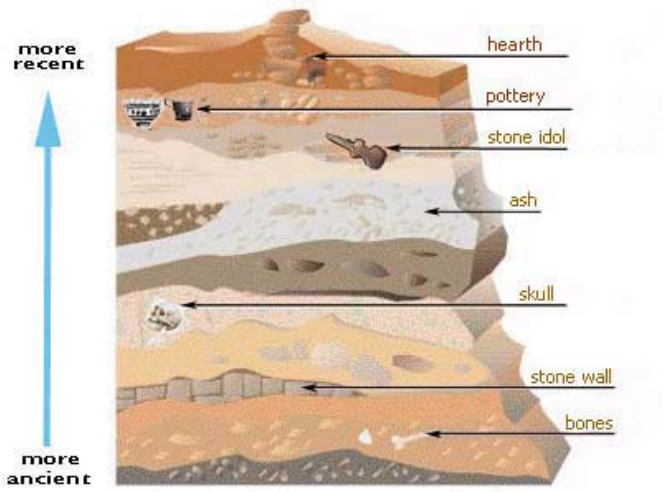
History and Ecology of R

Martyn Plummer

University of Warwick, UK

SPE 2023, Tartu

Abstract



Pre-history

Before there was R, there was S.

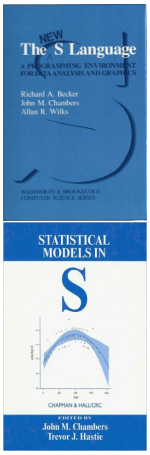
The S language

Developed at AT&T Bell laboratories by Rick Becker, John Chambers, Doug Dunn, Paul Tukey, Graham Wilkinson.

Version 1	1976–1980	Honeywell GCOS, Fortran-based
Version 2	1980–1988	Unix; Macros, Interface Language
	1981–1986	QPE (Quantitative Programming Environment)
	1984–	General outside licensing; books
Version 3	1988-1998	C-based; S functions and objects
	1991–	Statistical models; informal classes and methods
Version 4	1998	Formal class-method model; connections; large objects
	1991–	Interfaces to Java, Corba?

Source: Stages in the Evolution of S `http://ect.bell-labs.com/sl/S/history.html`

The "Blue Book" and the "White Book"



Key features of S version 3 outlined in two books:

- Becker, Chambers and Wilks, *The New S Language: A Programming Environment for Statistical Analysis and Graphics* (1988)
  - Functions and objects
- Chambers and Hastie (Eds), *Statistical Models in S* (1992)
  - Data frames, formulae

These books were later used as a prototype for R.

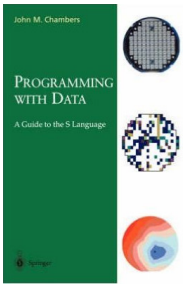
Programming with Data

> *"We wanted users to be able to begin in an interactive environment, where they did not consciously think of themselves as programming. Then as their needs became clearer and their sophistication increased, they should be able to slide gradually into programming." – John Chambers, Stages in the Evolution of S*

This philosophy was later articulated explicitly in *Programming With Data* (Chambers, 1998) as a kind of mission statement for S

> *To turn ideas into software, quickly and faithfully*

The "Green Book"



Key features of S version 4 were outlined in Chambers, *Programming with Data* (1998).

- S as a programming language
- Introduced formal classes and methods, which were later introduced into R by John Chambers himself.

S-PLUS

- AT&T was a regulated monopoly with limited ability to exploit creations of Bell Labs.
- S source code was supplied for free to universities
- After the break up of AT&T in 1984 it became possible for them to sell S.
- S-PLUS was a commercially available form of S licensed to Statistical Sciences (later Mathsoft, later Insightful) with added features:
  - GUI,survival analysis, non-linear mixed effects, Trellis graphics, ...

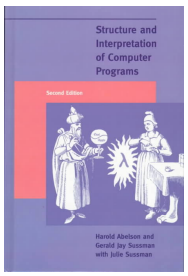
The Rise and Fall of S-PLUS

- 1988. Statistical Science releases first version of S-PLUS.
- 1993. Acquires exclusive license to distribute S. Merges with Mathsoft.
- 2001. Changes name to Insightful.
- 2004. Purchases S language for $2 million.
- 2008. Insightful sold to TIBCO. S-PLUS incorporated into TIBCO Spotfire.

History

How R started, and how it turned into an S clone

The Dawn of R



- Ross Ihaka and Robert Gentlemen at the University of Auckland
- An experimental statistical environment
- Scheme interpreter with S-like syntax
  - Replaced scalar type with vector-based types of S
  - Added lazy evaluation of function arguments
- Announced to *s-news* mailing list in August 1993.

A free software project

- June 1995. Martin Maechler (ETH, Zurich) persuades Ross and Robert to release R under GNU Public License (GPL)
- March 1996. Mailing list *r-testers* mailing list
  - Later split into three *r-announce*, *r-help*, and *r-devel*.
- Mid 1997. Creation of *core team* with access to central repository (CVS)
  - Doug Bates, Peter Dalgaard, Robert Gentleman, Kurt Hornik, Ross Ihaka, Friedrich Leisch, Thomas Lumley, Martin Maechler, Paul Murrell, Heiner Schwarte, Luke Tierney
- 1997. Adopted by the GNU Project as "GNU S".

The draw of S

"Early on, the decision was made to use S-like syntax. Once that decision was made, the move toward being more and more like S has been irresistible"
– Ross Ihaka, R: Past and Future History (Interface '98)

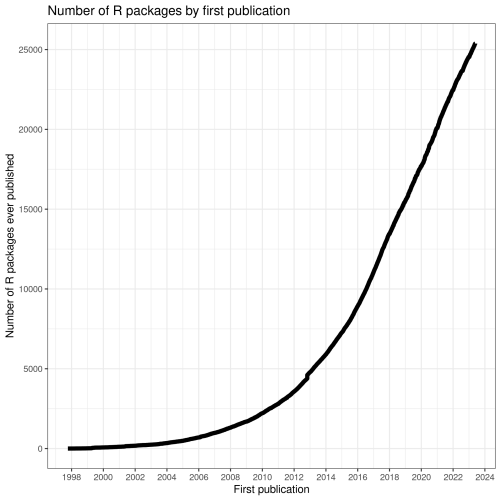
R 1.0.0, a complete and stable implementation of S version 3, was released in 2000.

A Souvenir

Packages

- Comprehensive R Archive Network (CRAN) started in 1997
  - Quality assurance tools built into R
  - Increasingly demanding with each new R release
- Recommended packages distributed with R
  - Third-party packages included with R distribution
  - Provide more complete functionality for the R environment
  - Starting with release 1.3.0 (completely integrated in 1.6.0)

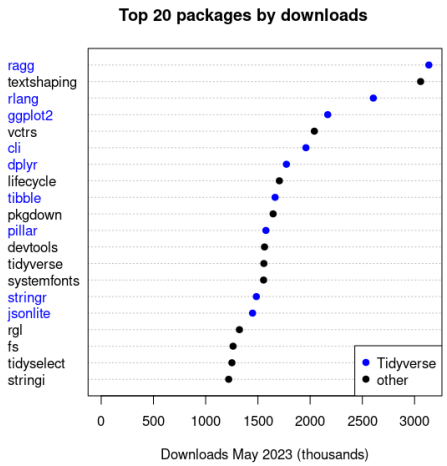
Growth of CRAN



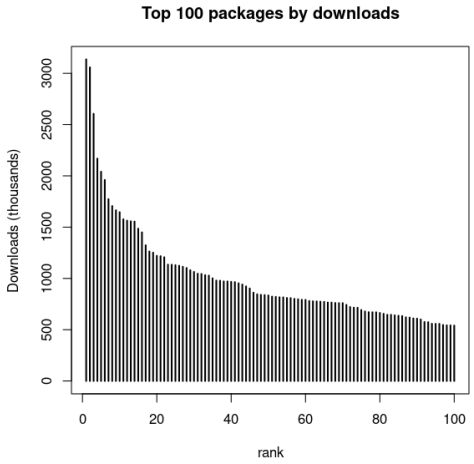
Community

- useR! Annual conference
  - Toulouse (2019), Online (2020, 2021), Nashville (2022)
- R Journal (`http://journal.r-project.org`)
  - Journal of record, peer-reviewed articles, indexed
  - Journal of Statistical Software (JSS) has many articles dedicated to R packages (`http://jstatsoft.org`)
- Migration to social media
  - Stack Exchange/Overflow, Github, Twitter, Mastodon (#rstats)
  - Follow @_R_Foundation on Twitter, or @R_Foundation@fosstodon.org on Mastodon

Much important R infrastructure is now in package space



Much important R infrastructure is now in package space



The tidyverse

- Many of the popular packages on CRAN come from the company Posit (formerly R Studio).
- These packages are known as the "tidyverse" (`www.tidyverse.org`).
- All packages in the tidyverse have a common design philosophy and work together. Common features are:
  - Non-standard evaluation rules for function calls.
  - Use of the pipe operator |> (or %>%) to pass data transparently from one function call to another.
- The CRAN meta-package `tidyverse` installs all of these packages.

The R Foundation for Statistical Computing

A non-profit organization working in the public interest, founded in 2002 in order to:

- Provide support for the R project and other innovations in statistical computing.
- Provide a reference point for individuals, institutions or commercial enterprises that want to support or interact with the R development community.
- Hold and administer the copyright of R software and documentation (This never happened)

The R Consortium

In 2015, a group of organizations created a consortium to support the R ecosystem. Current members (May 2023)

R Foundation A statutory member of The R Consortium

Platinum members Biogen, Genentech, Microsoft, Posit

Gold members ASA, Esri, Google, GSK, Janssen, Lander Analytics, Mango Solutions, Merck

Silver members Novo Nordisk, Oracle, Parexel, Pfizer, Procogia, Swiss Re

The Future

"Prediction is very difficult, especially about the future" – variously attributed to Niels Bohr, Piet Hein, Yogi Bera

Trends

We cannot make predictions, but some long-term trends are very visible:

- Average age of R Core Team?
- Younger R developers more closely associated with industry than academia
- Strong competition from Python

What does all of this mean for the course?

- R incorporates over 40 years of ideas in statistical computing from multiple contributors.
- There is usually more than one way to do something in R.
- Some of the peculiarities of the R language are there for historical reasons.
- The course does not cover some of the recent additions to the R ecosystem.

Resources

- Chambers J, Stages in the Evolution of S
- Becker, R, A Brief History of S
- Chambers R, Evolution of the S language
- Ihaka, R and Gentleman R, R: A language for Data Analysis and Graphics, *J Comp Graph Stat*, **5**, 299–314, 1996.
- Ihaka, R, R: Past and Future History, Interface 98.
- Ihaka, R, Temple Lang, D, Back to the Future: Lisp as a Base for a Statistical Computing System
- Fox, J, Aspects of the Social Organization and Trajectory of the R Project, R Journal, Vol 1/2, 5–13, 2009.

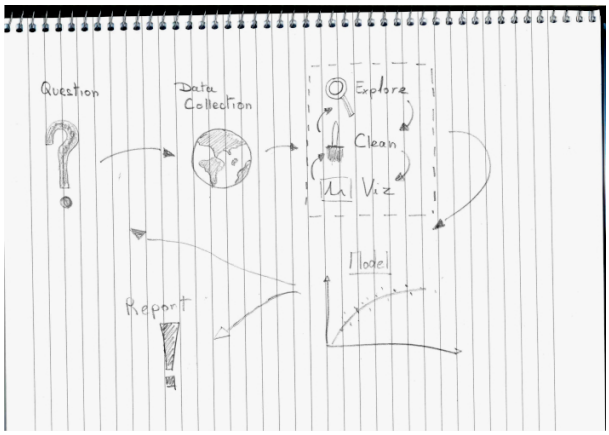
Data manipulation with tidyverse

Damien Georges

International Agency for Resarch on Cancer

June 2023 - Tartu

Epidemiological study workflow



Data manipulation tools in R

- `R` core function
- `tidyverse` / `dplyr`
- `data.table`
- ...

=> The best tool is the one you feel the most comfortable with

Tidyverse (from www.tidyverse.org)

R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

pipe functions %>% or |>

```
chill(fold(add(melt(add(chocolate, butter)),
               beat(add(eggs.white, cream)))))
```

pipe functions %>%

```
chill(fold(add(melt(add(chocolate, butter)),
               beat(add(eggs.white, cream)))))
```

```
chocolate %>%
  add(butter) %>%
  melt() %>%
  add(
    eggs.white %>%
      add(cream) %>%
      beat()
  ) %>%
  fold() %>%
  chill()
```

Non-standard evaluation rules for function calls

- used in different R packages
- provide flexibility and ease of use
- more concise and expressive programming in R

```
dat <- data.frame(x = 1:10)
## subset supports NSE
subset(dat, x < 5)
## and SE
subset(dat, data$x < 5)
```

data manipulation with 

Code as you speak: Data manipulation with `dplyr` is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- select rows (`slice`)

data manipulation with 

Code as you speak: Data manipulation with `dplyr` is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- select rows (`slice`)
- filter rows (`filter`)

data manipulation with 

Code as you speak: Data manipulation with `dplyr` is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- select rows (`slice`)
- filter rows (`filter`)
- arrange rows (`arrange`)

data manipulation with 

Code as you speak: Data manipulation with `dplyr` is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- select rows (`slice`)
- filter rows (`filter`)
- arrange rows (`arrange`)
- select columns (`select`)

data manipulation with 

Code as you speak: Data manipulation with `dplyr` is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- select rows (`slice`)
- filter rows (`filter`)
- arrange rows (`arrange`)
- select columns (`select`)
- create/modify columns (`mutate`)

data manipulation with 

Code as you speak: Data manipulation with `dplyr` is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- select rows (`slice`)
- filter rows (`filter`)
- arrange rows (`arrange`)
- select columns (`select`)
- create/modify columns (`mutate`)
- group and summarize data (`group_by` and `summarise`)

data manipulation with 

Code as you speak: Data manipulation with `dplyr` is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- select rows (`slice`)
- filter rows (`filter`)
- arrange rows (`arrange`)
- select columns (`select`)
- create/modify columns (`mutate`)
- group and summarize data (`group_by` and `summarise`)
- bind different tables (`bind_rows`, `bind_cols`)

data manipulation with 

Code as you speak: Data manipulation with `dplyr` is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- select rows (`slice`)
- filter rows (`filter`)
- arrange rows (`arrange`)
- select columns (`select`)
- create/modify columns (`mutate`)
- group and summarize data (`group_by` and `summarise`)
- bind different tables (`bind_rows`, `bind_cols`)
- merge different tables (`left_join`, `right_join`, `inner_join`, `full_join`)

discovering other `tidyverse` packages features

- data visualization with  (`ggplot`, `geom_bars`, ...)

discovering other `tidyverse` packages features

- data visualization with  (`ggplot`, `geom_bars`, ...)
- pivoting data with  (`pivot_wider`, `pivo_longer`)

discovering other `tidyverse` packages features

- data visualization with  (`ggplot`, `geom_bars`, ...)
- pivoting data with  (`pivot_wider`, `pivo_longer`)
- reading data with  (`read_table`, `read_csv`)

discovering other `tidyverse` packages features

- data visualization with  (`ggplot`, `geom_bars`, ...)
- pivoting data with  (`pivot_wider`, `pivo_longer`)
- reading data with  (`read_table`, `read_csv`)
- manipulating lists with  (`map`, `map_chr`, `reduce`, ...)

discovering other `tidyverse` packages features

- data visualization with  (`ggplot`, `geom_bars`, ...)
- pivoting data with  (`pivot_wider`, `pivo_longer`)
- reading data with  (`read_table`, `read_csv`)
- manipulating lists with  (`map`, `map_chr`, `reduce`, ...)
- manipulating strings with  (`str_length`, `str_remove`, ...)

Poisson and Binary Regression

Janne Pitkäniemi

Finnish Cancer Registry

Tampere university

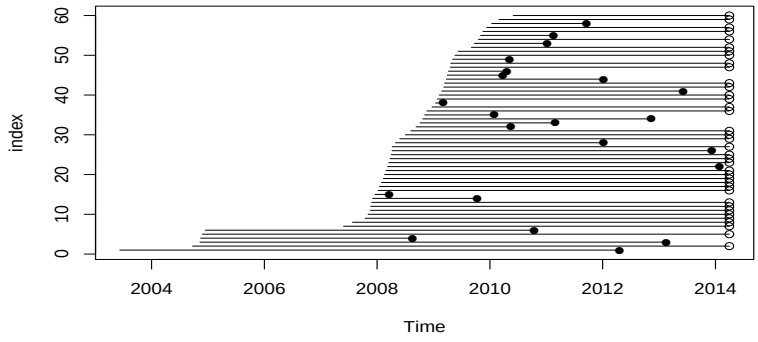
Statistical Practice in Epidemiology (2023, Tartu)

Points to be covered

- Incidence rates, rate ratios and rate differences from *follow-up studies* can be computed by fitting *Poisson regression models*.
- Risk ratios and differences can be computed from binary data by fitting *Logistic regression models*.
- Both models are special instances of *Generalized linear models*.
- There are various ways to do these tasks in R.

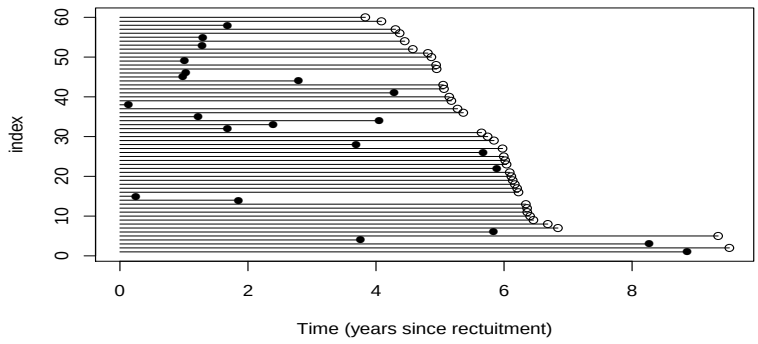
The Estonian Biobank cohort: survival among the elderly

Follow-up of 60 random individuals aged 75-103 at recruitment, until death (•) or censoring (o) in April 2014 (linkage with the Estonian Causes of Death Registry). (time-scale: calendar time).



The Estonian Biobank cohort: survival among the elderly

Follow-up time for 60 random individuals aged 75-103 at recruitment (time-scale: time in study).



Events, dates and risk time

- Mortality as the outcome:
  - `d`: indicator for **status** at exit:
    - **1**: death observed
    - **0**: censored alive
- Dates:

  `doe` = date of **E**ntry to follow-up,
  `dox` = date of e**X**it, end of follow-up.

- Follow-up time (years) computed as:

  `y = (dox - doe)/365.25`

Crude overall rate computed by hand and model

Total no. cases, person-years & rate (/1000 y):

```
> D <- sum( d ); Y <- sum(y) ; R <- D/(Y/1000)
>   round( c(D=D, Y=Y, R=R), 2)
     D        Y        R
884.00 11678.24    75.70
```

R-implementation of the rate estimation with Poisson regression:

A model with offset term

```
> m1 <- glm( D ~ 1, family=poisson,
    offset=log(Y))

> coef(m1)
(Intercept)
     -2.581
```

A model with poisreg−family (Epi package)

```
> glm(cbind(D, Y) ~1, family=poisreg)

 Coefficients :
(Intercept)
     -2.581
```

From the coefficient we get estimate of the rate $exp(-2.581) * 1000 = 75.70$

Constant hazard — Poisson model

Let $Y \sim exp(\lambda)$, then $f(y;\lambda) = \lambda e^{-\lambda y} I(y > 0)$
Constant rate model: $\lambda(y) = \frac{f(y;\lambda)}{S(y;\lambda)} = \lambda$ and observed data $\{(y_i, \delta_i); i = 1, ..., n\}$.

The likelihood $L(\lambda) = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda y_i}$ and
$log(L) = \sum_{i=1}^{n} [\delta_i log(\lambda) - \lambda y_i]$
Solving the *score equations*:

$\frac{\partial \log L(\lambda)}{\partial \lambda} = \sum \left[\frac{\delta_i}{\lambda} - y_i\right] = \frac{D}{\lambda} - Y = 0$ and $D - \lambda Y = 0$

$\rightarrow$ **maximum likelihood estimator** (MLE) of λ:

$$\widehat{\lambda} = \frac{D}{Y} = \frac{\text{number of cases}}{\text{total person-time}} = \text{ empirical rate!}$$

offset term — Poisson model

- Previous model without offset: Intercept 6.784=log(884)
- We should use an offset if we suspect that the underlying **population sizes (person-years) differ** for each of the observed counts – For example varying person-years by sex,age,treatment group,...
- We need a term in the model that "scales" the likelihood, but does not depend on model parameters (include a **term with reg. coef. fixed to 1**) – offset term is log(y)
- This is all taken care of by family=poisreg – recommend to use

$$log(\tfrac{\mu}{y}) = \beta_0 + \beta_1 x_1$$
$$log(\mu) = 1 \times log(y) + \beta_0 + \beta_1 x_1$$

Comparing rates: The Thorotrast Study

- Cohort of seriously ill patients in Denmark on whom angiography of brain was performed.
- Exposure: `contrast` medium used in angiography,
  1. `thor` = thorotrast (with ^{232}Th), used 1935-50
  2. `ctrl` = other medium (?), used 1946-63
- Outcome of interest: death

  `doe` = date of **E**ntry to follow-up,
  `dox` = date of e**X**it, end of follow-up.

- `data(thoro)` in the `Epi` package.

Tabulating rates: thorotrast vs. control

Tabulating cases, person-years & rates by group

```
> stat.table( contrast,
+             list ( N = count(),
+                    D = sum(d),
+                    Y = sum(y),
+                 rate = ratio(d,y,1000) ) )
```

contrast	N	D	Y	rate
ctrl	1236	797.00	30517.56	26.12
thor	807	748.00	19243.85	38.87

Rate ratio estimation with Poisson regression

- Include `contrast` as the explanatory variable (factor).
- Insert person years in units that you want rates in

```
> m2 <- glm( cbind(d,y/1000) ~ contrast,family = poisreg(link="log") )
> round( summary(m2)$coef, 4)[, 1:2]
```

	Estimate	Std. Error
(Intercept)	3.2626	0.0354
contrast thor	0.3977	0.0509

- Rate ratio and CI?
  Call function `ci.exp()` in `Epi`

```
> round( ci.exp( m2 ), 3 )
             exp(Est.)  2.5% 97.5%
(Intercept)     26.116 24.364 27.994
contrast thor    1.488  1.347  1.644
```


Rates in groups with Poisson regression

- Include `contrast` as the explanatory variable (factor).
- Remove the intercept (-1)
- Insert person-years in units that you want rates in

```
> m3 <- glm( cbind(d,y/1000) ~ factor(contrast)-1,family = poisreg)
> round( summary(m3)$coef, 4)[, 1:2]

              Estimate Std. Error
contrast ctrl  3.2626     0.0354
contrast thor  3.6602     0.0366

> round( ci.exp( m3 ), 3 )

              exp(Est.)  2.5% 97.5%
contrast ctrl    26.116 24.364 27.994
contrast thor    38.870 36.181 41.757
```

Rate difference estimation with Poisson regression

- The approach with d/y enables additive rate models too:

```
> contrast<-c(0,1)
> m5 <-glm(cbind(d,y/1000) ~contrast,
          family=poisreg(link="identity") )
> round( ci.exp(m5,Exp=F), 3 )


             Estimate  2.5% 97.5%
(Intercept)    26.116 24.303 27.929
contrast thor  12.753 9.430 16.077
```


Binary data: Treatment success Y/N

85 diabetes-patients with foot-wounds:

- Dalterapin (Dal)
- Placebo (Pl)

Treatment/Placebo given to diabetes patients, the design is prospective and outcome is measured better(Y)/worse(N). Is the probability of outcome more than 15% – yes, then use the risk difference or risk ratio (RR)

	Treatment group	
	Dalterapin	Placebo
Better	29	20
Worse	14	22
Total	43	42

$$\hat{p}_{\text{Dal}} = \frac{29}{43} = 67\% \qquad \hat{p}_{\text{Pl}} = \frac{20}{42} = 47\%$$

Binary data: Crosstabulation analysis of 2x2 table

```
> library(Epi)
> dlt <- rbind( c(29,14), c(20,22) )
> colnames( dlt ) <- c("Better","Worse")
> rownames( dlt ) <- c("Dal","Pl")
> kable(twoby2( dlt ),"latex")
```

```
2 by 2 table analysis:
    Better Worse  P(Better) 95% conf. interval
Dal     29    14      0.6744   0.5226  0.7967
Pl      20    22      0.4762   0.3316  0.6249
                             95% conf. interval
            Relative Risk: 1.4163  0.9694  2.0692
        Sample Odds Ratio: 2.2786 0.9456  5.4907
Conditional MLE Odds Ratio: 2.2560 0.8675 6.0405
    Probability difference: 0.1982 -0.0110 0.3850

            Exact P-value: 0.0808
       Asymptotic P-value: 0.0665
```

Binary regression – estimation of odds ratio

For grouped binary data, the response is a two-column matrix with columns (successes,failures).

```
> library(Epi)
> library(xtable)
> dlt <- data.frame(rbind( c(29,14),c(20,22) ))
> colnames( dlt ) <- c("Better","Worse")
> dlt$trt <- c(1,0)
> b2<-glm(cbind(Better,Worse)~trt,
+           family=binomial(link="logit"),
+           data=dlt)
> xtable(round( ci.exp( b2 ), digits=6 ))
```

	exp(Est.)	2.5%	97.5%
(Intercept)	0.91	0.50	1.67
trt	2.28	0.95	5.49

- The default parameters in logistic regression are **odds** (the intercept: $20/22 = 0.9090$) and the **odds-ratio** ($(29/14)/(20/22) = 2.28$).
- This is **NOT** what you want, because odds ratio is biased estimate of the risk ratio.(recall if p>10% $\frac{p}{1-p} \not\approx p$)

Binary regression - Estimation of risk ratio (Relative risk)

```
> library(Epi)
> library(xtable)
> dlt <- data.frame(rbind( c(29,14),c(20,22) ))
> colnames( dlt ) <- c("Better","Worse")
> dlt$trt <- c(1,0)
> b2<-glm(cbind(Better,Worse)~trt,
+           family=binomial(link="log"),
+           data=dlt)
> xtable(round( ci.exp( b2 ), digits=6 ))
```

	exp(Est.)	2.5%	97.5%
(Intercept)	0.48	0.35	0.65
trt	1.42	0.97	2.07

Diabetics with Dalterapin treatment are 1.4 times likely to get better than those treated with placebo

Binary regression - Estimation of risk difference

```
> library(Epi)
> library(xtable)
> dlt <- data.frame(rbind( c(29,14),c(20,22) ))
> colnames( dlt ) <- c("Better","Worse")
> dlt$trt <- c(1,0)
> b2<-glm(cbind(Better,Worse)~trt,
+           family=binomial(link="identity"),
+           data=dlt)
> xtable(round( ci.exp( b2,Exp=F ), digits=6 ))
```

	Estimate	2.5%	97.5%
(Intercept)	0.48	0.33	0.63
trt	0.20	-0.01	0.40

Twenty percent more of the Diabetics with Dalterapin treatment are getting better compared to Diabetics treated with placebo

Conclusion: What did we learn?

- Rates, their ratio and difference can be analysed by Poisson regression
- In Poisson models the response can be either:
  - case indicator `d` with `offset = log(y)`, or
  - case and person-years `cbind(d,y)` with poisreg-family (Epi-package)
- Both may be fitted on either grouped data, or individual records.
- Binary outcome can be modeled with binary regression.

Everything you ever wanted to know about splines but were too afraid to ask

Martyn Plummer

University of Warwick

03 June 2023

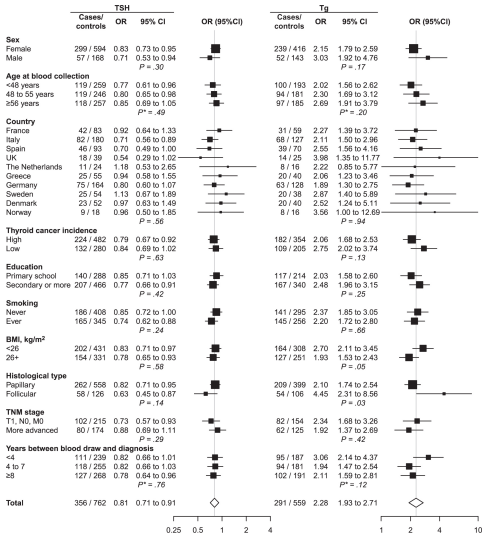
Overview

Introduction

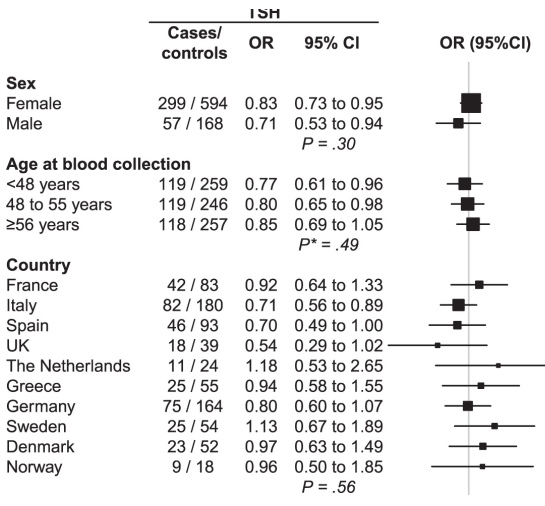
- Splines are a flexible class of models that can be helpful for representing dose-response relationships in epidemiology
- In this course we will be using spline models extensively.
- However, spline models are widely misunderstood.
- The purpose of this lecture is to give a conceptual background on where spline models come from.

Outline

Rinaldi et al, JNCI. 2014 Jun;106(6):dju097



Rinaldi et al, JNCI. 2014 Jun;106(6):dju097



Statisticians against categorization

- Greenland S (1995) Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis, Epidemiology, **6**, 450–454.
- Senn S (2005) Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials.
- Bennette C, and Vickers A, (2012), Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Medical Research Methodology 12:21

Epidemiologists against categorization

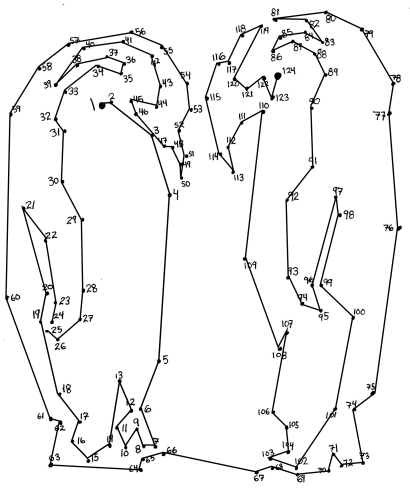
Rose, G. (1992) The Strategy of Preventive Medicine

- Many diseases are not discrete. Instead there is an underlying continuum of increasing severity (e.g. hypertension).
- In medicine, we tend to conflate a clinical action (treat vs. do not treat) with the presence/absence of disease.
- Disease prevention efforts are best targeted at shifting the distribution of risk for the whole population instead of trying to identify and target a "high risk" group.

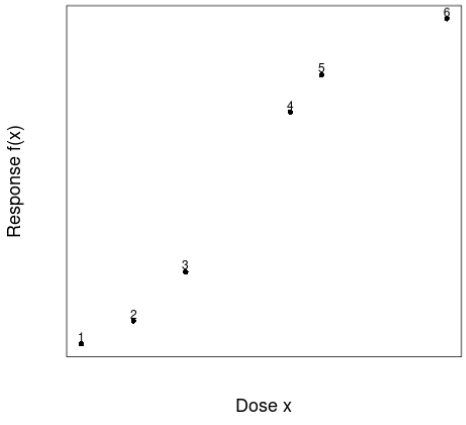
Outline

Join the dots

Join the dots

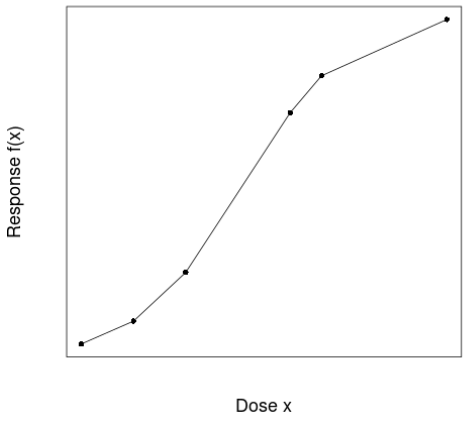


Linear interpolation



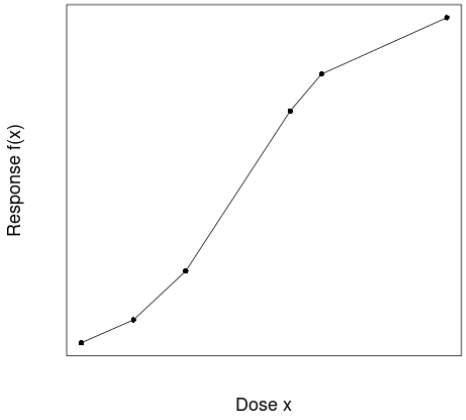
- Suppose a dose response curve is known exactly at certain points
- We can fill in the gaps (interpolate) by drawing a straight (linear) line between adjacent points
- This creates a mathematical function $f()$ which gives a response value $f(x)$ for every dose value x.

Linear interpolation



- Suppose a dose response curve is known exactly at certain points
- We can fill in the gaps (interpolate) by drawing a straight (linear) line between adjacent points
- This creates a mathematical function $f()$ which gives a response value $f(x)$ for every dose value x.

Linear interpolation



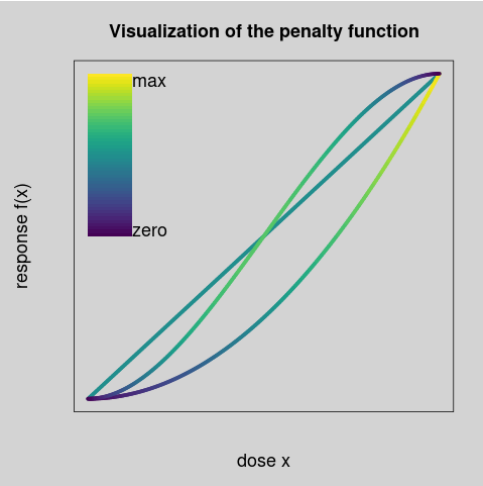
- Suppose a dose response curve is known exactly at certain points
- We can fill in the gaps (interpolate) by drawing a straight (linear) line between adjacent points
- This creates a mathematical function $f()$ which gives a response value $f(x)$ for every dose value x.

Why linear interpolation?

Out of all possible curves that go through the observed points, linear interpolation is the one that minimizes the penalty function

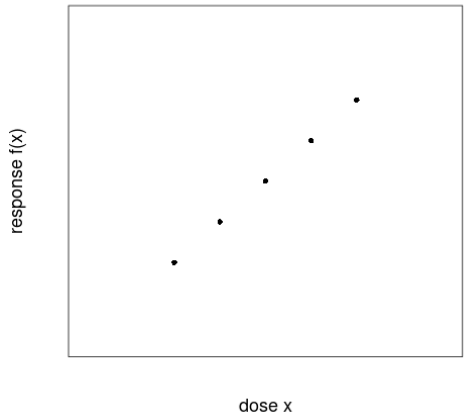
$$\int \left(\frac{\partial f}{\partial x}\right)^2 dx$$

What does the penalty mean?



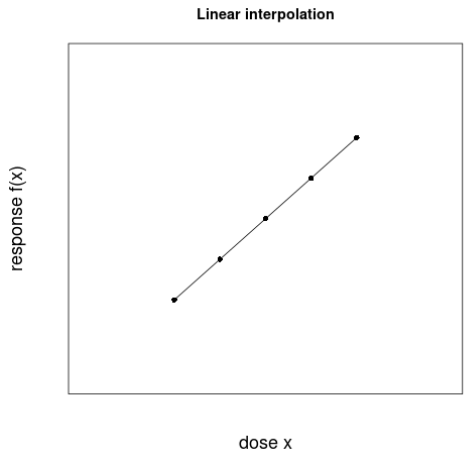
- The contribution to the penalty at each point depends on the steepness of the curve (represented by a colour gradient)
- Any deviation from a straight line between the two fixed points will incur a higher penalty overall.

Extrapolation



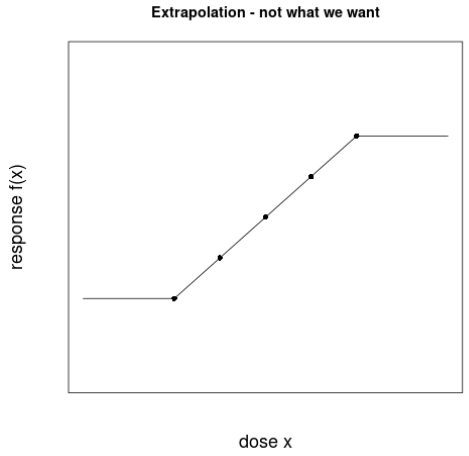
- Linear interpolation fits a linear dose-response curve exactly.
- But it breaks down when we extrapolate.

Extrapolation



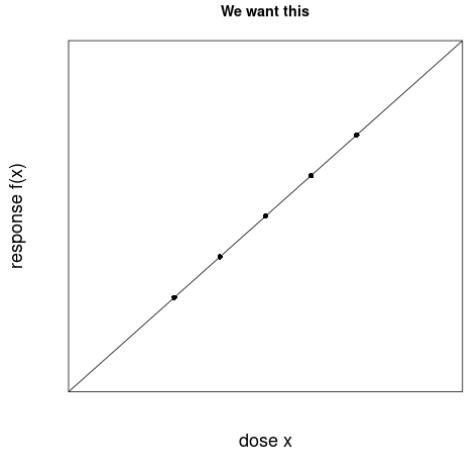
- Linear interpolation fits a linear dose-response curve exactly.
- But it breaks down when we extrapolate.

Extrapolation



- Linear interpolation fits a linear dose-response curve exactly.
- But it breaks down when we extrapolate.

Extrapolation



- Linear interpolation fits a linear dose-response curve exactly.
- But it breaks down when we extrapolate.

Why does linear interpolation break down?

- The penalty function
$$\int \left(\frac{\partial f}{\partial x}\right)^2 dx$$
penalizes the steepness of the curve
- Minimizing the penalty function gives us gives us the "flattest" curve that goes through the points.
  - In between two observations the flattest curve is a straight line.
  - Outside the range of the observations the flattest curve is completely flat.

A roughness penalty

- If we want a fitted curve that extrapolates a linear trend then we want to minimize the curvature.

$$\int \left(\frac{\partial^2 f}{\partial x^2} \right)^2 dx$$

- Like the first penalty function but uses the second derivative of f (i.e. the curvature).
- This is a roughness penalty.

Categorization and its discontents | Join the dots | Smoothing splines | Splines in R

A roughness penalty

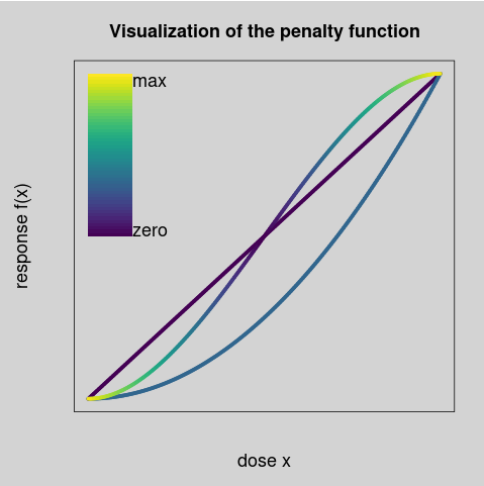
- If we want a fitted curve that extrapolates a linear trend then we want to minimize the curvature.

$$\int \left(\frac{\partial^2 f}{\partial x^2} \right)^2 dx$$

- Like the first penalty function but uses the second derivative of f (i.e. the curvature).
- This is a roughness penalty.

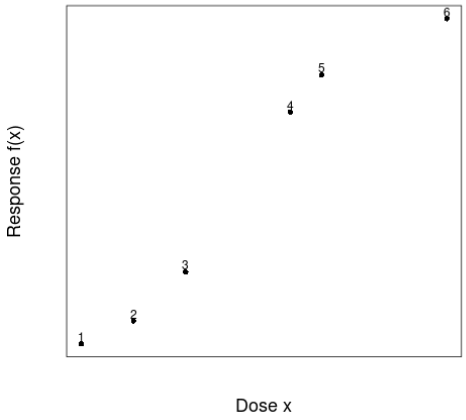
Categorization and its discontents | Join the dots | Smoothing splines | Splines in R

What does the roughness penalty mean?



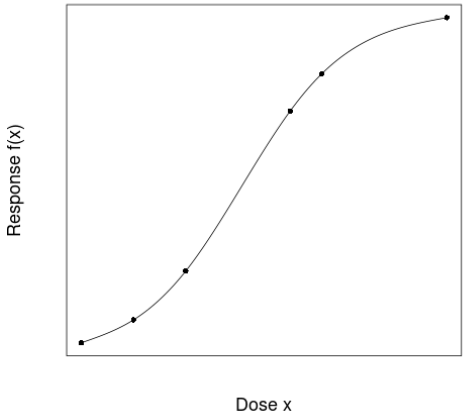
- The contribution to the penalty at each point depends on the curvature (represented by a colour gradient)
- A straight line has no curvature, hence zero penalty.
- Sharp changes in the slope are heavily penalized.

An interpolating cubic spline



- The smoothest curve that goes through the observed points is a cubic spline.

An interpolating cubic spline



- The smoothest curve that goes through the observed points is a cubic spline.

What is a cubic spline?

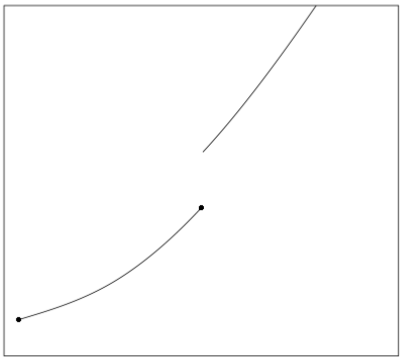
Splines are piecewise cubic curves

- Every observed point is a knot.
- The knots divide the curve into sections
- Each section is a cubic function

$$f(x) = a + bx + cx^2 + dx^3$$

- The parameters a, b, c, d are different for different sections

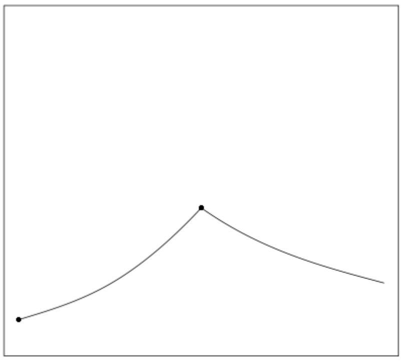
Boundary conditions



Sections need to join up smoothly.

- Both sides must go through the knot.
- The slope cannot change at a knot
- The curvature cannot change at a knot

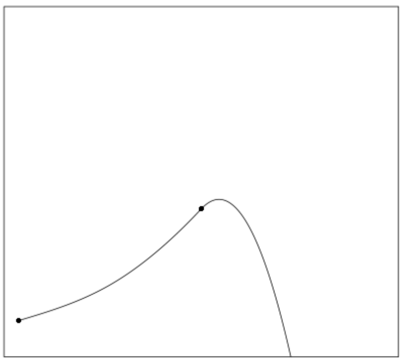
Boundary conditions



Sections need to join up smoothly.

- Both sides must go through the knot.
- The slope cannot change at a knot
- The curvature cannot change at a knot

Boundary conditions

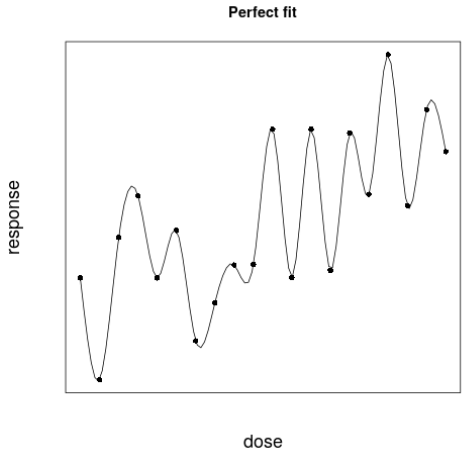


Sections need to join up smoothly.

- Both sides must go through the knot.
- The slope cannot change at a knot
- The curvature cannot change at a knot

Outline

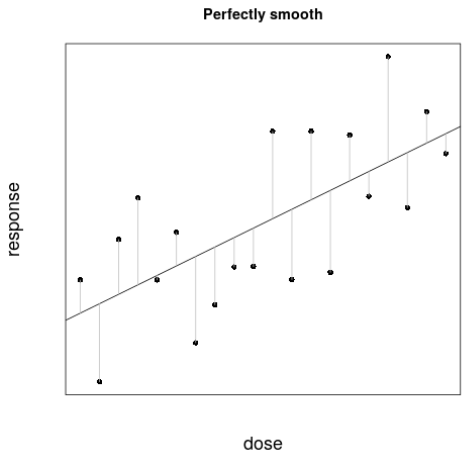
Dose response with error



In practice we never know the dose response curve exactly at any point but always measure with error. A spline model is then a compromise between

- Model fit
- Smoothness of the spline

Dose response with error



In practice we never know the dose response curve exactly at any point but always measure with error. A spline model is then a compromise between

- Model fit
- Smoothness of the spline

Fitting a smoothing spline

Minimize

$$\sum_i [y_i - f(x_i)]^2 + \lambda \int \left(\frac{\partial^2 f}{\partial x^2}\right)^2 dx$$

Or, more generally

$$\text{Deviance } + \lambda \times \text{Roughness penalty}$$

Size of tuning parameter λ determines compromise between model fit (small λ) and smoothness (large λ).

Smoothing and degrees of freedom

Software will choose the smoothing parameter λ for you automatically using cross-validation.

The smoothing parameter is adapted to the data.

Smoothness of the model can be measured with the *effective degrees of freedom* (EDF)

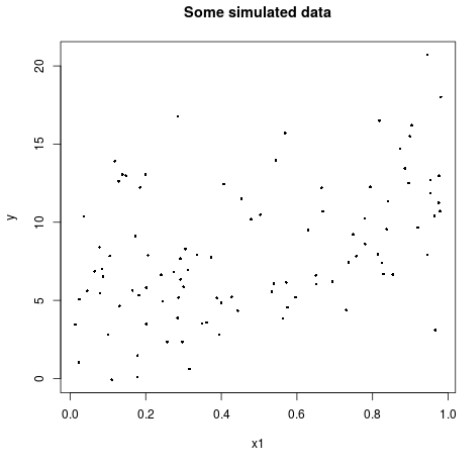
- Linear model: maximally smooth
  - EDF=2 (intercept + slope parameter)
- Intepolating mode: best fit
  - EDF=n (one parameter for every observation)

Outline

Spline models in R

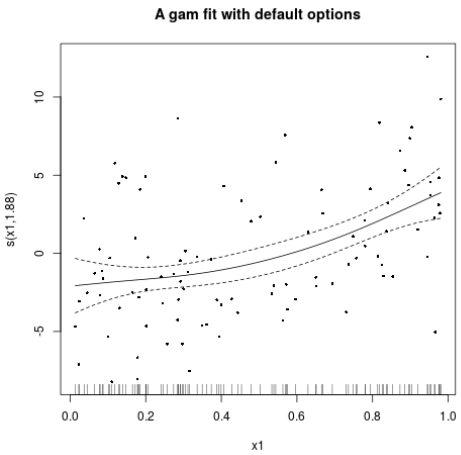
- Do not use the splines package.
- Use the `gam` function from the mgcv package to fit your spline models.
- The `gam` function chooses number and placement of knots for you and estimates the size of the tuning parameter λ automatically.
- You can use the `gam.check` function to see if you have enough knots. Also re-fit the model explicitly setting a larger number of knots (e.g. double) to see if the fit changes.

Penalized spline



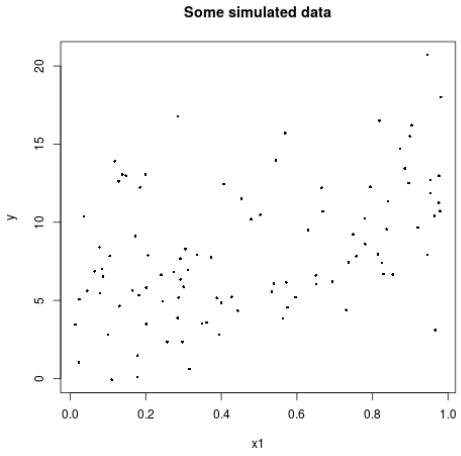
- A gam fit to some simulated data
- Model has 9 degrees of freedom
- Smoothing reduces this to 2.88 effective degrees of freedom

Penalized spline



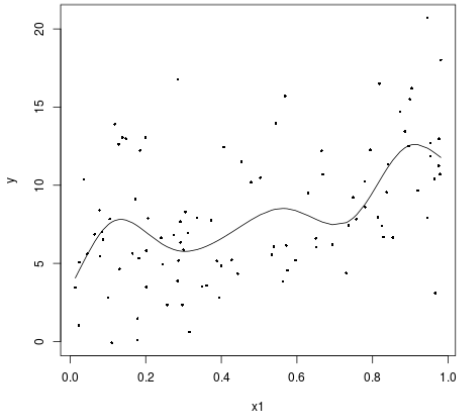
- A gam fit to some simulated data
- Model has 9 degrees of freedom
- Smoothing reduces this to 2.88 effective degrees of freedom

Unpenalized spline



- An unpenalized spline using the same spline basis as the gam fit.
- Model has 9 degrees of freedom

Unpenalized spline



- An unpenalized spline using the same spline basis as the gam fit.
- Model has 9 degrees of freedom

Conclusions

- Epidemiologists like to turn continuous variables into categories.
- Statisticians do not like categorization because it loses information.
- Splines are a flexible class of models that avoid categorization but also avoid making strong assumptions about the shape of a dose-response relationship.
- Penalized regression splines are based on compromise between goodness-of-fit and smoothness.
- Most of the decisions in fitting a penalized regression spline can be made for you
  - Degree of smoothing
  - Number of knots
  - Placement of knots

Linear and generalized linear models

Saturday 3 June, 2023

Esa Läärä

Statistical Practice in Epidemiology using **R**
2 to 7 June, 2023
University of Tartu, Estonia

Outline

- Simple linear regression.
- Fitting a regression model and extracting results.
- Predictions and diagnostics.
- Categorical factors and contrast matrices.
- Main effects and interactions.
- Modelling curved effects.
- Generalized linear models.
- Binary regression and Poisson regression.

Variables in generalized linear models

- The **outcome** or **response** variable must be numeric.
- Main types of response variables are
  - Metric or continuous (a measurement with units).
  - Binary ("yes" vs. "no", coded 1/0), or proportion.
  - Failure in person-time, or incidence rate.
- **Explanatory** variables or **regressors** can be
  - Numeric or quantitative variables
  - Categorical factors, represented by class indicators or contrast matrices.

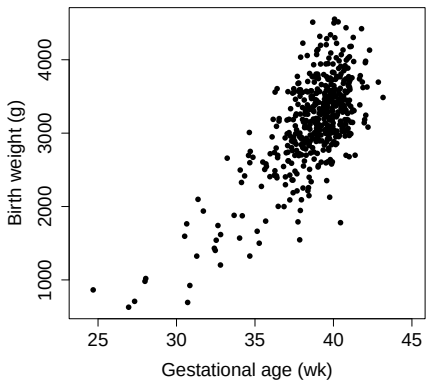
The `births` data in `Epi`

`id`:	Identity number for mother and baby.
`bweight`:	Birth weight of baby.
`lowbw`:	Indicator for birth weight less than 2500 g.
`gestwks`:	Gestation period in weeks.
`preterm`:	Indicator for gestation period less than 37 weeks.
`matage`:	Maternal age.
`hyp`:	Indicator for maternal hypertension (0 = no, 1 = yes).
`sex`:	Sex of baby (1 = male, 2 = female).

Declaring and transforming some variables as factors:

```
> library(Epi) ; data(births)
> births <- transform(births,
+    hyp = factor(hyp, labels=c("N", "H")),
+    sex = factor(sex, labels=c("M", "F")),
+    gest4 = cut(gestwks,breaks=c(20, 35, 37, 39, 45), right=FALSE)  )
> births <- subset(births, !is.na(gestwks))
```


Birth weight and gestational age




```
> with(births, plot(bweight ~ gestwks, xlim = c(24,45), pch = 16, cex.axis=1.5, cex.lab = 1.5,
+  xlab= "Gestational age (wk)", ylab= "Birth weight (g)" ) )
```


Metric response, numeric explanatory variable

Roughly linear relationship btw `bweight` and `gestwks`

→ Simple **linear regression model** fitted.

```
> m <- lm(bweight ~ gestwks, data=births)
```

- `lm()` is the function that fits linear regression models, assuming **Gaussian** distribution or **family** for **error** terms.
- `bweight ~ gestwks` is the **model formula**
- `m` is a **model object** belonging to **class** "`lm`".

`> coef(m)` – Printing the estimated regression coefficients

```
(Intercept)     gestwks
    -4489.1       197.0
```

Interpretation of **intercept** and **slope**?

Model object and extractor functions

Model object = **list** of different elements, each being separately accessible.
– See `str(m)` for the full list.

Functions that extract results from the fitted model object

- `summary(m)` – lots of output
- `coef(m)` – beta-hats only (see above)
- `ci.lin(m)[,c(1,5,6)]` – $\widehat{\beta}_j$s plus confidence limits

```
             Estimate    2.5%   97.5%
(Intercept)  -4489.1 -5157.3 -3821.0
gestwks        197.0   179.7   214.2
```

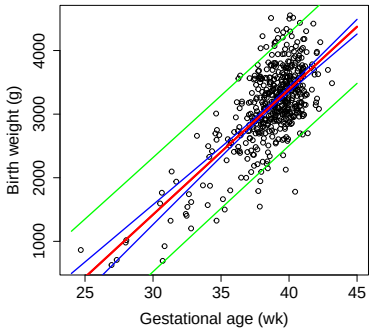
  Function `ci.lin()` is found in `Epi` package.
- `anova(m)` – Analysis of Variance Table

Other extractor functions, for example

- `fitted(m), resid(m), vcov(m), ...`
- `predict(m, newdata = ..., interval=...)`
  - Predicted responses for desired combinations of new values of the regressors – `newdata`
  - Argument `interval` specifies whether **confidence** intervals for the *mean* response or **prediction** intervals for *individual* responses are returned.
- `plot(m)` – produces various diagnostic plots based on residuals (raw, standardized or studentized residuals).

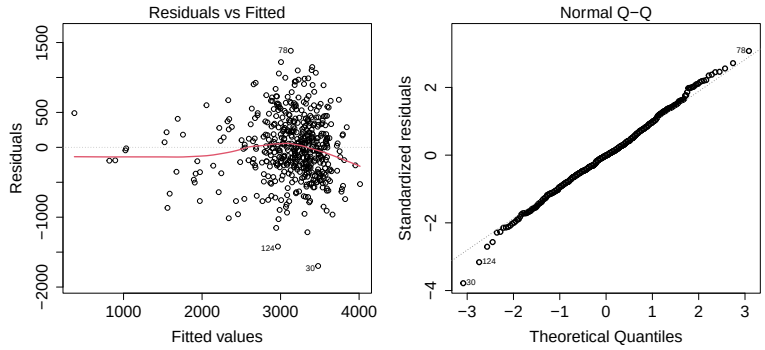
Many of these are special **methods** for certain **generic functions**, aimed at acting on objects of class "`lm`".

Fitted values, confidence & prediction intervals




```
> nd <- data.frame( gestwks = seq(24, 45, by = 0.25 ) )
> pr.c1 <- predict( m, newdata=nd, interval="conf" )
> pr.p1 <- predict( m, newdata=nd, interval="pred" )
> with(births, plot(bweight ~ gestwks, xlim = c(24,45), cex.axis=1.5, cex.lab = 1.5, xlab = 'Gestation
> matlines( nd$gestwks, pr.c1, lty=1, lwd=c(3,2,2), col=c('red','blue','blue'))
> matlines( nd$gestwks, pr.p1, lty=1, lwd=c(3,2,2), col=c('red','green','green'))
```

A couple of diagnostic plots




```
> par(mfrow=c(1,2))
> plot(m, 1:2, cex.lab = 1.5, cex.axis=1.5, cex.caption=1.5, lwd=2)
```

- Some deviation from linearity?
- Reasonable agreement with Gaussian error assumption?

Factor as an explanatory variable

- How `bweight` depends on maternal hypertension?

```
> mh <- lm( bweight ~ hyp, data=births)

             Estimate   2.5%  97.5%
(Intercept)   3198.9 3140.2 3257.6
hypH          -430.7 -585.4 -275.9
```

- Removal of intercept → mean `bweights` by `hyp`:

```
> mh2 <- lm( bweight ~ -1 + hyp, data = births)
> coef(mh2)
  hypN   hypH
3198.9 2768.2
```

- Interpretation: `-430.7 = 2768.2 - 3198.9`
  = difference between level 2 ("`H`") *vs.* reference level 1 ("`N`") of factor `hyp`.

Additive model with both gestwks and hyp

- Joint effect of `hyp` and `gestwks` is modelled e.g. by updating:

```
> mhg <- update(mh, . ~ . + gestwks)
             Estimate    2.5%   97.5%
(Intercept)  -4285.0 -4969.7 -3600.3
hypH          -143.7  -259.0   -28.4
gestwks        192.2   174.7   209.8
```

- The coefficient for `hyp`: `H` *vs.* `N` is attenuated (from -430.7 to -143.7).
- Does -143.7 estimate the **causal effect** of `hyp` **adjusted** for `gestwks`?
- No, as `gestwks` is most likely a **mediator**. – Much of the effect of `hyp` on `bweight` is mediated via shorter `gestwks` in hypertensive mothers.
- Instead, for **total causal effect** of `hyp`, adjustment for at least age is needed, but adjusting for `gestwks` is **overadjustment**.
- Yet, for **predictive modelling** it is OK to keep `gestwks`.

Model with interaction of `hyp` and `gestwks`

- `mhgi <- lm(bweight ~ hyp + gestwks + hyp:gestwks, ...)`
- Or with shorter formula: `bweight ~ hyp * gestwks`

```
              Estimate    2.5%   97.5%
(Intercept)    -3960.8 -4758.0 -3163.6
hypH           -1332.7 -2841.0   175.7
gestwks          183.9   163.5   204.4
hypH:gestwks      31.4    -8.3    71.1
```

- Estimated slope: 183.9 g/wk in reference group `N` of normotensive mothers and $183.9 + 31.4 =$ 215.3 g/wk in hypertensive mothers.

$\Leftrightarrow$ For each additional week the difference in mean `bweight` between `H` and `N` group increases by 31.4 g.

- *Interpretation of* `Intercept` *and "main effect"* `hypH`*?*

Model with interaction (cont'd)

More interpretable parametrization obtained if `gestwks` is **centered** at some reference value, using e.g. the **insulate** operator `I()` for explicit transformation of an original term.

- `mi2 <- lm(bweight ~ hyp*I(gestwks-40), ...)`

```
                     Estimate   2.5%  97.5%
(Intercept)            3395.6 3347.5 3443.7
hypH                    -77.3 -219.8   65.3
I(gestwks - 40)         183.9  163.5  204.4
hypH:I(gestwks - 40)     31.4   -8.3   71.1
```

- The "main effect" of `hyp` $= -77.3$ is the difference between `H` and `N` at the reference value `gestwks` $= 40$.
- `Intercept` $=$ 3395.6 is the estimated mean `bweight` at the reference value 40 of `gestwks` in group `N`.

Factors and contrasts in R

- A categorical explanatory variable or **factor** with L **levels** will be represented by $L - 1$ linearly independent columns in the **model matrix** of a linear model.
- These columns can be defined in various ways implying alternative **parametrizations** for the effect of the factor.
- Parametrization is defined by given type of **contrasts**.
- Default: **treatment** contrasts, in which 1st class is the **reference**, and regression coefficient β_k for class k is interpreted as $\beta_k = \mu_k - \mu_1$
- Own parametrization may be tailored by function `C()`, with the pertinent **contrast matrix** as argument.
- Or, use `ci.lin(mod, ctr.mat = CM)` after fitting.

Two factors: additive effects

- Factor X has 3 levels, Z has 2 levels – Model:

$$\mu = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 Z_1 + \gamma_2 Z_2$$

- X_1 (reference), X_2, X_3 are the indicators for X,
- Z_1 (reference), Z_2 are the indicators for Z.
- Omitting X_1 and Z_1 the model for mean is:

$$\mu = \alpha + \beta_2 X_2 + \beta_3 X_3 + \gamma_2 Z_2$$

with predicted means μ_{jk} $(j = 1, 2, 3; k = 1, 2)$:

		$Z = 1$	$Z = 2$
	1	$\mu_{11} = \alpha$	$\mu_{11} = \alpha + \gamma_2$
X	2	$\mu_{21} = \alpha + \beta_2$	$\mu_{22} = \alpha + \beta_2 + \gamma_2$
	3	$\mu_{31} = \alpha + \beta_3$	$\mu_{32} = \alpha + \beta_3 + \gamma_2$

Two factors with interaction

- Effect of Z differs at different levels of X:

		$Z = 1$	$Z = 2$
	1	$\mu_{11} = \alpha$	$\mu_{12} = \alpha + \gamma_2$
X	2	$\mu_{21} = \alpha + \beta_2$	$\mu_{22} = \alpha + \beta_2 + \gamma_2 + \delta_{22}$
	3	$\mu_{31} = \alpha + \beta_3$	$\mu_{32} = \alpha + \beta_3 + \gamma_2 + \delta_{32}$

- How much the effect of Z (level 2 *vs.* 1)
  changes when the level of X is changed from 1 to 3:

$$\begin{aligned}\delta_{32} &= (\mu_{32} - \mu_{31}) - (\mu_{12} - \mu_{11}) \\ &= (\mu_{32} - \mu_{12}) - (\mu_{31} - \mu_{11}),\end{aligned}$$

  $=$ how much the effect of X (level 3 *vs.* 1)
  changes when the level of Z is changed from 1 to 2.

- See the exercise: interaction of `hyp` and `gest4`.

Contrasts in R

- All contrasts can be implemented by supplying a suitable **contrast function** giving the **contrast matrix** e.g:

```
> contr.cum(3)        > contr.sum(3)
1 0 0                  1    1   0
2 1 0                  2    0   1
3 1 1                  3   -1  -1
```

- In model formula factor name `faktori` can be replaced by expression like `C(faktori, contr.cum)`.
- Function `ci.lin()` can calculate CI's for linear functions of the parameters of a fitted model `mall` when supplied by a relevant contrast matrix
  `> ci.lin(mall, ctr.mat = CM)[ , c(1,5,6)]`

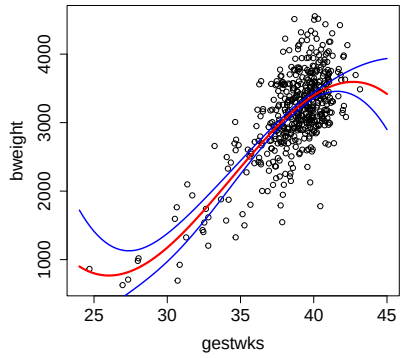
  $\rightarrow$ No need to specify contrasts in model formula!

More about numeric regressors

What if dependence of Y on X is non-linear?

- **Categorize** the values of X into a factor.
  - Continuous effects violently discretized by often arbitrary cutpoints. This is inefficient.
- Fit a low-degree (e.g. 2 to 4) **polynomial** of X.
  - Tail behaviour may be problematic.
- Use **fractional polynomials**.
  - Invariance problems. Only useful if $X = 0$ is well-defined.
- Use a **spline** model: smooth function $s(X; \beta)$. – See Martyn's lecture
  - More flexible models that act locally.
  - Effect of X reported by graphing $\widehat{s}(X; \beta)$ & its CI

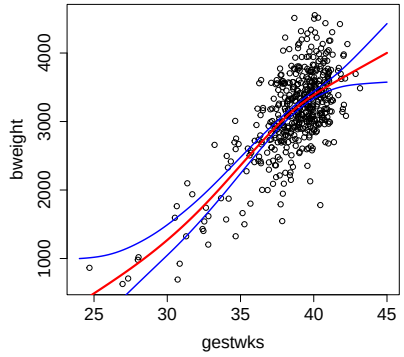
Mean `bweigth` as 3rd order polynomial of `gestwks`




```
> mp3 <- update( m, . ~ . - gestwks + poly(gestwks, 3) )
```

- The model is linear in parameters with 4 terms & 4 df.
- Otherwise good, but the tails do not behave well.

Penalized spline model with cross-validation




```
> library(mgcv)
> mpen <- gam( bweight ~ s(gestwks), data = births)
```

- Looks quite nice.
- Model df ≈ 4.2; close to 4, as in the 3rd degree polynomial model.

From linear to generalized linear models

- An alternative way of fitting our 1st Gaussian model:

```
> m <- glm(bweight ~ gestwks, family=gaussian, data=births)
```

- Function `glm()` fits **generalized linear models** (GLM).
- Requires specification of the
  - **family** – i.e. the assumed "error" distribution for Y_is,
  - **link** function – a transformation of the expected Y_i.
- Covers common models for other types of response variables and distributions, too, e.g. **logistic** regression for **binary** responses and **Poisson** regression for counts.
- Fitting: method of **maximum likelihood**.
- Many extractor functions for a `glm` object similar to those for an `lm` object.

Generalized linear models

Modelling how expected values, risks, rates, etc. depend on explanatory variables or regressors $X = (X_1, \ldots, X_p)$. – Common elements:

- Each subject i $(i = 1, \ldots, N)$ has an own **regressor profile**, i.e. vector $x_i^{\mathsf{T}} = (x_{i1}, \ldots, x_{ip})$ of values of X.
- Let vector $\beta^{\mathsf{T}} = (\beta_0, \beta_1, \ldots, \beta_p)$ contain regression coefficients. The **linear predictor** is a linear combination of β_js and x_{ij}s:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

- Some X_js can be **product terms** for interactions and modifications if needed, and **splines** may be used for continuous covariates.
- Further model specification depends on the type of outcome variable, assumed error distribution or family, desired interpretation of coefficients, and importance and choice of time scale(s).

Binary regression and interpretations of coefficients

- Basic model for risks $\pi(x_i) = P\{Y_i = 1 | X = x_i\} = E(Y_i | X = x_i)$ with fixed risk period, complete follow-up (no censoring, nor competing events):

$$g\{\pi(x_i)\} = \beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip}, \quad i = 1, \ldots, N.$$

- **Link** $g(\cdot)$ and interpretation of β_js, assuming the validity of model (including homogeneity or non-modification of the coefficent in question):
  - id $\Rightarrow$ β_j = adjusted **risk difference** (RD) for $X_j = 1$ vs. $X_j = 0$,
  - log $\Rightarrow$ β_j = adjusted log of **risk ratio** (RR) – " –
  - logit $\Rightarrow$ β_j = adjusted log of **odds ratio** (OR), – " –
- Fitting: `glm(..., family=binomial(link=...), ...)`
- Issues with id & log links in keeping predicted $\widehat{\pi}(\cdot)$ between 0 and 1.
  – A solution for RR: Doubling the cases & logit-link! (Ning et al. 2022).
  – A solution for RD exists, too (Battey et al. 2019).

Poisson regression – model for rates

- A common outcome variable is a pair (D, Y) = (no. of cases, person-time), from which the **incidence rate** $= D/Y$ (see Janne's lecture on Friday).
- **Poisson regression model** specifies, how theoretical rates or **hazards** $\lambda(x_i)$ are assumed to depend on values of X.
- Some components of X represent the relevant **time scales** (as in the exercise of today; more details in Bendix's lecture on Monday).
- Linear predictor as above – **Link** $g(\cdot)$ and interpretation of β_js:
  - id $\Rightarrow \beta_j =$ adjusted **rate difference** (RD) for $X_j = 1$ vs. $X_j = 0$,
  - log $\Rightarrow \beta_j =$ adjusted log of **rate ratio** (RR) – " –
- Fitting – our recommended approach using `Epi`:

```
glm(cbind(d,y) ~ ..., family=poisreg(link=...),...)
```


What was covered

- A wide range of models from simple linear regression to splines.
- Gaussian family for continuous outcomes, binomial for binary outcomes, and Poisson family for rates.
- Various link functions for different parametrizations.
- R functions fitting linear and generalized models: `lm()` and `glm()`.
- Parametrization of categorical explanatory factors; contrast matrices.
- Extracting results and predictions: `ci.lin()`, `fitted()`, `predict()`.
- Model diagnostics: `resid()`, `plot.lm()`, ...

Introduction to causal inference

Krista Fischer

Institute of Mathematics and Statistics, University of Tartu
Institute of Genomics, University of Tartu
Estonian Academy of Sciences

Statistical Practice in Epidemiology, Tartu 2023

Statistical associations vs causal effects in epidemiology

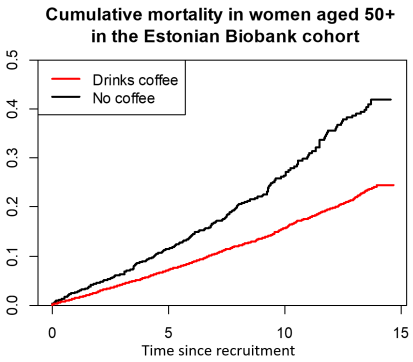
Does the exposure (smoking level, obesity, etc) have a causal effect on the outcome (blood pressure, cancer diagnosis, mortality, etc)?

is not the same question as

Is the exposure associated with the outcome?

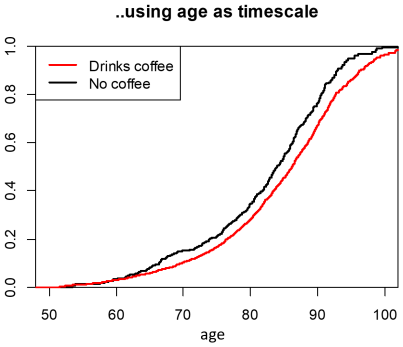
Conventional statistical analysis will answer the second one, but not necessarily the first.

Example



Does coffee-drinking prolong life?
(so drastically???)

Example (cont.)



Does coffee-drinking prolong life?
Or: do coffee-drinkers live longer (for several reasons)?

How to define causal effects (properly)?

- One can think of some basic guidelines (sometimes called as “criteria”) that must be satisfied for causal effect to be identifiable.
- Such principles may include temporality (cause preceding the outcome), consistency (reproducibility), monotonicity (dose-response), plausibility (e.g. biologically), etc. (Bradford Hill’s guidelines)
- However, although such general guidelines are useful, they are often not sufficient to establish causality

Causal effects and counterfactuals

- To define causal effects more properly, counterfactual (what-if) thinking is useful.
- Mathematically, the individual causal effect can be defined as the difference

$$Y^1 - Y^0,$$

where $Y^1 = Y(X = 1)$ and $Y^0 = Y(X = 0)$ are defined as individual's potential (counterfactual) outcomes if this individual's exposure level X were set to 1 or 0, respectively.

- Example: Y^1 individual's blood pressure, if he/she were a smoker; Y^0 individual's blood pressure, if he/she were a nonsmoker;
- For a particular individual, either Y^1 or Y^0 can be observed at any moment.

The "naïve" association analysis

- With a binary exposure X, compare average outcomes in exposed and unexposed populations:

$$E(Y|X = 1) - E(Y|X = 0)$$

Is cancer incidence different in smokers and nonsmokers?

- But mostly:

$$E(Y|X = 1) \neq E(Y^1)$$

Cancer risk in smokers is not the same as the potential cancer risk in the population if everyone were smoking

- Similarly:

$$E(Y|X = 0) \neq E(Y^0)$$

- In most cases there is always some unobserved confounding present and therefore the naïve analysis does not provide causal effect estimates.

Potential outcomes (counterfactuals) in different settings

- Randomized trials: probably the easiest setting to imagine Y^X for different X.
- "Actionable" exposures: smoking level, vegetable consumption, ... – potential interventions may alter exposure levels in future.
- Non-actionable exposures: e.g genotypes. It is difficult to ask *"What if I had different genes?"*. Still useful concept to formalize genetic effects (heritability, attributable risk).
- Combinations: With X– a behavioral intervention level, Z–smoking level and Y–a disease outcome, one could formalize the effect of intervention on outcome by using $Y^{X,Z(X)}$

A causal model in terms of potential outcomes

- More generally Y^x is defined as the potential outcome following the exposure level $X = x$
- A linear causal model can be specified as

$$Y_i^x - Y_i^0 = x\beta_1 + \varepsilon_i, \quad \text{with } E(\varepsilon_i|x) = 0$$

- Note that the observed outcome $Y_i = Y_i^x$ for individuals with $X_i = x$.
- The model could be generalized to include nonlinear terms or interactions with other covariates, or as a generalized linear model (logistic regression, survival model).
- However, as we don't observe Y^0 and Y^x (with $x > 0$) for the same individuals at the same time, thus it is not straightforward to actually fit the model on data.

Statistical model vs causal model

- More generally Y^x is defined as the potential outcome following the exposure level $X = x$
- A linear causal model can be specified as

$$Y_i^x - Y_i^0 = x\beta_1 + \varepsilon_i, \quad \text{with } E(\varepsilon_i|x) = 0$$

- Note that the observed outcome $Y_i = Y_i^x$ for individuals with $X_i = x$.
- A classical linear regression model:

$$Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i, \quad \text{with } E(\varepsilon_i|X_i) = 0$$

or

$$E(Y_i|X_i) = \beta_0 + X_i\beta_1.$$

- When are the two equivalent?

Statistical model vs causal model

- Rewrite the linear causal model as

$$Y_i^x = Y_i^0 + x\beta_1 + \varepsilon_i, \quad \text{with } E(\varepsilon_i|x) = 0$$

- Note that this would be equivalent with the classical linear model, if

$$E(Y_i^0 + \varepsilon_i|X_i) = \beta_0,$$

thus when the potential exposure-free outcome Y^0 is not associated with the exposure X

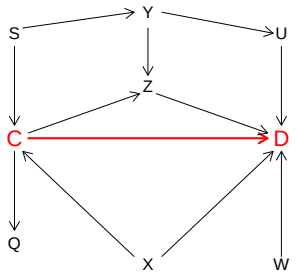
- For instance, this would mean that in the absence of smoking, the cancer risk for current smokers and current nonsmokers would be the same ($E(Y|X = 0) = E(Y^0)$).
- In other words, the two models are equivalent in the absence of confounding.

Classical/generalized regression estimates vs causal effects?

- In the presence of confounding, regression analysis provides a biased estimate for the true causal effect
- To reduce such bias, one needs to collect data on most important confounders and adjust for them
- However, too much adjustment may actually introduce more biases
- Causal graphs (Directed Acyclic Graphs, DAGs) may be extremly helpful in identifying the optimal set of adjustment variables

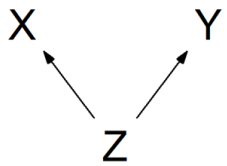
DAGs: directed acyclic graphs

- A Directed Acyclic Graph (DAG) is a graphical representation of the causal association structure in the data, where variables are presented as nodes (points) and the associations are presented as edges (lines, arrows);
- Thus an arrow pointing from variable X to a variable Y on such graph represents a causal effect of X on Y.



"Classical" confounding

Third factors Z influence both, X and Y

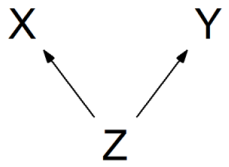


Also called as backdoor path between X and Y.
Implied statistical associations (Y is not independent of X in general, but it is independent of X, conditional on Z):

$$X \not\perp\!\!\!\perp Y \qquad X \perp\!\!\!\perp Y|Z$$

X and Y are independent, conditional on Z, but marginally dependent.

"Classical" confounding, mathematically



Now: $\mathrm{E}(Y|X) = b_{0y} + b_{zy}\mathrm{E}(Z|X)$.

If $b_{zx} \neq 0$, then also $r_{zx} \neq 0$ and so

$$\mathrm{E}(Z|X) = b_{0z} + b_{xz}X, \text{ where } b_{xz} \neq 0$$

. We see that:

$$\mathrm{E}(Y|X) = b^*_{0y} + b_{xz}b_{zy}X.$$

Assume:

$X = b_{0x} + b_{zx}Z + \varepsilon_x$, $\mathrm{E}(\varepsilon_x|Z) = 0$

$Y = b_{0y} + b_{zy}Z + \varepsilon_y$, $\mathrm{E}(\varepsilon_y|Z, X) = 0$.

One should adjust the analysis for Z, by fitting a regression model for Y with covariates X and Z. There is a causal effect between X and Y, if the effect of X is present in such model.

Example: COVID vaccination and Simpson's paradox

Suppose there are COVID infections in:

- 3000 unvaccinated individuals, 90 needing hospitalizations
- 1000 vaccinated individuals, 30 needing hospitalizations

No effect of vaccination?

More detailed data:

age	vaccination	total	hospitalized	% hospitalized
≥ 60	no	100	24	24%
	yes	300	24	8%
< 60	no	2900	66	2.3%
	yes	700	6	0.9%
all ages	no	3000	90	3%
	yes	1000	30	3%

Age is a confounder here!

COVID vaccination and Simpson's paradox

Real data from Estonia (August 2021):

age	vaccination	total	hospitalized	% hospitalized
≥ 60	no	186	50	26.9%
	yes	202	16	7.9%
< 60	no	3075	57	1.9%
	yes	666	5	0.8%
all ages	no	3261	107	3.3%
	yes	868	21	2.4%

Causal chain (mediation, front-door path):

The effect of X on Y is mediated by Z:

$$X \longrightarrow Z \longrightarrow Y$$

$Y = \beta_0 + \beta_{xy}X + \beta_{zy}Z + \varepsilon,$

- Don't adjust for Z, if you are interested in the total effect of X on Y
- Do adjust for Z, if you are interested in the direct effect of X on Y
- Adjusted analysis is valid only when the Z-Y association is unconfounded!

The case of a collider: adjustment is sometimes wrong!

X and Y have an effect on Z:

$$X \longrightarrow Z \longleftarrow Y$$

$Z = \beta_0 + \beta_{xz}X + \beta_{yz}Y + \varepsilon$, with $\beta_{xz} \neq 0$ and $\beta_{yz} \neq 0$

hence, there exist parameters $\beta_{xy} \neq 0$ and $\beta_{zy} \neq 0$, so that:
$Y = \beta_0^* + \beta_{xy}X + \beta_{zy}Z + \varepsilon^*$.

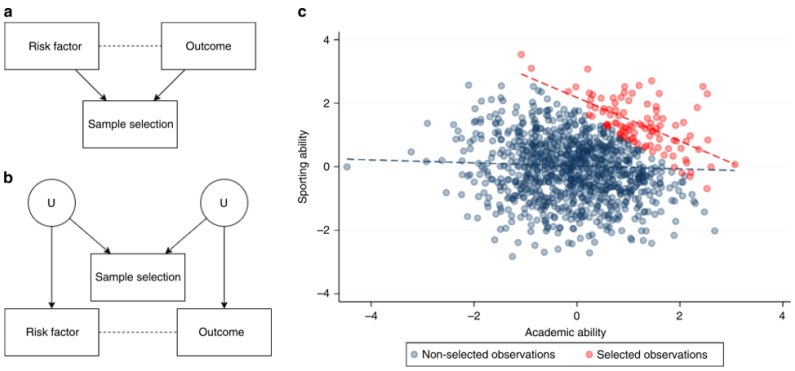
$$X \perp\!\!\!\perp Y \qquad X \not\perp\!\!\!\perp Y|Z$$

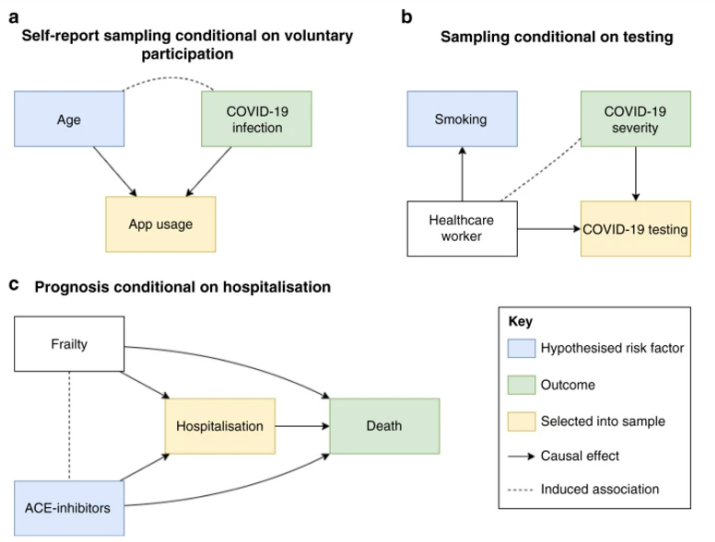
We see the association between X and Y only when the "effect" of Z has been taken into account.
But this is NOT a causal effect of X on Y.

One should NOT adjust the analysis for Z!

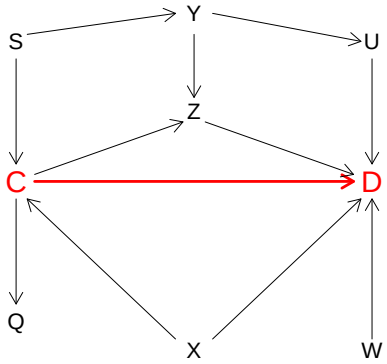
Selection bias: a special (but common) case of collider bias

- All analysis are done conditional on the selected sample
- However, selection itself might be a collider (Griffith et al. 2020, https://www.nature.com/articles/s41467-020-19478-2)





Actually there might be a complicated system of causal effects:



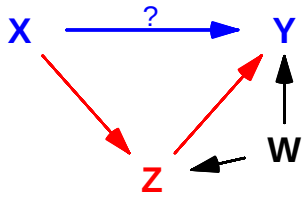
C-smoking; D-cancer
Q, S, U, W, X, Y, Z - other factors that influence cancer risks and/or smoking (genes, social background, nutrition, environment, personality, . . .)

What to do in complicated cases?

1. Sketch a causal graph
2. Identify all paths between the exposure and outcome (ways to go from X to Y regardless of the direction of the arrows).
3. Identify the closed paths that include colliders and open paths that don't.
4. You need to select adjustment variables that block all open paths.
5. Don't adjust for colliders (as they would open the closed paths)!
6. If you are looking for the total effects, you don't need to block the directed paths (that follow the directions of the arrows).
7. Often, there are unobserved confounders!

R package *dagitty* is useful for such tasks.

Example: mediation with confounding

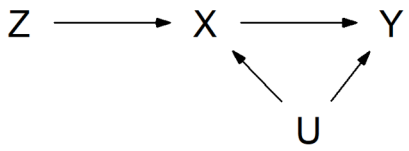


Paths: $X \to Z \to Y$ (open) and $X \to Z \leftarrow W \to Y$ (closed).

- The total effect of X on Y is estimable without any adjustment.
- For direct effect you need to adjust for Z, but that would open the closed path – to block that, you also need to adjust for W.
- If W is an unobserved confounder, direct effect of X on Y cannot be estimated.

Instrumental variables estimation: the idea

A DAG with the exposure X, outcome Y, confounder U and an instrument Z:



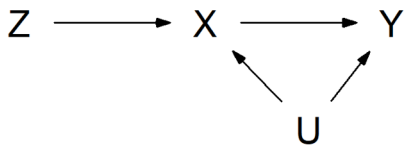
Assuming:
$Y = \alpha_y + \beta X + \gamma U + \epsilon,\ \mathrm{E}(\epsilon|X, U) = 0,$

simple regression will estimate:

$\mathrm{E}(Y|X) = \alpha_y + \beta X + \gamma \mathrm{E}(U|X).$

Thus the coefficient of X will be a biased estimate of β (as it also depends on γ).

Instrumental variables estimation: the idea



A variable Z is an instrument for the path $X \to Y$, if:

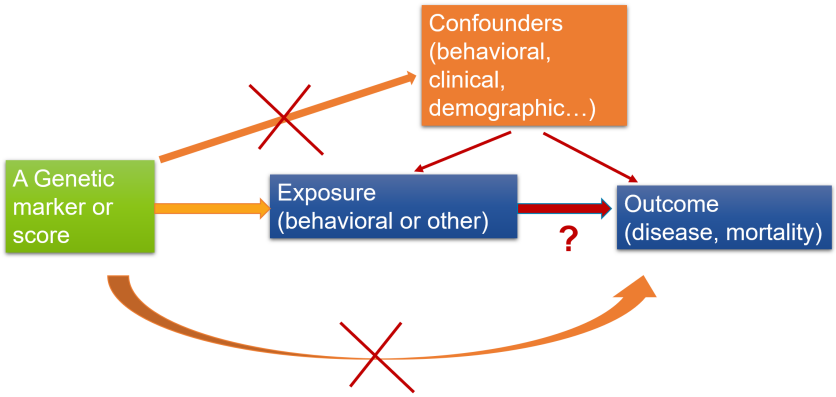
1. Z has a direct causal effect on X
2. Z does not have any direct or indirect causal effect on Y or the confounders U.

- It can be shown that the causal effect of X on Y equals:

$$\beta = \frac{cov(Z, Y)}{cov(Z, X)} = \frac{\beta_{ZY}}{\beta_{ZX}},$$

  where β_{ZY} and β_{ZX} are the coefficients of Z in a simple linear regression models for Y and X (with covariate Z).
- Replacing β_{ZY} and β_{ZX} by their estimates, we get the instrumental variables (IV) estimate of β.

Example
Mendelian randomisation

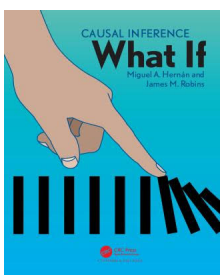


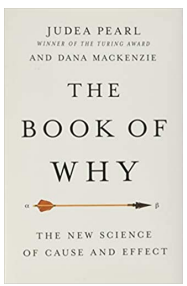
Summary

- There is no unique definition of “the causal effect”
- The validity of any causal effect estimates depends on the validity of the underlying assumptions.
- Adjustment for other available variables may remove (some) confounding, but it may also create more confounding. Do not adjust for variables that may themselves be affected by the outcome.
- Instrumental variables approaches can be helpful, but beware of assumptions!

Some references

- A webpage and a free online book by Miguel Hernan and Jamie Robins: http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
- Judea Pearl, “The Book of Why”
- Mendelian randomization: Sheehan, N., Didelez, V., et al., Mendelian Randomization and Causal Inference in Observational Epidemiology, PLoS Med. 2008; papers by G.D. Smith, J. Bowden, S. Burgess and others.





More Advanced Graphics in R

Martyn Plummer

University of Warwick, UK

SPE 2023, Tartu

Outline

Graphics Systems in R

R has several different graphics systems:

- Base graphics (the `graphics` package)
- Lattice graphics (the `lattice` package)
- Grid graphics (the `grid` package)
- Grammar of graphics (the `ggplot2` package)

Why so many? Which one to use?

Base Graphics

- The oldest graphics system in R.
- Based on S graphics (Becker, Chambers and Wilks, *The New S Language, 1988*)
- Implemented in the base package `graphics`
  - Loaded automatically so always available
- Ink on paper model; once something is drawn "the ink is dry" and it cannot be erased or modified.

Grid Graphics

- A complete rewrite of the graphics system of R, independent of base graphics.
- Programming with graphics:
  - Grid graphics commands create graphical objects (Grobs)
  - Printing a Grob displays it on a graphics device
  - Functions can act on grobs to modify or combine them
- Implemented in the base package `grid`, and extended by CRAN packages `gridExtra`, `gridDebug`, ...
- Described by the package author Paul Murrell in the book *R Graphics (2nd edition)*, 2011.

Grammar of Graphics

- Originally described by Leland Wilkinson in the book *The Grammar of Graphics*, 1999 and implemented in the statistical software nViZn (part of SPSS)
- Statistical graphics, like natural languages, can be broken down into components that must be combined according to certain rules.
- Provides a *pattern language* for graphics:
  - geometries, statistics, scales, coordinate systems, aesthetics, themes, ...
- Implemented in R in the CRAN package `ggplot2`
- Described more fully by the `ggplot2` package author Hadley Wickham in the book *ggplot2: Elegant Graphics for Data Analysis*, 2009.

Putting It All Together

- Base graphics are the default, and are used almost exclusively in this course
- `grid` provides alternate low-level graphics functions.
  - Experts only
- `ggplot2` is an alternate, high-level graphics package built on `grid`.
- All graphics packages take time to learn...

Graphics Devices

Graphics devices are used by all graphics systems.

- Plotting commands will draw on the current *graphics device*
- The *default* graphics device is a window on your screen:
  In RStudio RStudioGD()
  On Windows windows()
  On Unix/Linux x11()
  On Mac OS X quartz()
  It normally opens up automatically when you need it.
- You can have several graphics devices open at the same time (but only one is current)

Graphics Device in RStudio

RStudio has its own graphics device RStudioGD built into the graphical user interface

- You can see the contents in a temporary, larger window by clicking the zoom button.
- You can write the contents directly to a file with the export menu
- Sometimes the small size of the RStudioGD device causes problems. Open up a new device calling `RStudioGD()`. This will appear in its own window, free from the GUI.

Writing Graphs to Files

There are also non-interactive graphics devices that write to a file instead of the screen.

`pdf` produces Portable Document Format files

`win.metafile` produces Windows metafiles that can be included in Microsoft Office documents (windows only)

`postscript` produces postscript files

`png, bmp, jpeg` all produce bitmap graphics files

- Turn off a graphics device with `dev.off()`. Particularly important for non-interactive devices.
- Plots may look different in different devices

Types of Plotting Functions

- High level
  - Create a new page of plots with reasonable default appearance.
- Low level
  - Draw elements of a plot on an existing page:
    - Draw title, subtitle, axes, legend . . .
    - Add points, lines, text, math expressions . . .
- Interactive
  - Querying mouse position (`locator`), highlighting points (`identify`)

Base x-y Plots

- The `plot` function with one or two numeric arguments
- Scatterplot or line plot (or both) depending on `type` argument: `"l"` for lines, `"p"` for points (the default), `"b"` for both, plus quite a few more
- Also: formula interface, `plot(y~x)`, with arguments similar to the modeling functions like `lm`

Customizing Plots in Base

- Most plotting functions take optional parameters to change the appearance of the plot
  - e.g., `xlab, ylab` to add informative axis labels
- Most of these parameters can be supplied to the `par()` function, which changes the default behaviour of subsequent plotting functions
- Look them up via `help(par)`! Here are some of the more commonly used:
  - Point and line characteristics: `pch, col, lty, lwd`
  - Multiframe layout: `mfrow, mfcol`
  - Axes: `xlim, ylim, xaxt, yaxt, log`

Adding to Plots in Base

- `title()` add a title above the plot
- `points()`, `lines()` adds points and (poly-)lines
- `text()` text strings at given coordinates
- `abline()` line given by coefficients (*a* and *b*) or by fitted linear model
- `axis()` adds an axis to one edge of the plot region. Allows some options not otherwise available.

Strategy for Customization of Base Graphics

- Start with default plots
- Modify parameters (using `par()` settings or plotting arguments)
- Add more graphics elements. Notice that there are graphics parameters that turn things *off*, e.g. `plot(x, y, xaxt="n")` so that you can add completely customized axes with the `axis` function.
- Put all your plotting commands in a script or inside a function so you can start again

Demo 1

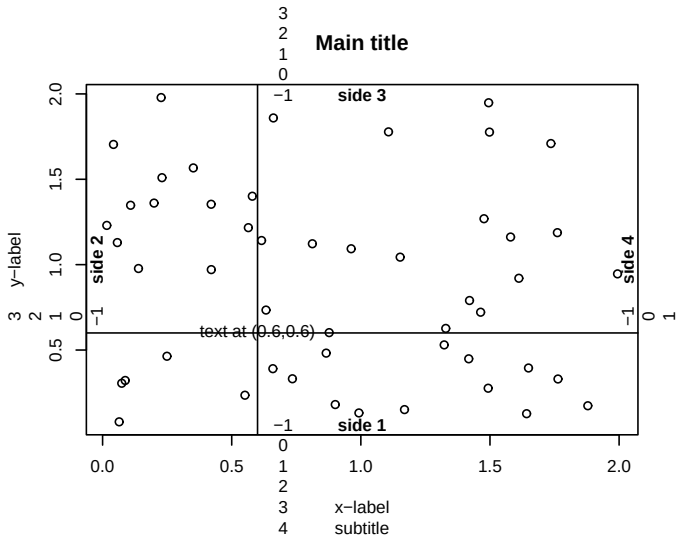
```
library(ISwR)
par(mfrow=c(2,2))
matplot(intake)
matplot(t(intake))
matplot(t(intake),type="b")
matplot(t(intake),type="b",pch=1:11,col="black",
        lty="solid", xaxt="n")
axis(1,at=1:2,labels=names(intake))
```


Margins

- R sometimes seems to leave too much empty space around plots (especially in multi-frame layouts).
- There is a good reason for it: You might want to put something there (titles, axes).
- This is controlled by the `mar` parameter. By default, it is `c(5,4,4,2)+0.1`
  - The units are *lines of text*, so depend on the setting of `pointsize` and `cex`
  - The sides are indexed in clockwise order, starting at the bottom (1=bottom, 2=left, 3=top, 4=right)
- The `mtext` function is designed to write in the margins of the plot
- There is also an *outer margin* settable via the `oma` parameter. Useful for adding overall titles etc. to multiframe plots

Demo 2

```
x <- runif(50,0,2)
y <- runif(50,0,2)
plot(x, y, main="Main title", sub="subtitle",
     xlab="x-label", ylab="y-label")
text(0.6,0.6,"text at (0.6,0.6)")
abline(h=.6,v=.6)
for (side in 1:4)
   mtext(-1:4,side=side,at=.7,line=-1:4)
mtext(paste("side",1:4), side=1:4, line=-1,font=2)
```



A Few Words on Grid Graphics

- Experts only, but . . .
- Recall that `ggplot2` uses `grid`
- The key concepts you need are *grobs* and *viewports*

Grobs: Graphical Objects

- Grobs are created by plotting functions in `grid`, and `ggplot2`
- Grobs are only displayed when they are printed
- Grobs can be modified or combined before being displayed
- The `ggplot2` package uses the `+` operator to combine grobs representing different elements of the plot

Viewports

- The plotting region is divided into viewports
- Grobs are displayed inside a viewport
  - Viewports can be different sizes (inches, centimetres, lines of text, or relative units)
  - Each viewport may have its own coordinate systems

Survival analysis with competing risks

Janne Pitkäniemi

Finnish Cancer Registry
Tampere University
Statistical Practice in Epidemiology (2023, Tartu)

Points to be covered

1. Survival or time to event data & censoring.
2. Competing risks: event-specific cumulative incidences & hazards.
3. Kaplan–Meier and Aalen–Johansen estimators.
4. Regression modelling of hazards: Cox model.
5. Packages `survival, mstate, Epi,(cmprisk)`.
6. Functions `Surv(), survfit(), plot.survfit(), coxph()`.

Survival time – time to event

Time spent (`lex.dur`) in a given **state** (`lex.Cst`) from its beginning till a certain *endpoint* or *outcome* **event** (`lex.Xst`) or *transition* occurs, changing the state to another.

Examples of such times and outcome events:

- lifetime: birth → death,
- duration of marriage: wedding → divorce,
- healthy exposure time:
  start of exposure → onset of disease,
- clinical survival time:
  diagnosis of a disease → death.

Ex. Survival of 338 oral cancer patients

Important variables:

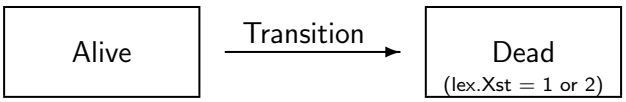
- `time` = duration of patientship from diagnosis (**entry**) till death (`death`) or censoring (`Alive`), (`lex.Cst` is (`Alive`))
- `event` = indicator for the outcome and its observation at the end of follow-up (**exit**):
  0 = censoring,
  1 = death from oral cancer

Special features:

- Two possible endpoints
- Censoring – incomplete observation of the survival time.

Set-up of classical survival analysis

- **Two-state model**: only one type of event changes the initial state.
- Major applications: analysis of lifetimes since birth and of survival times since diagnosis of a disease until death from any cause.



- **Censoring**: Death and final lifetime not observed for some subjects due to emigration or closing the follow-up while they are still alive

Distribution concepts: hazard function

The **hazard rate** or **intensity** function $\lambda(t)$

$$\lambda(t) = P(t < T \leq t + \Delta | T > t)/\Delta, \ \ forsmall\Delta$$

$\approx$ the conditional probability that the event occurs in a short interval $(t, t + \Delta]$, given that it does not occur before t, divided by interval length.

In other words, during a short interval

$$\text{risk of event} \approx \text{hazard} \times \text{interval length}$$

Distribution concepts: survival and cumulative hazard functions

Survival function

$$S(t) = P(T > t),$$

= probability of avoiding the event at least up to t (the event occurs only after t).

The **cumulative hazard** (or integrated intensity):

$$\Lambda(t) = \int_0^t \lambda(u)du$$

Connections between the functions:

$$S(t) \quad = \quad \exp\{-\Lambda(t)\}$$

Observed data on survival times

For individuals $i = 1, \ldots, n$ let
T_i = time to outcome event,
U_i = time to censoring.

Censoring is assumed **noninformative**, *i.e.*
independent from occurrence of events.
We observe

$y_i = \min\{T_i, U_i\}$, *i.e.* the exit time, and

$\delta_i = 1_{\{T_i < U_i\}}$, indicator (1/0) for the outcome event occurring first, before censoring.

Censoring must properly be taken into account in the statistical analysis.

Approaches for analysing survival time

- **Parametric model** (like Weibull, gamma, etc.) on hazard rate $\lambda(t) \rightarrow$ Likelihood:

$$L = \prod_{i=1}^{n} \lambda(y_i)^{\delta_i} S(y_i)$$

- **Piecewise constant rate** model on $\lambda(t)$
– see Bendix's lecture on time-splitting (Poisson likelihood).
- **Non-parametric** methods, like
Kaplan–Meier (KM) estimator of survival curve $S(t)$ and Cox proportional hazards model on $\lambda(t)$.

R package survival

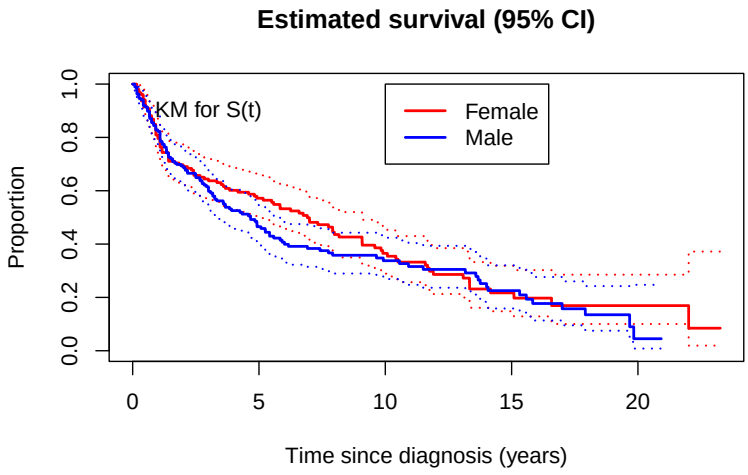
Tools for analysis with one outcome event.

- `Surv(time, event) -> sobj`
  creates a **survival object** `sobj` assuming that all start at 0, containing pairs (y_i, δ_i),
- `Surv(entry, exit, event) -> sobj2`
  creates a survival object from `entry` and `exit` times,
- `survfit(sobj ~ x) -> sfo`
  creates a **survfit** object `sfo` containing KM or other non-parametric estimates (also from a fitted Cox model),
- `plot(sfo), plotCIF(sobj)`
  plot method for survival curves and related graphs,
- `coxph(sobj ~ x1 + x2)`
  fits a Cox model with covariates `x1` and `x2`.
- `survreg()` – parametric survival models.

Ex. Oral cancer data (cont'd)

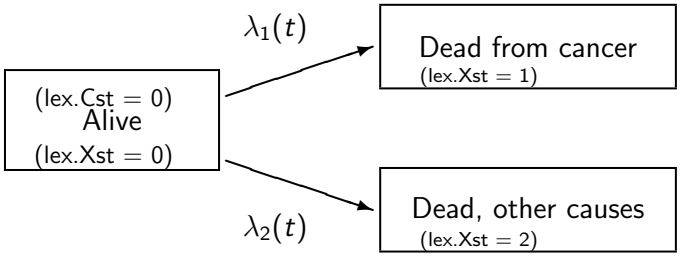
```
> orca$suob <- Surv(orca$time, 1*(orca$event > 0) )
> orca$suob[1:7]   #  + indicates censored observation
[1] 5.081+ 0.419  7.915  2.480  2.500  0.167  5.925+
> km1 <- survfit( suob ~ 1, data = orca)
> km1              #  brief  summary
Call: survfit(formula = suob ~ 1, data = orca)

      n events median 0.95LCL 0.95UCL
[1,] 338    229   5.42    4.33    6.92
> summary(km1)     #  detailed KM-estimate
Call: survfit(formula = suob ~ 1, data = orca)

   time n.risk n.event survival std.err lower 95% CI upper 95% CI
  0.085    338       2   0.9941 0.00417       0.9859        1.000
  0.162    336       2   0.9882 0.00588       0.9767        1.000
  0.167    334       4   0.9763 0.00827       0.9603        0.993
  0.170    330       2   0.9704 0.00922       0.9525        0.989
  0.246    328       1   0.9675 0.00965       0.9487        0.987
  0.249    327       1   0.9645 0.01007       0.9450        0.984
  0.252    326       3   0.9556 0.01120       0.9339        0.978
  0.329    323       1   0.9527 0.01155       0.9303        0.976
  0.334    322       1   0.9497 0.01189       0.9267        0.973
  0.413    321       1   0.9467 0.01221       0.9231        0.971
```


Oral cancer: Kaplan-Meier estimates



Competing risks model: causes of death

- Often the interest is focused on the risk or hazard of dying from one specific cause.
- That cause may eventually not be realized, because a **competing cause** of death hits first.



- Generalizes to several competing causes.

Competing events & competing risks

In many epidemiological and clinical contexts there are competing events that may occur before the target event and remove the person from the population at risk for the event, *e.g.*

- *target event*: occurrence of endometrial cancer, *competing events*: hysterectomy or death.
- *target event*: relapse of a disease (ending the state of remission), *competing event*: death while still in remission.
- *target event*: divorce, *competing event*: death of either spouse.

Event-specific quantities

Cumulative incidence function (CIF) or

$$F_c(t) = P(T \leq t \text{ and } C = c), \quad c = 1, 2,$$

From these one can recover

- $F(t) = \sum_c F_c(t)$, CDF of event-free survival time T, *i.e.* cumulative risk of any event by t.
- $S(t) = 1 - F(t)$, **event-free survival function**, *i.e.* probability of avoiding all events by t, but $S(t) \neq F_1(t) + F_2(t)$

Event-specific quantities (cont'd)

Event- or **cause-specific hazard function**

$$\begin{aligned}\lambda_c(t) &= \lim_{\Delta \to 0} \frac{P(t < T \leq t + \Delta \text{ and } C = c \mid T > t)}{\Delta} \\ &= \frac{f_c(t)}{1 - F(t)}\end{aligned}$$

CIF = risk of event c over risk period $[0, t]$ in the presence of competing risks, also obtained

$$F_c(t) = \int_0^t \lambda_c(v) S(v) dv, \quad c = 1, 2,$$

More on the technical definitions of relevant quantities:
http://bendixcarstensen.com/AdvCoh/papers/fundamentals.pdf

Warning of "net risk" and "cause-specific survival"

- The "**net risk**" of outcome c by time t, assuming hypothetical elimination of competing risks, is often defined as
$$F_1^*(t) = 1 - S_1^*(t) = 1 - \exp\{-\Lambda_1(t)\} \neq S(t)$$
- In clinical survival studies, function $S_1^*(t)$ is often called "**cause-specific survival**", or "**net survival**"
- Yet, these *-functions, $F_1^*(t)$ and $S_1^*(t)$, lack proper probability interpretation when competing risks exist.
- Hence, their use should be viewed critically (Andersen & Keiding, *Stat Med*, 2012)

Analysis with competing events

Let U_i = censoring time, T_i = time to first event, and C_i = variable for event 1 or 2. We observe

- $y_i = \min\{T_i, U_i\}$, *i.e.* the exit time, and
- $\delta_{ic} = 1_{\{T_i < U_i \ \& \ C_i = c\}}$, indicator (1/0) for event c being first observed, $c = 1, 2$.

Non-parametric estimation of CIF

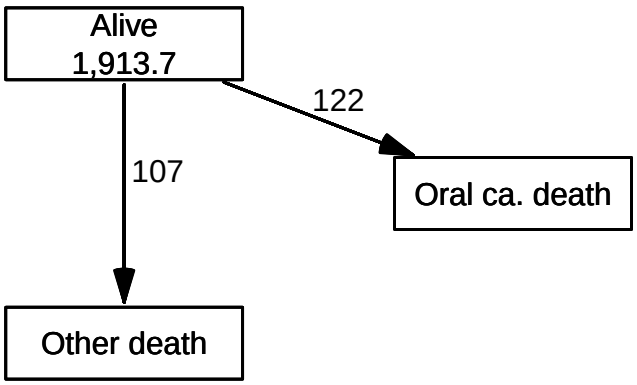
- Let $t_1 < t_2 < \cdots < t_K$ be the K distinct time points at which any outcome event was observed,
Let also $\widetilde{S}(t)$ be KM estimator for overall $S(t)$.
- **Aalen-Johansen estimator** (AJ) for the cumulative incidence function $F(t)$ should be used

R tools for competing risks analysis

- `survfit( Surv(...,type="mstate") )` in Survival-package can be fitted for any transition of a multistate model and to obtain A-J estimates.
- Package `cmprsk` – `cuminc(ftime, fstatus, ...)` computes CIF-estimates, and can be compared in more than two samples. `plot.cuminc()` plots them.
- Package `Epi` – `Lexis` tools for multistate analyses
  Will be advertised by Bendix!

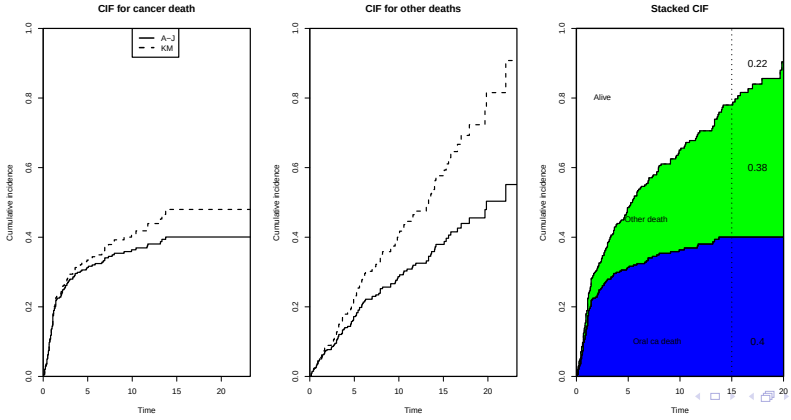
Box diagram for transitions

```
NOTE: entry.status has been set to "Alive" for all.
NOTE: entry is assumed to be 0 on the stime timescale.
```

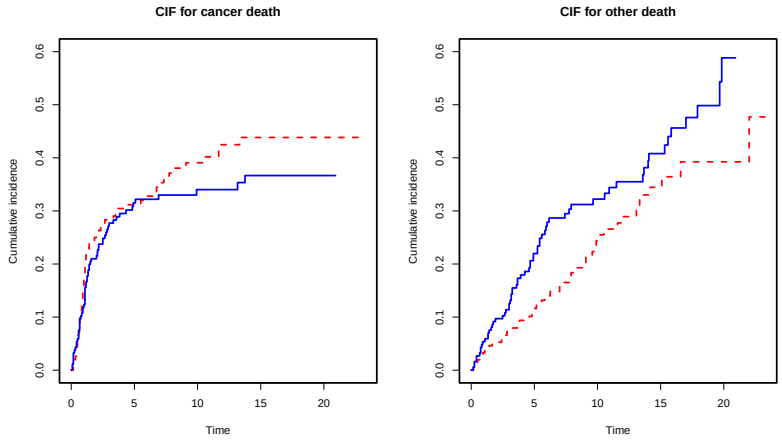


Ex. Survival from oral cancer

- AJ-estimates of CIFs (solid) for both causes.
- Naive KM-estimates of CIF (dashed) > AJ-estimates
- CIF curves may also be stacked (right).



Ex. CIFs by cause in men and women



CIF for cancer higher in women (chance?) but for other causes higher in men (no surprise).

Regression models for time-to-event data

Regression models for hazards can be defined *e.g.* for

(a) hazards, multiplicatively:

$$\lambda_i(t) = \lambda_0(t;\alpha) r(\eta_i), \quad \text{where}$$

$\lambda_0(t;\alpha) =$ baseline hazard and
$r(\eta_i) =$ relative rate function, typically $\exp(\eta_i)$

(b) hazards, additively:

$$\lambda_i(t) = \lambda_0(t;\alpha) + \eta_i.$$

Relative hazards model or Cox model

In model (b), the baseline hazard $\lambda_0(t,\alpha)$ may be given a parametric form (*e.g.* Weibull) or a piecewise constant rate (exponential) structure.

Often a parameter-free form $\lambda_0(t)$ is assumed. Then

$$\lambda_i(t) = \lambda_0(t)\exp(\eta_1),$$

specifies the **Cox model** or the **semiparametric proportional hazards model**. bigskip $\eta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ not depending on time.

Generalizations: **time-dependent** covariates $x_{ij}(t)$

PH model: interpretation of parameters

Present the model explicitly in terms of x's and β's.

$$\lambda_i(t) = \lambda_0(t)\exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

Consider two individuals, i and i', having the same values of all other covariates except the j^{th} one.

The ratio of hazards is constant:

$$\frac{\lambda_i(t)}{\lambda_{i'}(t)} = \frac{\exp(\eta_i)}{\exp(\eta_{i'})} = \exp\{\beta_j(x_{ij} - x_{i'j})\}.$$

Thus $e^{\beta_j} = \text{HR}_j =$ **hazard ratio** or relative rate associated with a unit change in covariate X_j.

Ex. Total mortality of oral ca. patients

Fitting Cox models with sex and sex + age.

```
> cm0 <- coxph( suob ~ sex, data = orca)
> ci.exp(cm0)

        exp(Est.)      2.5%    97.5%
sexMale  1.134004 0.8724905 1.473902
```

Total mortality in males is 13% higher in male than females, but not significant.

```
> cm0 <- coxph( suob ~ age+sex, data = orca)
> ci.exp(cm0)

        exp(Est.)     2.5%    97.5%
age      1.041914 1.030655 1.053296
sexMale  1.494305 1.139254 1.960009
```

The M/F contrast visible only after age-adjustment.(43% higher in males).

Predictions from the Cox model

- Individual survival *times* cannot be predicted but ind'l survival *curves* can. PH model implies:
$$S_i(t) = [S_0(t)]^{\exp(\beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$
- Having estimated β by partial likelihood, the baseline $S_0(t)$ is estimated by Breslow method
- From these, a survival curve for an individual with given covariate values is predicted.
- In R: `pred <- survfit(mod, newdata=...)` and `plot(pred)`, where `mod` is the fitted `coxph` object, and `newdata` specifies the covariate values. `newdata` is always needed for predictions.

Modelling with competing risks

Main options, providing answers to different questions.

(a) Cox model for event-specific hazards $\lambda_c(t) = f_c(t)/[1 - F(t)]$, when *e.g.* the interest is in the biological effect of the prognostic factors on the fatality of the very disease that often leads to the relevant outcome.

(b) **Fine–Gray model** for the hazard of the subdistribution $\gamma_c(t) = f_c(t)/[1 - F_c(t)]$ when we want to assess the impact of the factors on the overall cumulative incidence of event c.
– Function `crr()` in package `cmprsk`.

SMR

Relate population mortality to the mortality of your "exposed" cohort

Let

- $\lambda_(a)$ be the mortality in the cohort
- $\lambda_P(a)$ be the population mortality
- $\lambda_E(a)$ be the excess hazard of dying from the disease among cohort members
- SMR is the relative mortality in the cohort

$\lambda(a) = \lambda_E(a) + \lambda_P(a)$ (excess mortality)

$\lambda(a) = SMR \times \lambda_P(a)$ (standardized mortality ratio)

Representation of follow-up

Bendix Carstensen Steno Diabetes Center Copenhagen
Herlev, Denmark
http://BendixCarstensen.com

SPE, Tartu, Estonia,

June 2023

http://BendixCarstensen.com/SPE

Representation of follow-up

Bendix Carstensen

Representation of follow-up

SPE, Tartu, Estonia,

June 2023

http://BendixCarstensen.com/SPE time-split

- In follow-up studies we estimate rates from:
  - D — events, deaths
  - Y — person-years
  - $\hat{\lambda} = D/Y$ rates
  - . . . empirical counterpart of intensity — an **estimate**
- Rates differ between persons.
- Rates differ within persons:
  - by age
  - by calendar time
  - by disease duration
  - . . .
- Multiple timescales.
- Multiple states (little boxes — later)

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 pieces of information recorded:

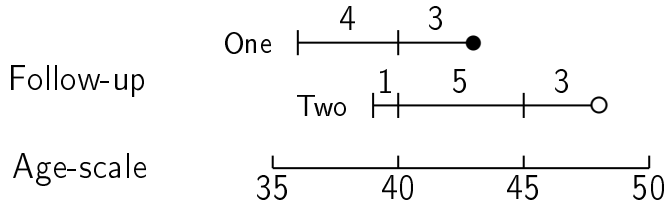
Date of entry	`entry`	date variable
Date of exit	`exit`	date variable
Status at exit	`fail`	indicator (mostly $0/1$)

These are specific for each **type** of outcome.

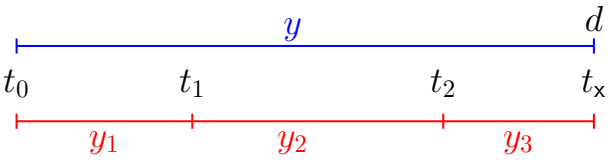
Stratification by age

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

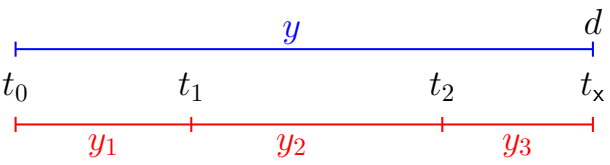
If follow-up is long, stratification by categories of **current age** is preferable.



— allowing rates to vary across age-bands
— how do we do the split and why is it OK?



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



Probability	log-Likelihood
$\mathrm{P}(d \text{ at } t_x \mid \text{entry } t_0)$	$d\log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_0 \to t_1 \mid \text{entry } t_0)$	$= 0\log(\lambda_1) - \lambda_1 y_1$
$\times \mathrm{P}(\text{surv } t_1 \to t_2 \mid \text{entry } t_1)$	$+ 0\log(\lambda_2) - \lambda_2 y_2$
$\times \mathrm{P}(d \text{ at } t_x \mid \text{entry } t_2)$	$+ d\log(\lambda_3) - \lambda_3 y_3$

— allows different rates (λ_i) in each interval

Dividing time into bands requires:

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, mimicking continuous time by using small intervals:
—time at the beginning of interval as quantitative variable.

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 (D) 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

	subj. 1		subj. 2		subj. 3		$\sum$	
Age	Y	D	Y	D	Y	D	Y	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?

Follow-up intervals 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
- . . . do not confuse **duration** and **timescale** !

Follow-up data in Epi — Lexis objects I

```
> thoro[1:6,1:8]
  id sex birthdat contrast injecdat volume  exitdat exitstat
1  1   2 1916.609        1 1938.791     22 1976.787        1
2  2   2 1927.843        1 1943.906     80 1966.030        1
3  3   1 1902.778        1 1935.629     10 1959.719        1
4  4   1 1918.359        1 1936.396     10 1977.307        1
5  5   1 1902.931        1 1937.387     10 1945.387        1
6  6   2 1903.714        1 1937.316     20 1944.738        1

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


Follow-up data in Epi — Lexis objects II

```
NOTE: entry.status has been set to 0 for all.
NOTE: Dropping  2  rows with duration of follow up < tol
> summary(thL, timeScales = TRUE)
Transitions:
     To
From   0    1  Records:  Events: Risk time:  Persons:
   0 504 1964      2468     1964   51934.08      2468

Timescales:
age dte tfi
 ""  ""  ""
```

Definition of Lexis object

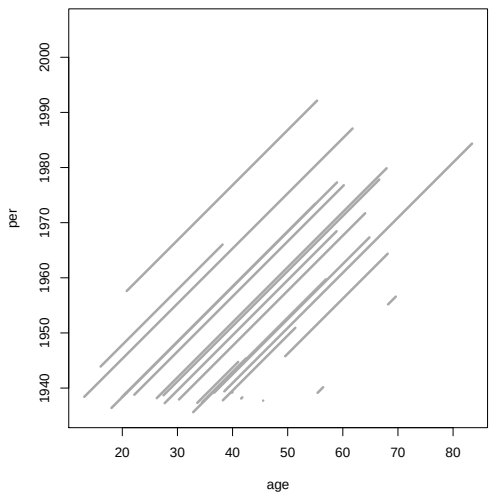
```
thL <- Lexis(entry = list(age = injecdat-birthdat,
                          dte = injecdat,
                          tfi = 0),
             exit = list(dte = exitdat),
        exit.status = as.numeric(exitstat==1),
             data = thoro)
```

entry is defined on **three** timescales,
but exit is only needed on **one** timescale (or vice versa):
Follow-up time is the same on all timescales: exitdat - injecdat

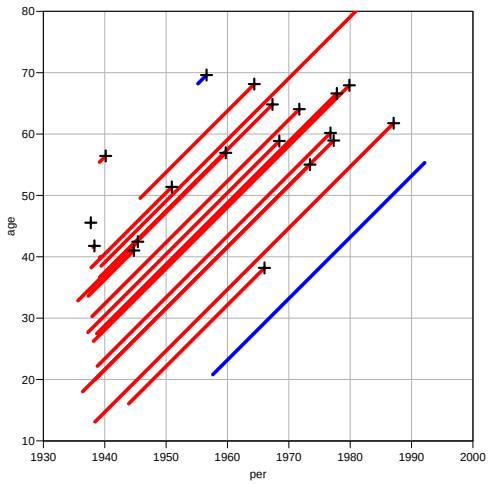
One element of entry and exit must have same name (dte).

The looks of a Lexis object

```
> thL[1:4,1:9]
    age      dte 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:
   0 504 1964      2468      1964    51934.08      2468
```

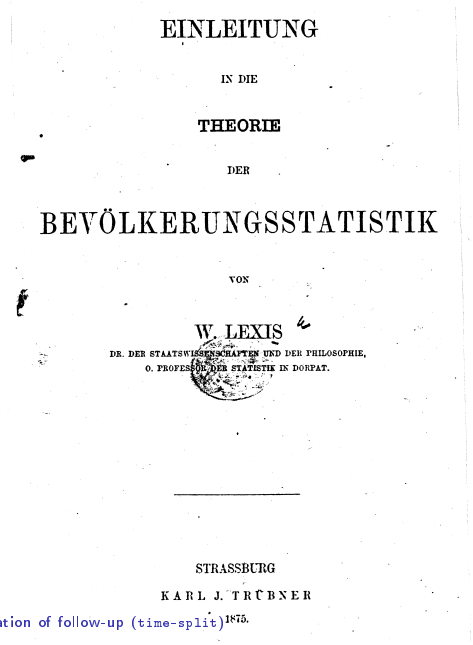




```
> plot( thL, lwd=3 )
```



Lexis diagram

```
> 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 )
```



EINLEITUNG

IN DIE

THEORIE

DER

BEVÖLKERUNGSSTATISTIK

VON

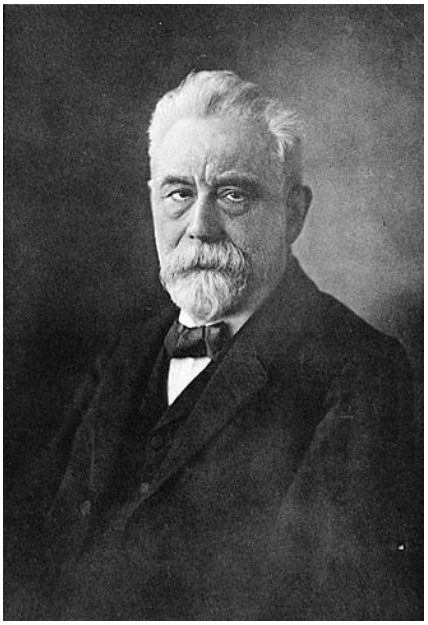
W. LEXIS

DR. DER STAATSWISSENSCHAFTEN UND DER PHILOSOPHIE,
O. PROFESSOR DER STATISTIK IN DORPAT.

STRASSBURG

KARL J. TRÜBNER

1875.



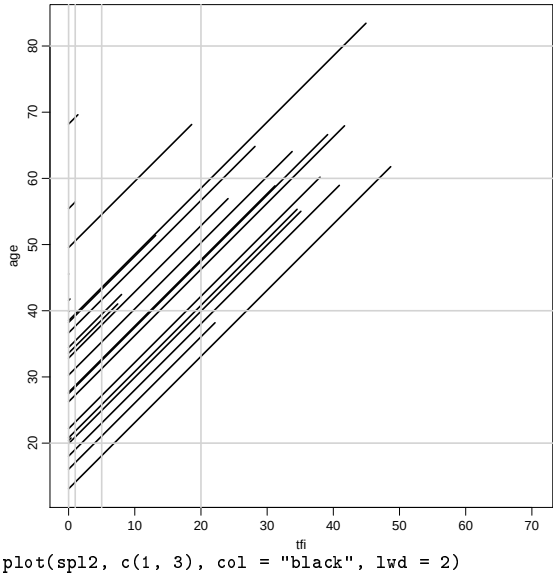
Splitting follow-up time

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

Split on another timescale

```
> spl2 <- splitLexis( spl1, time.scale="tfi", breaks=c(0,1,5,20,100) )
> round( spl2, 1 )
```

	lex.id	age	dte	tfi	lex.dur	lex.Cst	lex.Xst	id	sex	birthdat	contrast	inje
1	1	22.2	1938.8	0.0	1.0	0	0	1	2	1916.6	1	19
2	1	23.2	1939.8	1.0	4.0	0	0	1	2	1916.6	1	19
3	1	27.2	1943.8	5.0	12.8	0	0	1	2	1916.6	1	19
4	1	40.0	1956.6	17.8	2.2	0	0	1	2	1916.6	1	19
5	1	42.2	1958.8	20.0	17.8	0	0	1	2	1916.6	1	19
6	1	60.0	1976.6	37.8	0.2	0	1	1	2	1916.6	1	19
7	2	49.5	1945.8	0.0	1.0	0	0	640	2	1896.2	1	19
8	2	50.5	1946.8	1.0	4.0	0	0	640	2	1896.2	1	19
9	2	54.5	1950.8	5.0	5.5	0	0	640	2	1896.2	1	19
10	2	60.0	1956.2	10.5	8.1	0	1	640	2	1896.2	1	19
11	3	68.2	1955.2	0.0	1.0	0	0	3425	1	1887.0	2	19
12	3	69.2	1956.2	1.0	0.4	0	1	3425	1	1887.0	2	19
13	4	20.8	1957.6	0.0	1.0	0	0	4017	2	1936.8	2	19
14	4	21.8	1958.6	1.0	4.0	0	0	4017	2	1936.8	2	19
15	4	25.8	1962.6	5.0	14.2	0	0	4017	2	1936.8	2	19
16	4	40.0	1976.8	19.2	0.8	0	0	4017	2	1936.8	2	19
17	4	40.8	1977.6	20.0	14.5	0	0	4017	2	1936.8	2	19




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

age	tfi	lex.dur	lex.Cst	lex.Xst
22.2	0.0	1.0	0	0
23.2	1.0	4.0	0	0
27.2	5.0	12.8	0	0
40.0	17.8	2.2	0	0
42.2	20.0	17.8	0	0
60.0	37.8	0.2	0	1

Splitting on several timescales

```
> spl1 <- splitLexis(thL , time.scale = "age", breaks = seq(0, 100, 20))
> spl2 <- splitLexis(spl1, time.scale = "tfi", breaks = c(0, 1, 5, 20, 100))
> summary(spl2)

Transitions:
     To
From    0    1  Records:  Events: Risk time:  Persons:
   0 8250 1964     10214     1964   51934.08      2468

> library(popEpi)
> splx <- splitMulti(thL, age = seq(0, 100, 20), tfi = c(0, 1, 5, 20, 100))
> summary(splx)

Transitions:
     To
From    0    1  Records:  Events: Risk time:  Persons:
   0 8248 1964     10212     1964   51916.98      2468

> # NOTE: splitMulti excludes follow-up outside range of breaks
```

Likelihood for time-split data

- We assume that rates are constant in each (small) intervals
- Each observation in the dataset represents an interval, contributing a term to the (log-)likelihood for the rate
- Each **term** looks like a contribution from a Poisson variate (albeit with values only 0 or 1)
- So the likelihood from a single **person** looks like the likelihood from several independent Poisson variates
- . . . but the data are neither independent nor Poisson

Analysis of time-split data

Observations (records) classified by p—person and i—interval

- d_{pi} — events in the variable: `lex.Xst & lex.Xst!=lex.Cst`
- y_{pi} — risk time: `lex.dur` (duration)
- Covariates are:
  - timescales (age, period, time in study)
  - other variables for this person (constant in each interval).
- Likelihood for rates for one person is identical to a Poisson likelihood for many independent Poisson variates
- Modeling rates using `glm` or `gam`:
  time-scales and other covariates are treated alike

Fitting a simple model—data:

```
> 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           928.00 20094.74    4.62
 2          1036.00 31839.35    3.25

 Total      1964.00 51934.08    3.78
 ------------------------------------
```

Fitting a simple model

```
-------------------------------------
contrast         D        Y    Rate
-------------------------------------
1           928.00 20094.74    4.62
2          1036.00 31839.35    3.25
-------------------------------------
```

```
> m0 <- glm((lex.Xst==1) ~ 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      4.62 4.33  4.93
factor(contrast)2      3.25 3.06  3.46
```

. . . a Poisson model for mortality using log-person-years as offset

Fitting a simple model

```
-------------------------------------
contrast         D        Y    Rate
-------------------------------------
1           928.00 20094.74    4.62
2          1036.00 31839.35    3.25
-------------------------------------
```

```
> m0 <- glm(cbind(lex.Xst, lex.dur / 100) ~ factor(contrast) - 1,
+           family = poisreg,
+             data = spl2)
> round(ci.exp(m0), 2)
                  exp(Est.) 2.5% 97.5%
factor(contrast)1      4.62 4.33  4.93
factor(contrast)2      3.25 3.06  3.46
```

. . . a Poisson model for mortality rates based on deaths and person-years

Fitting a simple model

The wrapper `glm.Lexis` requires that `lex.Cst` and `lex.Xst` are factors —use `factorize` to make them:

```
> splf <- factorize(spl2)
> m0 <- glm.Lexis(splf, ~ factor(contrast) - 1, scale = 100)
stats::glm Poisson analysis of Lexis object splf with log link:
Rates for the transition:
0->1
, lex.dur (person-time) scaled by 100
> round(ci.exp(m0), 2)
                  exp(Est.) 2.5% 97.5%
factor(contrast)1      4.62 4.33  4.93
factor(contrast)2      3.25 3.06  3.46
```

. . . a Poisson model for mortality rates based on deaths and person-years in a `Lexis` object

Fitting a simple model — aggregate data

```
-----------------------------------
contrast         D        Y    Rate
-----------------------------------
1           928.00 20094.74    4.62
2          1036.00 31839.35    3.25
-----------------------------------
```

As long as we only use covariates that take only a few values, we can model the aggregate data directly:

```
> mx <- glm(cbind(c(928, 1036), c(20094.74, 31839.35) / 100) ~ factor(1:2) - 1,
+           family = poisreg )
> round(ci.exp(mx), 2)
             exp(Est.) 2.5% 97.5%
factor(1:2)1      4.62 4.33  4.93
factor(1:2)2      3.25 3.06  3.46
```


SMR

Bendix Carstensen

Representation of follow-up

SPE, Tartu, Estonia,

June 2023

http://BendixCarstensen.com/SPE 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)
- Hospital registers (disease rates)

Cohort rates vs. population rates: RSR

- **Additive:** $\lambda(a) = \delta(a) + \lambda_{\mathsf{p}}(a)$
- Note that the survival (since $a = a_0$, say) is:

$$\begin{aligned} S(a) &= \exp\Big(-\int_{a_0}^{a} \delta(a) + \lambda_{\mathsf{p}}(a)\,\mathrm{d}a\Big) \\ &= \exp\Big(-\int_{a_0}^{a} \delta(a)\,\mathrm{d}a\Big) \times S_{\mathsf{p}}(a) \\ \Rightarrow \quad r(a) = S(a)/S_{\mathsf{p}}(a) &= \exp\Big(-\int_{a_0}^{a} \delta(a)\,\mathrm{d}a\Big) \end{aligned}$$

- **Additive** model for **rates** $\Leftrightarrow$ **Relative survival** model.

Cohort rates vs. population rates: SMR

- **Multiplicative:** $\lambda(a) = \theta \times \lambda_{\mathsf{p}}(a)$
- Cohort rates proportional to reference rates, λ_{p}: $\lambda(a) = \theta \times \lambda_{\mathsf{p}}(a)$ — θ the same in all age-bands.
- D_a deaths during Y_a person-years an age-band a gives the likelihood:

$$\begin{aligned} D_a\log(\lambda(a)) - \lambda(a)Y_a &= D_a\log(\theta\lambda_{\mathsf{p}}(a)) - \theta\lambda_{\mathsf{p}}(a)Y_a \\ &= D_a\log(\theta) + D_a\log(\lambda_{\mathsf{p}}(a)) - \theta(\lambda_{\mathsf{p}}(a)Y_a) \end{aligned}$$

- The constant $D_a\log(\lambda_{\mathsf{p}}(a))$ does not involve θ, and so can be dropped.

- $\lambda_{\mathsf{p}}(a)Y_a = E_a$ is the "expected" number of cases in age a, so the log-likelihood contribution from age a is:

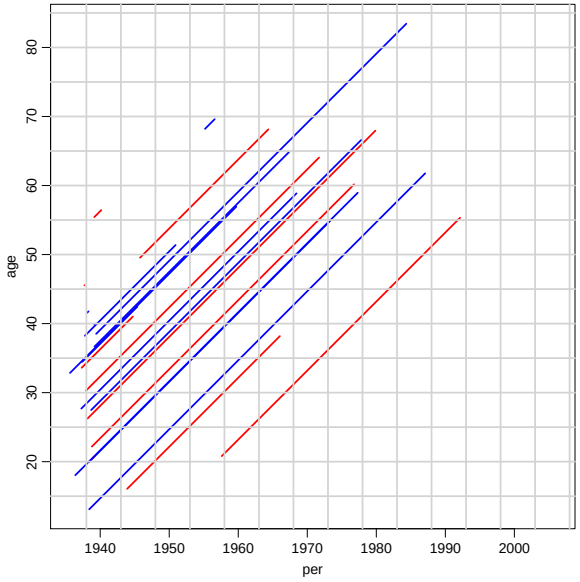
$$D_a\log(\theta) - \theta\big(\lambda_{\mathsf{p}}(a)Y_a\big) = D_a\log(\theta) - \theta(E_a)$$

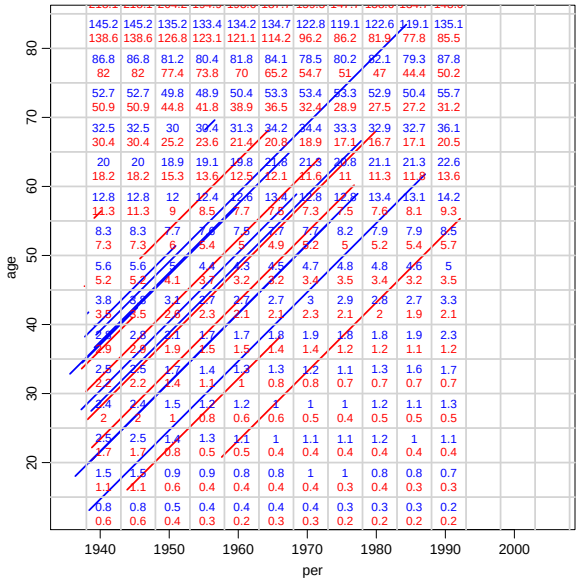
- The log-likelihood is similar to the log-likelihood for a rate, so:

$$\hat{\theta} = \sum_a D_a / \sum_a E_a = \text{Observed/Expected} = \text{SMR}$$

Modeling the SMR in practice

- 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, $e_i = y_i \times \lambda_{\text{p}}$ (`lex.dur * rate`)
- λ_{p} is the population rate corresponding to the age, period and sex of the follow-up period y_i.





Split the data to fit with population data

```
> thad <- splitMulti(thL, age=seq(0,90,5), dte=seq(1938,2038,5) )
> summary( thad )

Transitions:
     To
From     0    1  Records:  Events: Risk time:  Persons:
   0 21059 1939     22998     1939   51787.96      2463
```

Create variables to fit with the population data

```
> thad$agr <- timeBand( thad, "age", "left" )
> thad$per <- timeBand( thad, "dte", "left" )
> round( thad[1:5,c("lex.id","age","agr","dte","per","lex.dur","lex.Xst","sex")],

 lex.id   age     dte lex.dur lex.Xst agr  per sex
      1 22.18 1938.79    2.82       0  20 1938   2
      1 25.00 1941.61    1.39       0  25 1938   2
      1 26.39 1943.00    3.61       0  25 1943   2
      1 30.00 1946.61    1.39       0  30 1943   2
      1 31.39 1948.00    3.61       0  30 1948   2
```



```
> data( gmortDK )
> dim( gmortDK )

[1] 418  21

> gmortDK[1:6,1:6]

  agr per sex   risk    dt     rt
1   0  38   1 996019 14079 14.135
2   5  38   1 802334   726  0.905
3  10  38   1 753017   600  0.797
4  15  38   1 773393  1167  1.509
5  20  38   1 813882  2031  2.495
6  25  38   1 789990  1862  2.357

> gmortDK$per <- gmortDK$per+1900
> #
> thadx <- merge( thad, gmortDK[,c("agr","per","sex","rt")] )
> #
> thadx$E <- thadx$lex.dur * thadx$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 = thadx)

 ----------------------------------------
 contrast         D        Y       E     SMR
 ----------------------------------------
 1           917.00 20045.46  214.66    4.27
 2          1022.00 31742.51  447.21    2.29

 Total      1939.00 51787.96  661.87    2.93
 ----------------------------------------
```

```
-------------------------------------------
contrast         D        Y       E     SMR
-------------------------------------------
1           917.00 20045.46  214.66    4.27
2          1022.00 31742.51  447.21    2.29
-------------------------------------------
```

```
> m.SMR <- glm(cbind(lex.Xst, E) ~ factor(contrast) - 1,
+              family = poisreg,
+                data = thadx)
> round(ci.exp(m.SMR), 2)
                  exp(Est.) 2.5% 97.5%
factor(contrast)1      4.27 4.00  4.56
factor(contrast)2      2.29 2.15  2.43
```

- Analysis of SMR is like analysis of rates:
- Replace Y with E — that's all! (`glm.Lexis` not usable)
- . . . it's the calculation of E that is difficult

Nested case-control and case-cohort studies

Tuesday, 06 June, 2023

Esa Läärä & Martyn Plummer & Krista Fischer

Statistical Practice in Epidemiology with R
University of Tartu, Estonia
June, 2023

Points to be covered

- Outcome-dependent sampling designs a.k.a. **case-control** studies *vs.* **full cohort** design.
- **Nested case-control** study (NCC): sampling of controls from risk-sets during follow-up of study population.
- **Matching** in selection of control subjects in NCC.
- R tools for NCC: function `ccwc()` in `Epi` for sampling controls, and `clogit()` in `survival` for model fitting.
- **Case-cohort** study (CC): sampling a subcohort from the whole cohort as it is at the start of follow-up.
- R tools for CC model fitting: function `cch()` in `survival`

Example: Smoking and cervix cancer

Study population, measurements, follow-up, and sampling design

- Joint cohort of $N \approx$ 500 000 women from 3 Nordic biobanks.
- Follow-up: From variable entry times since 1970s till 2000.
- For each of 200 cases, 3 controls were sampled; matched for biobank, age (± 2 y), and time of entry (± 2 mo).
- Frozen sera of cases and controls analyzed for cotinine *etc.*

Main result: Adjusted OR = 1.5 (95% CI 1.1 to 2.3) for high (>242.6 ng/ml) *vs.* low (<3.0 ng/ml) cotinine levels.

Simen Kapeu *et al.* (2009) *Am J Epidemiol*

Example: USF1 gene and CVD

Study population, measurements, follow-up, and sampling design

- Two FINRISK cohorts, total $N \approx 14000$ M & F, 25-64 y.
- Baseline health exam, questionnaire & blood specimens at recruitment in the 1990s – Follow-up until the end of 2003.
- Subcohort of 786 subjects sampled.
- 528 incident cases of CVD; 72 of them in the subcohort.
- Frozen blood from cases and subchort members genotyped.

Main result: Female carriers of a high risk haplotype had a 2-fold hazard of getting CVD [95% CI: 1.2 to 3.5]

Komulainen *et al.* (2006) *PLoS Genetics*

Full cohort design & its simple analysis

- **Full cohort design**: Data on exposure variables obtained for all subjects in a large study population.
- Summary data for crude comparison:

	Exposed	Unexposed	Total
Cases	D_1	D_0	D
Non-cases	B_1	B_0	B
Group size at start	N_1	N_0	N
Follow-up times	Y_1	Y_0	Y

- Crude estimation of **hazard ratio** $\rho = \lambda_1/\lambda_0$: **incidence rate ratio** IR, with standard error of log(IR):

$$\widehat{\rho} = \text{IR} = \frac{D_1/Y_1}{D_0/Y_0} \qquad \text{SE}[\log(\text{IR})] = \sqrt{\frac{1}{D_1} + \frac{1}{D_0}}.$$

- More refined analyses: Poisson or Cox regression.

Problems with full cohort design

Obtaining exposure and covariate data

- Slow and expensive in a big cohort.
- Easier with questionnaire and register data,
- Extremely costly and laborious for *e.g.*
  - measurements from biological specimens, like genotyping, antibody assays, *etc.*
  - dietary diaries & other manual records

Can we obtain equally valid estimates of hazard ratios etc. with nearly as good precision by some other strategies?

Yes – we can!

Estimation of hazard ratio

The incidence rate ratio can be expressed:

$$\text{IR} = \frac{D_1/D_0}{Y_1/Y_0} = \frac{\text{cases: exposed / unexposed}}{\text{person-times: exposed / unexposed}}$$

$$= \frac{\textit{exp're odds} \text{ in cases}}{\textit{exp're odds} \text{ in p-times}} = \textbf{exposure odds ratio} \text{ (EOR)}$$

= Exposure distribution in cases *vs.* that in cohort!

Implication for more efficient design:

- *Numerator*: Collect exposure data on all cases.
- *Denominator*: Estimate the ratio of person-times Y_1/Y_0 of the exposure groups in the cohort by **sampling** "control" subjects, on whom exposure is measured.

Case-control designs

General principle: Sampling of subjects from a given study population is *outcome-dependent*.

Data on risk factors are collected separately from

(I) **Case group**: All (or high % of) the D subjects in the study population (total N) encountering the outcome event during the follow-up.

(II) **Control group**:

- Random **sample** (simple or stratified) of C subjects ($C << N$) from the population.
- Eligible controls must be bf risk (alive, under follow-up & free of outcome) at given time(s).

Study population in a case-control study?

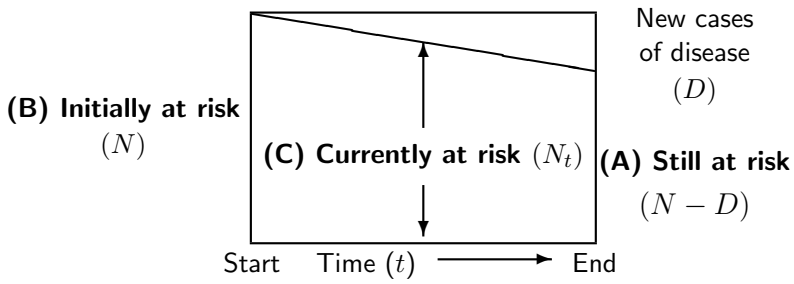
Ideally: The study population comprises subjects who *would be* included as cases, *if they got* the outcome in the study

- *Cohort-based studies*: **cohort** or **closed** population of well-identified subjects under intensive follow-up for outcomes (*e.g.* biobank cohorts).
- *Register-based studies*: **open** or **dynamic** population in a region covered by a disease register.
- *Hospital-based* studies: dynamic **catchment** population of cases – may be hard to identify (*e.g.* hospitals in US).

In general, the role of control subjects is to represent the distribution of person-times by exposure variables in the underlying population from which the cases emerge.

Sampling of controls – alternative frames

Illustrated in a simple longitudinal setting:
Follow-up of a cohort over a fixed risk period & no censoring.



Rodrigues, L. & Kirkwood, B.R. (1990). Case-control designs of common diseases . . . *Int J Epidemiol* **19**: 205-13.

Sampling schemes or designs for controls

(A) Exclusive or **traditional, "case-noncase" sampling**

- Controls chosen from those $N - D$ subjects still at risk (healthy) *at the end* of the risk period (follow-up).

(B) Inclusive sampling or **case-cohort design** (CC)

- The control group – *subcohort* – is a random sample of the cohort (N) *at start*.

(C) Concurrent sampling or **density sampling**

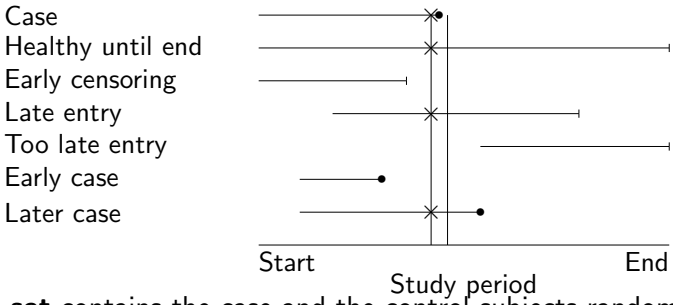
- Controls drawn *during the follow-up*
- **Risk-set** or **time-matched sampling:**
  A set of controls is sampled from the *risk set at each time t of diagnosis* of a new case – a.k.a. **nested case-control design** (NCC)

Nested case-control – two meanings

- In some epidemiologic books, the term "nested case-control study" (NCC) covers jointly all variants of sampling: **(A)**, **(B)**, and **(C)**, from a cohort.
  Rothman *et al.* (2008): *Modern Epidemology, 3rd Ed.*
  Dos Santos Silva (1999): *Cancer Epidemiology.* Ch 8-9
- In biostatistical texts NCC typically refers only to the variant of concurrent or density sampling **(C)**, in which *risk-set* or *time-matched* sampling is employed.
  Borgan & Samuelsen (2003) in *Norsk Epidemiologi*
  Langholz (2005) in *Encyclopedia of Biostatistics.*
- We shall follow the biostatisticians!

NCC: Risk-set sampling with staggered entry

Sampling frame to select controls for a given case:
Members ($\times$) of the **risk set** at t_k, *i.e.* the population at risk at the time of diagnosis t_k of case k.



Sampled risk set contains the case and the control subjects randomly sampled from the non-cases in the risk set at t_k.

Use of different sampling schemes

(A) Exclusive sampling, or "textbook" case-control design

- Almost exclusively(!) used in studies of epidemics.
- (Studies on birth defects with *prevalent* cases.)

(B) Inclusive sampling or case-cohort design

- Good esp. for multiple outcomes, if measurements of risk factors from stored material remain stable.

(C) Concurrent or density sampling (without or with time-matching)

- The only logical design in an open population.
- Most popular in chronic diseases (Knol *et al.* 2008).

Designs **(B)** and **(C)** allow valid estimation of hazard ratios ρ without any "rare disease" assumption.

Case-control studies: Textbooks vs. real life

- Many epi texts focus on the traditional design: **exclusive sampling** of controls, ignoring other designs.
- Claim: "*Odds ratio is the only estimable parameter.*"
- Yet, over 60% of published case-control studies apply **concurrent sampling** or **density sampling** of controls from an **open** or **dynamic** population.
- Thus, the parameter most often estimated is the **hazard ratio** (HR) or **rate ratio** ρ.
- Still, 90% of authors really estimating HR, reported as having estimated an OR (*e.g.* Simen Kapeu *et al.* 2009)

Knol *et al.* (2008). What do case-control studies estimate?
Am J Epidemiol **168**: 1073-81.

Exposure odds ratio – estimate of what?

- Crude summary of case-control data

	exposed	unexposed	total
cases	D_1	D_0	D
controls	C_1	C_0	C

- Depending on study base & sampling strategy, the **exposure odds ratio**

$$\text{EOR} = \frac{D_1/D_0}{C_1/C_0} = \frac{\text{cases: exposed / unexposed}}{\text{controls: exposed / unexposed}}$$

  is a consistent estimator of
  (a) hazard ratio, (b) risk ratio, (c) risk odds ratio,
  (d) prevalence ratio, or (e) prevalence odds ratio
- **NB.** In case-cohort studies with variable follow-up times C_1/C_0 is substituted by $\widehat{Y}_1/\widehat{Y}_0$, from estimated p-years.

Precision and efficiency

With exclusive **(A)** or concurrent **(C)** sampling of controls (unmatched), the estimated variance of log(EOR) is

$$\begin{aligned} \widehat{\text{var}}[\log(\text{EOR})] &= \frac{1}{D_1} + \frac{1}{D_0} + \frac{1}{C_1} + \frac{1}{C_0} \\ &= \text{cohort variance} + \text{sampling variance} \end{aligned}$$

- Depends basically on the numbers of cases, with ≥ 4 controls per case.
- Is not much bigger than $1/D_1 + 1/D_0 =$ variance in a full cohort study with same numbers of cases.

$\Rightarrow$ Usually < 5 controls per case is enough.

$\Rightarrow$ *These designs are very cost-efficient!*

Estimation in concurrent or density sampling

- Assume a simple situation: Prevalence of exposure in the study population stable over time.

$\Rightarrow$ The exposure odds C_1/C_0 among controls
$=$ a consistent estimator of exposure odds Y_1/Y_0 of person-times.

- Therefore, the crude EOR $= (D_1/D_0)/(C_1/C_0)$
$=$ a consistent estimator of hazard ratio $\rho = \lambda_1/\lambda_0$.
- Variance of log(EOR) estimated as above.
- Yet, stability of exposure distribution may be unrealistic, especially in a closed study population or cohort.
- Solution: **Time-matched** sampling of controls from **risk sets**, *i.e.* NCC, & matched EOR to estimate HR.

Prentice & Breslow (1978), Greenland & Thomas (1982).

Matching in case-control studies

= **Stratified sampling** of controls, *e.g.* from the same region, sex, and age group as a given case

- **Frequency matching** or **group matching:** For cases in a specific stratum (*e.g.* same sex and 5-year age-group), a set of controls from a similar subgroup.
- **Individual matching** (1:1 or 1:m matching): For each case, choose 1 or more (rarely > 5) closely similar controls (*e.g.* same sex, age within ± 1 year.
- **NCC**: Sampling from risk-sets implies time-matching at least. Additional matching for other factors possible.
- **CC**: Subcohort selection involves no matching with cases.

Virtues of matching

- Increases *efficiency*, if the matching factors are both
  (i) strong *risk factors* of the disease, and
  (ii) *correlated* with the main exposure.
  – Major reason for matching.
- *Confounding* due to poorly quantified factors (sibship, neighbourhood, *etc.*) may be removed by close matching – only if properly analyzed.
- Biobank studies: Matching for storage time, freeze-thaw cycle & analytic batch improves **comparability of measurements** from frozen specimens
  → Match on the time of baseline measurements within the case's risk set.

Warnings for overmatching

Matching a case with a control subject is a different issue than matching an unexposed subject to an exposed one in a cohort study – much trickier!

- Matching on an *intermediate* variable between exposure and outcome.
  ⇒ *Bias*!
- Matching on a *surrogate* or *correlate* of exposure, which is not a true risk factor.
  ⇒ *Loss of efficiency*.

→ **Counter-matching:** Choose a control which is not similar to the case w.r.t a correlate of exposure.

⇒ Increases efficiency!

- Requires appropriate weighting in the analysis.

Sampling matched controls for NCC using R

- Suppose key follow-up items are recorded for all subjects in a cohort, in which a NCC study is planned.
- Function `ccwc()` in package `Epi` can be used for risk-set sampling of controls. – Arguments:

`entry` :	Time of entry to follow-up
`exit` :	Time of exit from follow-up
`fail` :	Status on exit (1 for case, 0 for censored)
`origin` :	Origin of analysis time scale (e.g. time of birth)
`controls` :	Number of controls to be selected for each case
`match` :	List of matching factors
`data` :	Cohort data frame containing input variables

- Creates a data frame for a NCC study, containing the desired number of matched controls for each case.

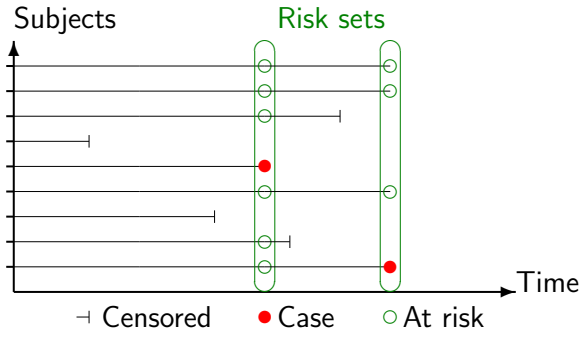
Analysis of matched studies

- Close matching induces a new parameter for each matched case-control set or stratum.
  $\Rightarrow$ **unconditional logistic regression** breaks down.
- Matching on well-defined variables (like age, sex)
  – include these factors as covariates.
- Matching on "soft" variables (like sibship) can be dealt with **conditional logistic regression**.
- Same method in matched designs **(A)**, exclusive, and **(C)**, concurrent, but interpretation of β_js differs:

  (A) β_j = log of risk odds ratio (ROR),
  (C) β_j = log of hazard ratio (HR).

Full cohort design: Follow-up & risk sets

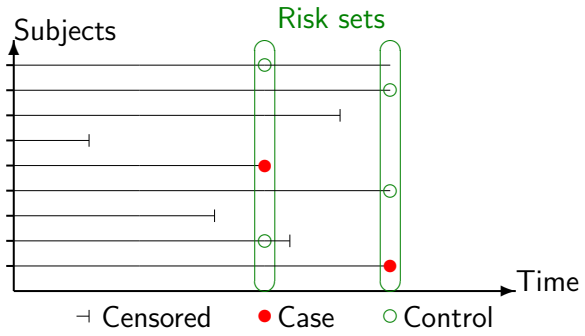
Each member of the cohort provides exposure data for all cases, as long as this member is at risk, *i.e.* (i) alive, (ii) not censored & (iii) free from outcome.



Times of new cases define the **risk-sets**.

Nested case-control (NCC) design

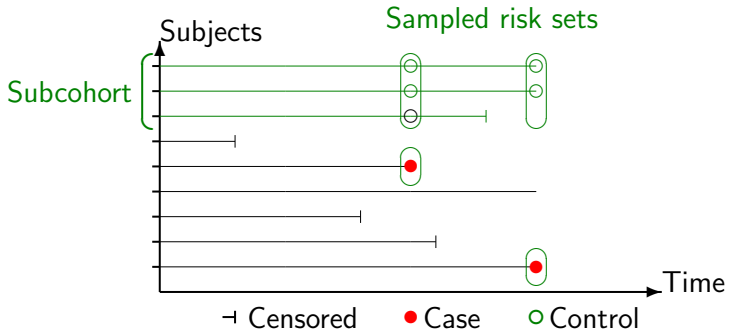
Whenever a new case occurs, a set of controls (here 2/case) are sampled from its risk set.



NB. A control once selected for some case can be selected as a control for another case, and can later on become a case, too.

Case-cohort (CC) design

Subcohort: Sample of the whole cohort randomly selected at the outset.
– Serves as a reference group for all cases.



NB. A subcohort member can become a case, too.

Modelling in NCC and other matched studies

Cox proportional hazards model:

$$\lambda_i(t, x_i; \beta) = \lambda_0(t) \exp(x_{i1}\beta_1 + \cdots + x_{ip}\beta_p),$$

Estimation: partial likelihood $L^P = \prod_k L_k^P$:

$$L_k^P = \exp(\eta_{i_k}) / \sum_{i \in \widetilde{R}(t_k)} \exp(\eta_i),$$

where $\widetilde{R}(t_k)$ = **sampled risk set** at observed event time t_k, containing the case + sampled controls $(t_1 < \cdots < t_D)$

$\Rightarrow$ Fit stratified Cox model, with $\widetilde{R}(t_k)$'s as the strata.

$\Leftrightarrow$ **Conditional logistic regression**
– function `clogit()` in `survival`, wrapper of `coxph()`.

Modelling case-cohort data

Cox's PH model $\lambda_i(t) = \lambda_0(t)\exp(\eta_i)$ again, but ...

- Analysis of survival data relies on the theoretical principle that *you can't know the future*.
- Case-cohort sampling breaks this principle:
  cases are sampled based on what *is known* to be happening to them during follow-up.
- The union of cases and subcohort is a mixture
  1. random sample of the population, and
  2. "high risk" subjects who are *certain* to become cases.

$\Rightarrow$ Ordinary Cox partial likelihood is wrong.

- Overrepresentation of cases must be corrected for, by (I) **weighting**, or (II) **late entry method**.

Correction method I – weighting

The method of **weighted partial likelihood** borrows some basics ideas from survey sampling theory.

- Sampled risk sets
  $\widetilde{R}(t_k) = \{\text{cases}\} \cup \{\text{subcohort members}\}$ at risk at t_k.
- Weights:
  - $w = 1$ for all cases (within and outside the subcohort),
  - $w = N_{\text{non-cases}}/n_{\text{non-cases}}$ = inverse of sampling-fraction f for selecting a non-case to the subcohort.
- Function `coxph()` with option `weights = w` would provide consistent estimation of β parameters.
- However, the SEs must be corrected!
- R solution: Function `cch()` – a wrapper of `coxph()` – in package `survival`, with `method = "LinYing"`.

Comparison of NCC and CC designs

- Statistical efficiency
  Broadly similar in NCC and CC with similar numbers of cases and controls.
- Statistical modelling and valid inference
  Straightforward for both designs with appropriate software, now widely available for CC, too
- Analysis of outcome rates on several time scales?

NCC: Only the time scale used in risk set definition can be the time variable t in the baseline hazard of PH model.

CC: Different choices for the basic time in PH model possible, because subcohort members are not time-matched to cases.

Comparison of designs (cont'd)

- Missing data

  NCC: With close 1:1 matching, a case-control pair is lost, if either of the two has data missing on key exposure(s).

  CC: Missingness of few data items is less serious.

- Quality and comparability of biological measurements

  NCC: Allows each case and its controls to be matched also for analytic batch, storage time, freeze-thaw cycle, $\rightarrow$ better comparability.

  CC: Measurements for subcohort performed at different times than for cases $\rightarrow$ differential quality & misclassification.

- Possibility for studying many diseases with same controls

  NCC: Complicated, but possible if matching is not too refined.

  CC: Easy, as no subcohort member is "tied" with any case.

Conclusion

- "Case-controlling" is very cost-effective.
- Case-cohort design is useful especially when several outcomes are of interest, given that the measurements on stored materials remain stable during the study.
- Nested case-control design is better suited *e.g.* for studies involving biomarkers that can be infuenced by analytic batch, long-term storage, and freeze-thaw cycles.
- Matching helps in improving effciency and in reducing bias – but only if properly done.
- Handy R tools are available for all designs.

Causal Inference 2: Model-based estimation of causal contrasts

Tuesday, 6 June, 2023

Esa Läärä

Statistical Practice in Epidemiology with **R**
2 to 7 June, 2023
University of Tartu, Estonia

Outline

- Causal questions
- Factual risks and associational contrasts
- Causal estimands: contrasts of counterfactual quantities
- Marginal and conditional contrasts, effect among treated, etc.
- Outcome regression models, standardization or g-formula
- Exposure modelling, propensity scores and weighting
- Double robust estimators and machine learning algorithms
- Time-to-event outcomes: hazards of hazard ratios and estimation of causal contrasts of cumulative risks.

Some literature

- Austin & Stuart (2015) *Stat Med* 34(28):3661-3679.
- Funk et al. (2011) *Am J Epidemiol* 173(7):761-767
- Hernan & Robins (2020). *Causal Inference: What if*. CRC Press.
- Luque Fernandez et al. (2018) *Stat Med* 2018;37(16):2530-2546
- Schuler & Rose (2017) *Am J Epidemiol* 185(1):65-73.
- Sjölander (2016) *Eur J Epidemiol* 31:563-574
- Smith et al. (2022) *Stat Med* 2022;41(2):407-432.
- Zhou et al. (2022) `PSweight` vignette.

Causal question in PECOT format & Example

P **Population**: 2900 women with breast cancer (Rotterdam study)

E **Exposure**: Hormonal treatment (HT)

C **Comparator**: Placebo, no HT

O **Outcome**: Recurrence or death

T **Time frame**: 10 y from surgery to outcome

Causal questions of interest – comparisons of counterfactuals:

- What is the 10-year risk π^1 of the outcome, if everybody in P were exposed to HT, as compared with π^0, the risk if nobody were exposed?
- What is the 10-year risk π_1^1 of the outcome, among those in P, who are factually exposed to HT, as compared with the risk π_1^0, if they were not exposed?

Risks by factual exposure and their associational contrasts

- Let Y be a binary indicator (1/0) for the outcome to occur within fixed risk period (assuming no censoring, nor competing events), and X be an exposure variable or risk factor.
- Let π_x = risk of the outcome to occur during the period in the subset of the target population factually exposed to level $X = x$:

$$\pi_x = P\{Y = 1 \mid X = x\} = E(Y|X = x).$$

- For simplicity, let X be binary, too: exposed vs. unexposed .
- Common **associational contrasts** of risks between exposure groups:
  - **Risk difference** $\tau = \pi_1 - \pi_0 = E(Y|X = 1) - E(Y|X = 0)$,
  - **Risk ratio** $\phi = \pi_1/\pi_0$,
  - **Risk odds ratio** $\psi = \frac{\omega_1}{\omega_0} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$.

Conditional associational contrasts

- The associational quantities above were **marginal**; not conditioned on (or stratified by) any **covariate** – such as sex, age, etc.
- Let now Z be a covariate (can be multivariable) and

$$\pi_{xz} = P\{Y = 1 \mid X = x, Z = z\} = E(Y|X = x, Z = z)$$

be the risk of outcome during risk period in a population group where both $X = x$ and $Z = z$, $x = 0, 1$.

- **Conditional associational contrasts** between exposed and unexposed among those with $Z = z$.
  - $\tau_z = \pi_{1z} - \pi_{0z}$ is the risk difference conditional on $Z = z$, i.e. z-**specific** risk difference.
  - $\phi_z = \pi_{1z}/\pi_{0z}$ and $\psi_z = \pi_{1z}(1 - \pi_{1z})/[\pi_{0z}(1 - \pi_{0z})]$ are the z-specific risk ratio and odds ratio, respectively.

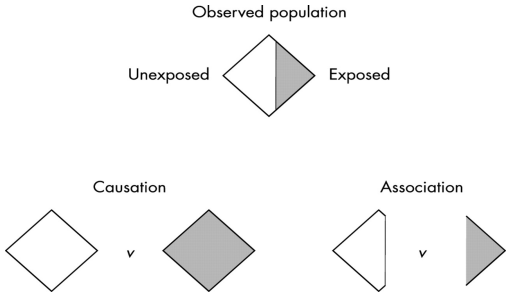
Example: Single binary covariate Z

- Let the prevalence of exposure be $P\{X = 1\} = 0.45$ in the population
- Let $P\{Z = 1\} = 1 - P\{Z = 0\} = 0.40$ in the population and
  $P\{Z = 1|X = 1\} = 0.667$ and $P\{Z = 1|X = 0\} = 0.182$
- Let also factual risks $\pi_{xz} = P\{Y = 1|X = x, Z = z\}$ $(x, z = 0, 1)$ by X and Z be as shown in the cells of the table below :

	$Z = 1$	$Z = 0$	π_x (obtained by formula of total probability)
$X = 1$	0.50	0.20	$\pi_1 =$ **0.40** (0.50× 0.667 + 0.20× 0.333)
$X = 0$	0.25	0.10	$\pi_0 =$ **0.13** (0.25× 0.182 + 0.10× 0.818)
Contrasts	$\tau_1 =$0.25	$\tau_0 =$ 0.10	$\tau =$ **0.27**

- Marginal risks, π_1, π_0, contrast $\tau = \pi_1 - \pi_0$, and conditional contrasts $\tau_z = \pi_{1z} - \pi_{0z}$ are shown in table margins.

Associational and causal contrasts



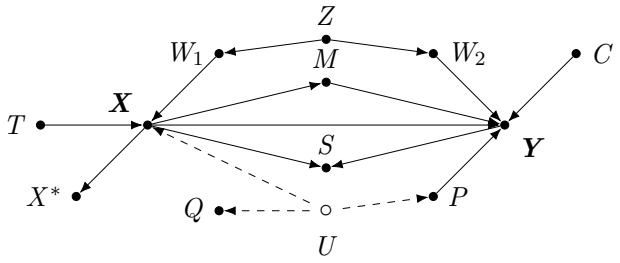
- **Associational**: Contrast of risks between the **subsets** of the population determined by the subjects' **factual** exposure value.
- **Causal**: Contrast of risks in the **entire population** under the alternative **potential** or **counterfactual** exposure values;
  see Hernan (2004), Hernan & Robins (2006), H&R (2020)

Causal estimands: contrasts of counterfactual risks

- Let $Y^{X=x} = Y^x$ indicate (1/0) the event to occur within the risk period, **if** exposure X were – **counterfactually** – forced to value x in the whole target population.
- The **counterfactual** risk if everybody had exposure level $X = x$
  $$\pi^x = P\{Y^{X=x} = 1\} = E(Y^{X=x}).$$
- **Marginal causal contrasts** of risk
  - risk difference (RD) $\tau^c = \pi^1 - \pi^0 = P\{Y^{X=1} = 1\} - P\{Y^{X=0} = 1\}$,
  - risk ratio (RR) $\phi^c = \pi^1/\pi^0$,
  - risk odds ratio (OR) $\psi^c = [\pi^1/(1 - \pi^1)]/[\pi^0/(1 - \pi^0)]$,

NB. Alternative notation: Judea Pearl's (2010) **do-operator**
$$P\{Y = 1|\text{do}(X = x)\} = P\{Y^{X=x} = 1\}.$$

Identifying causal contrasts from causal diagram



- **Causal paths** $X \to Y$ and $X \to M \to Y$: Don't block!
- **Non-causal paths** between X and Y: Block!
  – If already blocked, don't open (e.g. by conditioning on $\boxed{S}$).
- **Backdoor paths** $X \leftarrow W_1 \leftarrow Z \to W_2 \to Y$ and $X \leftarrow U \to P \to Y$: Block with minimal effort. – **Sufficient sets**: P plus one from $\{Z, W_1, W_2\}$. – If P unobserved, substitute by Q, proxy of U.
- No need to adjust for T. – Adjusting for C can improve precision.

Identifying causal contrasts from causal diagram

- Let Z' be a set of observed covariates that are **non-descendants** of X
- If $Z \subset Z'$ were sufficient to **block** all **open non-causal paths** btw X and Y, then counterfactuals are identified by **standardization** – or **g-formula**:

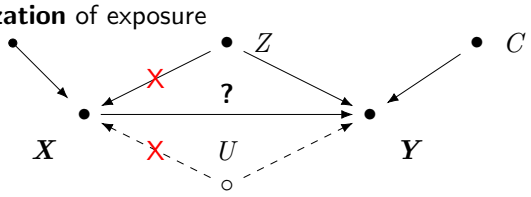
$$\begin{aligned}\pi^x &= E(Y^{X=x}) = E_Z[E_Y(Y|X=x,Z)] \\ &= \sum_z P\{Y=1 \mid X=x, Z=z\}P\{Z=z\}, \quad \text{for discrete } Z.\end{aligned}$$

- Causal contrasts τ^c, ϕ^c, ψ^c are obtained from π^1 and π^0 thus derived.
- If there are open paths btw X and Y, e.g. via unmeasured confounders U, the causal contrasts are not identified $\Leftrightarrow$ **residual confounding**.
- If X is **randomized**, then $X \perp\!\!\!\perp Z \cup U$, and it holds simply

$$\pi^x = P\{Y^{X=x}=1\} = P\{Y=1 \mid X=x\} = \pi_x, \ \forall x,$$

Randomized study and causal diagram

R = **Randomization** of exposure



- When $X \equiv R$, no arrow points to X, and X is independent of $Z, U, \ldots$, measured and unmeasured.

$\Rightarrow$ No confounding!

$\Rightarrow$ Estimation of causal effect: unadjusted, crude comparison is enough.

- Precision may be improved by including Z and C as covariates.
- Often realized exposure X is affected by Z and U, thus differing from R. Then, R may be utilized as an **instrumental variable**.

Example (cont'd): Single binary common cause Z

- Causal diagram $X \to Y$, $X \leftarrow Z \to Y$; classical confounding triangle.
- Counterfactual risks (from items on slide 6) are obtained by g-formula $\pi^x = \sum_z \pi_{xz} P\{Z = z\}$ with weights from total population:

$$\pi^1 = 0.50 \times 0.4 + 0.20 \times 0.6 = \mathbf{0.32},$$
$$\pi^0 = 0.25 \times 0.4 + 0.10 \times 0.6 = \mathbf{0.16}$$

- Marginal causal contrasts (vs. associational ones)

$$\tau^c = 0.32 - 0.16 = \mathbf{0.16} \neq 0.27 = 0.40 - 0.13 = \tau,$$
$$\phi^c = 0.32/0.16 = \mathbf{2} \neq 3.14 = 0.40/0.13 = \phi,$$
$$\psi^c = \frac{0.32/(1-0.32)}{0.16/(1-0.16)} = \mathbf{2.47} \neq 4.57 = \psi.$$

- Associational contrasts were clearly confounded by Z.

Conditional causal contrasts

- With covariate Z, counterfactual z-specific risks are defined

$$\pi_z^x = P\{Y^{X=x} = 1 \mid Z = z\}, \quad \text{for all } z \text{ and } x = 0, 1.$$

- These have their own identifiability conditions.
- **Conditional** or z**-specific causal contrasts** of risks are, for instance

$$\tau_z^c = \pi_z^1 - \pi_z^0 = P\{Y^{X=1} = 1 \mid Z = z\} - P\{Y^{X=0} = 1 \mid Z = z\},$$
$$\phi_z^c = \pi_z^1 / \pi_z^0 = P\{Y^{X=1} = 1 \mid Z = z\} / P\{Y^{X=0} = 1 \mid Z = z\}$$

- If τ_z^c has the same value for all z, the risk difference is **homogenous**. Otherwise it is **heterogenous** or **modified** by Z.
- These concepts are defined similarly for risk ratio and odds ratio.
- Homogeneity of one type of contrast implies heterogeneity of other types.

Causal contrasts in factual exposure groups

- Causal risk difference **among exposed** is defined

$$\tau_1^c = P\{Y^{X=1} = 1 \mid X = 1\} - P\{Y^{X=0} = 1 \mid X = 1\},$$

  also known as **average treatment effect among treated** (ATT).
  – The contrast **among unexposed** (ATU) is analogously defined.
- The effect often heterogenous, and groups noncomparable.
- If Z is a sufficient set, g-formulas for identifying these are

$$\text{ATT} = \pi_1 - \sum \pi_{0z} P\{Z = z | X = 1\} = \text{“observed} - \text{expected”},$$
$$\text{ATU} = \sum \pi_{1z} P\{Z = z | X = 0\} - \pi_0 = \text{“expected} - \text{observed”}.$$

- Different standard populations for ATT, ATU, and for marginal contrast, a.k.a. **average treatment effect in the whole population**:

$$\text{ATE} = \tau^c = \pi^{X=1} - \pi^{X=0} = \sum_z \pi_{1z} P\{Z = z\} - \sum_z \pi_{0z} P\{Z = z\}.$$

Example: Single binary Z (cont'd)

- z-specific risks, marginal assoc. & causal contrasts are on slides 6 & 11.
- For ATT, we have the observed risk $\pi_1^1 = \pi_1 = \mathbf{0.40}$, and the expected risk is $\pi_1^0 = \sum_z \pi_{0z} P\{Z = z|X = 1\} = 0.25{\times}0.667 + 0.10{\times}0.333 = \mathbf{0.20}$, so ATT $= 0.40 - 0.20 = \mathbf{0.20}$.
- For ATU, the expected risk is $\pi_0^1 = \sum_z \pi_{1z} P\{Z = z|X = 0\} = 0.50{\times}0.182 + 0.20{\times}0.818 = \mathbf{0.26}$, the observed risk is $\pi_0^0 = \pi_0 = \mathbf{0.13}$, and ATU $= 0.26 - 0.13 = \mathbf{0.13}$.
- Here, the causal risk difference is bigger among exposed. – Being exposed seems to be a modifier of the effect of exposure on this scale!
- Interestingly, the causal risk ratio $= 2$ is homogenous.

NB Popular design for estimating ATT: **matched cohort study**.

Outcome regression modelling (see lecture on Saturday)

Modelling how expected values, risks, hazards, etc. depend on exposure X and covariates Z (modifiers, and/or confounders). – Common elements:

- Each subject i $(i = 1, \ldots, n)$ has an own **profile**, i.e. vector (x_i, z_i^T) of values of X and covariates Z.
- In the spirit of **generalized linear models**, let vector $(\alpha, \beta, \gamma^\mathsf{T})$ contain regression coefficients, and specify the **linear predictor** – assuming so far no **interactions**, nor **effect modifications**

$$\eta_i = \alpha + \beta x_i + \gamma^\mathsf{T} z_i$$

- **Product terms** can be added for interactions and modifications if needed, and **splines** may be used for continuous covariates.
- Further model specification depends on the type of outcome variable, causal contrasts of interest, and importance and choice of time scale(s).

Binary outcome model and classical causal estimation

- Basic outcome regression model for risks π in fixed risk periods:

$$g\{\pi(x_i)\} = \alpha + \beta x_i + \gamma^\mathsf{T} z_i, \quad i = 1, \ldots, N.$$

- Link $g(\cdot)$ and causal interpretation of β, assuming the validity of model (including homogeneity or non-modification of the contrast in question) and that Z blocks all backdoor paths:
  - id $\Rightarrow \beta =$ risk difference (RD) τ^c for $X = 1$ vs. $X = 0$, adjusted for Z
  - log $\Rightarrow \beta =$ log of risk ratio (RR) ϕ^c – " –
  - logit $\Rightarrow \beta =$ log of conditional risk odds ratio (OR), ψ_z^c, – " – **NB.** This is different from marginal OR due to **non-collapsibility**.
- Random component: Binomial family – Fitting: `glm()`
- Problems with id & log in keeping predicted $\widehat{\pi}(\cdot)$ between 0 and 1.

Modern approach: Causal contrasts by g-formula

- Assuming that Z is sufficient to block non-causal paths, a logistic model is fitted, which may even contain product terms allowing modification
  $\text{logit}(\pi_i) = \log[\pi_i/(1-\pi_i)] = \alpha + \beta x_i + \gamma^{\mathsf{T}} z_i + \delta^{\mathsf{T}}(x_i z_i), \quad i = 1, \ldots, n.$
- For each individual i, **predicted risks** are computed for both possibilities of exposure: $X = 0$ and $X = 0$, but keeping $Z = z_i$ as it is
  $\widetilde{\pi}_i^{X_i=x} = \text{expit}\{\widehat{\alpha} + \widehat{\beta} x + \widehat{\gamma}^{\mathsf{T}} z_i + \widehat{\delta}^{\mathsf{T}}(x z_i)\}, \quad x = 0, 1.$
- Marginal counterfactual risks for $x = 1, 0$ are estimated applying **g-formula**:
  $$\widehat{\pi}^{X=x} = \widehat{E}_Z[E(Y|X=x, Z)] = \frac{1}{n}\sum_{i=1}^{n} \widetilde{E}(Y_i|X_i = x, Z = z_i) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\pi}_i^{X_i=x}$$
  as the data provide a non-parametric estimate of the joint distribution of Z.
- **Estimators marginal causal contrasts** of risks are now, e.g.
  $\widehat{\tau}^c = \widehat{\pi}^{X=1} - \widehat{\pi}^{X=0}, \qquad \widehat{\psi}^c = [\widehat{\pi}^1/(1-\widehat{\pi}^1)]/[\widehat{\pi}^0/(1-\widehat{\pi}^0)]$

Exposure modelling, propensity scores and weighting

Let X be a binary exposure variable. Assume again that Z is a sufficient set

- **Exposure model** predicting individual X_i:s by confounders is fitted
  $\text{logit}[P\{X_i = 1|Z = z_i\}] = \alpha^* + z_i^{\mathsf{T}}\gamma^*, \quad i = 1, \ldots, N.$
- **Propensity scores** PS_i, or fitted probabilities of exposure are obtained
  $\text{PS}_i = \widehat{P}\{X_i = 1|Z = z_i\} = \text{expit}(\widehat{\alpha}^* + z_i^{\mathsf{T}}\widehat{\gamma}^*).$
- Individual **weights** $W_i = w(\text{PS}_i, X_i)$ are computed (see next slide).
- Counterfactual risks are estimated as weighted averages of the outcome in the two exposure groups
  $$\widehat{\pi}^{X=x} = \frac{\sum_{i=1}^{n} \mathbf{1}_{\{X_i=x\}} W_i Y_i}{\sum_{i=1}^{n} \mathbf{1}_{\{X_i=x\}} W_i} = \frac{\sum_{X_i=x} W_i Y_i}{\sum_{X_i=x} W_i}, \quad x = 0, 1$$
- From these, marginal causal contrasts are estimated as before.

Exposure modelling, propensity scores and weighting (cont'd)

- **Inverse probability weights** (IPW) are used to estimate marginal causal contrasts (like ATE) in the whole population. They are based on inverses of the fitted probabilities of belonging to the realized exposure group:
  $$W_i = w(\text{PS}_i, X_i) = \frac{\mathbf{1}_{\{X_i=1\}}}{\text{PS}_i} + \frac{\mathbf{1}_{\{X_i=0\}}}{1-\text{PS}_i}, \quad i = 1, \ldots, n.$$
- If the interest is on causal contrasts among the treated (like ATT), the **treated weights** are used:
  $W_i = 1$ for $X_i = 1$, and $W_i = \text{PS}_i/(1-\text{PS}_i)$ for $X_i = 0$.
- Other: **overlap weights**, **matching weights**, **entropy weights**.
- The goodness-of-fit of the exposure model needs to be assessed.
  For that purpose, various measures of **covariate balance** are developed.

Double robust (DR) estimators and machine learning methods

- The validity of estimator utilizing g-formula or PS-based weighting depends on, how accurately the outcome model or exposure model is specified.
- **Double robust** (DR) estimation of causal contrasts:
  Combination of g-formula and IPW. – Alternatives
  - **Augmented IPW** (AIPW); see Jonsson Funk et al. (2011),
  - **Targeted maximum likelihood estimation** (TMLE);
    see Schuler & Rose (2017), Luque-Fernandez et al. (2018)

  Validity of a DR estimator requires that either the exposure model or the outcome model (or both) is correctly specified.

- Algorithms developed for **supervised learning** increase flexibility in modelling both outcome and exposure (see Bi et al. 2019, Blakely et al. 2020).

Interim conclusions

- Careful specification of causal question and estimands needed.
- Adjustment for confounding via efficient blocking of backdoor paths.
- Basic estimation methods: outcome regression & g-formula, exposure modelling & PS-weighting, double robust estimators.
- Sufficiently flexible models desirable to reduce misspecification bias.
- Statistical inference (ignored here): robust covariance matrix & delta method, bootstrapping, efficient influence curve, etc.
- We also limited to time-fixed exposure (binary) and confounders. Extensions exist for polytomous exposure, time-varying exposure and confounding.
- **Warning**: There can still remain open non-causal paths between X and Y inducing residual confounding and/or selection bias.

Time-to-event outcomes: associational hazard quantities

- Let T = time to outcome event from a defined zero time, and $Y(t) = \mathbf{1}_{\{T \leq t\}}$ indicator (1/0) for the outcome to occur by t.
- The risk of outcome by t conditional on exposure level x
$$\pi_x(t) = P\{Y(t) = 1 | X = x\} = P\{T \leq t | X = x\}, \quad x = 0, 1.$$
- The **hazard** of outcome at t for those exposed to level $X = x$ be
$$\lambda_x(t) = \lambda(t \mid X = x) = \lim_{h \to 0} \frac{P\{Y(t+h) = 1 \mid X = x\}/h}{P\{Y(t) = 1 \mid X = x\}}.$$
- Common associational contrasts:
  - **Hazard difference** $\delta(t) = \lambda_1(t) - \lambda_0(t)$,
  - **Hazard ratio** $\rho(t) = \lambda_1(t)/\lambda_0(t)$.
    This is often assumed constant ρ – as in **Cox regression**.

Causal contrasts of hazards

- Let $T^{X=x} = T^x$, be time to event, and $Y^{X=x}(t) = Y^x(t) = \mathbf{1}_{\{T^x \leq t\}}$ indicate the event occurring during risk period $(0, t]$, if exposure X were forced to value x in the whole target population.
- The **counterfactual hazard**, if everybody were exposed to $X = x$:

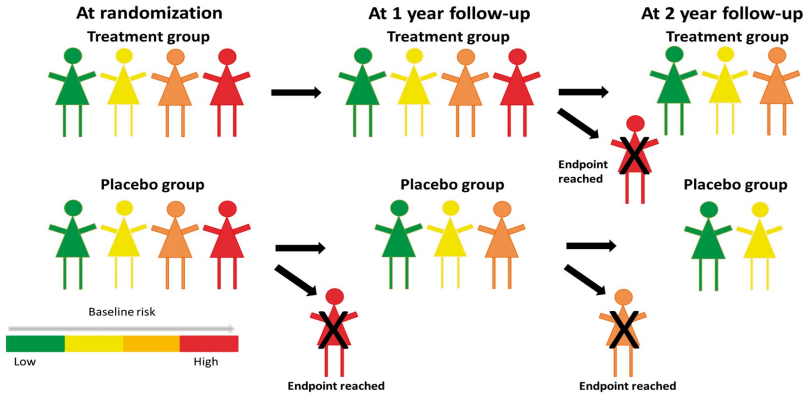
$$\lambda^x(t) = \lim_{h \to 0} \frac{1}{h} \frac{P\{Y^{X=x}(t+h) = 1\}}{P\{Y^{X=x}(t) = 0\}}, \quad x = 1, 0.$$

- Marginal causal contrasts:
  hazard difference (HD) $\lambda^1(t) - \lambda^0(t)$, and hazard ratio (HR) $\lambda^1(t)/\lambda^0(t)$.
- If X is randomized, these are identified by corresp. assoc. contrasts.
- Yet, hazard at any t is conditional on survival by t. If X has any effect, $Y^1(t) = 0$ and $Y^0(t) = 0$ imply different populations at risk for $t > 0$.

$\Rightarrow$ Even if exposure groups were comparable at $t = 0$, after that they aren't.

- Causal interpretation of HR problematic even in a randomized study.

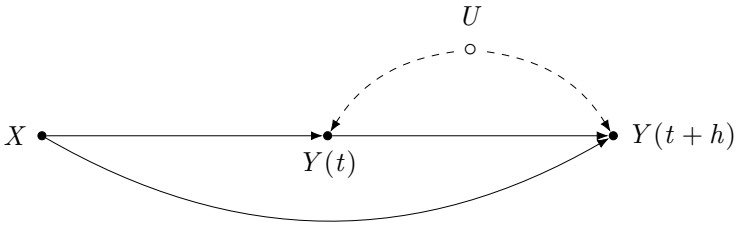
Example: The untreated have a higher hazard (Stensrud et al 2019)



- In the course of time, the prognostic profile of the remaining active treatment group will be worse than that in the remaining placebo group.

Hazard of hazard ratios (Hernan 2010, Aalen et al. 2015)

- The hazard at any time $t > 0$ is affected by known and unknown causes of the outcome $\Rightarrow$ individual **frailty** U varies in the population.



- $Y(t)$ is a **collider** on the path from X to $Y(t+h)$ via U.
  Conditioning on $Y(t) = 0$ opens this non-causal path $\Rightarrow$ **selection bias**.
- The observable hazards may behave strangely over time and lead to conclusions like "HR> 1 before t^* but HR< 1 after that".

Example: WHI Trial on MHT and CHD (Manson et al. 2003)

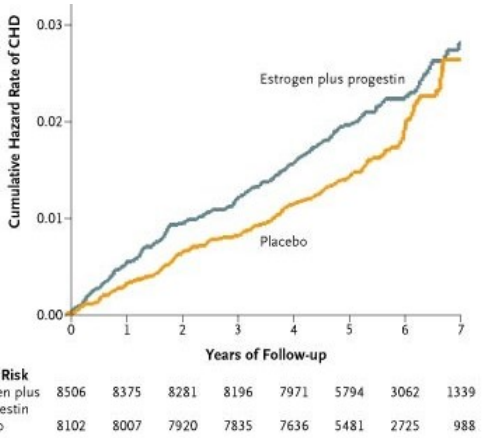
Table 2. Estrogen plus Progestin and the Risk of CHD, According to Year of Follow-up.*

Year of Follow-up	CHD		Hazard Ratio for CHD (95% CI)
	Estrogen-plus-Progestin Group	Placebo Group	
	no. of cases (annualized percentage)		
1	42 (0.50)	23 (0.29)	1.81 (1.09–3.01)
2	38 (0.45)	28 (0.35)	1.34 (0.82–2.18)
3	19 (0.23)	15 (0.19)	1.27 (0.64–2.50)
4	32 (0.39)	25 (0.32)	1.25 (0.74–2.12)
5	29 (0.41)	19 (0.28)	1.45 (0.81–2.59)
≥6	28 (0.37)	37 (0.56)	0.70 (0.42–1.14)

* CHD includes acute myocardial infarction (MI) necessitating hospitalization, silent myocardial infarction as determined by serial electrocardiography, and death due to CHD. There were nine silent myocardial infarctions (four in the estrogen-plus-progestin group and five in the placebo group). Hazard ratios are stratified according to age, presence or absence of a previous coronary event, and randomly assigned diet-modification group and are adjusted for previous coronary-artery bypass grafting or percutaneous transluminal coro-

- Women, 50-79 y, MHT: $N_1 = 8506$, placebo: $N_0 = 8102$
- Followed-up for max 8.6 y, mean 5.6 y.
- Cases & rates/10^4 y $D_1 = 188$, $I_1 = 39$, $D_0 = 147$, $I_0 = 33$.
- Crude IR $= 1.20$, adjusted 1.24 (1.00–1.54)

Example: WHI Trial (cont'd)



- Curves of cumulative hazard approximate the development of cumulative risks $\pi_x(t)$ over time.
- In early years, the curve of MHT runs on top, reflecting higher hazard in that period.
- By 6-7 year, cumulative risks appear to have reached same level.

From hazards to causal contrasts of risk

- A well-specified predictive model for hazards $\lambda(t|x,z)$ (e.g. Cox or Poisson; suitably flexible) can, however, be used to estimate counterfactual cumulative risks $\pi^x(t) = P\{Y^{X=x}(t)\} = P\{T^{X=x} \leq t\}, x = 0, 1.$
- Suppose Z blocks all non-causal paths. Then **counterfactual conditional hazards** $\lambda^x(t|Z=z)$ are identified by observable hazards $\lambda(t|x,z)$
- When no competing events exist, counterfactual z-specific risks $\pi^x(t|Z=z) = P\{Y^{X=x}(t)|Z=z\}$ are identified from factual z-conditional hazards
$$\pi^x(t|Z=z) = 1 - \exp\left\{-\int_0^t \lambda(v|x,z)dv\right\}.$$
- Counterfactual marginal risks are obtained using the g-formula:

$$\pi^x(t) = \sum_z \pi_x(t|Z=z)P\{Z=z\}.$$

Estimation of causal contrasts from time-to-event data

- Various methods to estimate counterfactual risks $\pi^{X=x}(t)$ and their contrasts (see Denz et al. 2023) – For instance

(a) Fit a Cox model $\lambda(t|x_i, z_i) = \lambda_0(t)\exp(\beta x_i + \gamma^T z_i)$, take estimates of coefficients and baseline cumulative hazard $\widehat{\Lambda}_0(t)$ from which:

$$\widetilde{\pi}_i^{X_i=x}(t) = 1 - \exp\{-\widehat{\Lambda}_0(t)\exp(\widehat{\beta}x + \widehat{\gamma}^{\mathsf{T}} z_i)\}.$$

Counterfactuals $\pi^{X=x}(t)$ and contrasts are then estimated by g-formula.

(b) Get weights W_i from an exposure model, fit Cox with "intercept only" specifying X as a `strata()` variable and W_i:s as `weights`, and estimate $\widehat{\pi}^{X=x}(t)$ using `survfit()`, etc.

- Other: IPW Kaplan-Meier, use of pseudo-values, DR methods, ...
- Competing event setting: additional complexities in defining and analysing causal contrasts (see Rudolph et al. 2020, Young et al. 2020).

Multistate models

Bendix Carstensen Steno Diabetes Center Copenhagen
Herlev, Denmark
http://BendixCarstensen.com

SPE, Tartu, Estonia,

June 2023

http://BendixCarstensen.com/SPE

Multistate models

Bendix Carstensen, Martyn Plummer

Multistate models

SPE, Tartu, Estonia,

June 2023

http://BendixCarstensen.com/SPE ms-Markov

Common assumptions in survival analysis

1. Subjects are **either** "healthy" **or** "diseased", with no intermediate state.
2. The disease is **irreversible**, or requires intervention to be cured.
3. The time of disease incidence is known **exactly**.
4. The disease is **accurately** diagnosed.

These assumptions are true for death and many chronic diseases.

A question of definition:
– consider occurrence of **recording of** a given disease

A model for cervical cancer

Invasive squamous cell cancer of the cervix is preceded by cervical intraepithelial neoplasia (CIN)

Purpose of a screening programme is to detect and treat CIN — status of persons obtained at screening dates

Aim of the modeling the transition rates between states, is to be able predict how population moves between states

- Transition rates between states
- Probability of state occupancy

Markov models for multistate processes

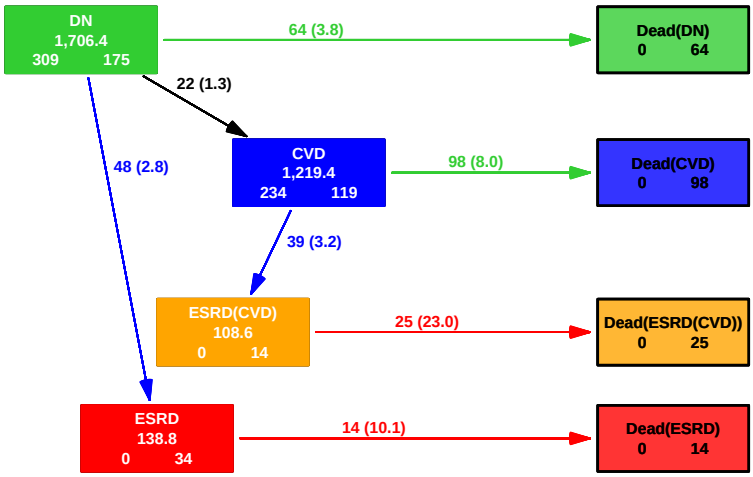
The natural generalization of Poisson regression to multiple disease states:

- transition between states depends **only** on current state
- — this is the **Markov** property
- $\Rightarrow$ transition rates are constant over time
- (time-fixed) covariates may influence transition rates
- the formal Markov property is **very** restrictive
- in the clinical litterature "Markov model" is often used about any type of multistate model

Components of a multistate (Markov) model

- Define the disease states
- Define which transitions between states are allowed
- Select covariates influencing transition rates (may be different between transitions)
- Not a trivial task — do we want *e.g.*
  - cause of death (CVD, Cancer, Other)
  - disease status at death (prev.CVD, prev.Can, neither)

A more complicated multistate model



Likelihood for a multistate model

- The likelihood of the model depends on the probability of being in state j at time t_1, given that you were in state i at time t_0.
- Assume transition rates constant in small time intervals
- $\Rightarrow$ each interval for a person contributes term(s) to the likelihood
- one term for each possible transition between states
- the total likelihood for person p in intervals i is a product of these terms, $d_{pi}\log(\lambda_{pi}) - \lambda_{pi}y_{pi}$
- $\Rightarrow$ each term has the form of the likelihood for a Poisson variate d with mean λy

Likelihood for a multistate model

- each term has the form of the likelihood for a Poisson variate d with mean λy
- terms are **not** independent, but the total likelihood is a product; hence of the same form as the likelihood from independent Poisson variates
- but observations from intervals from one person are neither Poisson nor independent

Realms of multistate modeling

- intensities — dimension time^{-1}
- state probabilities — dimensionless, time^{0}
  integral of intensities w.r.t. to time
- sojourn times — dimension time^{1}
  integral of state probabilities w.r.t. to time

Classes of multistate models

- Markov model: transition between states depends **only** on current state $\Rightarrow$ transition rates are constant
  time-homogeneous Markov model
- If transition rates depend on the **same timescale** only we have a time-**in**homogeneous Markov model
- If transition rates depend on the time since entry to the current state we have a **semi**-Markov model
- If transition rates depend on several timescales we have a general multistate model (there is no name for this)

Computing state probabilities from intensities in multistate models

- time-homogeneous Markov model:
  closed-form formulae exist
- time-**in**homogeneous Markov model:
  closed-form formulae exist (a bit more complicated)
- **semi**-Markov model:
  no closed form formulae exist
- general multistate model:
  no closed form formulae exist

No formulae means that any inference on state probabilities and sojourn times must be based on **simulation** from the model.

Multistate models with Lexis

Bendix Carstensen

Multistate models

SPE, Tartu, Estonia,

June 2023

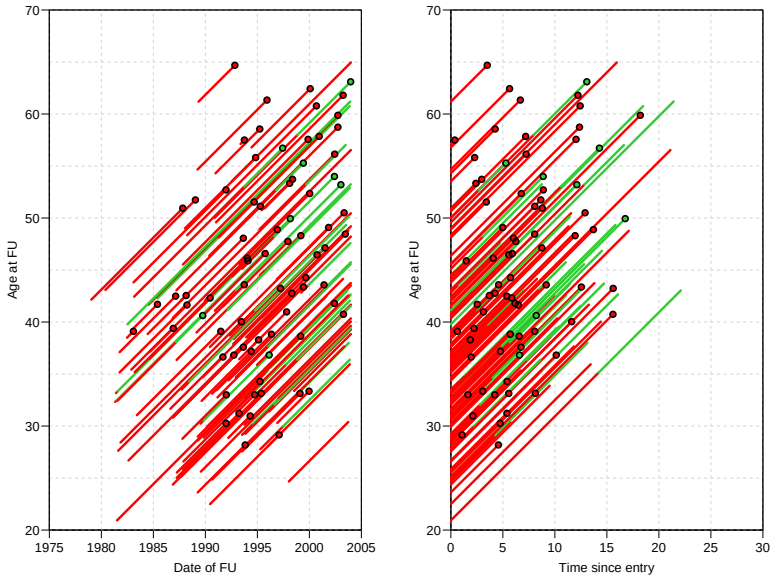
http://BendixCarstensen.com/SPE ms-Lexis

Example: Renal failure data from Steno

Hovind P, Tarnow L, Rossing P, Carstensen B, and Parving H-H: Improved survival in patients obtaining remission of nephrotic range albuminuria in diabetic nephropathy. *Kidney Int.*, 66(3):1180–1186, 2004.

- Endpoint of interest: Death or end stage renal disease (ESRD), i.e. dialysis or kidney transplant.
- 96 patients entering at nephrotic range albuminuria (NRA), i.e. U-alb>300mg/day.
- Is remission from this condition (i.e return to U-alb<300mg/day) predictive of the prognosis?

		Remission	
	Total	Yes	No
No. patients	125	32	93
No. events	77	8	69
Follow-up time (years)	1084.7	259.9	824.8
Cox-model:			
Timescale:	Time since nephrotic range albuminuria (NRA)		
Entry:	2.5 years of GFR-measurements after NRA		
Outcome:	ESRD or Death		
Estimates:	RR	95% c.i.	p
Fixed covariates:			
Sex (F vs. M):	0.92	(0.53,1.57)	0.740
Age at NRA (per 10 years):	1.42	(1.08,1.87)	0.011
Time-dependent covariate:			
Obtained remission:	0.28	(0.13,0.59)	0.001



Features of the analysis

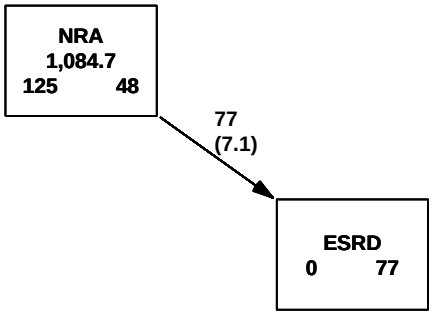
- Remission is included as a time-dependent variable.
- Age at entry is included as a fixed variable.

```
renal[1:5,]
id       dob       doe       dor       dox event
17 1967.944 1996.013        NA 1997.094     2
26 1959.306 1989.535 1989.814 1996.136     1
27 1962.014 1987.846        NA 1993.239     3
33 1950.747 1995.243 1995.717 2003.993     0
42 1961.296 1987.884 1996.650 2003.955     0
```

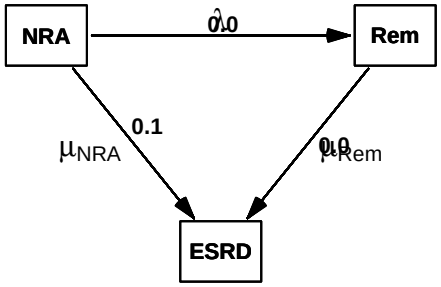
Note patient 26, 33 and 42 obtain remission.


```
> Lr <- Lexis(entry = list(per = doe,
+                          age = doe-dob,
+                          tfi = 0),
+             exit = list(per = dox),
+       exit.status = event>0,
+            states = c("NRA", "ESRD"),
+              data = renal)
> summary(Lr)

Transitions:
     To
From  NRA ESRD  Records:  Events: Risk time:  Persons:
  NRA  48   77       125       77    1084.67       125
```

```
> boxes(Lr, boxpos = list(x = c(25, 75),
+                         y = c(75, 25)),
+           scale.R = 100, show.BE = TRUE )
```



Illness-death model



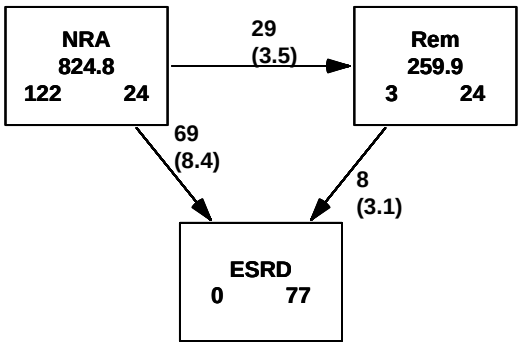
λ: remission rate.
μ_{NRA}: mortality/ESRD rate **before** remission.
μ_{rem}: mortality/ESRD rate **after** remission.

Cutting follow-up at remission: `cutLexis`

```
> Lc <- cutLexis(Lr, cut = Lr$dor,
+                timescale = "per",
+                new.state = "Rem",
+          precursor.states = "NRA")
> summary(Lc)

Transitions:
     To
From  NRA Rem ESRD  Records:  Events: Risk time:  Persons:
  NRA  24  29   69       122       98     824.77       122
  Rem   0  24    8        32        8     259.90        32
  Sum  24  53   77       154      106    1084.67       125
```

Showing states and FU: boxes.Lexis

```
> boxes(Lc, boxpos = list(x = c(15, 85, 50),
+                         y = c(85, 85, 20)),
+           scale.R = 100, show.BE = TRUE)
```

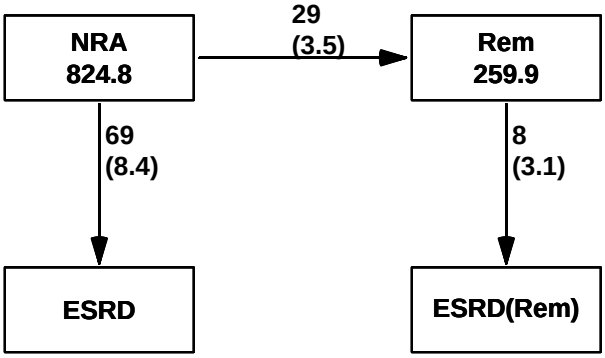


Cutting follow up at events: cutLexis

```
> Lc <- cutLexis( Lr, cut = Lr$dor,
+                 timescale = "per",
+                 new.state = "Rem",
+          precursor.states = "NRA",
+             split.states = TRUE )
> summary( Lc )

Transitions:
     To
From  NRA Rem ESRD ESRD(Rem)  Records:  Events: Risk time:  Persons:
  NRA  24  29   69         0       122      98     824.77       122
  Rem   0  24    0         8        32       8     259.90        32
  Sum  24  53   69         8       154     106    1084.67       125
```

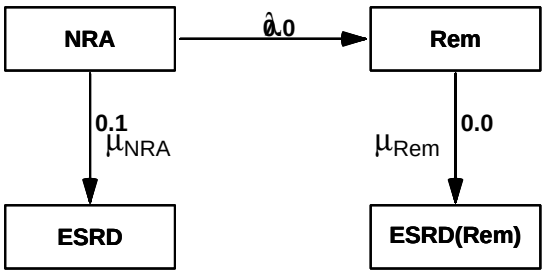

Showing states and FU: boxes.Lexis

```
> boxes(Lc, boxpos = list(x = c(15, 85, 15, 85),
+                         y = c(85, 85, 20, 20)),
+           scale.R = 100)
```



Likelihood for a general MS-model

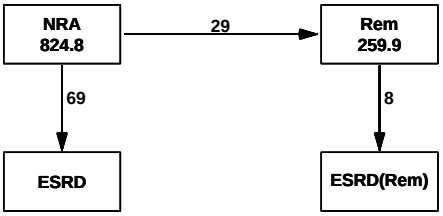
- Product of likelihoods for each transition
  — each one as for a survival model
- **Risk time** is the risk time in the "From" state
- **Events** are transitions to the "To" state
- All other transitions out of "From" are treated as **censorings**
- Possible to fit models separately for each transition



Cox-analysis with remission as time-dependent covariate:

- Ignores λ, the remission rate.
- Assumes μ_{NRA} and μ_{rem} use the same timescale.

Model for all transitions



Cox-model:

- Different timescales for transitions possible
- . . . only one per transition
- No explicit representation of estimated rates.

Poisson-model:

- Timescales can be different
- Multiple timescales can be accomodated simultaneously
- Explicit representation of all transition rates

Calculating state probabilities

$$\begin{aligned}&\mathrm{P}\left\{\text{Remission } \textbf{before} \text{ time } t\right\}\\ &= \int_0^t \lambda(u)\exp\left(-\int_0^u \lambda(s)+\mu_{\mathsf{NRA}}\,\mathrm{d}s\right)\,\mathrm{d}u\end{aligned}$$

$$\begin{aligned}&\mathrm{P}\left\{\text{Being in remission } \textbf{at} \text{ time } t\right\}\\ &= \int_0^t \lambda(u)\exp\left(-\int_0^u \lambda(s)+\mu_{\mathsf{NRA}}(s)\,\mathrm{d}s\right)\times\\ &\qquad\qquad \exp\left(-\int_u^t \mu_{\mathsf{rem}}(s)\,\mathrm{d}s\right)\,\mathrm{d}u\end{aligned}$$

Note μ_{rem} could also depend on u, time since obtained remission.

Sketch of programming, assuming that λ (`lambda`), μ_{NRA} (`mu.nra`) and μ_{rem} (`mu.rem`) are known at any age (stored in vectors)

```
c.rem      <- cumsum(lambda)
c.mort.nra <- cumsum(mu.nra)
c.mort.rem <- cumsum(mu.rem)
pr1 <- cumsum(lambda * exp(-(c.rem + c.mort.nra)))

intgr(t,s) <-
function(t,s){
lambda[s] * exp(-(c.rem[s] + c.mort.nra[s])) *
            exp(-(c.mort.rem[t] - c.mort.rem[s]))
             }
for(t in 1:100) p2[t] <- sum(intgr(t,1:t))
```

If μ_{rem} also depends on time since remission, then `c.mort.rem` should have an extra argument—technically very complicated