Modern Demographic Methods in Epidemiology with R

Bendix Carstensen Steno Diabetes Center, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk
http://BendixCarstensen.com

University of Edinburgh
26–29 August 2014
http://BendixCarstensen/AdvCoh/Scot-2014

Introducing R

Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
http://BendixCarstensen/AdvCoh/Scot-2014

Data

The best way to learn R

- The best way to learn **R** is to use it!
- This is a very short introduction before you sit down in front of a computer.
- **R** is a little different from other packages for statistical analysis.
- These differences make **R** very powerful, but for a new user they can sometimes be confusing.
- Our first job is to help you up the initial learning curve so that you can be comfortable with R.

Nothing is lost or hidden

- Statistical software provides "canned" procedures to address common statistical problems.
- Canned procedures are useful for routine analysis, but they are also limiting.
  - You can only do what the programmer lets you do.
- In R, the results of statistical calculations are always accessible.
  - You can use them for further calculations.
  - You can always see how the calculations were done.

R Packages

- The capabilities of **R** can be extended using "packages".
- Distributed over the Internet *via* **CRAN**: (the **C**omprehensive **R** **A**rchive **N**etwork) and can be downloaded directly from an **R** session.
- There is an **R** package developed during the annual course on "Statistical Practice in Epidemiology using **R**, called "`Epi`".
- Contains special functions for epidemiologists and some data sets that .
- There are 5,825 other user contributed packages on CRAN.

Objects and functions

R allows you to build powerful procedures from simple building blocks. These building blocks are **objects** and **functions**.

- All data in **R** is represented by **objects**, for example:
  - A dataset (called data frame in R)
  - A vector of numbers
  - The result of fitting a model to data
- You, the user, call **functions**
- Functions act on objects to create **new objects**:
- Using `glm` on a dataframe (an object) produces a fitted model (another object).

Because all is functions. . .

- You will always (almost) use parentheses:
  ```
  > res <- FUN( x, y )
  ```
- . . . which is pronounced
- `res` **gets** ("`<-`") `FUN` **of** `x,y` ("`(x,y)`")

Vectors

One of the simplest objects in **R** is a sequence of numbers, called a **vector**.

You can create a vector in **R** with the collection (`c`) function:

```
> c(1,3,2)
[1] 1 3 2
```

You can save the results of any calculation using the left arrow:

```
> x <- c(1,3,2)
> x
[1] 1 3 2
```

The workspace

- Every time you use `<-`, you create a new object in the **workspace** (or overwrite an old one).
- A list of objects in the workspace can be seen with the `objects` function (synonym: `ls()`):

```
> objects()
 [1] "a"  "aa"  "acz2"  "alpha"  "b"
 [6] "bar"  "bb"  "bdendo"  "beta"  "cc"
[11] "Col"
```

- In `Epi` is a function `lls()` that gives a bit more information on the objects.
- The workspace is held entirely in (volatile) computer memory and will be lost at the end of the session unless you explicitly save it.

Working Directory

Every **R** session has a **current working directory**, which is the location on the hard disk where files are saved, and the default location from which files are read into R.

- `getwd()` Prints the current working directory
- `setwd("c:/Users/Martyn/Project")` sets the current working directory.
- You may also use a Graphical User Interface (GUI) to change directory.

Ending an R session

- To end an **R** session, call the `quit()` function
  - Every time you want to do something in R, you call a function.
- You will be asked "Save workspace image?"

  Yes saves the workspace to the file ".RData" in your current working directory. It will be automatically loaded into **R** the next time you start an **R** session.

  No does not save the workspace.

  Cancel continues the current **R** session without saving anything.
- It is recommended you just say "No".

Always start with a clean workspace

Keeping objects in your workspace from one session to another can be dangerous:

- You forget how they were made.
- You cannot easily recreate them if your data changes.
- They may not even be from the same project

It is almost always best to start with an empty workspace and use a script file to create the objects you need from scratch.

Rectangular Data

Rectangular data sets are common to most statistical packages

"id"	"visit"	"time"	"status"
1	1	0.0	0
1	2	1.5	0
2	1	0.0	0
2	2	1.1	0
2	3	2.3	1

Columns represent variables.
Rows represent individual records.

The world is not a rectangle!

- Most statistical packages used by epidemiologists assume that **all data** can be represented as a rectangular data set.
- R allows a much richer set of data structures, represented by *objects* of different *classes*.
- Rectangular data sets are just one type of object that may be in your workspace. This class of object is called a *data frame*.

Data Frames

Each column of a data frame is a variable.

Variables may be of different types:

- **vectors:**
  - **numeric:** `c(1,2,3)`
  - **character:** `c("John","Paul","George","Ringo")`
  - **logical:** `c(FALSE,FALSE,TRUE)`
- **factors:** `factor(c("low","medium","high","low","low"))`

Building your own data frame

Data frames can be constructed from a list of vectors

```
> mydata <- data.frame(x=c(3,6,7),f=c("a","b","a"))
> mydata
  x f
1 3 a
2 6 b
3 7 a
```

Character vectors are automatically converted to factors.

Inspecting data frames

Most data frames are too large to inspect by printing them to the screen, so use:

- `names` returns a vector of variable names.
  - You can use `sort(names(x))` to get them in alphabetical order.
- `head` prints the first few lines, and `tail`...
- `str` prints a brief overview of the **str**ucture of the data frame. Can be used on any object.
- `summary` prints a more comprehensive summary
  - Quantiles for numeric variables
  - Tables for factors

Extracting values from a data frame

Use square brackets to take **subsets** of a data frame

- `mydata[1,2]`. The value in row 1, column 2.
- `mydata[1,]`. The whole of the first row.
- `mydata[,2]`. The whole of the second column.

You can also extract a column from a data frame by name:

- `mydata$age`. The column, or variable, named "age"
- `mydata[,"age"]`. The same.

Importing data

- R has good facilities for importing data from other applications:
  - `read.dta` for reading Stata datasets.
  - `read.spss` for reading SPSS datasets.
  - `read.xport` and `read.ssd` for reading SAS-datasets.

Reading Text Files

The function `read.table` reads data from a text file and returns a data frame.

- `mydata <- read.table("myfile")`
- `myfile` could be
  - A file in the **current working directory**: `fem.dat`
  - A path to a file: `c:/rex/fem.dat`
  - A URL: `http://BendixCarstensen.com/AdvCoh/Scot-2014/data/bogus.txt`
- Note: `myfile` must be enclosed in quotes.

`write.table` does the opposite.

R uses a forward slash `/` for file paths. If you want to use backslash, you have to double it:

`c:\\rex\\fem.dat`

Some useful arguments to `read.table`

- `header = TRUE` if first line contains variable names
- `sep=","` if values are comma-separated instead of being space-delimited.
- `as.is = TRUE` to stop strings being converted to factors
- `na.strings = "99"` to denote that 99 means "missing". Default values are:
  - `NA` "Not Available"
  - `NaN` "Not a Number"
- For comma-separated files there is coderead.csv

Reading Binary Data

- **R** can read in data in binary (non-text) format from other statistical systems using the foreign extension package.
- R is an open source project, and relies on the format for binary files to be well-documented.
- Example: SAS `XPORT` format has been adopted as a data exchange standard by the US Food and Drug Administration. SAS CPORT format remains a proprietary format.

Some functions in the foreign package

- `read.dta` for Stata (also `write.dta`)
- `read.xport` for SAS XPORT format (not CPORT)
- `read.epiinfo` for EPIINFO
- `read.mtp` for MiniTab Portable Worksheet
- `read.spss` for SPSS

See the "R Data Import/Export manual" for more details. `RShowDoc("R-data")`

Accessing databases systems

Microsoft **Access**:

```
> library(RODBC)
> ch <- odbcConnectAccess("../data/theData.mdb")
> bd <- sqlFetch(ch, "aTable" )
```

Microsoft **Excel**:

```
> library( RODBC )
> cnc <- odbcConnectExcel(paste("../theXel.xls",sep=""))
> sht <- sqlFetch( cnc, "theSheet" )
> close( cnc )
```

Other databases

```
> ?odbcConnect
```

Summary - data

- You can use a data frame to organize your variables
- You can extract variables from a data frame using `$`.
- You can extract variables and observation using indecing `[,]`
- You can read in data using
  - `read.table`
  - tailored function from the `foreign` package
  - database interface from the `RODBC` package

Summary - when it goes wrong

When somthing is fishy with an object `obj`, try to find out what you (accidentally) got, by using:

```
> lls()
> str( obj )
> dim( obj )
> length( obj )
> names( obj )
> head( obj )
> class( obj )
> mode( obj )
```


R language

Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
`http://BendixCarstensen/AdvCoh/Scot-2014`

`lang`

Language

- R is a programming language – also on the command line
- (This means that there are *syntax rules*)
- Print an object by typing its name
- Evaluate an expression by entering it on the command line
- Call a function, giving the arguments in parentheses – possibly empty
- Notice `ls` vs. `ls()`

Objects

- The simplest object type is *vector*
- Modes: numeric, integer, character, generic (list)
- Operations are vectorized: you can add entire vectors with `a + b`
- Recycling of objects: If the lengths don't match, the shorter vector is reused

R expressions

```
x <- rnorm(10, mean=20, sd=5)
m <- mean(x)
sum((x - m)^2)
```

- Object names
- Explicit constants
- Arithmetic operators
- Function calls
- Assignment of results to names

Function calls

Lots of things you do with R involve calling functions.
For instance

```
mean(x, na.rm=TRUE)
```

The important parts of this are

- The name of the function
- Arguments: input to the function
- Sometimes, we have named arguments

Function arguments

```
rnorm(10, mean=m, sd=s)
hist(x, main="My histogram")
mean(log(x + 1))
```

Items which may appear as arguments:

- Names of an R objects
- Explicit constants
- Return values from another function call or expression
- Some arguments have their *default values*.
- Use `help(function)` or `args(function)` to see the arguments (and their order and default values) that can be given to any function.

Creating simple functions

```
logit <- function(p) log(p/(1-p))
logit(0.5)

simpsum <-
function(x, dec=5)
{
  # produces mean and SD of a variable
  # default value for dec is 5
round(c(mean=mean(x),sd=sd(x)),dec)
}

x <- rnorm(100)
simpsum(x)
simpsum(x,2)
```


Indexing

- **R** has several useful indexing mechanisms:
- `a[5]` single element
- `a[5:7]` several elements
- `a[-6]` all except the 6th
- `a[c(1,1,2,1,2)]` some elements repeated
- `a[b>200]` logical index
- `a[■well■]` indexing by name

Lists

- Lists are vectors where the elements can have different types
- Functions often return lists
- `lst <- list(A=rnorm(5),B="hello",K=12)`
- Special indexing:
- `lst$A`
- `lst[1:2]` a list with first two first elements (`A` and `B` — NB: single brackets)
- `lst[1]` a list of length 1 which is the first element (codeA — NB: single brackets)
- `lst[[1]]` first element (NB: double brackets) — a vector of length 5.

Classes, generic functions

- R objects have *classes*
- Functions can behave differently depending on the class of an object
- E.g. `summary(x)` or `print(x)` does different things if `x` is numeric, a factor, or a linear model fit

The workspace

- The *global environment* contains R objects created on the command line.
- There is an additional *search path* of loaded packages and attached data frames.
- When you request an object by name, R looks first in the global environment, and if it doesn't find it there, it continues along the search path.
- The search path is maintained by `library()`, `attach()`, and `detach()`
- List the search path by `search()`
- Notice that objects in the global environment may mask objects in packages and attached data frames

Data manipulation and `with`

```
bmi <- with(stud, weight/(height/100)^2)
```

uses variables weight and height in the data frame `stud` (not the variables with the same name in the workspace), but creates the variable `bmi` in the global environment (not in the data frame).

To create a new variable in the data frame, you can use:

```
stud$bmi <- with( stud, weight/(height/100)^2 )
```


Constructors

- Matrices and arrays, constructed by the (surprise) `matrix` and `array` functions.
- You can extract and set names with `names(x)`; for matrices and data frames also `colnames(x)` and `rownames(x)`
- You can also construct a matrix from its columns using `cbind`, whereas joining two matrices with equal no of columns (with the same column names) can be done using `rbind`.

Factors (class variables)

- Factors are used to describe groupings.
- Basically, these are just integer codes plus a set of names for the *levels*
- They have class `"factor"` making them (a) print nicely and (b) maintain consistency
- A factor can also be *ordered* (class `"ordered"`), signifying that there is a natural sort order on the levels
- In model specifications, factors play a fundamental role by indicating that a variable should be treated as a classification rather than as a quantitative variable (similar to a CLASS statement in SAS)

The `factor` function

- This is typically used when `read.table` gets it wrong,
- e.g. group codes read as numeric
- or read as factors, but with levels in the wrong order (e.g. `c("rare", "medium", "well-done")` sorted alphabetically.)
- Notice that there is a slightly confusing use of `levels` and `labels` arguments:
  - `levels` are the value codes *on input*
  - `labels` are the value codes *on output* (and becomes the levels of the resulting factor)
  - The levels of a factor is shown by the `levels()` function.

Working with Dates

- Dates are usually read as character or factor variables
- Use the `as.Date` function to convert them to objects of class `"Date"`
- If data are not in the default format (`yyyy-mm-dd`) you need to supply a format specification

```
> as.Date("11/3-1959",format="%d/%m-%Y")
[1] "1959-03-11"
```


Working with Dates

- Computing the differences between `Date` objects gives an object of class `"difftime"`, which is number of days between the two dates:

```
> as.numeric(as.Date("2007-5-25")-
             as.Date("1959-3-11"),"days")
[1] 17607
```

- In the `Epi` package is a function that converts dates to calendar years with decimals:

```
> as.Date("1952-07-14")
[1] "1952-07-14"
> cal.yr( as.Date("1952-07-14") )
[1] 1952.533
attr(,"class")
[1] "cal.yr"  "numeric"
```


Basic graphics

The `plot()` function is a generic function, producing different plots for different types of arguments. For instance, `plot(x)` produces:

- a plot of observation index against the observations, when `x` is a numeric variable
- a bar plot of category frequencies, when `x` is a factor variable
- a time series plot (interconnected observations) when `x` is a time series
- a set of diagnostic plots, when `x` is a fitted regression model
- ...

Basic graphics

Similarly, the `plot(x,y)` produces:

- a scatter plot of `x` is a numeric variable
- a bar plot of category frequencies, when `x` is a factor variable

Basic graphics

Examples:

```
x <- c(0,1,2,1,2,2,1,1,3,3)
plot(x)
plot(factor(x))
plot(ts(x))   # ts() defines x as time series
y <- c(0,1,3,1,2,1,0,1,4,3)
plot(x,y)
plot(factor(x),y)
```


Basic graphics

More simple plots:

- `hist(x)` produces a histogram
- `barplot(x)` produces a bar plot (useful when x contains counts – often one uses `barplot(table(x))`)
- `boxplot(y x)` produces a box plot of `y` by levels of a (factor) variable `x`.

Rates and Survival

Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
`http://BendixCarstensen/AdvCoh/Scot-2014`

`surv-rate`

Survival data

Persons enter the study at some date.

Persons exit at a later date, either dead or alive.

Observation:
Actual time span to death ("event")
or
Some time alive ("at least this long")

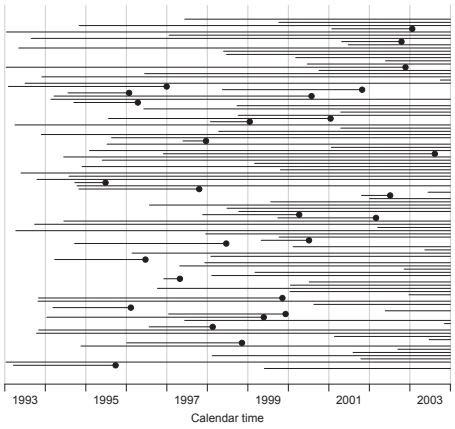
Examples of time-to-event measurements

- Time from diagnosis of cancer to death.
- Time from randomisation to death in a cancer clinical trial
- Time from HIV infection to AIDS.
- Time from marriage to 1st child birth.
- Time from marriage to divorce.
- Time to re-offending after being released from jail

Each line a person

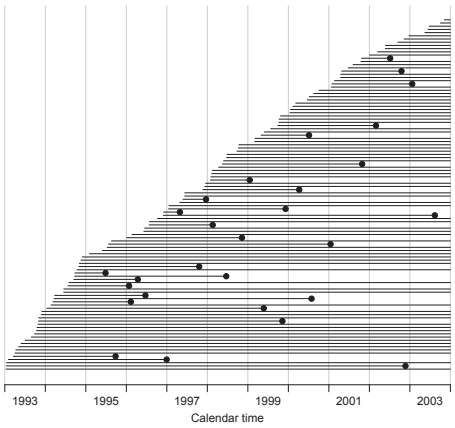
Each blob a death

Study ended at 31 Dec. 2003

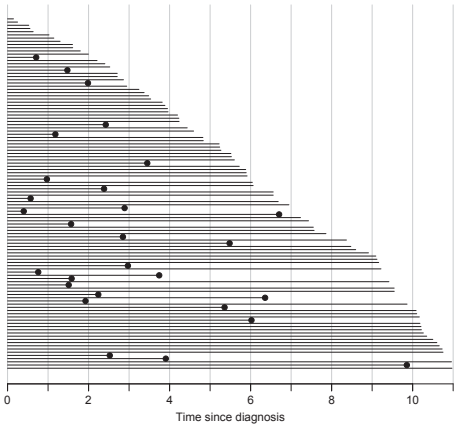


Ordered by date of entry

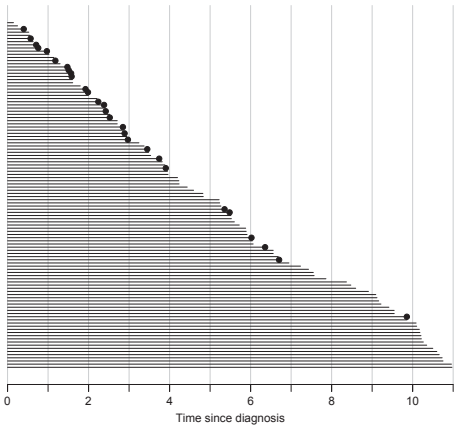
Most likely the order in your database.



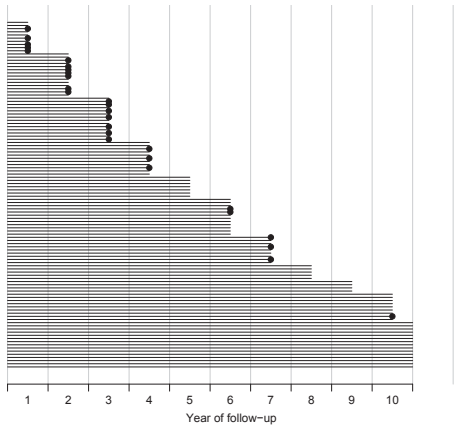
Timescale changed to "Time since diagnosis".



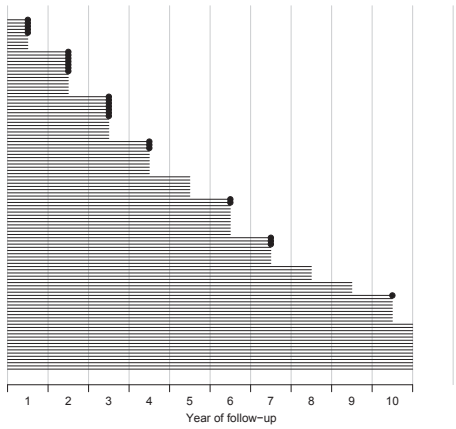
Patients ordered by survival time.



Survival times grouped into bands of survival.



Patients ordered by survival status within each band.



Survival after Cervix cancer

	Stage I			Stage II		
Year	N	D	L	N	D	L
1	110	5	5	234	24	3
2	100	7	7	207	27	11
3	86	7	7	169	31	9
4	72	3	8	129	17	7
5	61	0	7	105	7	13
6	54	2	10	85	6	6
7	42	3	6	73	5	6
8	33	0	5	62	3	10
9	28	0	4	49	2	13
10	24	1	8	34	4	6

Estimated risk in year 1 for Stage I women is $5/107.5 = 0.0465$

Estimated 1 year survival is $1 - 0.0465 = 0.9535$

Life-table estimator.

Survival function

Persons enter at time 0:
Date of birth, date of randomization, date of diagnosis.

How long do they survive?
Survival time T — a stochastic variable.

Distribution is characterized by the survival function:

$$\begin{aligned} S(t) &= \mathrm{P}\{\text{survival at least till } t\} \\ &= \mathrm{P}\{T > t\} = 1 - \mathrm{P}\{T \leq t\} = 1 - F(t) \end{aligned}$$

$F(t)$ is the cumulative risk of death before time t.

Intensity or rate

$$\begin{aligned} &\mathrm{P}\{\text{event in } (t, t+h] \mid \text{alive at } t\}/h \\ &= \frac{F(t+h) - F(t)}{S(t) \times h} \\ &= -\frac{S(t+h) - S(t)}{S(t)h} \xrightarrow[h\to 0]{} -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t} \\ &= \lambda(t) \end{aligned}$$

This is the **intensity** or **hazard function** for the distribution. Characterizes the survival distribution as does f or F.

Theoretical counterpart of a **rate**.

Relationships

$$\begin{aligned} -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t} &= \lambda(t) \\ &\Updownarrow \\ S(t) &= \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right) = \exp\left(-\Lambda(t)\right) \end{aligned}$$

$\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ is called the **integrated intensity**. **Not** an intensity, it is dimensionless.

$$\lambda(t) = -\frac{\mathrm{d}\log(S(t))}{\mathrm{d}t} = -\frac{S'(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

Rate and survival

$$S(t) = \exp\left(-\int_0^t \lambda(s)\,\mathrm{d}s\right) \qquad \lambda(t) = \frac{S'(t)}{S(t)}$$

Survival is a *cumulative* measure, the rate is an *instantaneous* measure.

Note: A cumulative measure requires an origin!

Observed survival and rate

- **Survival studies**: Observation of (right censored) survival time:

  $$X = \min(T, Z), \quad \delta = 1\{X = T\}$$

  — sometimes conditional on $T > t_0$ (left truncation, delayed entry).
- **Epidemiological studies**: Observation of (components of) a rate:

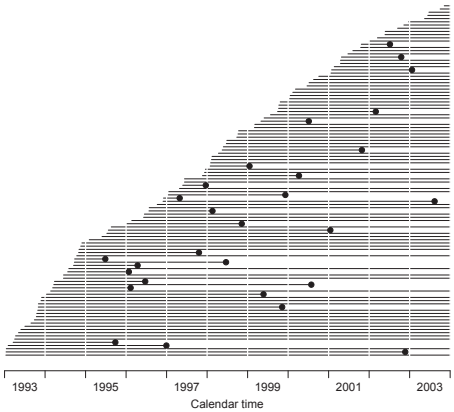
  $$D/Y$$

  D: no. events, Y no of person-years, in a prespecified time-frame.

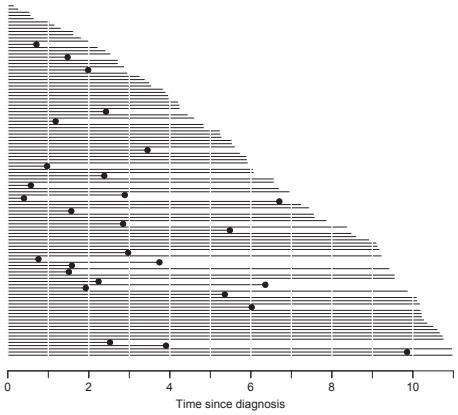
Empirical rates for individuals

- At the *individual* level we introduce the **empirical rate:** (d, y), — number of events ($d \in \{0, 1\}$) during y risk time.
- A person contributes several observations of (d, y), with associated covariate values.
- Empirical rates are **responses** in survival analysis.
- The timescale t is a **covariate** — varies within each individual:
  t: age, time since diagnosis, calendar time.
- Don't confuse with y — difference between two points on **any** timescale we may choose.

Empirical rates by calendar time.



Empirical rates by time since diagnosis.



Statistical inference: Likelihood

Two things needed:

- **Data** — what did we actually observe
  Follow-up for each person:
  Entry time, exit time, exit status, covariates
- **Model** — how was data generated
  Rates as a function of time:
  Probability machinery that generated data

Likelihood is the probability of observing the data, assuming the model is correct.

Maximum likelihood estimation is choosing parameters of the model that makes the likelihood maximal.

Likelihood from one person

The likelihood from several empirical rates from one individual is a product of conditional probabilities:

$$\begin{aligned}\mathrm{P}\{\text{event at } t_4|t_0\} &= \mathrm{P}\{\text{survive } (t_0,t_1)|\text{ alive at } t_0\}\times\\ &\quad \mathrm{P}\{\text{survive } (t_1,t_2)|\text{ alive at } t_1\}\times\\ &\quad \mathrm{P}\{\text{survive } (t_2,t_3)|\text{ alive at } t_2\}\times\\ &\quad \mathrm{P}\{\text{event at } t_4|\text{ alive at } t_3\}\end{aligned}$$

Log-likelihood from one individual is a sum of terms.

Each term refers to one empirical rate (d, y) — $y = t_i - t_{i-1}$ and mostly $d = 0$.

t_i is the timescale (covariate).

Likelihood for an empirical rate

Model: the rate is constant in the interval we are looking at.

The interval should sufficiently small for this assumption to be reasonable:

$$\mathrm{P}\{\text{event in } (t,t+h] \mid \text{alive at } t\}/h = \lambda(t)$$

$$\begin{aligned}&\mathrm{P}\{\text{survive a timespan of } y\} =\\ &\mathrm{P}\{\text{survive } n \text{ int's of length } y/n\} = \big(1-\lambda(t)\frac{y}{n}\big)^n\end{aligned}$$

now, since: $\lim_{n\to\infty}(1+x/n)^n = \exp(x)$

$$\Rightarrow \quad (1-\lambda(t)\times y/n)^n \approx \exp\big(\lambda(t)y\big)$$

Likelihood for an empirical rate

Death probability is: $\pi = 1 - \mathrm{e}^{-\lambda y}$, so for $d = 0, 1$:

$$\begin{aligned}L(\lambda) &= \mathrm{P}\{d \text{ events during } y \text{ time}\} = \pi^d(1-\pi)^{1-d}\\ &= (1-\mathrm{e}^{-\lambda y})^d(\mathrm{e}^{-\lambda y})^{1-d}\\ &= \left(\frac{1-\mathrm{e}^{-\lambda y}}{\mathrm{e}^{-\lambda y}}\right)^d(\mathrm{e}^{-\lambda y}) \approx (\lambda y)^d\mathrm{e}^{-\lambda y}\end{aligned}$$

since the first term is equal to $\mathrm{e}^{\lambda y} - 1 \approx \lambda y$.

Log-likelihood:

$$\ell(\lambda) = d\log(\lambda y) - \lambda y = d\log(\lambda) + d\log(y) - \lambda y$$

The term $d\log(y)$ does not include λ, so the relevant part of the log-likelihood is:

$$\ell(\lambda) = d\log(\lambda) - \lambda y$$

Poisson likelihood

The likelihood contributions from follow-up of **one** individual:

$$d_t\log\big(\lambda(t)\big) - \lambda(t)y_t, \quad t = t_1,\ldots,t_n$$

is also the log-likelihood from several independent Poisson observations with mean $\lambda(t)y_t$, i.e. log-mean $\log\big(\lambda(t)\big) + \log(y_t)$

Analysis of the rates, (λ) can be based on a Poisson model with log-link applied to empirical rates where:

- d is the response variable.
- $\log(\lambda)$ is modelled by covariates
- $\log(y)$ is the offset variable.

Likelihood for follow-up of many subjects

Adding empirical rates over the follow-up of persons:

$$D = \sum d \qquad Y = \sum y \quad \Rightarrow \quad D\log(\lambda) - \lambda Y$$

- Persons are assumed independent
- Contribution from the same person are **conditionally** independent, hence give separate contributions to the log-likelihood.
- No need to correct for dependent observations; the likelihood is a product.

Likelihood theory

- Likelihood depends on **data** (X) and model **parameters** (λ):

$$L(\lambda, X) = \mathrm{P}\{X|\lambda\}, \quad \ell(\lambda, X) = \log\big(\mathrm{P}\{X|\lambda\}\big)$$

- Choose the value of λ that makes the (log-)likelihood as large as possible, $\hat{\lambda}$:

$$\ell(\hat{\lambda}, X) \geq \ell(\lambda, X), \quad \forall\lambda$$

- Standard error of $\hat{\lambda}$:

$$\mathrm{s.e.}(\hat{\lambda}) = 1/\sqrt{-\ell''(\lambda, X)|_{\lambda=\hat{\lambda}}}$$

- $\ell''(\lambda, X)|_{\lambda=\hat{\lambda}}$: **observed information** on λ

Likelihood theory in practise

- Derivatives of the log-likelihood, for a rate λ, w.r.t. $\theta = \log(\lambda)$:

$$\ell(\theta|D, Y) = D\theta - \mathrm{e}^{\theta} Y, \quad \ell'_\theta = D - \mathrm{e}^{\theta} Y, \quad \ell''_\theta = -\mathrm{e}^{\theta} Y$$

- Likelihood maximal if:

$$\ell' = 0 \quad \Leftrightarrow \quad \hat{\lambda} = \mathrm{e}^{\hat{\theta}} = D/Y$$

- Information about $\theta = \log(\lambda)$:

$$-I(\hat{\theta}) = \mathrm{e}^{\hat{\theta}} Y = \hat{\lambda} Y = D \Rightarrow \mathrm{s.e.}(\hat{\theta}) = 1/\sqrt{D}$$

- Note that this only depends on the no. events, **not** on the follow-up time.

Likelihood

Probability of the data and the parameter:

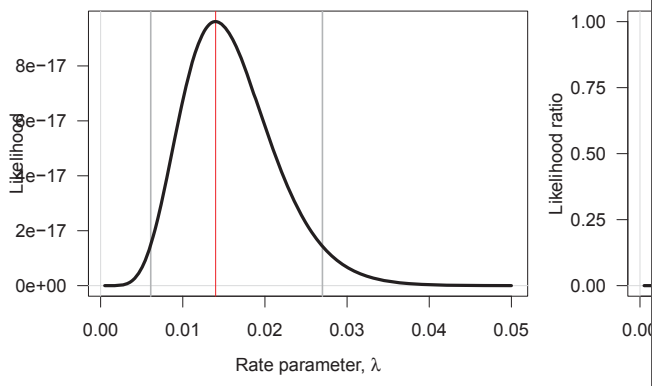
Assuming the rate (intensity) is constant, λ, the probability of observing 7 deaths in the course of 500 person-years:

$$\begin{aligned} \mathrm{P}\{D = 7, Y = 500|\lambda\} &= \lambda^D \mathrm{e}^{\lambda Y} \times K \\ &= \lambda^7 \mathrm{e}^{\lambda 500} \times K \\ &= L(\lambda|\text{data}) \end{aligned}$$

Best guess of λ is where this function is as large as possible.

Confidence interval is where it is not too far from the maximum

Likelihood function



Confidence interval for a rate

A 95% confidence interval for the log of a rate is:

$$\hat{\theta} \pm 1.96/\sqrt{D} = \log(\lambda) \pm 1.96/\sqrt{D}$$

Take the exponential to get the confidence interval for the rate:

$$\lambda \stackrel{\times}{\div} \underbrace{\exp(1.96/\sqrt{D})}_{\text{error factor,erf}}$$

Example

Suppose we have 17 deaths during 843.6 years of follow-up.

The rate is computed as:

$\hat{\lambda} = D/Y = 17/843.7 = 0.0201 = 20.1$ per 1000 years

The confidence interval is computed as:

$$\hat{\lambda} \stackrel{\times}{\div} \mathrm{erf} = 20.1 \stackrel{\times}{\div} \exp(1.96/\sqrt{D}) = (12.5, 32.4)$$

per 1000 person-years.

Ratio of two rates

If we have observations two rates λ_1 and λ_0, based on (D_1, Y_1) and (D_0, Y_0), the variance of the difference of the log-rates, the $\log(\mathrm{RR})$, is:

$$\begin{aligned} \mathrm{var}(\log(\mathrm{RR})) &= \mathrm{var}(\log(\lambda_1/\lambda_0)) \\ &= \mathrm{var}(\log(\lambda_1)) + \mathrm{var}(\log(\lambda_0)) \\ &= 1/D_1 + 1/D_0 \end{aligned}$$

As before a 95% c.i. for the RR is then:

$$\mathrm{RR} \stackrel{\times}{\div} \underbrace{\exp\left(1.96\sqrt{\frac{1}{D_1} + \frac{1}{D_0}}\right)}_{\text{error factor}}$$

Example

Suppose we in group 0 have 17 deaths during 843.6 years of follow-up in one group, and in group 1 have 28 deaths during 632.3 years.

The rate-ratio is computed as:

$$\begin{aligned} \mathrm{RR} &= \hat{\lambda}_1/\hat{\lambda}_0 = (D_1/Y_1)/(D_0/Y_0) \\ &= (28/632.3)/(17/843.7) = 0.0443/0.0201 = 2.19 \end{aligned}$$

The 95% confidence interval is computed as:

$$\begin{aligned} \hat{\mathrm{RR}} \stackrel{\times}{\div} \mathrm{erf} &= 2.198 \stackrel{\times}{\div} \exp\big(1.96\sqrt{1/17 + 1/28}\big) \\ &= 2.198 \stackrel{\times}{\div} 1.837 = (1.20, 4.02) \end{aligned}$$

Example using R

Poisson likelihood, for one rate,
based on 17 events in 843.7 PY:

```
library( Epi )
D <- 17 ; Y <- 843.7
m1 <- glm( D ~ 1, offset=log(Y/1000), family=poisson)
ci.exp( m1 )
```

```
             exp(Est.)     2.5%    97.5%
(Intercept)  20.14934 12.52605 32.41213
```

Poisson likelihood, two rates, or one rate and RR:

```
D <- c(17,28) ; Y <- c(843.7,632.3) ; gg <- factor(0:1)
m2 <- glm( D ~ gg, offset=log(Y/1000), family=poisson)
ci.exp( m2 )
```

```
             exp(Est.)      2.5%     97.5%
(Intercept) 20.149342 12.526051 32.412130
gg1          2.197728  1.202971  4.015068
```


Example using R

Poisson likelihood, two rates, or one rate and RR:

```
D <- c(17,28) ; Y <- c(843.7,632.3) ; gg <- factor(0:1)
m2 <- glm( D ~ gg, offset=log(Y/1000), family=poisson)
ci.exp( m2 )
```

```
             exp(Est.)      2.5%     97.5%
(Intercept) 20.149342 12.526051 32.412130
gg1          2.197728  1.202971  4.015068
```

```
m3 <- glm( D ~ gg - 1, offset=log(Y/1000), family=poisson)
ci.exp( m3 )
```

```
     exp(Est.)     2.5%    97.5%
gg0  20.14934 12.52605 32.41213
gg1  44.28278 30.57545 64.13525
```

You do it!

Survival analysis

- Response variable: Time to event, T
- Censoring time, Z
- We observe $(\min(T, Z), \delta = 1\{T < Z\})$.
- This gives time a special status, and mixes the response variable (risk)time with the covariate time(scale).
- Originates from clinical trials where everyone enters at time 0, and therefore $Y = T - 0 = T$

The life table method

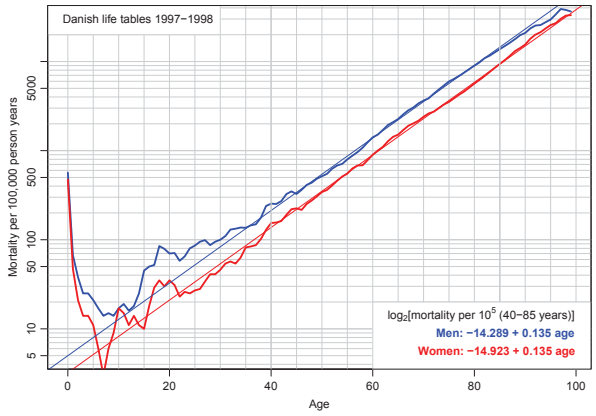
The simplest analysis is by the "life-table method":

interval i	alive n_i	dead d_i	cens. l_i	p_i
1	77	5	2	$5/(77-2/2) = 0.066$
2	70	7	4	$7/(70-4/2) = 0.103$
3	59	8	1	$8/(59-1/2) = 0.137$

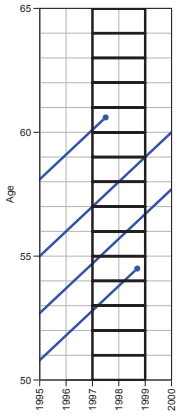
$$
\begin{aligned}
p_i &= \mathrm{P}\{\text{death in interval } i\} = 1 - d_i/(n_i - l_i/2) \\
S(t) &= (1-p_1) \times \cdots \times (1-p_t)
\end{aligned}
$$

Population life table, DK 1997–98

	Men			Women		
a	$S(a)$	$\lambda(a)$	$\mathrm{E}[\ell_{\mathrm{res}}(a)]$	$S(a)$	$\lambda(a)$	$\mathrm{E}[\ell_{\mathrm{res}}(a)]$
0	1.00000	567	73.68	1.00000	474	78.65
1	0.99433	67	73.10	0.99526	47	78.02
2	0.99366	38	72.15	0.99479	21	77.06
3	0.99329	25	71.18	0.99458	14	76.08
4	0.99304	25	70.19	0.99444	14	75.09
5	0.99279	21	69.21	0.99430	11	74.10
6	0.99258	17	68.23	0.99419	6	73.11
7	0.99242	14	67.24	0.99413	3	72.11
8	0.99227	15	66.25	0.99410	6	71.11
9	0.99213	14	65.26	0.99404	9	70.12
10	0.99199	17	64.26	0.99395	17	69.12
11	0.99181	19	63.28	0.99378	15	68.14
12	0.99162	16	62.29	0.99363	11	67.15
13	0.99147	18	61.30	0.99352	14	66.15
14	0.99129	25	60.31	0.99338	11	65.16
15	0.99104	45	59.32	0.99327	10	64.17
16	0.99059	50	58.35	0.99317	18	63.18
17	0.99009	52	57.38	0.99299	29	62.19
18	0.98957	85	56.41	0.99270	35	61.21
19	0.98873	79	55.46	0.99235	30	60.23
20	0.98795	70	54.50	0.99205	35	59.24
21	0.98726	71	53.54	0.99170	31	58.27



Observations for the lifetable



Life table is based on person-years and deaths accumulated in a short period.

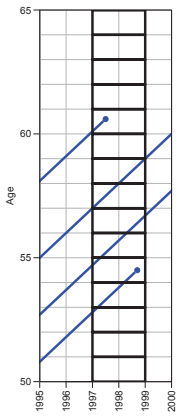
Age-specific rates — cross-sectional!

Survival function:

$$S(t) = \mathrm{e}^{-\int_0^t \lambda(a)\,\mathrm{d}a} = \mathrm{e}^{-\sum_0^t \lambda(a)}$$

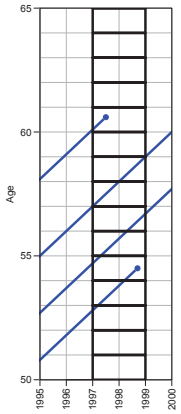
— assumes stability of rates to be interpretable for actual persons.

Observations for the lifetable

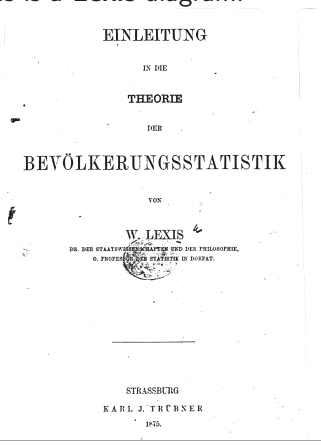


This is a **Lexis** diagram.

Observations for the lifetable



This is a **Lexis** diagram.


EINLEITUNG

IN DIE

THEORIE

DER

BEVÖLKERUNGSSTATISTIK

VON

W. LEXIS

STRASSBURG

KARL J. TRÜBNER

Life table approach

The observation of interest is **not** the survival time of the **individual**.

- It is the **population** experience:
  - D: Deaths (events).
  - Y: Person-years (risk time).
- The classical lifetable analysis compiles these for prespecified intervals of age, and computes age-specific mortality **rates**.
- Data are collected crossectionally, but interpreted longitudinally.
- The **rates** are the basic building bocks — used for construction of:
  - RRs
  - cumulative measures (survival and risk)

Summary

- Follow-up studies observe time to event
- — in the form of **empirical rates**, (d, y) for small interval
- each interval (empirical rate) has covariates attached
- each interval contribute $d\log(\lambda) - \lambda y$
- — like a Poisson observation d with mean λy
- identical covariates: pool obervations to $D = \sum D, Y = \sum y$
- — like a Poisson obervation D with mean λY
- the result is an **estimate** of the rate λ
- from a **model** where rates are constant within intervals — but varies between intervals.

Classical estimators

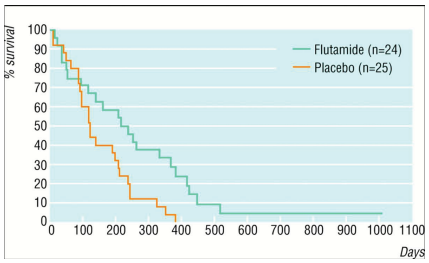
Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
http://BendixCarstensen/AdvCoh/Scot-2014

km-na

The Kaplan-Meier Method

- The most common method of estimating the survival function.
- A non-parametric method.
- Divides time into small intervals where the intervals are defined by the unique times of failure (death).
- Based on conditional probabilities as we are interested in the probability a subject surviving the next time interval given that they have survived so far.

Example of KM Survival Curve from BMJ

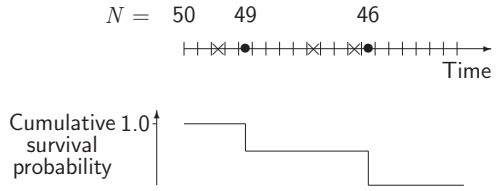


BMJ 1998;316:1935-1938

Kaplan-Meier curve from an RCT of patients with pancreatic cancer

Kaplan–Meier method illustrated

(• = failure and × = censored):



- Steps caused by multiplying by $(1 - 1/49)$ and $(1 - 1/46)$ respectively
- Late entry can also be dealt with

Using R: `Surv()`

```
library( survival )
data( lung )
head( lung, 3 )
```

```
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal
1    3  306      2  74   1       1       90       100     1175
2    3  455      2  68   1       0       90        90     1225
3    3 1010      1  56   1       0       90        90       NA
```

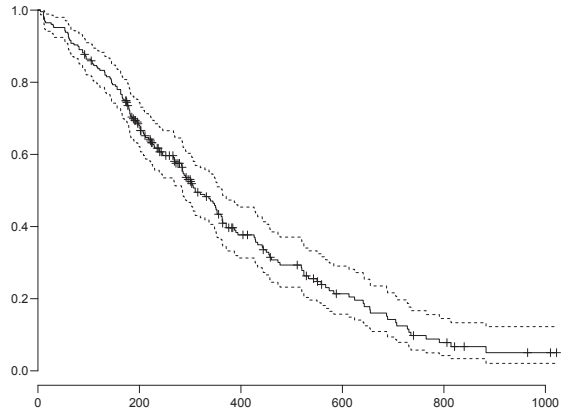
```
with( lung, Surv( time, status==2 ) )[1:10]
```

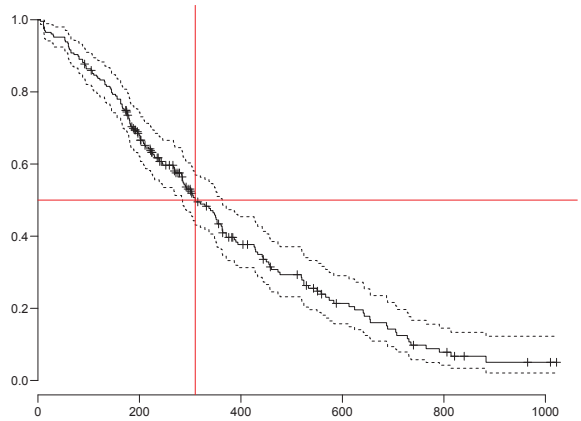
```
  [1]  306   455  1010+  210   883  1022+  310   361   218   16
```

```
( s.km <- survfit( Surv( time, status==2 ) ~ 1 , data=lung ) )
```

```
 Call: survfit(formula = Surv(time, status == 2) ~ 1, data = lu

 records   n.max n.start  events  median 0.95LCL 0.95UCL
     228     228     228     165     310     285     363
```

```
plot( s.km )
abline( v=310, h=0.5, col="red" )
```





The Cox model

Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
http://BendixCarstensen/AdvCoh/Scot-2014

cox

Proportional Hazards model

Model hazard rate as function of time (t) and covariates $(\mathbf{x})$

$$\lambda_i(t, \mathbf{x}_i) = \lambda_0(t)\exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\right)$$

- $\lambda_i(t, \mathbf{x}_i)$ is the hazard rate for the i^{th} person.
- $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})$ are covariate values for ith person.
- $\lambda_0(t)$ is the **baseline hazard** function - a non-linear effect of the **covariate** t.
- $\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$ is the linear predictor.

The proportional hazards model

$$\lambda(t, x) = \lambda_0(t) \times \exp(x'\beta)$$

A model for the rate as a function of t and x.

The covariate t has a special status:

- Computationally, because all individuals contribute to (some of) the range of t.
- Conceptually it is less clear — t is but a covariate that varies **within** each individual.

Cox-likelihood

The partial likelihood for the regression parameters:

$$\ell(\beta) = \sum_{\text{death times}} \log\left(\frac{\mathrm{e}^{x_{\text{death}}\beta}}{\sum_{i\in\mathcal{R}_t}\mathrm{e}^{x_i\beta}}\right)$$

- This is David Cox's invention.
- Extremely efficient from a computational point of view.
- The baseline hazard is bypassed (profiled out).

Proportional Hazards model

- The baseline hazard rate, $\lambda_0(t)$, is the hazard rate when all the covariates are 0.
- The form of the above equation means that covariates act **multiplicatively** on the baseline hazard rate.
- Time is a covariate (albeit with special status).
- The baseline hazard is a function of time and thus varies with time.
- No assumption about the shape of the underlying hazard function.
- — but you will never see the shape. . .

The Cox Proportional Hazards likelihood

- By far the most common model applied to time-to-event outcomes.
- The proportionality assumption means that the difference between two groups can be summarised by one number. This is because the (relative) effect of a covariate is assumed to be the same throughout the time-scale.
- However, it does make the assumption that the hazard rates for patient subgroups are proportional over time.
- The Cox model models the hazard function, $\lambda_i(t; x_i)$ where x_i denotes the covariate vector.

Proportional Hazards Model

- Parameters are estimated on log scale:

$$\lambda_i(t) = \lambda_0(t)\exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\right)$$

$$\log\left(\lambda_i(t)\right) = \log\left(\lambda_0(t)\right) + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

- The baseline hazard is the hazard rate when all covariate values are equal to zero.
- Estimates of the parameters, β, are obtained by maximizing the partial likelihood.

Interpreting Regression Coefficients

- How do we interpret the parameters of interest?
- In a Cox model the baseline hazard $\lambda_0(t)$ is not included in the partial likelihood and so we only obtain estimates of the regression coefficients associated with each of the covariates.
- Consider a binary covariate x_1 which takes the values 0 and 1.

Interpreting Regression Coefficients

- The model is

$$\lambda_i(t) = \lambda_0(t)\exp\left(\beta_1 x_{1i}\right)$$

- The hazard rate when $x_1 = 0$ is $\lambda_0(t)$.
- The hazard rate when $x_1 = 1$ is $\lambda_0(t)\exp(\beta_1)$.
- The hazard ratio is therefore

$$\frac{\lambda_0(t)\exp(\beta)}{\lambda_0(t)}$$

- The $\lambda_0(t)$ cancels: β_1 is the log hazard ratio.
- Exponentiate β_1 to get the hazard ratio.

Interpreting Regression Coefficients

- If x_j is binary $\exp(\beta_j)$ is the estimated hazard ratio for subjects corresponding to $x_j = 1$ compared to those where $x_j = 0$.
- If x_j is continuous $\exp(\beta_j)$ is the estimated increase/decrease in the hazard rate for a unit change in x_j.
- With more than one covariate interpretation is similar, i.e. $\exp(\beta_j)$ is the hazard ratio for subjects who **only** differ with respect to covariate x_j.

Fitting a Cox- model in R

```
library( survival )
data(bladder)
bladder <- subset( bladder, enum<2 )
head( bladder)
```

```
   id rx number size stop event enum
1   1  1      1    3    1     0    1
5   2  1      2    1    4     0    1
9   3  1      1    1    7     0    1
13  4  1      5    1   10     0    1
17  5  1      4    1    6     1    1
21  6  1      1    1   14     0    1
```


Fitting a Cox-model in R

```
c0 <- coxph( Surv(stop,event) ~ number + size, data=bladder )
c0
```

```
Call:
coxph(formula = Surv(stop, event) ~ number + size, data = blad
```

```
         coef exp(coef) se(coef)     z      p
number 0.2049      1.23   0.0704 2.912 0.0036
size   0.0613      1.06   0.1033 0.594 0.5500

Likelihood ratio test=7.04  on 2 df, p=0.0296  n= 85, number
```

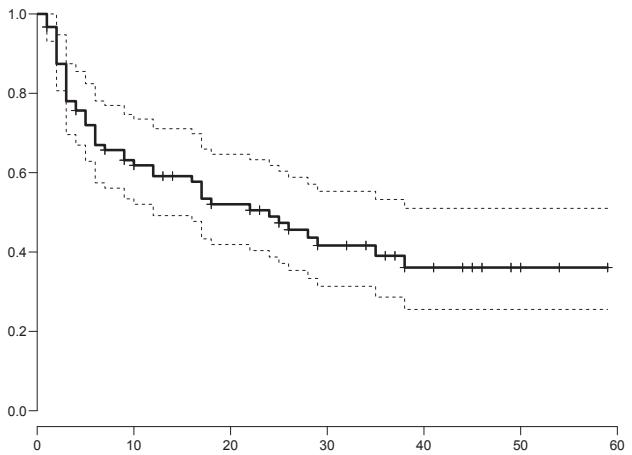

Plotting the base survival in R

```
plot( survfit(c0) )
lines( survfit(c0), conf.int=F, lwd=3 )
```

The `plot.coxph` plots the survival curve for a person with an average covariate value

— which is **not** the average survival for the population considered...

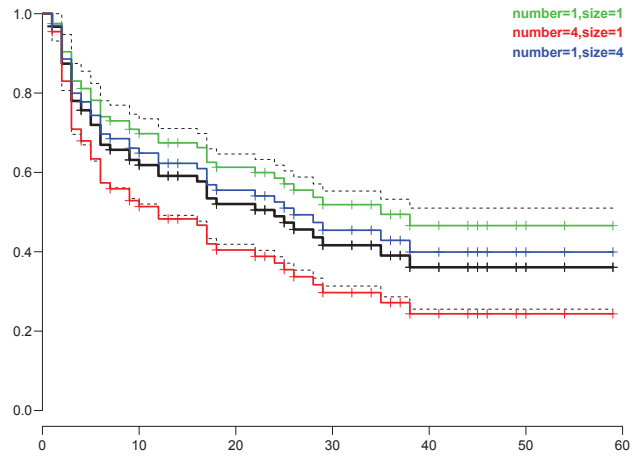
— and not necessarily meaningful



Plotting the base survival in R

You can plot the survival curve for specific values of the covariates, using the `newdata=` argument:

```
plot( survfit(c0) )
lines( survfit(c0), conf.int=F, lwd=3 )
lines( survfit(c0, newdata=data.frame(number=1,size=1)),
       lwd=2, col="limegreen" )
text( par("usr")[2]*0.98, 1.00, "number=1,size=1",
      col="limegreen", font=2, adj=1 )
```

Follow-up data

Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
`http://BendixCarstensen/AdvCoh/Scot-2014`

`time-split`

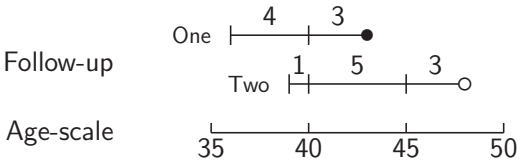
Follow-up and rates

- Follow-up studies:
  - D — events, deaths
  - Y — person-years
  - $\lambda = D/Y$ rates
- Rates differ between persons.
- Rates differ within persons:
  - By age
  - By calendar time
  - By disease duration
  - ...
- Multiple timescales.
- Multiple states (little boxes — later)

Stratification by age

If follow-up is rather short, age at entry is OK for age-stratification.

If follow-up is long, use stratification by categories of **current age**, both for:
No. of events, D, and Risk time, Y.



Representation of follow-up data

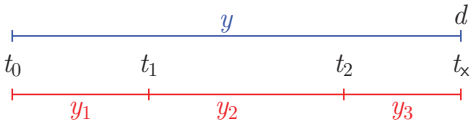
A cohort or follow-up study records:
Events and **Risk time**.

The outcome is thus **bivariate**: (d, y)

Follow-up **data** for each individual must therefore have (at least) three variables:

Date of entry	`entry`	date variable
Date of exit	`exit`	date variable
Status at exit	`fail`	indicator (0/1)

Specific for each **type** of outcome.

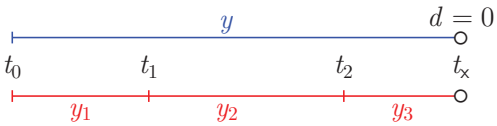


Probability

$$\begin{aligned}&\mathrm{P}(d \text{ at } t_x|\text{entry } t_0)\\ &= \mathrm{P}(\text{surv } t_0 \to t_1|\text{entry } t_0)\\ &\times \mathrm{P}(\text{surv } t_1 \to t_2|\text{entry } t_1)\\ &\times \mathrm{P}(d \text{ at } t_x|\text{entry } t_2)\end{aligned}$$

log-Likelihood

$$\begin{aligned}&d\log(\lambda) - \lambda y\\ &= 0\log(\lambda) - \lambda y_1\\ &+ 0\log(\lambda) - \lambda y_2\\ &+ d\log(\lambda) - \lambda y_3\end{aligned}$$

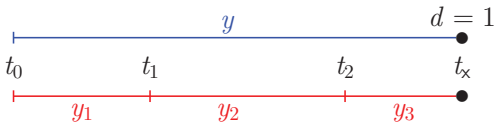


Probability

$$\begin{aligned}&\mathrm{P}(\text{surv } t_0 \to t_x|\text{entry } t_0)\\ &= \mathrm{P}(\text{surv } t_0 \to t_1|\text{entry } t_0)\\ &\times \mathrm{P}(\text{surv } t_1 \to t_2|\text{entry } t_1)\\ &\times \mathrm{P}(\text{surv } t_2 \to t_x|\text{entry } t_2)\end{aligned}$$

log-Likelihood

$$\begin{aligned}&0\log(\lambda) - \lambda y\\ &= 0\log(\lambda) - \lambda y_1\\ &+ 0\log(\lambda) - \lambda y_2\\ &+ 0\log(\lambda) - \lambda y_3\end{aligned}$$



Probability	log-Likelihood
$\mathrm{P}(\text{event at } t_x \mid \text{entry } t_0)$	$1\log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_0 \to t_1 \mid \text{entry } t_0)$	$= 0\log(\lambda) - \lambda y_1$
$\times\, \mathrm{P}(\text{surv } t_1 \to t_2 \mid \text{entry } t_1)$	$+ 0\log(\lambda) - \lambda y_2$
$\times\, \mathrm{P}(\text{event at } t_x \mid \text{entry } t_2)$	$+ 1\log(\lambda) - \lambda y_3$

Dividing time into bands:

If we want to put D and Y into intervals on the timescale we must know:

Origin: The date where the time scale is 0:

- Age — 0 at date of birth
- Disease duration — 0 at date of diagnosis
- Occupation exposure — 0 at date of hire

Intervals: How should it be subdivided:

- 1-year classes? 5-year classes?
- Equal length?

Aim: Separate rate in each interval

Example: cohort with 3 persons:

Id	Bdate	Entry	Exit	St
1	14/07/1952	04/08/1965	27/06/1997	1
2	01/04/1954	08/09/1972	23/05/1995	0
3	10/06/1987	23/12/1991	24/07/1998	1

- Age bands: 10-years intervals of current age.
- Split Y for every subject accordingly
- Treat each segment as a separate unit of observation.
- Keep track of exit status in each interval.

Splitting the follow up

	subj. 1	subj. 2	subj. 3
Age at **E**ntry:	13.06	18.44	4.54
Age at e**X**it:	44.95	41.14	11.12
Status at exit:	Dead	Alive	Dead
Y	31.89	22.70	6.58
D	1	0	1

Age	subj. 1 Y	subj. 1 D	subj. 2 Y	subj. 2 D	subj. 3 Y	subj. 3 D	$\sum$ Y	$\sum$ D
0–	0.00	0	0.00	0	5.46	0	5.46	0
10–	6.94	0	1.56	0	1.12	1	8.62	1
20–	10.00	0	10.00	0	0.00	0	20.00	0
30–	10.00	0	10.00	0	0.00	0	20.00	0
40–	4.95	1	1.14	0	0.00	0	6.09	1
$\sum$	31.89	1	22.70	0	6.58	1	60.17	2

Splitting the follow-up

id	Bdate	Entry	Exit	St	risk	int
1	14/07/1952	03/08/1965	14/07/1972	0	6.9432	10
1	14/07/1952	14/07/1972	14/07/1982	0	10.0000	20
1	14/07/1952	14/07/1982	14/07/1992	0	10.0000	30
1	14/07/1952	14/07/1992	27/06/1997	1	4.9528	40
2	01/04/1954	08/09/1972	01/04/1974	0	1.5606	10
2	01/04/1954	01/04/1974	31/03/1984	0	10.0000	20
2	01/04/1954	31/03/1984	01/04/1994	0	10.0000	30
2	01/04/1954	01/04/1994	23/05/1995	0	1.1417	40
3	10/06/1987	23/12/1991	09/06/1997	0	5.4634	0
3	10/06/1987	09/06/1997	24/07/1998	1	1.1211	10

Keeping track of calendar time too?

Timescales

- A timescale is a variable that varies **deterministically** *within* each person during follow-up:
  - Age
  - Calendar time
  - Time since treatment
  - Time since relapse
- All timescales advance at the same pace (1 year per year . . .)
- Note: Cumulative exposure is **not** a timescale.

Follow-up on several timescales

- The risk-time is the same on all timescales
- Only need the entry point on each time scale:
  - Age at entry.
  - Date of entry.
  - Time since treatment at entry.
    — if time of treatment is the entry, this is 0 for all.
- Response variable in analysis of rates:

$$(d, y) \qquad (\text{event}, \text{duration})$$

- Covariates in analysis of rates:
  - timescales
  - other (fixed) measurements

Follow-up data in Epi — Lexis objects

A follow-up study:

```
> round( th, 2 )
    id sex birthdat contrast injecdat volume exitdat ex
1    1   2  1916.61        1  1938.79     22 1976.79
2  640   2  1896.23        1  1945.77     20 1964.37
3 3425   1  1886.97        2  1955.18      0 1956.59
4 4017   2  1936.81        2  1957.61      0 1992.14
...
```

Timescales of interest:

- Age
- Calendar time
- Time since injection

Definition of Lexis object

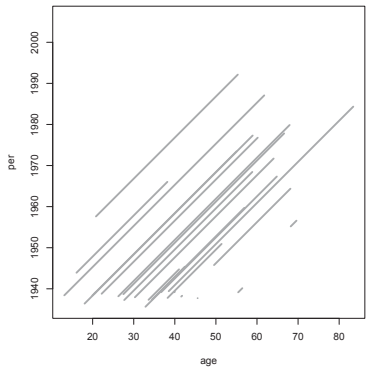
```
> thL <- Lexis( entry = list( age = injecdat-birthdat,
+                             per = injecdat,
+                             tfi = 0 ),
+               exit = list( per = exitdat ),
+         exit.status = as.numeric(exitstat==1),
+               data = th )
```

entry is defined on **three** timescales,
but exit is only defined on **one** timescale:
Follow-up time is the same on all timescales:

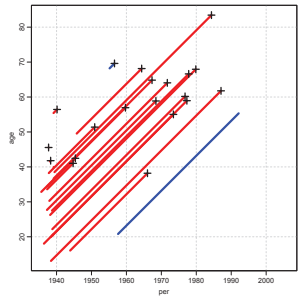
exitdat - injecdat

The looks of a Lexis object

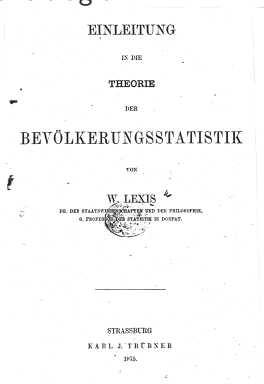
```
> thL[,1:9]
    age     per tfi lex.dur lex.Cst lex.Xst lex.id
1 22.18 1938.79   0   37.99       0       1      1
2 49.54 1945.77   0   18.59       0       1      2
3 68.20 1955.18   0    1.40       0       1      3
4 20.80 1957.61   0   34.52       0       0      4
...

> summary( thL )
Transitions:
     To
From 0  1 Records:  Events:  Risk time:  Persons:
   0 3 20        23       20      512.59        23
```

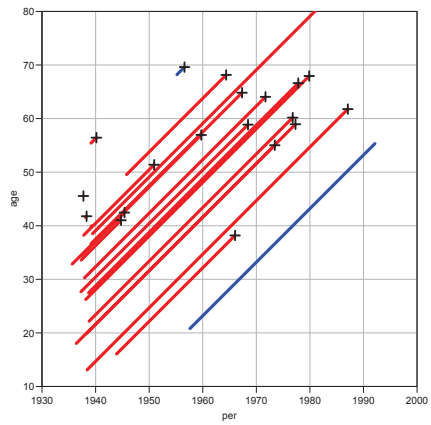

```
> plot( thL, lwd=3 )
```

Lexis diagram


EINLEITUNG
IN DIE
THEORIE
DER
BEVÖLKERUNGSSTATISTIK
VON
W. LEXIS
STRASSBURG
KARL J. TRÜBNER
1875.

```
> plot( thL, 2:1, lwd=5, col=c("red","blue")[thL$contrast], grid=T )
> points( thL, 2:1, pch=c(NA,3)[thL$lex.Xst+1],lwd=3, cex=1.5 )
```

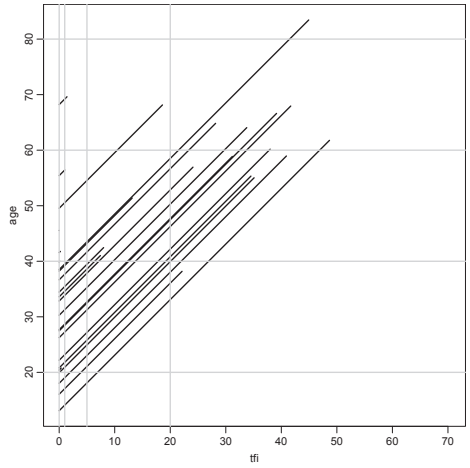

```
> plot( thL, 2:1, lwd=5, col=c("red","blue")[thL$contrast],
+       grid=TRUE, lty.grid=1, col.grid=gray(0.7),
+       xlim=1930+c(0,70), xaxs="i", ylim=  10+c(0,70), yaxs="i", las=1 )
> points( thL, 2:1, pch=c(NA,3)[thL$lex.Xst+1],lwd=3, cex=1.5 )
```


Splitting follow-up time

```
> spl1 <- splitLexis( thL, breaks=seq(0,100,20),
>                          time.scale="age" )
> round(spl1,1)
   age     per  tfi lex.dur lex.Cst lex.Xst   id sex birthdat con
1 22.2 1938.8  0.0    17.8       0       0    1   2    1916.6
2 40.0 1956.6 17.8    20.0       0       0    1   2    1916.6
3 60.0 1976.6 37.8     0.2       0       1    1   2    1916.6
4 49.5 1945.8  0.0    10.5       0       0  640   2    1896.2
5 60.0 1956.2 10.5     8.1       0       1  640   2    1896.2
6 68.2 1955.2  0.0     1.4       0       1 3425   1    1887.0
7 20.8 1957.6  0.0    19.2       0       0 4017   2    1936.8
8 40.0 1976.8 19.2    15.3       0       0 4017   2    1936.8
...
```


Split on another timescale

```
> spl2 <- splitLexis( spl1, time.scale="tfi",
                           breaks=c(0,1,5,20,100) )
> round( spl2, 1 )
   lex.id  age     per  tfi lex.dur lex.Cst lex.Xst   id sex birt
1       1 22.2 1938.8  0.0     1.0       0       0    1   2   19
2       1 23.2 1939.8  1.0     4.0       0       0    1   2   19
3       1 27.2 1943.8  5.0    12.8       0       0    1   2   19
4       1 40.0 1956.6 17.8     2.2       0       0    1   2   19
5       1 42.2 1958.8 20.0    17.8       0       0    1   2   19
6       1 60.0 1976.6 37.8     0.2       0       1    1   2   19
7       2 49.5 1945.8  0.0     1.0       0       0  640   2   18
8       2 50.5 1946.8  1.0     4.0       0       0  640   2   18
9       2 54.5 1950.8  5.0     5.5       0       0  640   2   18
10      2 60.0 1956.2 10.5     8.1       0       1  640   2   18
11      3 68.2 1955.2  0.0     1.0       0       0 3425   1   18
12      3 69.2 1956.2  1.0     0.4       0       1 3425   1   18
13      4 20.8 1957.6  0.0     1.0       0       0 4017   2   19
14      4 21.8 1958.6  1.0     4.0       0       0 4017   2   19
15      4 25.8 1962.6  5.0    14.2       0       0 4017   2   19
16      4 40.0 1976.8 19.2     0.8       0       0 4017   2   19
17      4 40.8 1977.6 20.0    14.5       0       0 4017   2   19
...
```




```
  age  tfi lex.dur
 22.2  0.0     1.0
 23.2  1.0     4.0
 27.2  5.0    12.8
 40.0 17.8     2.2
 42.2 20.0    17.8
 60.0 37.8     0.2
```

```
plot( spl2, c(1,3), col="black", lwd=2 )
```


Likelihood for a constant rate

- This setup is for a situation where it is assumed that rates are constant in each of the intervals.
- Each observation in the dataset contributes a term to a "Poisson" likelihood.
- Rates can vary along several timescales simultaneously.
- Models can include fixed covariates, as well as the timescales (the left end-points of the intervals) as continuous variables.

The Poisson likelihood for split data

- Split records (one per **p**erson-**i**nterval (p, i)):

$$D\log(\lambda) - \lambda Y = \sum_{p,i} \left(d_{pi}\log(\lambda) - \lambda y_{pi} \right)$$

- Assuming that the death indicator $(d_{pi} \in \{0, 1\})$ is Poisson, with $\log$-offset y_{pi} will give the same result.
- Model assumes that rates are constant.
- But the split data allows models that assume different rates for different (d_{pi}, y_{pi}), so rates can vary **within** a person's follow-up.

Where is (d_{pi}, y_{pi}) in the split data?

```
> round( spl2, 1 )
   lex.id  age    per  tfi lex.dur lex.Cst lex.Xst   id sex birt
1       1 22.2 1938.8  0.0     1.0       0       0    1   2   19
2       1 23.2 1939.8  1.0     4.0       0       0    1   2   19
3       1 27.2 1943.8  5.0    12.8       0       0    1   2   19
4       1 40.0 1956.6 17.8     2.2       0       0    1   2   19
5       1 42.2 1958.8 20.0    17.8       0       0    1   2   19
6       1 60.0 1976.6 37.8     0.2       0       1    1   2   19
7       2 49.5 1945.8  0.0     1.0       0       0  640   2   18
8       2 50.5 1946.8  1.0     4.0       0       0  640   2   18
9       2 54.5 1950.8  5.0     5.5       0       0  640   2   18
10      2 60.0 1956.2 10.5     8.1       0       1  640   2   18
...
```

— and what are covariates for the rates?

Analysis of results

- d_{pi} — events in the variable: `lex.Xst`:
  In the model as response: `lex.Xst==1`
- y_{pi} — risk time: `lex.dur` (duration):
  In the model as offset $\log(y)$, `log(lex.dur)`.
- Covariates are:
  - timescales (age, period, time in study)
  - other variables for this person (constant or *assumed* constant in each interval).
- Model rates using the covariates in `glm`:
  — no difference between time-scales and other covariates.

Fitting a simple model

```
> stat.table( contrast,
+             list( D = sum( lex.Xst ),
+                   Y = sum( lex.dur ),
+                Rate = ratio( lex.Xst, lex.dur, 100 ) )
+             margin = TRUE,
+             data = spl2 )
 ----------------------------------
 contrast         D       Y    Rate
 ----------------------------------
 1            19.00  476.67    3.99
 2             1.00   35.93    2.78

 Total        20.00  512.59    3.90
 ----------------------------------
```


Fitting a simple model

```
 ----------------------
 contrast         D       Y    Rate
 ----------------------
 1            19.00  476.67    3.99
 2             1.00   35.93    2.78

 Total        20.00  512.59    3.90
 ----------------------
```

```
> m0 <- glm( lex.Xst ~ factor(contrast) - 1,
+                      offset=log(lex.dur/100),
+                      family=poisson, data=spl2 )
> round( ci.exp( m0 ), 2 )

                  exp(Est.) 2.5% 97.5%
factor(contrast)1      3.99 2.54  6.25
factor(contrast)2      2.78 0.39 19.74
```


Who needs the Cox-model anyway?

Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
http://BendixCarstensen/AdvCoh/Scot-2014

WntCma

The proportional hazards model

$$\lambda(t, x) = \lambda_0(t) \times \exp(x'\beta)$$

A model for the rate as a function of t and x.

The covariate t has a special status:

- Computationally, because all individuals contribute to (some of) the range of t.
- Conceptually it is less clear — t is but a covariate that varies within individual.

Cox-likelihood

The (partial) log-likelihood for the regression parameters:

$$\ell(\beta) = \sum_{\text{death times}} \log\left(\frac{\mathrm{e}^{\eta_{\text{death}}}}{\sum_{i\in\mathcal{R}_t}\mathrm{e}^{\eta_i}}\right)$$

is also a **profile likelihood** in the model where observation time has been subdivided in small pieces (empirical rates) and each small piece provided with its own parameter:

$$\log\big(\lambda(t, x)\big) = \log\big(\lambda_0(t)\big) + x'\beta = \alpha_t + \eta$$

The Cox-likelihood as profile likelihood

- Regression parameters describing the effect of covariates (other than the chosen underlying time scale).
- One parameter per death time to describe the effect of time (i.e. the chosen timescale).

$$\log\big(\lambda(t, x_i)\big) = \log\big(\lambda_0(t)\big) + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} = \alpha_t +$$

- Profile likelihood:
  - Derive estimates of α_t as function of data and βs
  - Insert in likelihood, now only a function of data and βs
  - Turns out to be Cox's partial likelihood

- Suppose the time scale has been divided into small intervals with at most one death in each.
- Assume w.l.o.g. the ys in the empirical rates all are 1.
- Log-likelihood contributions that contain information on a specific time-scale parameter α_t will be from:
- the (only) empirical rate $(1, 1)$ with the death at time t.
- all other empirical rates $(0, 1)$ from those who were at risk at time t.

Note: There is one contribution from each person at risk to this part of the log-likelihood:

$$\begin{aligned}
\ell_t(\alpha_t, \beta) &= \sum_{i\in\mathcal{R}_t} d_i \log(\lambda_i(t)) - \lambda_i(t) y_i \\
&= \sum_{i\in\mathcal{R}_t} \left\{ d_i(\alpha_t + \eta_i) - \mathrm{e}^{\alpha_t+\eta_i} \right\} \\
&= \alpha_t + \eta_{\text{death}} - \mathrm{e}^{\alpha_t} \sum_{i\in\mathcal{R}_t} \mathrm{e}^{\eta_i}
\end{aligned}$$

where η_{death} is the linear predictor for the person that died.

The derivative w.r.t. α_t is:

$$\mathrm{D}_{\alpha_t}\ell(\alpha_t, \beta) = 1 - \mathrm{e}^{\alpha}_t \sum_{i\in\mathcal{R}_t} \mathrm{e}^{\eta_i} = 0 \quad \Leftrightarrow \quad \mathrm{e}^{\alpha}_t = \frac{1}{\sum_{i\in\mathcal{R}_t}\mathrm{e}^{\eta_i}}$$

If this estimate is fed back into the log-likelihood for α_t, we get the **profile likelihood** (with α_t "profiled out"):

$$\log\left(\frac{1}{\sum_{i\in\mathcal{R}_t}\mathrm{e}^{\eta_i}}\right) + \eta_{\text{death}} - 1 = \log\left(\frac{\mathrm{e}^{\eta_{\text{death}}}}{\sum_{i\in\mathcal{R}_t}\mathrm{e}^{\eta_i}}\right) - 1$$

which is the same as the contribution from time t to Cox's partial likelihood.

What the Cox-model really is

Taking the life-table approach *ad absurdum* by:

- dividing time very finely,
- modelling one covariate, the time-scale, with one parameter per distinct value,
- profiling these parameters out and maximizing the profile likelihood,
- regression parameters are the same as in the full model with all the interval-specific parameters
- Subsequently, one may recover the effect of the timescale by smoothing an estimate of the cumulative sum of these.

Sensible modelling

Replace the α_ts by a parmetric function $f(t)$ with a limited number of parameters, for example:

- Piecewise constant
- Splines (linear, quadratic or cubic)
- Fractional polynomials

Use Poisson modelling software on a dataset of empirical rates for small intervals (ys).

Splitting the dataset

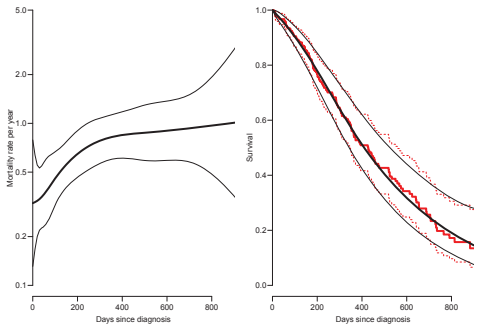
- The Poisson approach needs a dataset of empirical rates with small values of y.
- Larger than the original: each individual contributes many empirical rates. From each empirical rate we get:
  - Poisson-response d
  - Risk time y
  - Covariate value for the timescale (time since entry, current age, current date, ...)
  - other covariates

Example: Mayo Clinic lung cancer

```
> library( survival ) ; library( Epi )
> data( lung )
> head( lung )
  inst time status age sex ph.ecog ph.karno pat.karno meal.ca
1    3  306      2  74   1       1       90        100     117
2    3  455      2  68   1       0       90         90     122
3    3 1010      1  56   1       0       90         90      NA
4    5  210      2  57   1       1       90         60     115
5    1  883      2  60   1       0      100         90      NA
6   12 1022      1  74   1       1       50         80     513
> Lx <- Lexis( exit=list( tfd=time), exit.status=(status==2), da
  NOTE: entry is assumed to be 0 on the tfd timescale.
> summary( Lx, scale=365.25 )
  Transitions:
      To
  From    FALSE TRUE  Records:  Events: Risk time:  Persons:
    FALSE    63  165       228      165     190.54       228
> Sx <- splitLexis( Lx, "tfd", breaks=c(0,unique(Lx$time)) )
> summary( Sx, scale=365.25 )
```


Mayo clinic lung cancer data

Smoothing by natural splines with 5 parameters, knots at $0, 25, 100, 500, 1000$ days:

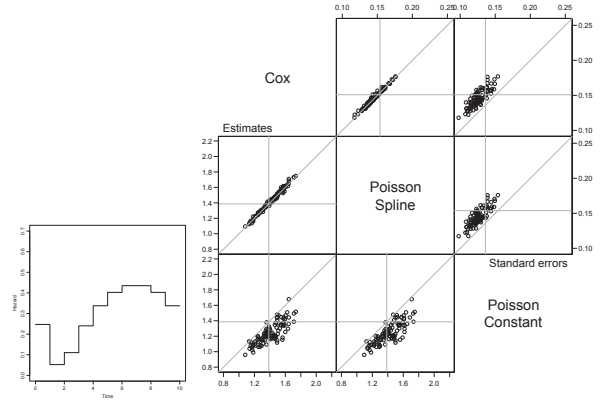


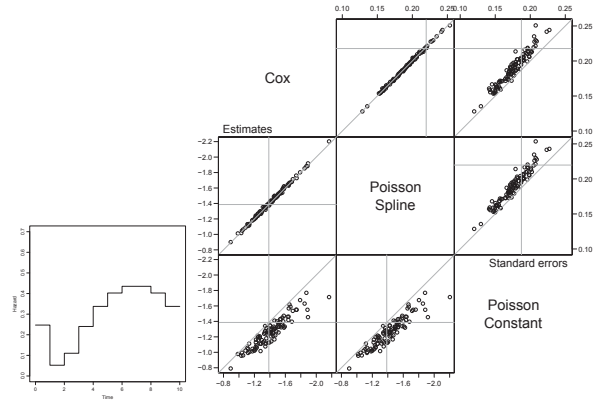
Practical: Cox and Poisson modelling

Any difference in covariate effects?

Simulation study:
100 survival datasets, 200 individuals in each.
Baseline hazard varying, censoring at time 10.
Two covariates, one standard normal with rate-ratio of 4 and the other log-normal with rate-ratio of 0.25.

For each dataset three models fitted:
1. standard Cox-model.
2. Poisson model using natural splines, 6 baseline parameters.
3. Poisson-model using constant baseline, 1 parameter.

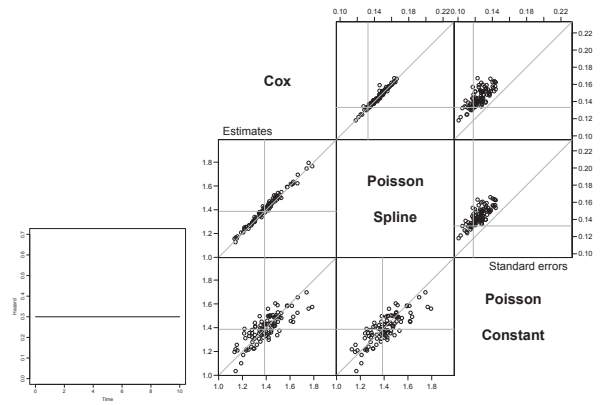


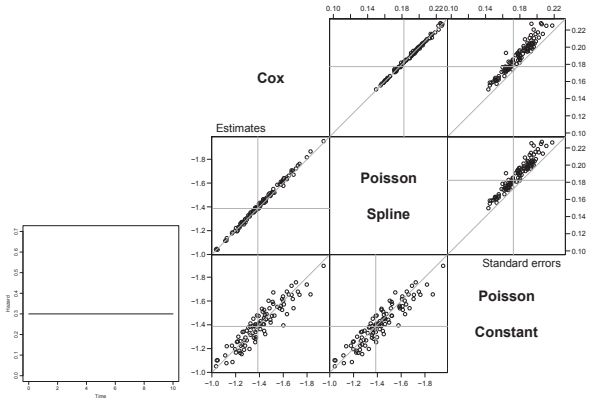


Modelling rates

Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
http://BendixCarstensen/AdvCoh/Scot-2014

rate-model





Computational aspects

- Cox model:
  - Only one timescale.
  - Each person contributes one (or very few) records.
  - Computationally simple, because time (risk / covariate) is profiled out in the estimation.
- Poisson modelling:
  - Many records per person.
  - Very large datasets.
  - Any number of timescales.
  - Timeconsuming due to the full modelling of the rates.

Historical aspects

Whitehead J: Fitting Cox's regression model to survival data using GLIM. Applied Statistics, 29(3):268–275, 1980.[1]

Set up tables of event counts and person-years, classified by event times and covariate patterns.

Even with moderate datasets this can be large, albeit smaller than some 100 separate records per person.

[1] Recall **Keiding's law**: "Any result was published earlier than you think, even if you take Keiding's law into account."

Computational practicalities

Early 1980s: Fitting of Poisson models on datasets with 50,000 records were out of the question. In particular with 100+ parameters.

Computationally feasible approaches to cohort studies were:

- Cox modelling — tanks to computational elegance.
- Time-splitting and tabulation before modelling.

Time-splitting and tabulation.

Man-years and PYRS programs:

Follow-up of each person was put into a table of (current) age-class by calendar time: Cut by the grid in a Lexis diagram. Possibly also classified by time since entry.

The tables of (D, Y) generated directly (disk space limitations prevented storage of the split dataset).

Used for SMR analysis, by merging with tables of population mortailty rates. Analyses based on a manageable number of analytical units.

The tabulation legacy (curse)

The **computational** need for tabulation has influenced thinking in epidemiology / demography:

- Life-tables in 1-year intervals.
- Rates are regarded in 5-year age by period intervals. Used for analysis of mortality and incidence rates based on registers. Age-period-cohort models with one parameter per level of the age/period factor.
- Yet, survival analysis is largely based on "time to event" methods (Kaplan-Meier, Cox), even from cancer registries.

The period method for survival analysis

H. Brenner, O. Gefeller & T. Hakulinen: Period analysis for 'up-to-date' cancer survival data: theory, empirical evaluation, computational realisation and applications European Journal of Cancer **40**, (2004), pp. 326–335

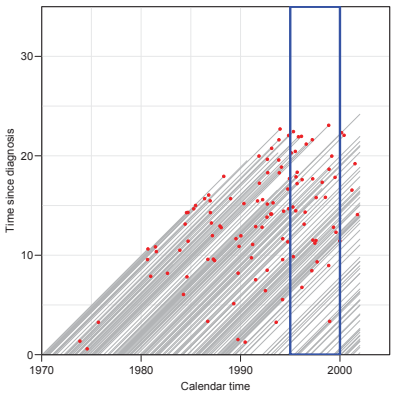
This method of survival analysis is designed to take interactions between two time-scale into account:

Mortality rates at a given time since entry into the study (usually diagnosis of cancer) depends on the current calendar time.

Brenner *et al.* propose to restrict analysis to the most recent period and then report results by survival curves.

Period analysis reports survival curve based on data from the blue rectangle.

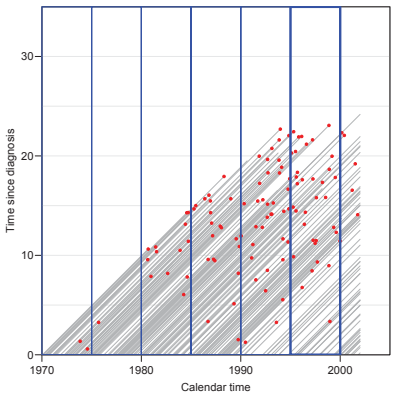
Interaction between current date and time since diagnosis.



Interaction between current date and time since diagnosis.

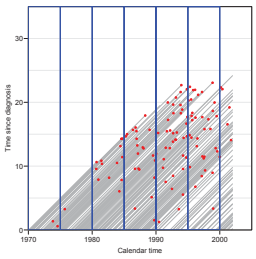
Separate survival curves for each period.

Period analysis reports the last set of parameters, because it is *clinically* the most relevant.



Interaction between current date and time since diagnosis:

- Separate survival curves for each period.
- Stratified Cox-model with time-dependent strata.
- In practical terms, data are split by (current) calendar time (period), and interactions with this are introduced throughout the model.



Using the Lexis diagram today

Rates are observed as little *empirical rates* (d, y), several per individual.

These vary by several *timescales*

- current age
- calendar time
- time since entry

and fixed covariates

- age at entry
- date of entry
- date of birth
- sex
- ...

Stratified Cox-model

$$\lambda(t, x) = \lambda_s(t) \times \exp(x'\beta)$$

The key is the "s" — separate baseline for each stratum.

In plain words:
The effect of time depends on s — an interaction between time and stratum.

Test of "proportionality" is merely a test of interaction between time and some (categorical) covariate.

Age at entry as covariate

t: time since entry
e: age at entry
$a = e + t$: current age

$$\log\big(\lambda(a, t)\big) = f(t) + \beta e = (f(t) - \beta t) + \beta a$$

Immaterial whether a or e is used as (log)-*linear* covariate as long as t is in the model.

In a Cox-model with time since entry as time-scale, only the baseline hazard will change if age at entry is replaced by current age (a time-dependent variable).

Non-linear effects of time-scales

Arbitrary effects of the three variables t, a and e:
$\Longrightarrow$ genuine extension of the model.

$$\log\big(\lambda(a, t, x_i)\big) = f(t) + g(a) + h(e) + \eta_i$$

Three quantities can be arbitrarily moved between the three functions:

$$\begin{aligned}\tilde{f}(t) &= f(a) - \mu_a - \mu_e + \gamma t \\ \tilde{g}(a) &= g(p) + \mu_a \qquad - \gamma a \\ \tilde{h}(e) &= h(c) \qquad + \mu_a + \gamma e\end{aligned}$$

because $t - a + e = 0$.
This is the age-period-cohort modelling problem again.

"Controlling for age"

— is not a well defined statement.

Mostly it means that age *at entry* is included in the model.

But ideally one would check whether there were non-linear effects of age at entry and current age.

This would require modelling of multiple timescales.

Which is best accomplished by splitting time.

SMR

Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
`http://BendixCarstensen/AdvCoh/Scot-2014`

`SMR`

Cohorts where all are exposed

When there is no comparison group we may ask:
Do mortality rates in cohort differ from those of an **external** population, for example:

Rates from:

- Occupational cohorts
- Patient cohorts

compared with reference rates obtained from:

- Population statistics (mortality rates)
- Disease registers (hospital discharge registers)

Log-likelihood

Cohort rates proportional to reference rates:
$\lambda(a) = \theta \times \lambda_R(a)$ — the same in all age-bands.

D_a deaths during Y_a person-years an age-band a gives the likelihood:

$$\begin{aligned} D_a \log(\lambda(a)) - \lambda(a)\, Y_a &= D_a \log(\theta \lambda_R(a)) - \theta \lambda_R(a)\, Y_a \\ &= D_a \log(\theta) + D_a \log(\lambda_R(a)) \\ &\quad -\theta(\lambda_R(a)\, Y_a) \end{aligned}$$

The constant $D_a \log(\lambda_R(a))$ does not involve θ, and so can be dropped.

The term $\lambda_R(a)\, Y_a = E_a$ is the "expected" number of cases in age a, so the log-likelihood for age a is:

$$D_a \log(\theta) - \theta(\lambda_R(a)\, Y_a) = D_a \log(\theta) - \theta(E_a)$$

Note: $\lambda_R(a)$ is known for all values of a. The total log-likelihood is:

$$D \log(\theta) - \theta E$$

Therefore:

$$\hat{\theta} = \frac{D}{\lambda_R\, Y} = \frac{D}{E} = \frac{\text{Observed}}{\text{Expected}} = \text{SMR}$$

SMR is the maximum likelihood estimator of the relative mortality in the cohort.

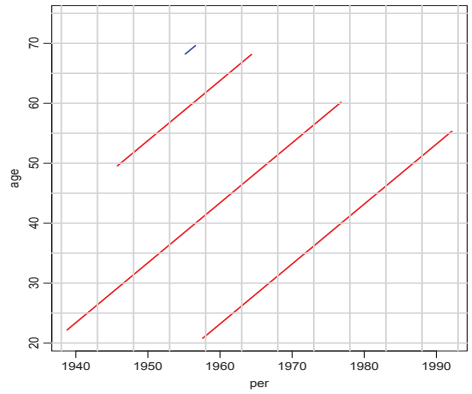
Accounting for age composition

- Compare rates in a study group with a standard set of age–specific rates.
- Reference rates are normally based on large numbers of cases, — assumed known.
- Calculate "expected" number of cases, $E_a = \lambda_R(a)\, Y_a$, and compare this with the observed number of cases, D:
- SMR is based on a log-likelihood similar to that for a rate — Y is replaced by E:

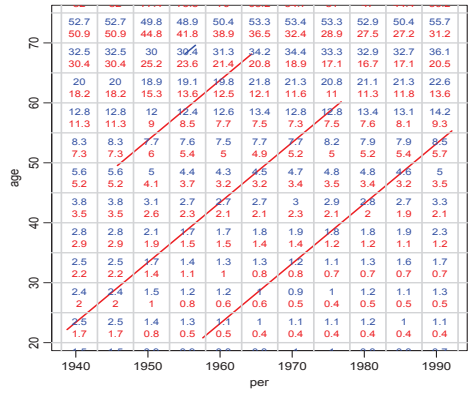
$$\text{SMR} = \frac{D}{E}, \qquad \text{s.d.}\big(\log(\text{SMR})\big) = \frac{1}{\sqrt{D}}$$

Modelling the SMR

- As for the rates, the SMR can be modelled using individual data.
- Response is d_i, the event indicator (`lex.Xst`).
- log-offset is the expected value for each piece of follow-up, $e_i = y_i \times \lambda_R$.
- λ_R is the population rate corresponding to the age, period and sex of the follow-up period y_i.




```
plot( thap, 2:1, col=c("blue","red")[thap$sex], lwd=2 )
```


```
plot( thap, 2:1, col=c("blue","red")[thap$sex], lwd=2 )
...
```


Split the data to fit with population data

```
> # Split the data for SMR-analysis
> tha  <- splitLexis(thL, "age", breaks=seq(0,90,5) )
> thap <- splitLexis(tha, "per", breaks=seq(1938,2038,5) )
> dim( thap )
[1] 41 15
> # Create variables to fit with the population data
> thap$agr <- timeBand( thap, "age", "left" )
> thap$cal <- timeBand( thap, "per", "left" )
> round( thap[,c("lex.id","age","agr","per","cal","lex.dur","lex
   lex.id   age agr     per  cal lex.dur lex.Xst sex
1       1 22.18  20 1938.79 1938    2.82       0   2
2       1 25.00  25 1941.61 1938    1.39       0   2
3       1 26.39  25 1943.00 1943    3.61       0   2
4       1 30.00  30 1946.61 1943    1.39       0   2
5       1 31.39  30 1948.00 1948    3.61       0   2
6       1 35.00  35 1951.61 1948    1.39       0   2
7       1 36.39  35 1953.00 1953    3.61       0   2
8       1 40.00  40 1956.61 1953    1.39       0   2
9       1 41.39  40 1958.00 1958    3.61       0   2
```

Merge with population data

```
> thapx <- merge( thap, gmortDK[,c("agr","cal","sex","rt")] )
> str( thapx )
Classes 'Lexis' and 'data.frame': 41 obs. of  18 variables:
 $ sex     : num  1 2 2 2 2 2 2 2 2 2 ...
 $ agr     : num  65 20 20 20 25 25 25 25 30 30 ...
 $ cal     : num  1953 1938 1953 1958 1938 ...
 $ lex.id  : int  3 1 4 4 1 1 4 4 1 1 ...
 $ age     : num  68.2 22.2 20.8 21.2 25.0 ...
 $ per     : num  1955 1939 1958 1958 1942 ...
 $ tfi     : num  0.000 0.000 0.000 0.389 2.818 ...
 $ lex.dur : num  1.405 2.818 0.389 3.806 1.391 ...
 $ lex.Cst : num  0 0 0 0 0 0 0 0 0 0 ...
 $ lex.Xst : num  1 0 0 0 0 0 0 0 0 0 ...
 $ id      : num  3425    1 4017 4017    1 ...
 $ birthdat: num  1887 1917 1937 1937 1917 ...
 $ contrast: num  2 1 2 2 1 1 2 2 1 1 ...
 $ injecdat: num  1955 1939 1958 1958 1939 ...
 $ volume  : num  0 22 0 0 22 22 0 0 22 22 ...
 $ exitdat : num  1957 1977 1992 1992 1977 ...
```


Calculation of the SMR

```
> thapx$E <- thapx$lex.dur * thapx$rt / 1000
> stat.table( contrast,
+             list( D = sum( lex.Xst ),
+                   Y = sum( lex.dur ),
+                   E = sum( E ),
+                 SMR = ratio( lex.Xst, E ) ),
+              margin = TRUE,
+                data = thapx )
 ----------------------------------------
 contrast         D       Y       E     SMR
 ----------------------------------------
 1             2.00   56.59    0.33    6.02
 2             1.00   35.93    0.11    8.70

 Total         3.00   92.52    0.45    6.71
 ----------------------------------------
```


Modelling the SMR

```
> m.SMR <- glm( lex.Xst ~ factor(contrast)-1+offset(log(E)),
+               family=poisson, data=thapx )
> round( ci.lin( m.SMR, Exp=TRUE )[,5:7], 3 )
                  exp(Est.)  2.5%  97.5%
factor(contrast)1     6.023 1.506 24.082
factor(contrast)2     8.698 1.225 61.745
```

- Analysis of SMR is like analysis of rates:
- Replace Y with E — that's all!

Likelihood for multistate follow-up

Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
http://BendixCarstensen/AdvCoh/Scot-2014

ms-lik

Likelihood for transition through states

$$\mathbf{A} \longrightarrow \mathbf{B} \longrightarrow \mathbf{C} \longrightarrow$$

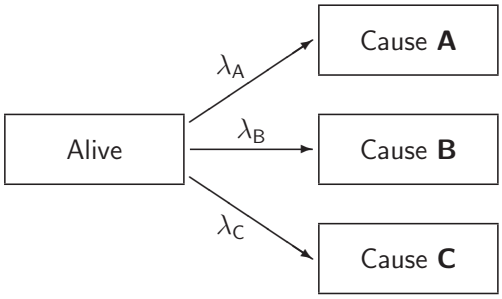
- given start of observation in **A** at time t_0
- transitions at times t_B and t_C
- survival in **C** till (at least) time t_x:

$$\begin{aligned} L = & \mathrm{P}\{\text{survive } t_0 \to t_B \text{ in } \mathbf{A}\} \\ & \times \mathrm{P}\{\text{transition } \mathbf{A} \to \mathbf{B} \text{ at } t_B \mid \text{ alive in } \mathbf{A}\} \\ & \times \mathrm{P}\{\text{survive } t_B \to t_C \text{ in } \mathbf{B} \mid \text{ entered } \mathbf{B} \text{ at } t_B\} \\ & \times \mathrm{P}\{\text{transition } \mathbf{B} \to \mathbf{C} \text{ at } t_C \mid \text{ alive in } \mathbf{B}\} \\ & \times \mathrm{P}\{\text{survive } t_C \to t_x \text{ in } \mathbf{C} \mid \text{ entered } \mathbf{C} \text{ at } t_C\} \end{aligned}$$

- Product of likelihoods for each transition
  — each one as for a survival model

Competing risks

But you may die from more than one cause
(or move to more than one state):



Cause-specific intensities

$$\begin{aligned} \lambda_A(t) &= \lim_{h\to 0} \frac{\mathrm{P}\{\text{death from cause A in } (t, t+h] \mid \text{alive at } t\}}{h} \\ \lambda_B(t) &= \lim_{h\to 0} \frac{\mathrm{P}\{\text{death from cause B in } (t, t+h] \mid \text{alive at } t\}}{h} \\ \lambda_C(t) &= \lim_{h\to 0} \frac{\mathrm{P}\{\text{death from cause C in } (t, t+h] \mid \text{alive at } t\}}{h} \end{aligned}$$

Total mortality rate:

$$\lambda_{\text{Total}}(t) = \lim_{h\to 0} \frac{\mathrm{P}\{\text{death from any cause in } (t, t+h] \mid \text{alive at } t\}}{h}$$

Cause-specific intensities

For small h, $\mathrm{P}\{2 \text{ events in } (t, t+h]\} \approx 0$, so:

$$\begin{aligned} & \mathrm{P}\{\text{death from any cause in } (t, t+h] \mid \text{alive at } t\} \\ = \; & \mathrm{P}\{\text{death from cause A in } (t, t+h] \mid \text{alive at } t\} + \\ & \mathrm{P}\{\text{death from cause B in } (t, t+h] \mid \text{alive at } t\} + \\ & \mathrm{P}\{\text{death from cause C in } (t, t+h] \mid \text{alive at } t\} \end{aligned}$$

$$\Longrightarrow \qquad \lambda_{\text{Total}}(t) = \lambda_A(t) + \lambda_B(t) + \lambda_C(t)$$

Intensities are additive,
if they all refer to the
same risk set, in this case "Alive".

Likelihood for competing risks

Data:

Y - person years in "Alive"
D_A - deaths from cause A
D_B - deaths from cause B
D_C - deaths from cause C

Now, assume for a start that transition rates between states are constant.

Likelihood for competing risks

A survivor contributes to the log-likelihood:

$$\log(\mathrm{P}\{\text{Survival for a time of } y\}) = -(\lambda_A+\lambda_B+\lambda_C)y$$

A death from cause **A** contributes an additional $\log(\lambda_A)$, from cause **B** an additional $\log(\lambda_B)$ etc.

The total log-likelihood is then:

$$\begin{aligned}\ell(\lambda_A,\lambda_B,\lambda_C) =& D_A\log(\lambda_A) + D_B\log(\lambda_B) + D_C\log(\lambda_C)\\ &- (\lambda_A+\lambda_B+\lambda_C)\,Y\\ =&[D_A\log(\lambda_A) - \lambda_A\,Y]+\\ &[D_B\log(\lambda_B) - \lambda_B\,Y]+\\ &[D_C\log(\lambda_C) - \lambda_C\,Y]\end{aligned}$$

Components of the likelihood

The log-likelihood is made up of three contributions:

- one for cause A,
- one for cause B and
- one for cause C

Deaths are the cause-specific deaths,

but the **person-years** are the same in all contributions.

Likelihood for multiple states

- **Product** of likelihoods for each transition — each one as for a survival model
- **conditional** on being alive at (observed) entry to current state
- **Risk time** is the risk time in the current ("From", `lex.Cst`) state
- **Events** are transitions to the "To" state (`lex.Xst`)
- All other transitions out of "From" are treated as **censorings** (but they are not)
- Fit models separately for each transition or jointly for all

Time varying rates:

- The same type of analysis as with a constant rates, but data must be
- split in intervals sufficiently small to justify an assumption of constant rate (intensity),
- the model should allow for a separate rate for each interval,
- but constrained to follow model with a smooth effect of the time-scale values allocated to each interval.

Practical implications

- Empirical rates ((d, y) from each individual) will be the same for all analyses except for those where deaths occur.
- Analysis of cause **A**:
  - Contributions $(1, y)$ only for those intervals where a cause **A** death occurs.
  - Intervals with cause **B** or **C** deaths (or no deaths) contribute only $(0, y)$ treated as censorings.

original

id	time	cause	xx	d.A	d.B	d.C
1	1	B	0.50	0	1	0
2	1	NA	1.00	0	0	0
3	8	B	-1.74	0	1	0
4	3	A	-0.55	1	0	0
5	7	NA	-0.58	0	0	0
6	7	C	-0.04	0	0	1

expanded

id	time	dd	xx	Tr
1	1	0	0.50	A
2	1	0	1.00	A
3	8	0	-1.74	A
4	3	1	-0.55	A
5	7	0	-0.58	A
6	7	0	-0.04	A
1	1	1	0.50	B
2	1	0	1.00	B
3	8	1	-1.74	B
4	3	0	-0.55	B
5	7	0	-0.58	B
6	7	0	-0.04	B
1	1	0	0.50	C
2	1	0	1.00	C
3	8	0	-1.74	C
4	3	0	-0.55	C
5	7	0	-0.58	C
6	7	1	-0.04	C

. . . accomplished by `stack.Lexis`

`Lexis` objects (data frame)

- Represents the **follow-up**
- `lex.dur` contains the total time at risk for (any) event
- `lex.Cst` is the state in which this time is spent
- `lex.Xst` is the state to which a transition occurs
  — if no transition, the same as `lex.Cst`.

This is used for modelling of single transitions between states — and multiple transitions with no two originating in the same state.

`stacked.Lexis` objects (data frame)

- Represents the **likelihood** contributions
- `lex.dur` contains the total time at risk for (any) event
- `lex.Tr` is the transition to which the record contributes
- `lex.Fail` is the event (failure) indicator for the transition in question.

This is used for joint modelling of **all** transition in a multistate set-up.

Particularly with several rates originating in the **same** state (competing risks).

Implemented in the `stack.Lexis` function:

```
> options( digits=3, width=200 )
> st.dmi <- stack( dmi )
> print( st.dmi[1:6,], row.names=F )

     Per  Age DMdur lex.dur lex.Cst lex.Xst  lex.Tr lex.Fail lex.
    1999 58.7     0  11.080      DM      DM DM->Ins    FALSE
    2003 64.1     0   6.689      DM      DM DM->Ins    FALSE
    2005 86.3     0   5.446      DM      DM DM->Ins    FALSE
    2009 44.0     0   0.736      DM      DM DM->Ins    FALSE
    2009 75.8     0   1.344      DM      DM DM->Ins    FALSE
    2008 80.0     0   2.037      DM    Dead DM->Ins    FALSE

> str( st.dmi )

   Classes 'stacked.Lexis' and 'data.frame': 21589 obs. of  16 va
    $ Per     : num  1999 2003 2005 2009 2009 ...
    $ Age     : num  58.7 64.1 86.3 44 75.8 ...
    $ DMdur   : num  0 0 0 0 0 0 0 0 0 0 ...
    $ lex.dur : num  11.08 6.689 5.446 0.736 1.344 ...
    $ lex.Cst : Factor w/ 3 levels "DM","Ins","Dead": 1 1 1 1 1 1
    $ lex.Xst : Factor w/ 3 levels "DM","Ins","Dead": 1 1 1 1 1 3
    $ lex.Tr  : Factor w/ 3 levels "DM->Ins","DM->Dead",..: 1 1 1
    $ lex.Fail: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
    $ lex.id  : int  1 2 3 4 5 6 7 8 9 10 ...
```


Implemented in the `stack.Lexis` function:

```
> library( Epi )
> data(DMlate)
> head(DMlate)

         sex    dobth     dodm    dodth    dooad doins      dox
   50185   F 1940.256 1998.917       NA       NA    NA 2009.997
   307563  M 1939.218 2003.309       NA 2007.446    NA 2009.997
   294104  F 1918.301 2004.552       NA       NA    NA 2009.997
   336439  F 1965.225 2009.261       NA       NA    NA 2009.997
   245651  M 1932.877 2008.653       NA       NA    NA 2009.997
   216824  F 1927.870 2007.886 2009.923       NA    NA 2009.923

> dml <- Lexis( entry = list(Per = dodm,
+                            Age = dodm-dobth,
+                          DMdur = 0 ),
+               exit = list(Per = dox ),
+         exit.status = factor(!is.na(dodth),
+                              labels=c("DM","Dead")),
+               data = DMlate )

   NOTE: entry.status has been set to "DM" for all.
```


Implemented in the `stack.Lexis` function:

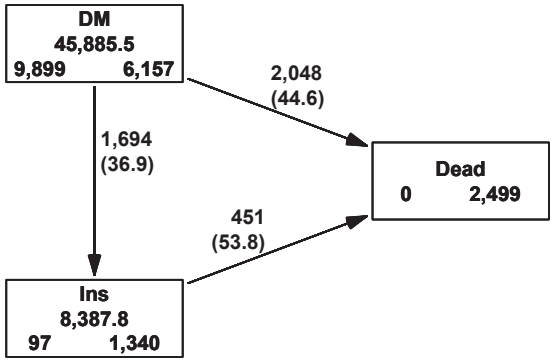
```
> print( subset(    dmi, lex.id %in% c(13,15,28) ), row.names=FA

     Per  Age DMdur lex.dur lex.Cst lex.Xst lex.id sex dobth dodm
    1997 59.4   0.0   0.890      DM    Dead     13   M  1938 1997
    2003 58.1   0.0   2.804      DM     Ins     15   M  1944 2003
    2005 60.9   2.8   4.643     Ins     Ins     15   M  1944 2003
    1999 73.7   0.0   8.701      DM     Ins     28   F  1925 1999
    2007 82.4   8.7   0.977     Ins    Dead     28   F  1925 1999

> print( subset( st.dmi, lex.id %in% c(13,15,28) ), row.names=FA

     Per  Age DMdur lex.dur lex.Cst lex.Xst    lex.Tr lex.Fail le
    1997 59.4   0.0   0.890      DM    Dead   DM->Ins    FALSE
    2003 58.1   0.0   2.804      DM     Ins   DM->Ins     TRUE
    1999 73.7   0.0   8.701      DM     Ins   DM->Ins     TRUE
    1997 59.4   0.0   0.890      DM    Dead  DM->Dead     TRUE
    2003 58.1   0.0   2.804      DM     Ins  DM->Dead    FALSE
    1999 73.7   0.0   8.701      DM     Ins  DM->Dead    FALSE
    2005 60.9   2.8   4.643     Ins     Ins Ins->Dead    FALSE
    2007 82.4   8.7   0.977     Ins    Dead Ins->Dead     TRUE
```


Implemented in the `stack.Lexis` function:

```
> dmi <- cutLexis( dml, cut = dml$doins,
+                  new.state = "Ins",
+                  precursor = "DM" )
> summary( dmi )

   Transitions:
        To
   From    DM  Ins Dead  Records:  Events: Risk time:  Persons:
     DM  6157 1694 2048      9899      3742   45885.49      9899
     Ins    0 1340  451      1791       451    8387.77      1791
     Sum 6157 3034 2499     11690      4193   54273.27      9996

> boxes( dmi, boxpos = list(x=c(20,20,80),
+                           y=c(80,20,50)),
+             scale.R=1000, show.BE=TRUE, hmult=1.2, wmult=1.1 )
```


Analysis of rates in multistate models

- Interactions between all covariates (including time) and state (`lex.Cst`):
  ⇒ separate analyses of all transition rates.
- Only interaction between state (`lex.Cst`) and time(scales):
  ⇒ same covariate effects for all causes transitions, but separate baseline hazards — "stratified model".
- Main effect of state only (`lex.Cst`):
  ⇒ proportional hazards
- No effect of state:
  ⇒ identical baseline hazards — hardly ever relevant.



Analysis approaches and data representation

- `Lexis` objects represents the precise follow-up in the cohort, in states and along timescales
- — used for analysis of single transition rates.
- `stacked.Lexis` objects represents contributions to the total likelihood
- — used for joint analysis of (all) rates in a multistate setup
- . . . which is the case if you want to specify common effects between different transitions.

Assumptions in competing risks

"Classical" way of looking at survival data: description of the distribution of time to death.

For competing risks that would require three variables:
T_A, T_B and T_C, representing times to death from each of the three causes.
But at most one of these is observed.

Often it is stated that these must be assumed independent in order to make the likelihood machinery work

1. It is not necessary.
2. Independence can never be assessed from data.

An account of these problems is given in:

PK Andersen, SZ Abildstrøm & S Rosthøj:
Competing risks as a multistate model,
Statistical Methods in Medical Research; **11**, 2002: pp. 203–215

Per Kragh Andersen, Ronald B Geskus, Theo de Witte & Hein Putter:
Competing risks in epidemiology: possibilities and pitfalls,
International Journal of Epidemiology; 2012: pp. 1–10

Contains examples where both dependent and independent "cause specific survival times" gives rise to the same set of cause specific rates.

Lifetime risk

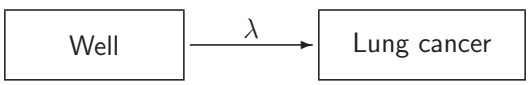
Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
`http://BendixCarstensen/AdvCoh/Scot-2014`

`DK-lung`

Competing risk interpretation

The problems with competing risk models **only** comes when estimated intensities (rates) are used to produce probability statements.

Classical set-up in cancer-registries:

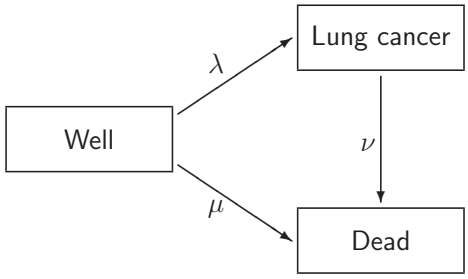


Common statement:

$$\mathrm{P}\{\text{Lung cancer before age 75}\} = 1 - \mathrm{e}^{-\Lambda(75)}$$

This is not quite right.

How the world really looks



Illness-death model, mortality of lung cancer patients (ν) not relevant here, we only want to find out how many pass through "Lung cancer"

How many get lung cancer before age a?

$$\mathrm{P}\{\text{Lung cancer before age 75}\} \neq 1 - \mathrm{e}^{-\Lambda(75)}$$

the r.h.s. does not take the possibility of death prior to lung cancer into account.

- $1 - \mathrm{e}^{-\Lambda(75)}$ often stated as the probability of lung cancer before age 75, assuming all other acuses of death absent.
- Lung cancer rates are however observed in a mortal population.
- If all other causes of death were absent, this would assume that lung cancer rates remained the same.

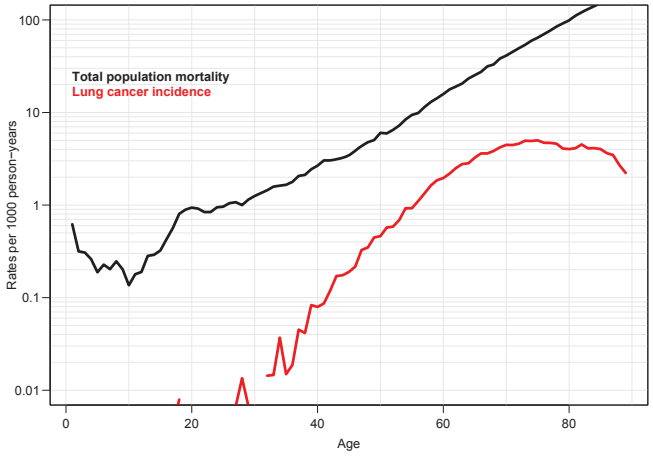
How it really is:

$$\begin{aligned}
&\mathrm{P}\{\text{Lung cancer diagnosis before age } a\} \\
&= \int_0^a \mathrm{P}\{\text{Lung cancer at age } u\}\,\mathrm{d}u \\
&= \int_0^a \mathrm{P}\{\text{Lung cancer in age } (u, u+\mathrm{d}u] \mid \text{alive at } u\} \\
&\qquad \times \mathrm{P}\{\text{alive at } u \text{ without lung cancer}\}\,\mathrm{d}u \\
&= \int_0^a \lambda(u)\exp\left(-\int_0^u \mu(s)+\lambda(s)\,\mathrm{d}s\right)\mathrm{d}u
\end{aligned}$$

Probability of lungcancer

The rates are easily plotted for inspection in R:

```
matplot( age, 1000*cbind( D/Y, lung/Y ),
         log="y", type="l", lty=1, lwd=3,
         ylim=c(0.01,100), xlab="Age",
         ylab="Rates per 1000 person-years" )
```



The probablility that a person contracts lung cancer before age a is:

$$\int_0^a \lambda(u) \exp\left(-\int_0^u \mu(s) + \lambda(s)\,\mathrm{d}s\right)\,\mathrm{d}u$$

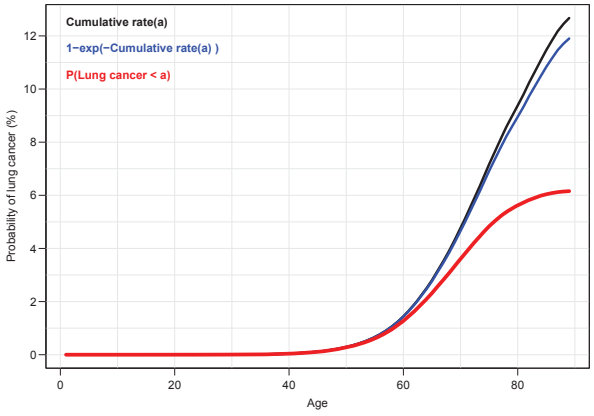
$$= \int_0^a \lambda(u) \exp\Big(-\big(\mathrm{M}(u) + \Lambda(u)\big)\Big)\,\mathrm{d}u$$

$\mathrm{M}(u)$ is the cumulative mortality rate.

$\Lambda(u)$ is the cumulative lung cancer incidence rate.

R-commands needed to do the calculations:

```
cr.death <- cumsum( D/Y )
cr.lung <- cumsum( lung/Y )
p.simple <- 1 - exp( -cr.lung )
p.lung <- cumsum( lung/Y *
                  exp( -(cr.death+cr.lung) ) )
matlines( age, 100*cbind( cr.lung, p.simple, p.lung ),
          type="l", lty=1, lwd=2*c(2,2,3),
          col=c("black","blue","red") )
```

Assumptions

- The calculation and the statement "6% of Danish males will get lung cancer" assumess that the lung cancer rates and the mortality rates in the file apply to a cohort of men.
- But they are cross-sectional rates, so the assumption is one of steady state of:
  1. mortality rates (which is dubious)
  2. lung cancer incidence rates (which is appalling).
- However, the machinery can be applied to any set of rates for competing risks, regardless of how they were estimated.

Interactions and timescales

Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
`http://BendixCarstensen/AdvCoh/Scot-2014`

`timescales`

Computational aspects of fitting models

- Cox model:
  - Only one timescale.
  - Each person contributes one (or very few) records.
  - Computationally simple, because time (risk / covariate) is profiled out in the estimation.
- Poisson modelling:
  - Many records per person.
  - Very large datasets.
  - Any number of timescales.
  - Timeconsuming due to the full modelling of the rates.

Historical aspects

Whitehead J: Fitting Cox's regression model to survival data using GLIM. Applied Statistics, 29(3):268–275, 1980.[**?**][2]

Set up tables of event counts and person-years, classified by event times and covariate patterns.

Even with moderate datasets this can be large, albeit smaller than some 100 separate records per person.

[2] Recall **Keiding's law**: "Any result was published earlier than you think, even if you take Keiding's law into account."

Computational practicalities

Early 1980s: Fitting of Poisson models on datasets with 50,000 records were out of the question.
In particular with 100+ parameters.

Computationally feasible approaches to cohort studies were:

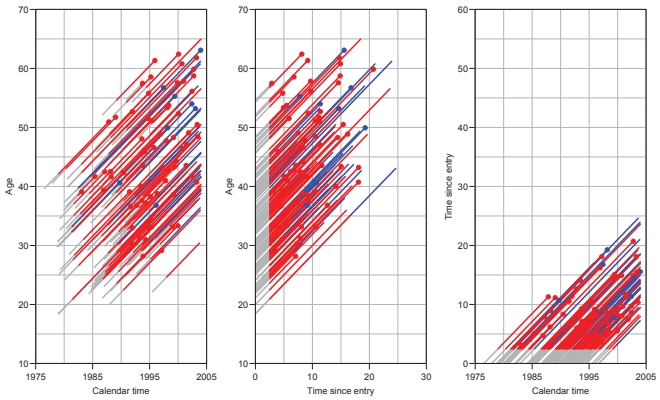
- Cox modelling — thanks to computational elegance.
- Time-splitting and tabulation before modelling.

The tabulation legacy (curse)

The **computational** need for tabulation has influenced thinking in epidemiology / demography:

- Life-tables in 1-year intervals.
- Rates are regarded in 5-year age by period intervals. Used for analysis of mortality and incidence rates based on registers.
  Age-period-cohort models with one parameter per level of the age/period factor.
- Yet, survival analysis is largely based on "time to event" methods (Kaplan-Meier, Cox), even from cancer registries — only one timescale.

Representation of follow-up



Age at entry as covariate

t: time since entry
e: age at entry
$a = e + t$: current age

$$\log\big(\lambda(a,t)\big) = f(t) + \beta e = (f(t) - \beta t) + \beta a$$

Immaterial whether a or e is used as (log)-*linear* covariate as long as t is in the model.

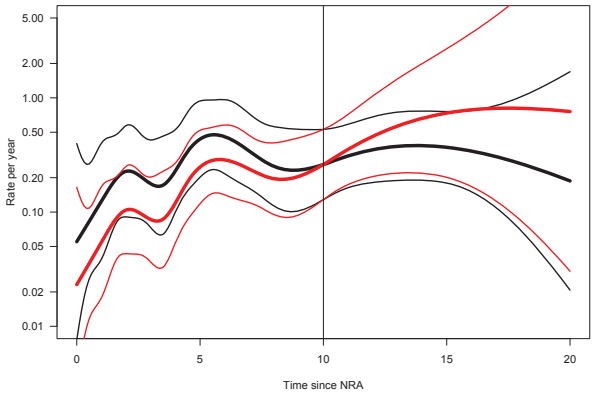
In a Cox-model with time since entry as time-scale, only the baseline hazard will change if age at entry is replaced by current age (a time-dependent variable).

"Controlling for age"

Including age at entry:

- Linear effect.
- Grouped variable.
- Parametric function.

— still only controls for the *linear* effect of *current age*.



Current age as covariate
Age at entry as covariate

Non-linear effects of time-scales

Arbitrary effects of the three variables t, a and e: Genuine extension of the model.

$$\log\big(\lambda(a,t,x_i)\big) = f(t) + g(a) + h(e) + \eta_i$$

Three quantities can be arbitrarily moved between the three functions:

$$\begin{aligned}\tilde{f}(t) &= f(a) - \mu_a - \mu_e + \gamma t\\ \tilde{g}(a) &= g(p) + \mu_a \qquad\quad - \gamma a\\ \tilde{h}(e) &= h(c) \qquad\quad + \mu_a + \gamma e\end{aligned}$$

because $t - a + e = 0$.
How many timescales in this model?

"Controlling for age"

— is not a well defined statement.

Mostly it means that age *at entry* is included in the model.

But ideally one would check whether there were non-linear effects of age at entry and current age.

This would require modelling of multiple timescales.

Which is best accomplished by splitting time and modelling the timescales explicitly.

Several timescales: Caveat

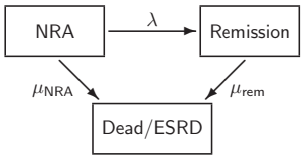
As an example, consider:
t: time since entry
e: age at entry
$a = e + t$: current age

The relation: $a = t + e$ must hold for all units of analysis.

In general: The difference between two time-scales must be constant within individuals.

The Boyle-Robertson fallacy from age-period-cohort models, where units with identical values of (current) age, a, and (current) period p had varying values of cohort, date of birth $c = p - a$! [**?**].

Several timescales



Cox-model:
— One dataset per transition.
— Combine datasets and make relevant interactions.
— Timescale must be the same.

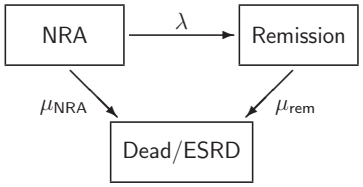
Poisson-model:
— One time-split dataset per transition.
— Combine datasets and make relevant interactions.
— Timescales can be different, and multiple timsecales can be accomodated simultaneously; duration of NRA, for example.

Time dependent variable

How does remission influence the mortality?

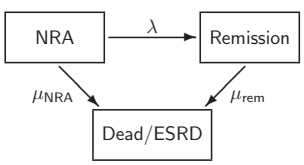
$$\lambda(t) = \lambda_0(t)\exp\big(1\{\text{remission}\}(t) \times \beta\big)$$

i.e. when remission occurs, mortality increase by e^{β}.



What transitions are modelled here?

Time-dependent variable



If we take

$$1\{\text{remission}\}(t)$$

as time-dependent variable, we assume that μ_{NRA} and μ_{rem} are proportional on the same timescale — no disease duration!.

— and λ is not modelled at all.

Stratified model

A popular version of the Cox-model allowing for non-proportionality is the **stratified model**:

$$\lambda(t, x) = \lambda_s(t) \times \exp(x'\beta)$$

where s refers to levels of a factor S.

This is but a completely general **interaction** between the factor S and the chosen timescale.

A better approach to interactions would be to specify a clinically founded form of interaction, so that test for interaction is against a specific (and sensible) alternative.

Time varying coefficients

This is a concept introduced by letting (some of) the parameters depend on time:

$$\lambda(t, x) = \lambda_0 \times \exp\big(x'\beta(t)\big)$$

This is also an interaction, but restricted:
The effect of a covariate is linear for any value of t.

If the covariate is a factor, then we just have a reparametrization of the stratified model.

Poisson modelling of interactions

When interactions are needed (or desired):

- use the familiar terminology of interaction as known from (generalized) linear models.
- use clinical judgement of which interactions are relevant.
- use clinical judgement of which forms of interaction are relevant.
- are interactions with time of special interest?

Poisson model for time-split data

- Clarifies the destinction between (risk) time as response variable and time(scales) as covariates.
- Multiple timescales easily handled.
- Hazard rates by standard methods.
- More credible estimates of survival functions.
- Sensible modelling of interactions between timescales and other variables (and between timescales).
- Interactions are called interactions.

Scottish diabetes data

Modern Demographic
Methods in Epidemiology
with R
26–29 August 2014
University of Edinburgh
http://BendixCarstensen/AdvCoh/Scot-2014

Scot-DM

Scottish DM data

- Population data as of 1 July and deaths during the year, by:
  - Year (2005–2012)
  - Age (0–90)
  - Sex
  - Deprivation index (1–10 (11))
  - `pop <- read.csv( "../data/PopulationSIMD2009.csv" )`
- Anonymized diabetes records, one per person:
  - Date of birth
  - Date of diabetes
  - Date of death
  - Sex
  - Deprivation index (1-10)
  - `DM <- read.csv( "../data/dm-data.csv" )`

Types of analyses

- Prevalence of diabetes
- Incidence rates of diabetes
- Mortality rates among diabetes patients
- SMR

Analyses from the special chapter in the practicals.