
# Survival Analysis

## Practicals: R version

# Table of Contents

# Practical 1

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

R packages required: survival, flexsurv.

## Introduction

This first practical is in three parts.

A Looks at some fundamentals of setting up 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 $\lambda$ 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
- Be able to derive the formula for the maximum likelihood estimator of the hazard rate parameter $\lambda$ 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

## Part A: Using survival data in 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 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 R stores 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:

   ```
   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 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. 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 |

5. 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")
   ```

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

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

   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 `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 `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 $\lambda$.

   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 R to help you calculate, by hand, the maximum likelihood estimate for $\lambda$ for these data.
4. We will fit an exponential model to these data to confirm the maximum likelihood estimate for $\lambda$ we calculated above. Compare the results with what you found by hand.

   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 $\lambda$.

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 (**??**). Here we will investigate graphically how the values of the parameters $\lambda, \kappa$ affect the hazard function. Play around with the values of the two parameters to investigate the effect on the shape of the hazard. In R we can plot the hazard function for example values of $\lambda, \kappa$ 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 $\theta$ and $\kappa$. What feature does the log-logistic distribution have that the Weibull distribution does not have?

# Practical 2

Datasets required: `pbcbase_2021` and `whitehall`

R packages required: 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
- 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

## 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 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 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 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 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 R code from previous questions (and Practical 1) to answer these questions.

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:

| 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 |

(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. In R you can add the option `risk.table = T` into the `ggsurvplot` function to see the number of individuals at risk at selected time points.

**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?**

# Practical 3

Datasets required: `whitehall`

R packages required: 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 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

## Questions

1. Load the data and explore the grade variable. Summarize the numbers and timings of CHD deaths and censorings by job grade.
   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 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?
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 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 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 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 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 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 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.

# Practical 4

Datasets required: `pbcbase_2021` & `alloauto`

R packages required: 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 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

## 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 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 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 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 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.

# Practical 5

Datasets required: `alloauto` and `pbcbase_2021`

R packages required: '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

## 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 R and identify the outcome variables (event/censoring time and event indicator).

   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 no association between the residuals and time.

   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

```
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

```
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

```
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.

   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 `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:

```
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 `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.

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`:

   ```
   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.

   (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.

```
#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?**

# Practical 6

Dataset required: `aaatrial_2016`

R packages required: survival, Epi, cmprsk.

## Introduction

This session looks at the issue of competing risks in survival analysis. You will use 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

## 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 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 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 `Surv` in R when there are competing risks.

   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 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 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 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 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 **??** 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 `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?**

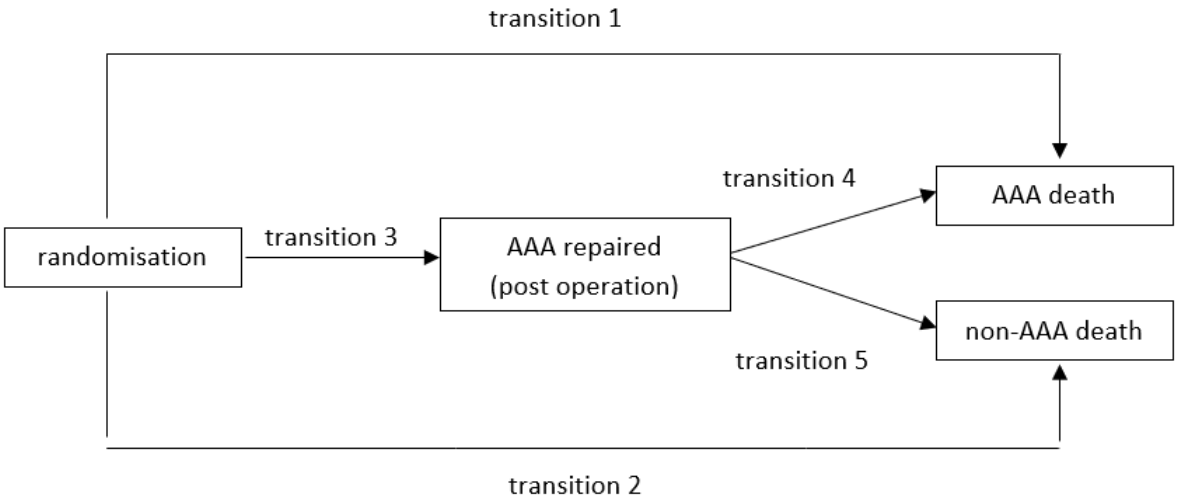
7. Repeat the investigations for non-AAA-death as the cause of interest, with AAA-death treated as a competing risk. How does randomisation arm affect non-AAA-deaths.

**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 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 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?

# Practical 7

Datasets required: `hip4` for part A. For part B: `kidney_frailty`.

R packages required: 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
- Interpret the results from fitting a survival model with time-dependent explanatory variables
- Fit and interpret frailty models

## 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) |

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 R, you should use the 'counting process' notation for the `Surv` object.

   ```
   Surv(hip$time0, hip$time1, hip$fracture, type="counting")
   ```

   (a) 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 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 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 `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 commands:

```
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 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. 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.

The 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)
```

# 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.

## 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 R and remind yourself of the key variables. 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 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 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 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 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 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 R.
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.