual in the treatment group tends to have the event of interest relative to an individual in the control group.

8.4 The accelerated failure time (AFT) model: Further details

Binary explanatory variable

In a proportional hazards model, where $h_1(t) = \psi_{PH} h_0(t)$, we replace the multiplying factor ψ_{PH} by $e^{\beta_{PH}}$. This is so that the parameter being estimated (β_{PH}) can take any value, positive or negative, and ensures that the estimated hazard is non-negative. It is convenient to do the same thing in an AFT model and to replace ψ_{AFT} by $e^{-\beta_{AFT}}$ giving

$$T_1 = e^{-\beta_{AFT}} T_0 \tag{8.3}$$

The reason for using the minus sign will become clearer in a minute.

We have not yet written down the form of the survivor function or the hazard. In an accelerated failure time model, the effect of the exposure (or treatment) on survival times is best represented by looking at the survivor function. Denote the survivor function in the control group by $S_0(t)$. The survivor function in the treatment group, $S_1(t)$, can be written

$$\begin{aligned} S_1(t) &= \Pr(T_1 > t) \\ &= \Pr(e^{-\beta_{AFT}} T_0 > t) \\ &= \Pr(T_0 > t e^{\beta_{AFT}}) \\ &= S_0(t e^{\beta_{AFT}}) \end{aligned} \tag{8.4}$$

The survivor function for the treatment group is the same as that for control group, but with t replaced by $te^{\beta_{AFT}}$. Under this model, the distribution of survival times in the treatment group has the same *shape* as the distribution as the survival times in the control group, except that time travels at a different speed (faster or slower) for those in the treatment group, i.e. those in the treatment group are going to move along the survival curve on a different time scale (faster or slower) compared with those in the control group.

The factor $e^{\beta_{AFT}}$ is called the *acceleration factor*. If $e^{\beta_{AFT}} > 1$ then survival times in the treatment group will tend to be earlier than survival times in the control group. If $e^{\beta_{AFT}} < 1$ then survival times in the treatment group will tend to be later than survival times in the control group. It can also be shown that under the AFT model the median survival time in the treatment group is equal to the median survival time in the control group multiplied by $e^{-\beta_{AFT}}$. In fact, it can be shown that the pth percentile of the distribution of the survival times in the treatment group is equal to the pth percentile of the distribution of the survival times in the control group multiplied by $e^{-\beta_{AFT}}$.

Extensions to a more general situation

So far we have just considered a single binary exposure. More generally, for explanatory variable x the survivor function under the AFT model of the form

$$S(t|x) = S_0(te^{\beta_{AFT} x}) \tag{8.5}$$

where $S_0(\cdot)$ denotes the survivor function for an individual with baseline explanatory variable ($x = 0$). This extends also to the situation with several explanatory variables, in which case x is a vector.

For an individual with explanatory variable(s) x_i the AFT model can also be expressed in terms of a log-linear model for the event time T_i:

$$\log T_i = \mu + \alpha x_i + \sigma\epsilon_i \tag{8.6}$$

where μ is called the intercept or location parameter and σ is called the scale parameter, and where $\alpha = -\beta_{AFT}$. Different distributions can be used for the random component ϵ_i, which results in different AFT models. Below we will discuss the Weibull and log-logistic models.

Comparison with the proportional hazards model

Under the proportional hazards model the hazard function and the survivor function are

$$h(t|x) = h_0(t)e^{\beta_{PH}x}, \qquad S(t|x) = \exp\{-e^{\beta_{PH}x}\int_0^t h_0(u)du\} \tag{8.7}$$

and under the AFT model the hazard function and survivor function are

$$h(t|x) = h_0(te^{\beta_{AFT}x})e^{\beta_{AFT}x}, \qquad S(t|x) = S_0(te^{\beta_{AFT}x}) \tag{8.8}$$

Exercise 8.1

- Consider a binary variable x, taking value 1 for a treatment group and 0 for a control group.
- What is the interpretation of $e^{\beta_{AFT}} = 1.5$?
- What is the interpretation of $e^{\beta_{AFT}} = 0.7$?

8.5 The Weibull model as an AFT model

Above we considered the features of an AFT model in a general way. Like the proportional hazards model the AFT model is really a family of models, which take the general form given in equation 8.6 . In this section we consider the specific case of the Weibull model. By now we are familiar with the general form for a Weibull model. The baseline survivor function, that is the survivor function for an individual with $x = 0$, is

$$S_0(t) = \exp(-\lambda t^\kappa) \tag{8.9}$$

Under the proportional hazards model the survivor function for a person with explanatory variable x is $S(t|x) = S_0(x)^{\exp(\beta_{PH}x)}$. Hence if the baseline survivor function is of Weibull form then we can write

$$S(t|x) = \exp(-\lambda t^\kappa e^{\beta_{PH}x}) \tag{8.10}$$

Under the AFT model, the survivor function for a person with explanatory variable x is $S(t|x) = S_0(te^{\beta_{AFT}x})$. Hence if the baseline survivor function is of Weibull form then we can write

$$S(t|x) = \exp\{-\lambda(te^{\beta_{AFT}x})^{\kappa}\} = \exp\{-\lambda t^{\kappa} e^{\kappa\beta_{AFT}x}\} \tag{8.11}$$

This survivor function is also in Weibull form – with λ replaced by $\lambda e^{\kappa\beta_{AFT}x}$. What this means is that all individuals (not just those from the baseline group) have survival times from a Weibull distribution, with the only difference being that one of the parameters changes depending on the explanatory variables. It can therefore be said that the Weibull distribution is an AFT model.

The Weibull model is in both the proportional hazards family of models and the AFT family of models. The AFT and proportional hazards models are simply two different ways of parameterizing the Weibull model and we can move between one and the other. The Weibull model is the only distribution that can be formulated as both a proportional hazards model and an AFT model. The exponential distribution, being a special case of the Weibull, also has this property.

The above results show the following relationship between the formulation of the Weibull model as a proportional hazards model and as an AFT model:

$$\exp(\beta_{PH}) = \exp(\kappa\beta_{AFT}) \tag{8.12}$$

Exercise 8.2

The Weibull AFT model is fitted by writing down the likelihood for the data and estimating the parameters using maximum likelihood, just like the parametric models we have considered in earlier lectures. Write down the likelihood we would use if we wanted to fit a Weibull model in the accelerated failure time form to our survival data. Include the possibility of censoring.

Example 8.1

Leukaemia patients data: Comparison of estimates from Weibull PH and Weibull AFT

The results from fitting a Weibull model using the PH parametrization and using the AFT parametrization are given in the table below. The output from fitting this model in Stata is given below to make clear where the numbers come from. Under the AFT parametrization Stata reports $-\beta_{AFT}$, κ, and $-\kappa^{-1}\log\lambda$. Using the result in (8.12) we can see that $\kappa\beta_{AFT} = -1.366 \times 1.267 = -1.731 = \beta_{PH}$.

```
. streg group, d(weibull) nohr

Weibull regression -- log relative-hazard form
```

	Estimate	SE	95% CI
PH Model			
β_{PH}	-1.731	0.413	(-2.54,-0.921)
κ	1.366	0.201	(1.023,1.823)
$\log \lambda$	-3.071	0.558	(-4.165,-1.977)
AFT Model			
β_{AFT}	1.267	0.311	(0.658,1.876)
κ	1.366	0.201	(1.023,1.823)
$-\kappa^{-1} \log \lambda$	2.248	0.166	(1.923,2.574)

Table 8.1: Leukemia data: Results from fitting a proportional hazards Weibull model and an AFT Weibull model.

```
No. of subjects =           42                     Number of obs   =        42
No. of failures =           30
Time at risk    =          541
                                                   LR chi2(1)      =     19.65
Log likelihood  =   -47.064102                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |  -1.730872   .4130819    -4.19   0.000    -2.540497   -.9212463
       _cons |  -3.070704   .5580701    -5.50   0.000    -4.164501   -1.976907
-------------+----------------------------------------------------------------
       /ln_p |   .3117092   .1472919     2.12   0.034     .0230224     .600396
-------------+----------------------------------------------------------------
           p |   1.365757    .201165                      1.02329     1.82284
         1/p |   .7321944   .1078463                      .5485944    .9772406
------------------------------------------------------------------------------

. streg group, d(weibull) time

Weibull regression -- accelerated failure-time form

No. of subjects =           42                     Number of obs   =        42
No. of failures =           30
Time at risk    =          541
                                                   LR chi2(1)      =     19.65
Log likelihood  =   -47.064102                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |   1.267335   .3106399     4.08   0.000     .6584916    1.876178
       _cons |   2.248352   .1659718    13.55   0.000     1.923054    2.573651
```

```
-------------+----------------------------------------------------------------
       /ln_p |   .3117092   .1472919     2.12   0.034     .0230224     .600396
-------------+----------------------------------------------------------------
           p |   1.365757    .201165                      1.02329     1.82284
         1/p |   .7321944   .1078463                      .5485944    .9772406
------------------------------------------------------------------------------
```

Under the proportional hazards model we have $\hat{\beta}_{PH} = -1.731$, so an estimated hazard ratio of $e^{\hat{\beta}_{PH}} = 0.177$. The hazard in the treatment group is about 18% of that in the control group. The event is 'remission', so remission tends to occur later for those in the treatment group.

Under the AFT model we have $-\hat{\beta}_{AFT} = 1.267$, giving $e^{-\hat{\beta}_{AFT}} = 3.550$ and $e^{\hat{\beta}_{AFT}} = 0.282$. Using the formula $T_1 = e^{-\hat{\beta}_{AFT}} T_0$ we can say that survival times (times to remission) in the treatment group tend to be over 3.5 times those in the control group, meaning that remission tends to occur later for those in the treatment group. In other words the time scale for remission in the treatment group is decelerated by a factor of 3.55 in the treatment group compared with the control group, or equivalently accelerated by a factor 0.28.

8.6 Other parametric AFT models: the log-logistic model

Other distributions for survival data which are in the AFT family of models include the following:

- Log-logistic distribution
- Log-normal distribution
- Gamma distribution
- Inverse gamma

We give some specific details about the log-logistic model. The hazard and survivor functions for the log-logistic model are

$$h(t) = \frac{e^{\theta}\kappa t^{\kappa-1}}{1 + e^{\theta} t^{\kappa}}, \qquad S(t) = \frac{1}{1 + e^{\theta} t^{\kappa}} \tag{8.13}$$

Figure 8.1 shows the shape of the hazard and survivor functions for the log-logistic distribution with specified values for the parameters θ and κ.

Explanatory variables can be incorporated into the log-logistic model as follows:

$$h(t|x) = \frac{e^{\theta} e^{\kappa \beta_{AFT} x} \kappa t^{\kappa-1}}{1 + e^{\theta} e^{\kappa \beta_{AFT} x} t^{\kappa}}, \qquad S(t|x) = \frac{1}{1 + e^{\theta} e^{\kappa \beta_{AFT} x} t^{\kappa}} \tag{8.14}$$

These functions accommodating explanatory variables are also of accelerated failure time form. Hence the log-logistic model is in the AFT family of models.

Assessing whether a log-logistic model may be appropriate

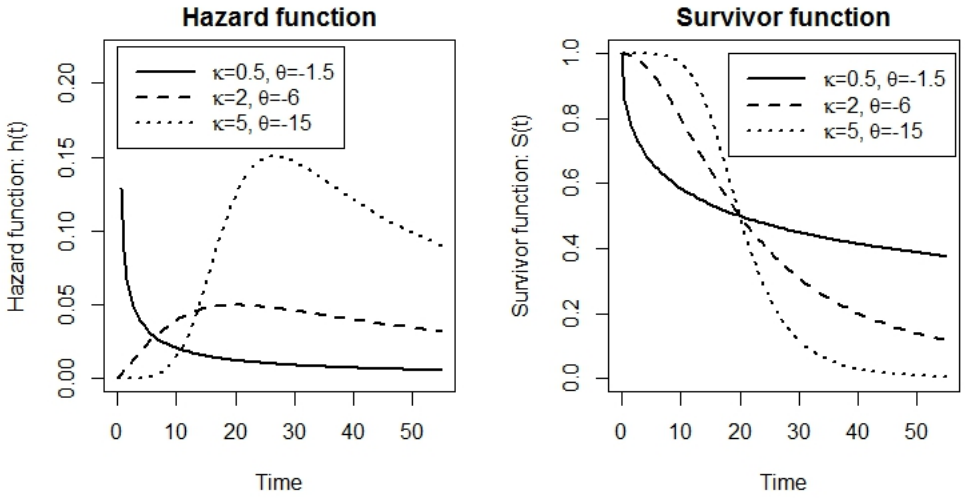


Figure 8.1: Hazard and survivor curves under the log-logistic model with different values for parameters.

In simple situations (binary or categorical covariate), the suitability of a log-logistic model can be assessed graphically. To see how graphs can be used we first note that under the log-logistic model the log odds of survival beyond time t is

$$\log \frac{S(t|x)}{1-S(t|x)} = -\theta - \kappa\beta_{AFT}x - \kappa \log t \tag{8.15}$$

For a binary covariate x we have

$$\log \frac{S(t|0)}{1-S(t|0)} = -\theta - \kappa \log t. \qquad \log \frac{S(t|1)}{1-S(t|x1} = -\theta - \kappa\beta_{AFT} - \kappa \log t \tag{8.16}$$

It follows that under the log-logistic model a plot of $\log \frac{S(t|x)}{1-S(t|x)}$ $(x = 0, 1)$ against $\log t$ will show straight and parallel lines in the two groups.

AFT models and proportional hazards models are not in general nested models (with the exception of the exponential model being nested within the Weibull model). Therefore likelihood ratio tests cannot be used to compare the fit of different models (e.g. Weibull and log-logistic). One way of comparing the fit of non-nested models is to use the Akaike Information Criterion (AIC). This was introduced in the GLM module. There are some different definitions of the AIC, but one is

$$AIC = -2 \times \text{loglikelihood} + 2p \tag{8.17}$$

where p is the number of parameters in the model in question. A lower AIC value indicates a better fit.

Example 8.2

Leukaemia patient data: Fitting the log-logistic model

In Example 8.1 we fitted a Weibull model to the leukaemia patient data. Here we fit a log-logistic model. This can be done in Stata as follows:

```
. streg group, d(loglogistic)

Loglogistic regression -- accelerated failure-time form

No. of subjects =           42                     Number of obs   =        42
No. of failures =           30
Time at risk    =          541
                                                   LR chi2(1)      =     15.38
Log likelihood  =   -48.146027                     Prob > chi2     =    0.0001

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |   1.265463   .3256613     3.89   0.000      .627179    1.903748
       _cons |   1.892691   .2076196     9.12   0.000     1.485764    2.299618
-------------+----------------------------------------------------------------
     /ln_gam |  -.6041234   .1501221    -4.02   0.000    -.8983572   -.3098896
-------------+----------------------------------------------------------------
       gamma |   .5465533   .0820497                      .4072381     .733528
------------------------------------------------------------------------------
```

The parameterization of the log-logistic model used in Stata is different to the parameterization we have used in the lecture notes. The survivor function used is Stata is of the form

$$S(t|x) = \frac{1}{1 + e^{\beta_0/\gamma} e^{\beta_1 x/\gamma} t^{1/\gamma}} \tag{8.18}$$

The equivalencies between the parameters we have used and those used by Stata are shown in the table below.

Our model	Stata model	Stata output
β_{AFT}	$-\beta_1$	`-group`
κ	$1/\gamma$	`1/gamma`
λ	$e^{\beta_0/\gamma}$	`exp(-_cons)/gamma`

Exercise 8.3

(a) Using the above Stata output, write down the parameter estimates. What is the interpretation of the β_{AFT} parameter?

(b) The AIC is 100.128 using the Weibull model and 102.292 using the log-logistic model. Which would be your preferred model?

8.7 Aalen's additive hazards model

The additive hazards model was introduced in 1980 by Aalen. For a single covariate x, under this model the hazard at time t is

$$h(t|x) = \beta_0(t) + \beta_1(t)x \tag{8.19}$$

where $\beta_0(t)$ is the baseline hazard (when $x = 0$) and $\beta_1(t)$ is the increase in the hazard at time t corresponding to a unit increase in x. $\beta_1(t)$ therefore has an interpretation as the excess hazard at time t for a unit increase in x. The additive increment to the hazard can be different at any time t, and hence the impact of x on the hazard is allowed to depend on t. The additive hazards model extends in a straightforward way to accommodate several covariates $x_1, \ldots, x_p$:

$$h(t|\mathrm{x}) = \beta_0(t) + \beta_1(t)x_1 + \beta_2(t)x_2 + \cdots + \beta_p(t)x_p \tag{8.20}$$

It also accommodates time-dependent covariates:

$$h(t|\mathrm{x}(t)) = \beta_0(t) + \beta_1(t)x_1(t) + \beta_2(t)x_2(t) + \cdots + \beta_p(t)x_p(t). \tag{8.21}$$

In this session we focus on time-fixed covariates for simplicity.

Under the additive hazards model the parameters $\beta_j(t)$ $(j = 0, 1, \ldots, p)$ are left unspecified and the model is fully non-parametric. The additive hazards model allows for time-varying effects through the time-dependent parameters $\beta_j(t)$, and because the form of these is left unspecified, we don't have to make assumptions about how they depend on time.

The additive hazards model has not been used anywhere near as commonly as the Cox model in 'routine' analysis of survival data. A criticism of the additive hazards model is that it does not constrain the hazard to be non-negative. However, its attractive properties are increasingly being recognised. Advantages of the additive hazards model include:

- It handles time-varying effects of covariates more easily than the extended Cox model, in which we have to specify the functional form of time-varying parameters
- An additive model may sometimes better describe the underlying relationships
- Additive hazards models are collapsible, unlike multiplicative hazard models such as the Cox model
- Without going into details, it has properties that make it attractive for use in causal inference and in studies of direct and indirect effects (mediation analysis).

8.8 Fitting the additive hazards model

Estimation of the additive hazards model focuses on estimating the cumulative regression coefficients

$$B_j(t) = \int_0^t \beta_j(u)du \qquad (j = 0, 1) \tag{8.22}$$

The estimation involves fitting a linear regression model at each event time among the individuals at risk at that time. To explain this it is necessary to introduce some new notation and concepts. Here we try to give an overview of this without going into too much detail. Much of the formal theory behind the analysis of survival data is based on 'counting processes'. A counting process counts the number of events as they occur over time. We provide a very brief introduction to counting processes, enabling an explanation of how the additive hazards model is estimated. See Aalen et al (2008) for more details.

Considering a study of some event of interest (e.g. death), let $N(t)$ denote the number of events that have occurred up to and including time t. $N(t)$ is a counting process. The number of jumps in the counting process $N(t)$ in the short time period $[t, t + dt)$ is denoted $dN(t)$. It is assumed that the period $[t, t + dt)$ is small such that $dN(t)$ is equal to either 0 or 1, i.e. there is at most one event in the period $[t, t + dt)$. The 'intensity process' $h(t)dt$ is defined as the conditional probability that an event occurs in the short time period $[t, t + dt)$ given everything that has occurred prior to time t, divided by the length of the interval, giving

$$h(t)dt = \Pr(dN(t) = 1|\text{past}) = E(dN(t)|\text{past}) \tag{8.23}$$

The above can be rearranged as $E(dN(t)|\text{past}) - h(t)dt = 0$. This motivates the equation

$$dN(t) = h(t)dt + dM(t) \tag{8.24}$$

where $dM(t)$ is a noise term (analogous to the residual in a linear regression model), where $M(t) = N(t) - \int_0^t \lambda(u)du$ is a Martingale, with $E(dM(t)|\text{past}) = 0$. The Martingale has an interpretation as the cumulative noise.

For a single explanatory variable x the hazard under an additive model for individual i takes the form $h(t|x_i) = \beta_0(t) + \beta_1(t)x_i$ (equation 8.19). Let $Y_i(t)$ denote the indicator of individual i being at risk at time t, taking value 1 if individual i remains at risk at time t and value 0 otherwise. Using this notation, together with the result in equation (8.24) gives the equation

$$dN_i(t) = Y_i(t)dB_0(t) + Y_i(t)x_i dB_1(t) + dM_i(t). \tag{8.25}$$

where $dB_0(t)$ and $dB_1(t)$ denote the increments in the cumulative regression coefficients $B_0(t)$ and $B_1(t)$ at time t. This is a linear regression model where the outcome is $dN_i(t)$, which takes value 0 or 1. The model is fitted separately at each time t using data on all individuals at risk at time t ($Y_i(t) = 1$). For individuals with $Y_i(t) = 1$ the outcome $dN_i(t)$ is equal to 1 for at most one individual (the person who has the event, if an event occurs at time t) and value 0 for everyone else. So we are fitting a linear regression model at each time t, and the outcome is binary. The estimate of $dB_0(t)$ is the increment in the baseline hazard at time t and the estimate of $dB_1(t)$ is the additional increment in the hazard at time t for each unit increase in x. Because at most one individual has the event at any time t, the components $dB_0(t)$ and $dB_1(t)$ are imprecisely estimated. Hence the focus is typically on the cumulative coefficients

$B_0(t)$ and $B_1(t)$ which are estimated as the cumlative sums of the , i.e.

$$\hat{B}_0(t) = \sum_{t_j \leq t} d\hat{B}_0(t_j), \qquad \hat{B}_1(t) = \sum_{t_j \leq t} d\hat{B}_1(t_j) \tag{8.26}$$

where the sum is over event times up to and including time t.

The above results extend easily to several explanatory variables x and to time-dependent explanatory variables. For a single binary explanatory variable x, the estimates of the cumulative hazard in the two groups as obtained from the additive hazards model are equivalent to the Nelson-Aalen estimates of the cumulative hazards. The Nelson-Aalen estimator for the cumulative hazard was introduced in lecture 2.

Asymptotic theory for the additive hazards model, fitted using the linear model in (8.25), follows from that of linear regression models and also relies of properties of Martingales. These can be used to obtain 95% CIs for the cumulative regression coefficients. Goodness of fit tests can also be performed using the residuals. Note that martingale residuals are not normally distributed. Details are not given here.

Example 8.3

Leukemia example: additive hazards model

We investigate fitting the additive hazards model to the leukemia data used in previous examples. There is only one covariate of interest in this example: 'group' (denoting treatment group).

The additive hazards model can be estimated in Stata using the `stlh` command, for example:

```
stlh group, xlabel(0,5,10,15,20) l1title("Cumulative regression function")
    b1title(" ") b2title("Time") gen(B)
```

In R the additive hazards model can be fitted using the `aalen` function in the `timereg` package, and there are ways of fitting the model in other R packages too.

The results from the additive hazards model are typically presented graphically. The plots of the cumulative regression coefficients from fitting the additive hazards model to the leukemia data are shown in Figure 8.2.

The plot for the intercept is showing how the baseline hazard is increasing over time in an approximately linear way. The plot for the treatment group variable shows that the impact of treatment is to reduce the hazard in an approximately linear way over time. The 95% CIs on the plot for 'group' indicate evidence at the 5% level of an impact of treatment group on the hazard. A linear association and a slope of 1 would indicate that the coefficient for 'group' in the additive hazards model does not depend on time. Tests of the hypothesis of a constant coefficient can be performed, but we do not give the details. Extensions to Aalen's additive hazards model have been made,

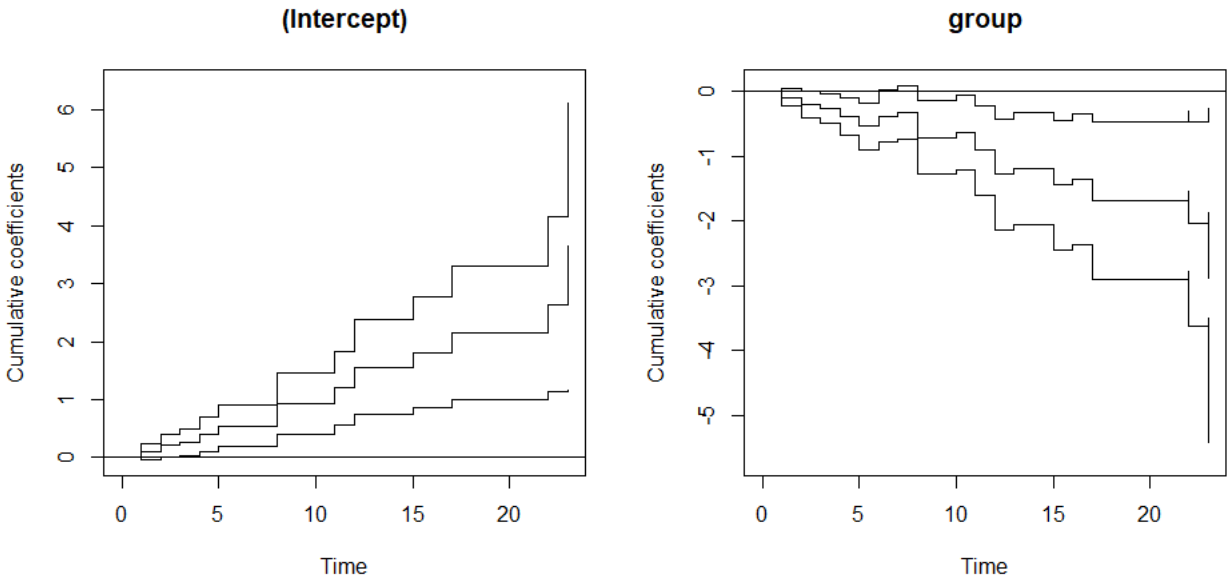


Figure 8.2: Results from fitting the additive hazards model to the leukemia data: estimates of cumulative regression coefficients

in which some or all regression coefficients are constant over time - those models are semi-parametric rather than non-parametric.

Example 8.4

Acute graft versus host disease example: additive hazards model

In Lecture (example 4.4) we looked at data from a trial of people who received a bone marrow transplant. The event of interest was diagnosis of a life-threatening stage of acute graft versus host disease (AGVHD). Figure is a repeat of Figure 4.8, were we looked at the survival curves in two treatment groups. One group received methotrexate alone (MTX) (treatment=0) and the other group received methotrexate plus cyclosporine (CSP+MTX) (treatment=1).

There is clear evidence of non-proportional hazards from these plots. An additive hazards model was fitted to these data. Figure 8.4 shows the resulting estimated cumulative coefficients. The cumulative coefficient for treatment group is around 0 up until about time 20, and then decreases. This indicates no differences in the hazard in the two treatment groups p to around time 20, after which the CSP+MTX treatment group had a lower hazard.

The additive hazards model becomes more useful for detecting time-varying associations between treatment and the hazard when the hazard model includes several covariates, and when there are continuous covariates. The additive hazards model can be used to obtain estimated survival curves at chosen values for the explanatory variables.

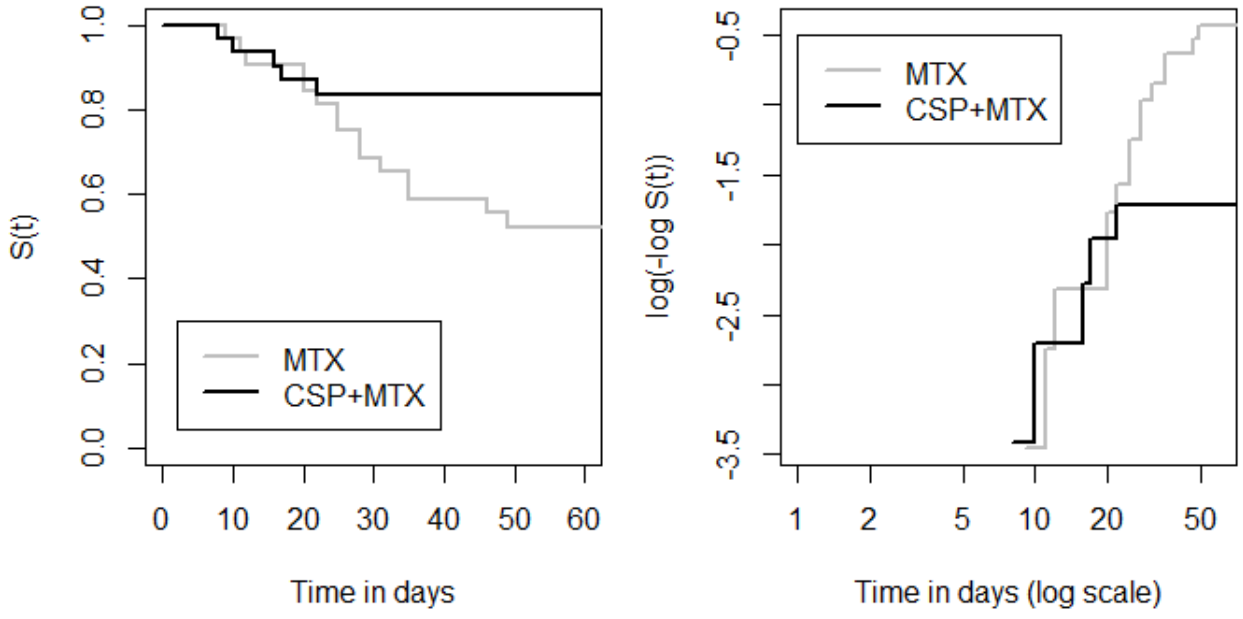


Figure 8.3: Acute graft versus host disease (AGVHD) example: Plot of a Kaplan-Meier estimate of $\log\{-\log S(t|x)\}$ against $\log t$ in the two treatment groups ($x = 0, 1$).

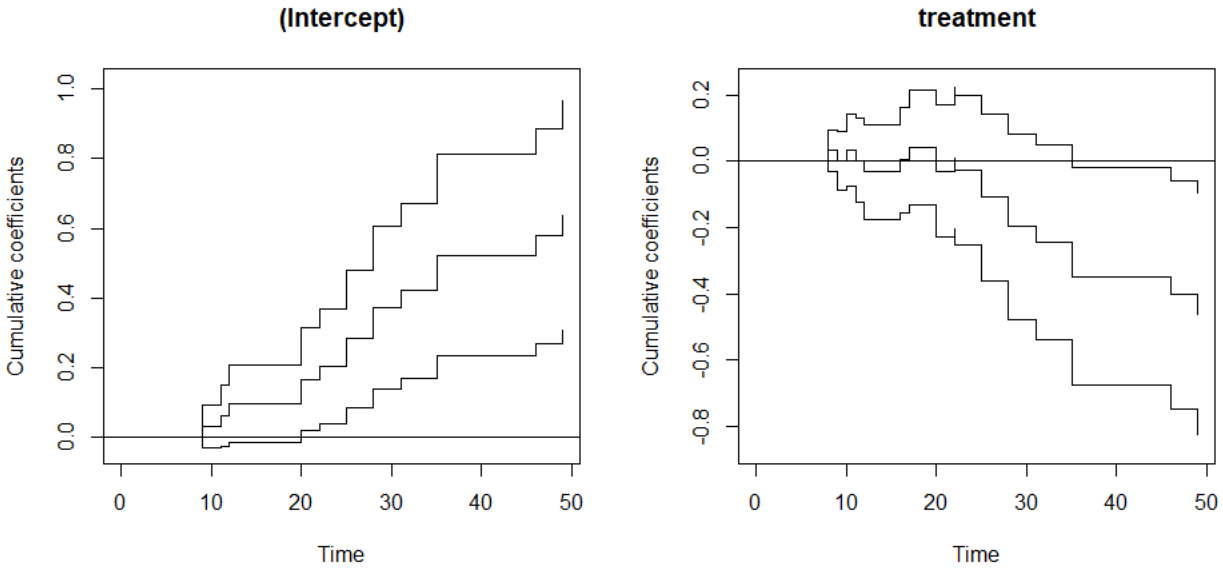


Figure 8.4: Results from fitting the additive hazards model to the Acute graft versus host disease data: estimates of cumulative regression coefficients

The Stata Journal paper by Hosmer and Royston (2002) on the additive hazards model and the `stlh` command provides a nice overview of the model and some examples.

References

Aalen OO. A model for non-parametric regression analysis of life times. In W. Klonecki, A. Kozek, and J. Rosinski (Eds), *Mathematical Statistics and Probability Theory*, Volume 2 of *Lecture Notes in Statistics*, pp 1-25. 1980. New York: Spinger-Verlag.

Aalen OO. A linear regression model for the analysis of life times. *Statistics in Medicine* 1989; 8: 907-925.

Aalen OO, Borgan O, Gjessing HK. *Survival and event history analysis: A process point of view.* Springer. 2008

Collett D, Kimber A. *Modelling Survival Data in Medical Research.* CRC Press. 2014.

Hosmer HW, Royston P. Using Aalen's linear hazards model to investigate time-varying effects in the proportional hazards regression model. *The Stata Journal* 2002; 2: 331-350.

Practical 8

Datasets required: `pbcbase_2021` and `alloauto`

R packages required: survival, flexsurv, eha, timereg.

Introduction

There are three parts to this session

- Part A of the practical investigates two types of **Accelerated Failure Time models**: the Weibull model and the log-logistic model. We will use the familiar PBC base dataset.
- In Part B we look at an additive model: **Aalen's additive hazard model**. For this we will again use the PBC data.
- The final, optional, section of this session offers extra practice on the models used in Parts A and B, using the data on bone marrow transplants among people with advanced acute myelogenous leukemia.

Details are given below for how to carry out the practical in Stata. A solutions Do file will also be made available An R script file is provided on Moodle (from the start of the practical), which shows how the same analyses can be performed using R.

Aims

By the end of this session you should be able to

- Interpret the acceleration factor in an accelerated failure time model
- Show how the Weibull model can be parameterized as a proportional hazards model or an accelerated time model and compare the parameters under each formulation
- Interpret the results from fitting a log-logistic distribution model to survival data
- Interpret the results from Aalen's additive model to survival data
- Fit accelerated failure time models and additive hazards models in the software of your choice and interpret the results appropriately

Part A: AFT models

We will consider a survival model including treatment group and baseline bilirubin measurement (`bil0`) as explanatory variables, as in Practicals 4 and 5.

1. Read the pbcbase data into Stata or R, remind yourself of the key variables, and apply `stset` if you're using Stata. We will be using the time-in-study timescale.
2. Fit a Weibull model with treatment group as the explanatory variable, using the proportional hazards parameterization. Interpret your results.

   In Stata: Try running the `streg` analysis with and without the `nohr` option (this might help you in a later question).

In R use the `weibreg` function as in earlier practicals..

3. Fit the Weibull model again, but this time use the accelerated failure time (AFT) parameterization.

   In Stata using the `time` option in `streg` indicates the AFT model: `streg i.treat, distribution(weibull) time`

   In R `survreg` (shown below) uses the AFT parametrization for the Weibull, as does `flexsurvreg` (see R script):

   ```
   mod.weib.aft=survreg(Surv(time,d) as.factor(treat),data=pbc,dist="weibull")
   summary(mod.weib.aft)
   ```

   What is the estimate of the acceleration factor $e^{\beta_{AFT}}$? Interpret the estimate.

4. What is the relationship between your estimates of $e^{\beta_{PH}}$ and $e^{\beta_{AFT}}$?

5. In Practicals 4 and 5 you also investigated adding bilirubin measured at the start of the trial as an explanatory variable. In practical 5 we found that using a log transformation of this variable was a good idea.

   (a) Create the log-transformed bilirubin variable.

   (b) Fit a Weibull model including treatment and log bil0 as explanatory variables using the proportional hazards parametrization and interpret the results.

   (c) Fit the Weibull model again using the AFT parameterization, find the estimates of the acceleration factors, and interpret the results.

   Discuss: Interpret the associations between treatment and bilirubin and survival.

6. A log-logistic model is another type of AFT model. Fit a log-logistic model with treatment and log bilirubin as explanatory variables. Find the estimates of the acceleration factors, and interpret the results.

   In Stata: use `streg` with the option `dist(loglogistic)`

   In R: use `survreg` with the option `dist="loglogistic"`

7. Compare the fit of the Weibull model and the log-logistic model using the AIC.

   In Stata, one way to do this is as follows:

   ```
   streg i.treat logbil0, distribution(weibull)
   estimates store A
   streg i.treat logbil0, distribution(loglogistic)
   estimates store B
   estimates table _all, stats(aic)
   ```

   In R you can get the AIC from a fitted model using `AIC(model)`

   Discuss: Which model provides the better fit?

Part B: Aalen's additive hazards model

8. Fit the additive hazards model using treatment as the only explanatory variable:

   In Stata:

   To fit an additive hazard model in Stata you must first install the `stlh` command:

   ```
   net install st0024.pkg
   ```

   This user-written command fits an Aalen's linear hazard model to the data, and produces plots of the estimated cumulative regression coefficients. It can, optionally, generate new variables of these estimates. See the help file for more details. Note that the `stlh` command doesn't like it when we try to use `i.treat`, so binary variables should all be coded as 0/1. Recode the `treat` variable as a 0/1 variable, so that the intercept is interpretable.

   To fit the model:

   ```
   stlh treat
   ```

   Note that well as creating graphical output, `stlh` can also be used to record the estimated cumulative regression coefficients in the data.

   ```
   stlh treat, gen(cumulative_coef)
   ```

   In R:

   First, recode the `treat` variable as a 0/1 variable, so that the intercept is interpretable. We will use the `aalen` function from the 'timereg' package to fit the additive hazards model. There are other functions available in other packages too.

   ```
   mod.aalen=aalen(Surv(time,d) treat,data=pbc) plot(mod.aalen)
   ```

   The cumulative regression coefficients can be seen using `View(mod.aalen$cum)`

   Discuss: Interpret the graphical output.

   The following advanced questions will help extend your understanding of Aalen additive hazards models.

9. EXTRA:

   (a) How would you estimate the survival probability from the additive hazards model?

   (b) What is the probability of survival beyond time 5 for an individual in the placebo group and for an individual in the active treatment group?

10. EXTRA:

    (a) Add log bilirubin as an explanatory variable into the additive hazards model. Interpret the graphical output.

(b) Do you see anything surprising?

(c) Centre the log bilirubin variable by subtracting the mean (which is approximately 3.5), and run the additive hazards model again using the centered variable. Compare the plots with what was found using the uncentered variable.

11. EXTRA: What is the probability of survival beyond time 5 for an individual in the placebo group with log bilirubin equal to 5?

Part C: Extra exercises

This section is a chance to practice fitting and interpreting a log-logistic model and an additive hazards model, using the alloauto dataset introduced in practicals 4 and 5.

12. Read the alloauto data into Stata or R. In Stata use `stset` to prepare the data for survival analysis (time-in-study is the only available timescale with these data).

13. Plot the Kaplan-Meier survival curves by treatment type. Is the proportional hazards assumption met? You may wish to use other plots or tests used in earlier practicals.

14. Use a suitable plot to investigate whether a log-logistic model would be appropriate. If so, fit the log-logistic model and interpret your results.

15. To investigate the Aalen additive hazards model, we must first recode the `type` variable as a 0/1 variable, as on the earlier exercise. Fit an additive hazards model with treatment type as the only covariate and interpret the results.