Statistical Analysis in the Lexis Diagram:

Age-Period-Cohort models — and some cousins

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

European Doctoral School of Demography, Odense, April 2019

Introduction

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

http://BendixCarstensen/APC/EDSD-2019 intro

Welcome

- Purpose of the course:
  - knowledge about APC-models
  - technical knowledge of handling them
  - insight in the basic concepts of analysis of rates
  - handling observation in the Lexis diagram
- Remedies of the course:
  - Lectures with handouts (BxC)
  - Practicals with suggested solutions (BxC)
  - Assignment for Thursday

Scope of the course

- Rates as observed in populations
  — disease registers for example.
- Understanding of survival analysis (statistical analysis of rates)
  — this is the content of much of the first day.
- Besides concepts, practical understanding of the actual computations (in **R**) are emphasized.
- There is a section in the practicals:
  "Basic concepts of rates and survival"
  — read it; use it as reference.
- If you are not quite familiar with matrix algebra in **R**, there is an intro on the course homepage.

About the lectures

- Please interrupt:
  Most likely I did a mistake or left out a crucial argument.
- The handouts are not perfect
  — please comment on them,
  prospective students would benefit from it.
- Time-schedule:
  Two lectures ($\approx$ 2 hrs)
  one practical ($\approx$ 1 hr)

About the practicals

- You should use you preferred **R**-environment.
- `Epi`-package for **R** is needed, check that you have version `2.35`
- Data are all on the course website.
- Try to make a text version of the answers to the exercises —
  it is more rewarding than just looking at output.
  The latter is soon forgotten — `Rmd` is a possibility.
- An opportunity to learn `emacs`, `ESS` and `Sweave`?

Rates and Survival

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

http://BendixCarstensen/APC/EDSD-2019 surv-rate

Survival data

- Persons enter the study at some date.
- Persons exit at a later date, either dead or alive.
- Observation:
  - Actual time span to death ("event")
  - . . . or . . .
  - Some time alive ("at least this long")

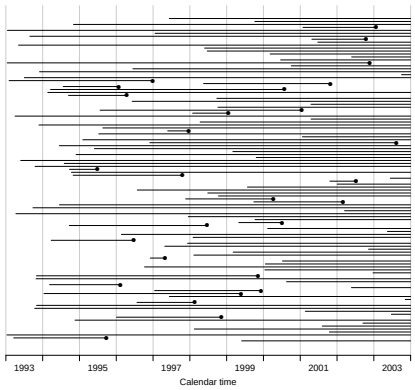
Examples of time-to-event measurements

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

Each line a person

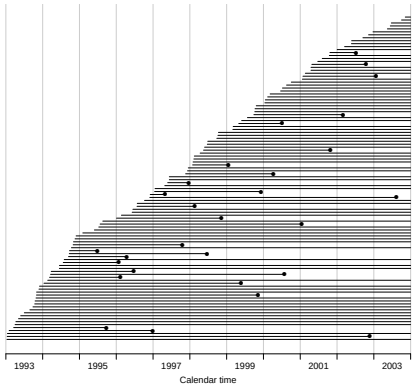
Each blob a death

Study ended at 31 Dec. 2003

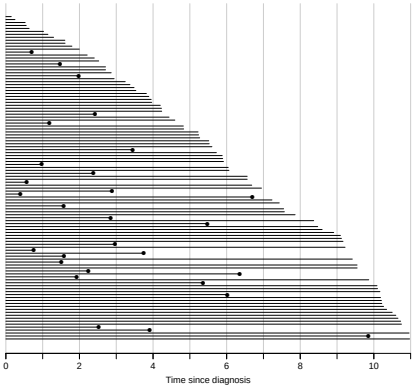


Ordered by date of entry

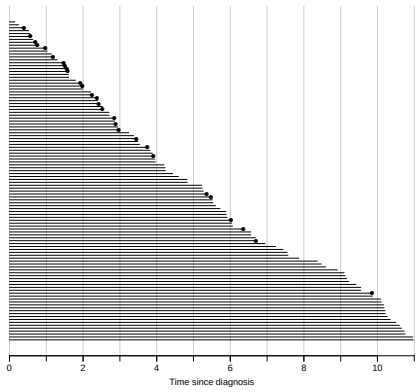
Most likely the order in your database.



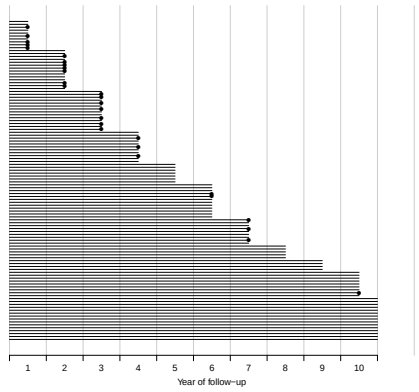
Timescale changed to "Time since diagnosis".



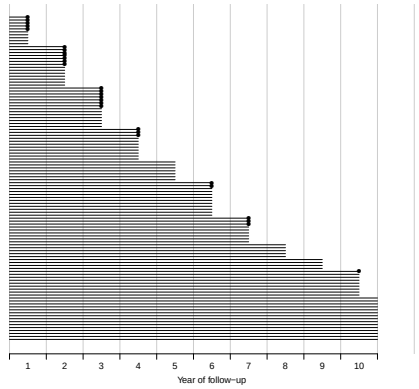
Patients ordered by survival time.



Survival times grouped into bands of survival.



Patients ordered by survival status within each band.



Survival after Cervix cancer

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

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

Estimated 1 year survival is $1 - 0.0465 = 0.9535$ — Life-table estimator.

Survival function

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

How **long** they survive, survival time T — a stochastic variable.

Distribution is characterized by the survival function:

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

Intensity or rate

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

This is the **intensity** or **hazard function** for the distribution.

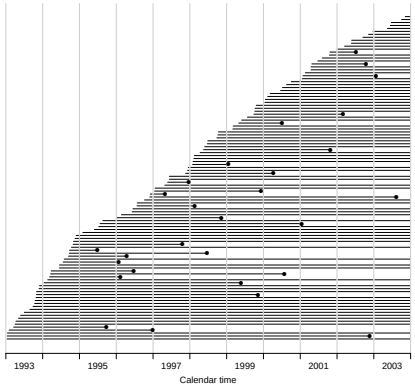
Characterizes the survival distribution as does f or F.

Theoretical counterpart of a **rate**.

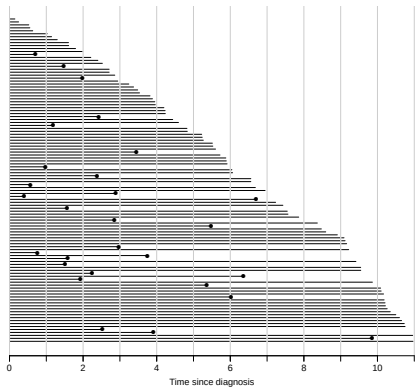
Empirical rates for individuals

- At the **individual** level we introduce the **empirical rate:** (d, y),
  — no. of events $(d \in \{0, 1\})$ during y risk time
- Each person may contribute several empirical (d, y)
- Empirical rates are **responses** in survival analysis
- The timescale is a **covariate**:
  — that varies between empirical rates from one individual:
  Age, calendar time, time since diagnosis
- Do not confuse timescale with
  y — risk time (called exposure in demography)
  a **difference** between two points on **any** timescale

Empirical rates by calendar time.



Empirical rates by time since diagnosis.



Two timescales

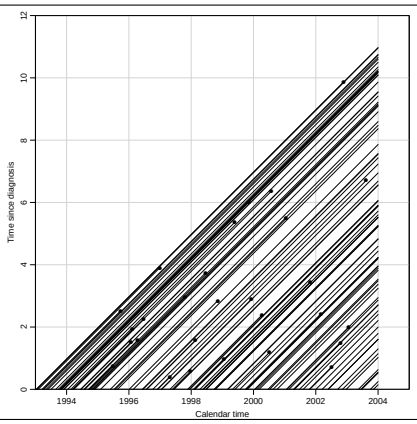
Note that we actually have two timescales:

- Time since diagnosis (*i.e.* since entry into the study)
- Calendar time.

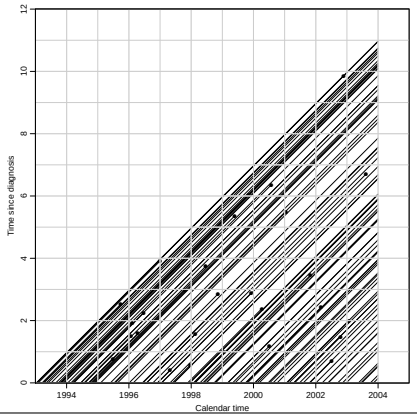
These can be shown **simultaneously** in a Lexis diagram.

Follow-up by calendar time *and* time since diagnosis:

A Lexis diagram!



Empirical rates by calendar time *and* time since diagnosis



So what's the purpose

- form the basis for statistical inference about occurrence rates:
- response: observed rates of events and person-time (d, y)
- covariates:
  - **A**: Age at follow-up
  - **P**: Period (date) of follow-up
  - **C**: (=P−A) Cohort (date of birth)

Likelihood for rates

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019` `likelihood`

Likelihood contribution from one person

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

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

Likelihood contribution from one individual is a **product** of terms.

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

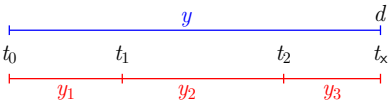
t_i is a **covariate**

Likelihood for an empirical rate

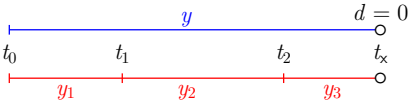
- Likelihood depends on **data** and the **model**
- Model: the rate (λ) is constant in the interval.
- The interval should be sufficiently small for this assumption to be reasonable.

$$\begin{aligned}L(\lambda|y, d) &= \mathrm{P}\{\text{survive } y\} \times \mathrm{P}\{\text{event}\}^d \\ &= \mathrm{e}^{-\lambda y} \times (\lambda\,\mathrm{d}t)^d \\ &= \lambda^d \mathrm{e}^{-\lambda y}\end{aligned}$$

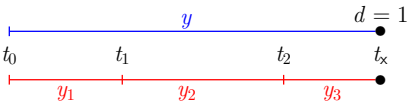
$$\ell(\lambda|y, d) = d\log(\lambda) - \lambda y$$



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



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



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

Aim of dividing time into bands:

- Compute rates in different bands of:
  - age
  - calendar time
  - disease duration
  - ...
- Allow rates to vary along the timescale:

$$\begin{aligned}&0\log(\lambda) - \lambda y_1 \\ &+ 0\log(\lambda) - \lambda y_2 \\ &+ d\log(\lambda) - \lambda y_3\end{aligned} \quad \to \quad \begin{aligned}&0\log(\lambda_1) - \lambda_1 y_1 \\ &+ 0\log(\lambda_2) - \lambda_2 y_2 \\ &+ d\log(\lambda_3) - \lambda_3 y_3\end{aligned}$$

Log-likelihood from more persons

- One person p, different times t: $\sum_t \left(d_{pt}\log(\lambda_t) - \lambda_t y_{pt}\right)$
- More persons: $\sum_p \sum_t \left(d_{pt}\log(\lambda_t) - \lambda_t y_{pt}\right)$
- Collect terms with identical values of λ_t:

$$\sum_t \sum_p \left(d_{pt}\log(\lambda_t) - \lambda_t y_{pt}\right) = \sum_t \left(\left(\sum_p d_{pt}\right)\log(\lambda_t) - \lambda_t \sum_p y_{pt} \right)$$
$$= \sum_t \left(D_t \log(\lambda_t) - \lambda_t Y_t \right)$$

- All events in interval t ("at" time t), D_t
- All exposure time in interval t ("at" time t), Y_t

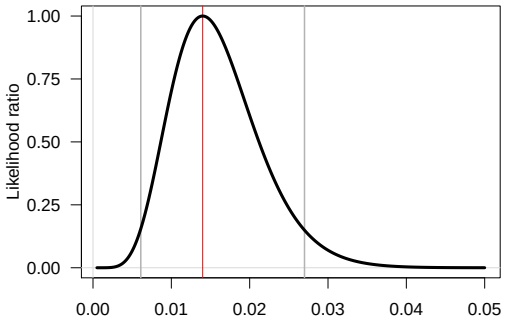
Likelihood example

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

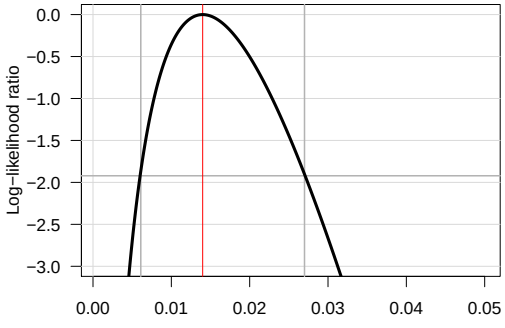
$$\begin{aligned} \mathrm{P}\left\{D = 7, Y = 500|\lambda\right\} &= \lambda^D \mathrm{e}^{\lambda Y} \times K \\ &= \lambda^7 \mathrm{e}^{\lambda 500} \times K \\ &= L(\lambda|\text{data}) \end{aligned}$$

- Best guess of λ is where this function is as large as possible.
- Confidence interval is where it is not too far from the maximum

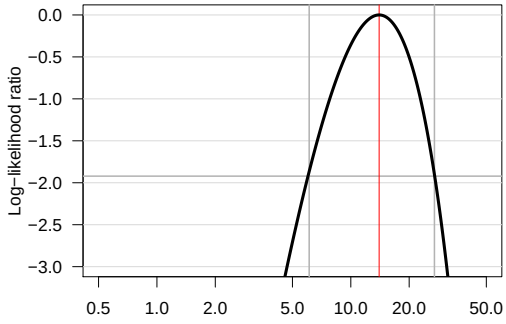
Likelihood-ratio function



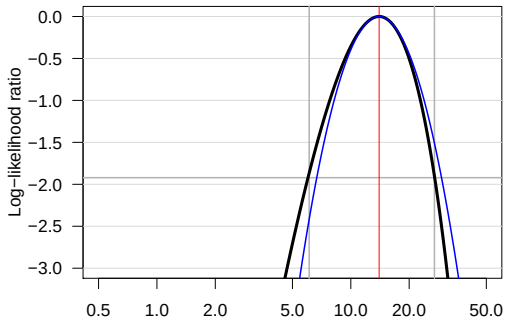
Log-likelihood ratio



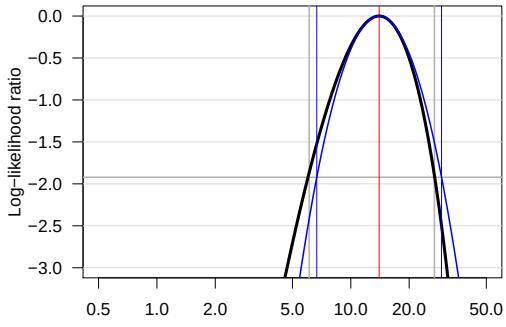
Log-likelihood ratio



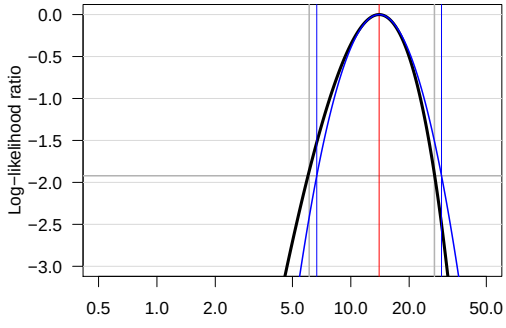
Log-likelihood ratio



Log-likelihood ratio



Log-likelihood ratio



$\hat{\lambda} = 7/500 = 14$

$\hat{\lambda} \overset{\times}{\div} \exp(1.96/\sqrt{7}) = (6.7, 29.4)$

Poisson likelihood

Log-likelihood from **follow-up** of **one individual**, p, in interval t:

$$\ell_{\text{FU}}(\lambda|d,y) = d_{pt}\log\big(\lambda(t)\big) - \lambda(t)y_{pt}, \quad t = 1,\ldots,t_p$$

Log-likelihood from a **Poisson observation** d_{pt} with mean $\mu = \lambda(t)y_{pt}$:

$$\begin{aligned}\ell_{\text{Poisson}}(\lambda y|d) &= d_{pt}\log\big(\lambda(t)y_{pt}\big) - \lambda(t)y_{pt}\\ &= \ell_{\text{FU}}(\lambda|d,y) + d_{pt}\log\big(y_{pt}\big)\end{aligned}$$

Extra term does not depend on the rate parameter λ.

Poisson likelihood

Log-likelihood contribution from **one** individual, p, say, is:

$$\ell_{\text{FU}}(\lambda|d,y) = \sum_t \big(d_{pt}\log\big(\lambda(t)\big) - \lambda(t)y_{pt}\big)$$

- The terms in the sum are **not** independent,
- but the log-likelihood is a **sum** of Poisson-like terms,
- the **same** as a likelihood for **independent** Poisson variates, d_{pt}
- with mean $\mu = \lambda_t y_{pt} \Leftrightarrow \log\mu = \log(\lambda_t) + \log(y_{pt})$

$\Rightarrow$ Analyze rates λ based on empirical rates (d,y) as a Poisson model for independent variates where:

- d_{pt} is the response variable.
- $\log(y_{pt})$ is the offset variable.

Likelihood for follow-up of many subjects

Adding empirical rates over the follow-up of persons:

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

- Persons are assumed independent
- Contribution from the same person are **conditionally** independent, hence give separate contributions to the log-likelihood.
- Follow-up **model** and Poisson **model** are different
- . . . but the **likelihoods** are the same.

The log-likelihood is maximal for:

$$\frac{\mathrm{d}\ell(\lambda)}{\mathrm{d}\lambda} = \frac{D}{\lambda} - Y = 0 \quad \Leftrightarrow \quad \hat\lambda = \frac{D}{Y}$$

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

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

so $I(\hat\theta) = \mathrm{e}^{\hat\theta} Y = \hat\lambda Y = D$, hence $\mathrm{var}(\hat\theta) = 1/D$

Standard error of log-rate: $1/\sqrt{D}$.

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

The log-likelihood is maximal for:

$$\frac{\mathrm{d}\ell(\lambda)}{\mathrm{d}\lambda} = \frac{D}{\lambda} - Y = 0 \quad \Leftrightarrow \quad \hat\lambda = \frac{D}{Y}$$

Information about the rate itself, λ:

$$\ell(\lambda|D,Y) = D\log(\lambda) - \lambda Y \quad \ell'_\lambda = \frac{D}{\lambda} - Y \quad \ell''_\lambda = -\frac{D}{\lambda^2}$$

so $I(\hat\lambda) = D/\hat\lambda^2 = Y^2/D$, hence $\mathrm{var}(\hat\lambda) = D/Y^2$

Standard error of a rate: $\sqrt{D}/Y$.

Confidence interval for a rate

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

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

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

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

Alternatively do the c.i. directly on the rate scale:

$$\lambda \pm 1.96\sqrt{D}/Y$$

Exercise

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

Calculate the mortality rate with a 95% c.i.

Rates with `glm`

```
> library(Epi)
> D <- 17
> Y <- 843.6/1000
> round( ci.exp( glm( D   ~ 1, offset=log(Y), family=poisson ) ), 2 )

            exp(Est.)  2.5% 97.5%
(Intercept)     20.15 12.53 32.42

> round( ci.exp( glm( D/Y ~ 1, weight=    Y , family=poisson ) ), 2 )

            exp(Est.)  2.5% 97.5%
(Intercept)     20.15 12.53 32.42

> round( ci.exp( glm( D/Y ~ 1, weight=    Y , family=poisson(link="identity")),
+               Exp=FALSE), 2 )

            Estimate  2.5% 97.5%
(Intercept)    20.15 10.57 29.73
```

Ratio of two rates

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

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

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

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

Lifetables

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

http://BendixCarstensen/APC/EDSD-2019

Exercise

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

Calculate the rate-ratio between group 1 and 0 with a 95% c.i.

The life table method

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

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

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

Rate-ratio with `glm`

```
> library(Epi)
> D <- c(17,28)
> Y <- c(843.6,632.3)/1000
> F <- factor(0:1)
> round( ci.exp( glm( D   ~ F,     offset=log(Y), family=poisson ) ), 2 )

            exp(Est.)  2.5% 97.5%
(Intercept)     20.15 12.53 32.42
F1               2.20  1.20  4.01

> round( ci.exp( glm( D   ~ F - 1, offset=log(Y), family=poisson ) ), 2 )

   exp(Est.)  2.5% 97.5%
F0     20.15 12.53 32.42
F1     44.28 30.58 64.14
```


The life table method

The life-table method computes survival probabilities for each time interval, in demography normally one year.

The rate is the number of deaths d_i divided by the risk time $(n_i - d_i/2 - l_i/2)\times\ell_i$:

$$\lambda_i = \frac{d_i}{(n_i - d_i/2 - l_i/2)\times\ell_i}$$

and hence the death probability:

$$p_i = 1 - \exp{-\lambda_i\ell_i} = 1 - \exp\left(-\frac{d_i}{(n_i - d_i/2 - l_i/2)}\right)$$

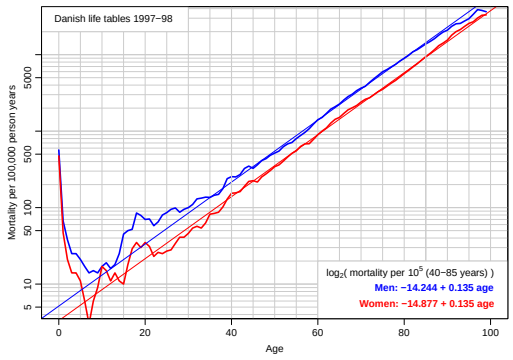
The modified life-table estimator.

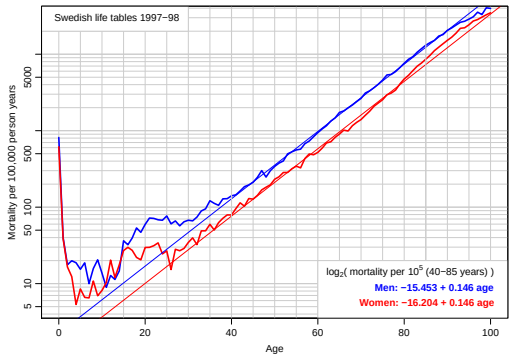
Rate-ratio and -difference with `glm`

```
> round( ci.exp( glm( D/Y ~ F    , weight=Y, family=poisson ) ), 2 )

            exp(Est.)  2.5% 97.5%
(Intercept)     20.15 12.53 32.42
F1               2.20  1.20  4.01

> round( ci.exp( glm( D/Y ~ F    , weight=Y, family=poisson(link="identity")),
+                Exp=FALSE), 2 )

            Estimate  2.5% 97.5%
(Intercept)    20.15 10.57 29.73
F1             24.13  5.14 43.13

> round( ci.exp( glm( D/Y ~ F - 1, weight=Y, family=poisson(link="identity")),
+                Exp=FALSE), 2 )

   Estimate  2.5% 97.5%
F0    20.15 10.57 29.73
F1    44.28 27.88 60.69
```


Population life table, DK 1997–98

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





Practical

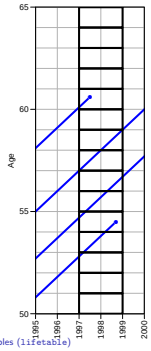
Based on the previous slides answer the following for both Danish and Swedish life tables:

- What is the doubling time for mortality?
- What is the rate-ratio between males and females?
- How much older should a woman be in order to have the same mortality as a man?

Denmark	Males	Females
$\log_2(\lambda(a))$	$-14.244 + 0.135$ age	$-14.877 + 0.135$ age
Doubling time	$1/0.135 = 7.41$ years	
M/F rate-ratio	$2^{-14.244+14.877} = 2^{0.633} = 1.55$	
Age-difference	$(-14.244 + 14.877)/0.135 = 4.69$ years	

Sweden:	Males	Females
$\log_2(\lambda(a))$	$-15.453 + 0.146$ age	$-16.204 + 0.146$ age
Doubling time	$1/0.146 = 6.85$ years	
M/F rate-ratio	$2^{-15.453+16.204} = 2^{0.751} = 1.68$	
Age-difference	$(-15.453 + 16.204)/0.146 = 5.14$ years	

Observations for the lifetable



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

Age-specific rates — cross-sectional!

Survival function:

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

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

Life table approach

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

It is the **population** experience:

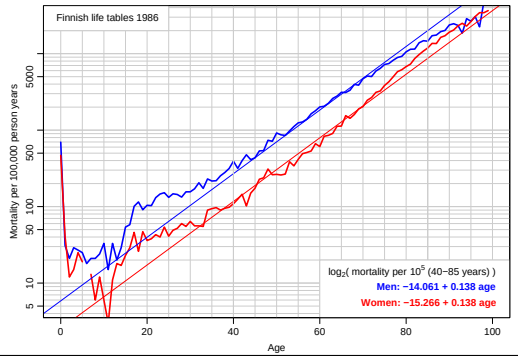
D: Deaths (events).

Y: Person-years (risk time).

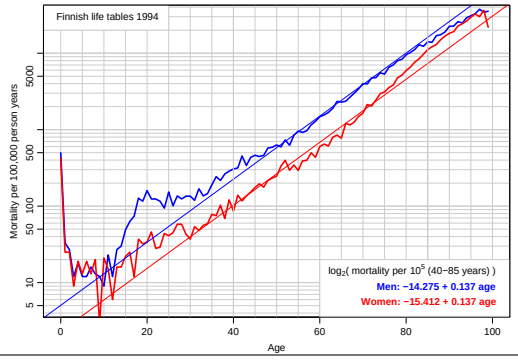
The classical lifetable analysis compiles these for prespecified intervals of age, and computes age-specific mortality **rates**.

Data are collected cross-sectionally, but interpreted longitudinally.

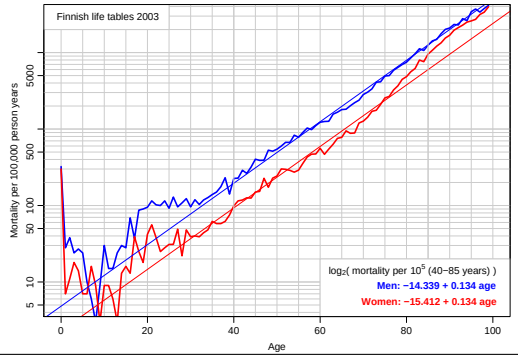
Rates vary over time:



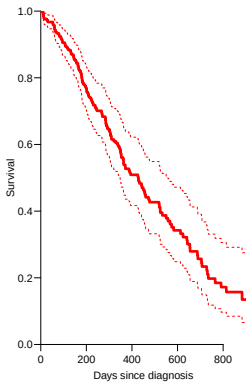
Rates vary over time:



Rates vary over time:



Mayo Clinic lung cancer data: 60 year old woman



Who needs the Cox-model anyway?

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019` `KMCox`

Splitting the dataset a priori

- The Poisson approach needs a dataset of empirical rates (d, y) with suitably small values of y.
- — each individual contributes many empirical rates
- (one per risk-set contribution in Cox-modeling)
- From each empirical rate we get:
  - Poisson-response d
  - Risk time $y \rightarrow \log(y)$ as offset
  - time scale covariates: current age, current date, ...
  - other covariates
- Contributions not independent, but likelihood is a product
- Same likelihood as for independent Poisson variates
- Poisson `glm` with spline/factor effect of time

A look at the Cox model

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

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

Covariates:

- x
- t
- ... often the effect of t is ignored (forgotten?)
- *i.e.* left unreported

Example: Mayo Clinic lung cancer

- Survival after lung cancer
- Covariates:
  - Age at diagnosis
  - Sex
  - Time since diagnosis
- Cox model
- Split data:
  - Poisson model, time as factor
  - Poisson model, time as spline

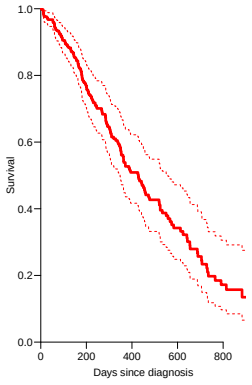
The Cox-likelihood as profile likelihood

- One parameter per death time to describe the effect of time (i.e. the chosen timescale).

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

- Profile likelihood:
  - Derive estimates of α_t as function of data and βs — assuming constant rate between death/censoring times
  - Insert in likelihood, now only a function of data and βs
  - This turns out to be Cox's partial likelihood
- Cumulative intensity $(\Lambda_0(t))$ obtained via the Breslow-estimator

Mayo Clinic lung cancer 60 year old woman



Example: Mayo Clinic lung cancer I

```
> library( survival )
> library( Epi )
> Lung <- Lexis( exit = list( tfe=time ),
+                exit.status = factor(status,labels=c("Alive","Dead")),
+                data = lung )

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

> summary( Lung )

Transitions:
     To
From    Alive Dead  Records:  Events: Risk time:  Persons:
  Alive    63  165       228      165      69593       228
```


Example: Mayo Clinic lung cancer II

```
> system.time(
+ mL.cox <- coxph( Surv( tfe, tfe+lex.dur, lex.Xst=="Dead" ) ~
+                   age + factor( sex ),
+                   method="breslow", data=Lung ) )

   user  system elapsed
  0.037   0.033   0.024

> Lung.s <- splitLexis( Lung,
+                       breaks=c(0,sort(unique(Lung$time))),
+                       time.scale="tfe" )
> summary( Lung.s )

Transitions:
     To
From    Alive Dead  Records:  Events: Risk time:  Persons:
  Alive 19857  165     20022      165      69593       228

> subset( Lung.s, lex.id==96 )[,1:11] ; nlevels( factor( Lung.s$tfe ) )
```


Example: Mayo Clinic lung cancer III

```
      lex.id tfe lex.dur lex.Cst lex.Xst inst time status age sex ph.ecog
9235      96   0       5   Alive   Alive   12   30      2  72   1       2
9236      96   5       6   Alive   Alive   12   30      2  72   1       2
9237      96  11       1   Alive   Alive   12   30      2  72   1       2
9238      96  12       1   Alive   Alive   12   30      2  72   1       2
9239      96  13       2   Alive   Alive   12   30      2  72   1       2
9240      96  15      11   Alive   Alive   12   30      2  72   1       2
9241      96  26       4   Alive    Dead   12   30      2  72   1       2

[1] 186

> system.time(
+ mLs.pois.fc <- glm( lex.Xst=="Dead" ~ - 1 + factor( tfe ) +
+                                age + factor( sex ),
+                                offset = log(lex.dur),
+                      family=poisson, data=Lung.s, eps=10^-8, maxit=25 )
+             )

   user  system elapsed
 12.297  17.636   8.537
```


Example: Mayo Clinic lung cancer IV

```
> length( coef(mLs.pois.fc) )

[1] 188

> t.kn <- c(0,25,100,500,1000)
> dim( Ns(Lung.s$tfe,knots=t.kn) )

[1] 20022     4

> system.time(
+ mLs.pois.sp <- glm( lex.Xst=="Dead" ~ Ns( tfe, knots=t.kn ) +
+                                age + factor( sex ),
+                      offset = log(lex.dur),
+                      family=poisson, data=Lung.s, eps=10^-8, maxit=25 )
+             )

   user  system elapsed
  0.306   0.431   0.209
```

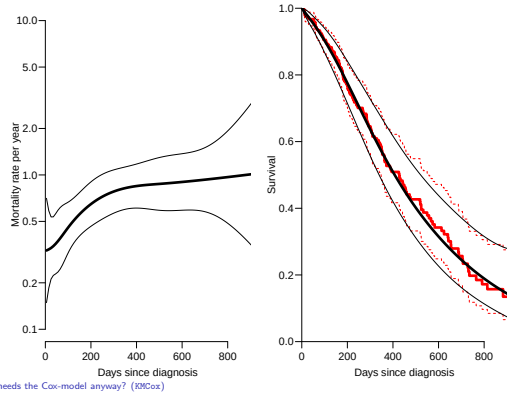

Example: Mayo Clinic lung cancer V

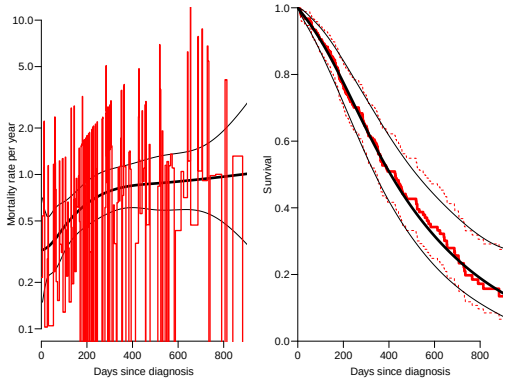
```
> library( mgcv )
> system.time(
+ mLs.pois.ps <- gam( lex.Xst=="Dead" ~ s( tfe ) +
+                                age + factor( sex ),
+                     offset = log(lex.dur),
+                     family=poisson, data=Lung.s, eps=10^-8, maxit=25 )
+             )

   user  system elapsed
  0.808   1.143   0.558

> ests <-
+ rbind( ci.exp(mL.cox),
+         ci.exp(mLs.pois.fc,subset=c("age","sex")),
+         ci.exp(mLs.pois.sp,subset=c("age","sex")),
+         ci.exp(mLs.pois.ps,subset=c("age","sex")) )
> cmp <- cbind( ests[c(1,3,5,7)  ,],
+               ests[c(1,3,5,7)+1,] )
> rownames( cmp ) <- c("Cox","Poisson-factor","Poisson-spline","Poisson-Pspline")
> colnames( cmp )[c(1,4)] <- c("age","sex")
```

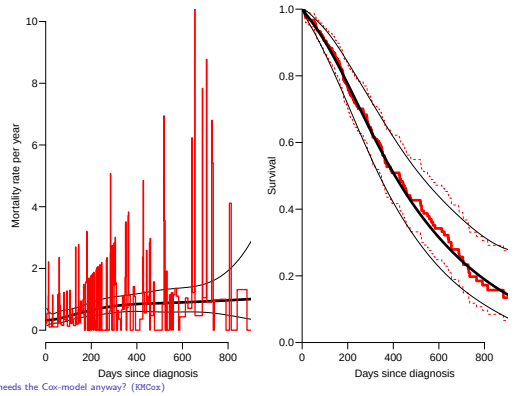

Example: Mayo Clinic lung cancer VI

```
> round( cmp, 7 )
```

	age	2.5%	97.5%	sex	2.5%	97.5%
Cox	1.017158	0.9989388	1.035710	0.5989574	0.4313720	0.8316487
Poisson-factor	1.017158	0.9989388	1.035710	0.5989574	0.4313720	0.8316487
Poisson-spline	1.016189	0.9980329	1.034676	0.5998287	0.4319932	0.8328707
Poisson-Pspline	1.016418	0.9982551	1.034912	0.6032132	0.4345782	0.8372858







Follow-up data

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019`

Deriving the survival function

```
> mLs.pois.sp <- glm( lex.Xst=="Dead" ~ Ns( tfe, knots=t.kn ) +
+                                age + factor( sex ),
+                     offset = log(lex.dur),
+                     family=poisson, data=Lung.s, eps=10^-8, maxit=25 )
```

```
> CM <- cbind( 1, Ns( seq(10,1000,10)-5, knots=t.kn ), 60, 1 )
> lambda <- ci.exp( mLs.pois.sp, ctr.mat=CM )
> Lambda <- ci.cum( mLs.pois.sp, ctr.mat=CM, intl=10 )[,-4]
> survP  <- exp(-rbind(0,Lambda))
```

Code and output for the entire example available in
`http://bendixcarstensen.com/AdvCoh/WNtCMa/`

Follow-up and rates

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

What the Cox-model really is

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

- dividing time very finely and
- modeling one covariate, the time-scale, with one parameter per distinct value.
- the **model** for the time scale is really with exchangeable time-intervals.

$\Rightarrow$ difficult to access the baseline hazard (which looks terrible)

$\Rightarrow$ uninitiated tempted to show survival curves where irrelevant

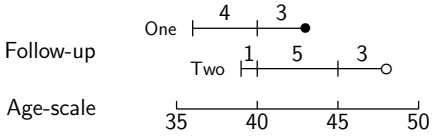
Code and output for the entire example available in
`http://bendixcarstensen.com/AdvCoh/WNtCMa/`

Examples: stratification by age

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

If follow-up is long, use stratification by categories of **current age**, both for:

No. of events, D, and Risk time, Y.



— assuming a constant rate λ throughout.

Models of this world

- Replace the α_is by a parametric function $f(t)$ with a limited number of parameters, for example:
  - Piecewise constant
  - Splines (linear, quadratic or cubic)
  - Fractional polynomials
- the two latter brings model into "this world":
  - smoothly varying rates
  - parametric closed form representation of baseline hazard
  - finite no. of parameters
- Makes it really easy to use rates directly in calculations of
  - expected residual life time
  - state occupancy probabilities in multistate models
  - . . .

Representation of follow-up data

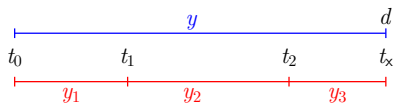
A cohort or follow-up study records:
Events and **Risk time**.

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

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

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

Specific for each **type** of outcome.



Probability	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

	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

Dividing time into bands:

If we want to compute D and Y in intervals on some timescale we must decide on:

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

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?

Example: cohort with 3 persons:

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

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

Timescales

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

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

Follow-up on several timescales

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

  (d, y) (event, duration)
- Covariates in analysis of rates:
  - timescales
  - other (fixed) measurements
- . . . do not confuse **duration** and **timescale** !

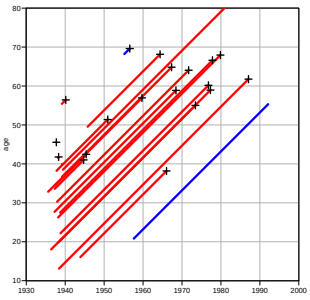
Follow-up data in Epi — Lexis objects

```
> 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

Timescales of interest:

- Age
- Calendar time
- Time since injection




```
> 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 )
```

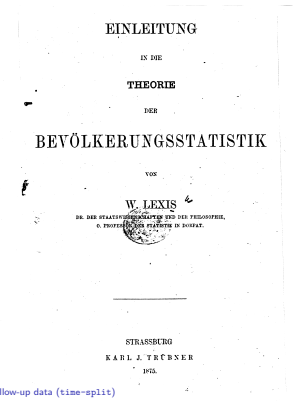

Definition of Lexis object

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

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

exitdat - injecdat

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


EINLEITUNG

IN DIE

THEORIE

DER

BEVÖLKERUNGSSTATISTIK

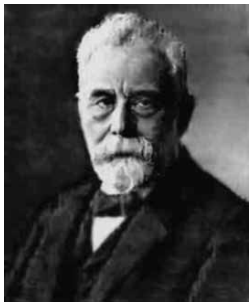
VON

W. LEXIS

STRASSBURG

KARL J. TRÜBNER

1875.

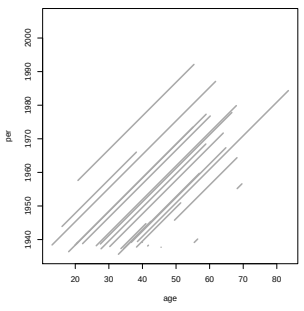


The looks of a Lexis object

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

> summary( thL )
Transitions:
     To
From    0    1 Records:  Events:  Risk time:  Persons:
   0 504 1964      2468     1964    51934.08      2468
```


Splitting follow-up time

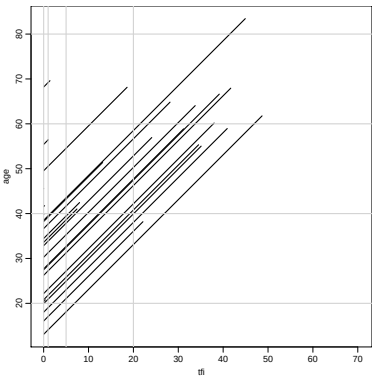
```
> spl1 <- splitLexis( thL, breaks=seq(0,100,20),
>                          time.scale="age" )
> round(spl1,1)
   age    per  tfi lex.dur lex.Cst lex.Xst   id sex birthdat contrast injecdat vol
1 22.2 1938.8  0.0    17.8       0       0    1   2    1916.6        1   1938.8
2 40.0 1956.6 17.8    20.0       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
...
```


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


Split on another timescale

```
> spl2 <- splitLexis( spl1, time.scale="tfi",
                            breaks=c(0,1,5,20,100) )
> round( spl2, 1 )
   lex.id  age    per  tfi lex.dur lex.Cst lex.Xst   id sex 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

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

```
> library( popEpi )
> spl1 <- splitMulti( thL , age=seq(0,100,20)   )
> spl2 <- splitMulti( spl1, tfi=c(0,1,5,20,100) )
> options( digits=5 )
> spl2[1:10,1:11]
```

	lex.id	age	per	tfi	lex.dur	lex.Cst	lex.Xst	id	sex	birthdat	contrast
1:	1	22.182	1938.8	0.000	1.00000	0	0	1	2	1916.6	1
2:	1	23.182	1939.8	1.000	4.00000	0	0	1	2	1916.6	1
3:	1	27.182	1943.8	5.000	12.81793	0	0	1	2	1916.6	1
4:	1	40.000	1956.6	17.818	2.18207	0	0	1	2	1916.6	1
5:	1	42.182	1958.8	20.000	17.81793	0	0	1	2	1916.6	1
6:	1	60.000	1976.6	37.818	0.17796	0	1	1	2	1916.6	1
7:	2	16.063	1943.9	0.000	1.00000	0	0	2	2	1927.8	1
8:	2	17.063	1944.9	1.000	2.93703	0	0	2	2	1927.8	1
9:	2	20.000	1947.8	3.937	1.06297	0	0	2	2	1927.8	1
10:	2	21.063	1948.9	5.000	15.00000	0	0	2	2	1927.8	1

— not the printing: it's a `data.table`

Likelihood for a constant rate

- This setup is for a situation where it is assumed that rates are constant in each of the intervals.
- Each observation in the dataset contributes a term to the likelihood.
- Each term looks like a contribution from a Poisson variate (albeit with values only 0 or 1)
- Rates can vary along several timescales simultaneously.
- Models can include fixed covariates, as well as the timescales (the left end-points of the intervals) as **quantitative** variables.
- The latter is where we will need splines.

Analysis of results

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

The Poisson likelihood for split data

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

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

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

Fitting a simple model

```
> stat.table( contrast,
+             list( D = sum( lex.Xst ),
+                   Y = sum( lex.dur ),
+              Rate = ratio( lex.Xst, lex.dur, 100 ) ),
+             margin = TRUE,
+             data = spl2 )
```

contrast	D	Y	Rate
1	928.00	20094.74	4.62
2	1036.00	31822.24	3.26
Total	1964.00	51916.98	3.78

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

```
> spl1 <- splitLexis( thL , breaks=seq(0,100,20)  , time.scale="age" )
> spl2 <- splitLexis( spl1, breaks=c(0,1,5,20,100), time.scale="tfi" )
> options( digits=5 )
> spl2[1:10,1:11]
```

	lex.id	age	per	tfi	lex.dur	lex.Cst	lex.Xst	id	sex	birthdat	contrast
1	1	22.182	1938.8	0.000	1.00000	0	0	1	2	1916.6	1
2	1	23.182	1939.8	1.000	4.00000	0	0	1	2	1916.6	1
3	1	27.182	1943.8	5.000	12.81793	0	0	1	2	1916.6	1
4	1	40.000	1956.6	17.818	2.18207	0	0	1	2	1916.6	1
5	1	42.182	1958.8	20.000	17.81793	0	0	1	2	1916.6	1
6	1	60.000	1976.6	37.818	0.17796	0	1	1	2	1916.6	1
7	2	16.063	1943.9	0.000	1.00000	0	0	2	2	1927.8	1
8	2	17.063	1944.9	1.000	2.93703	0	0	2	2	1927.8	1
9	2	20.000	1947.8	3.937	1.06297	0	0	2	2	1927.8	1
10	2	21.063	1948.9	5.000	15.00000	0	0	2	2	1927.8	1

— and what are covariates for the rates?

Fitting a simple model

contrast	D	Y	Rate
1	928.00	20094.74	4.62
2	1036.00	31822.24	3.26

```
> 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	3.26	3.06	3.46

Models for tabulated data

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019` `tab-mod`

Conceptual set-up

Follow-up of the entire (male) population from 1943–2006 w.r.t. occurrence of testis cancer:

- Split follow-up time for all about 4 mil. men in 1-year classes by age and calendar time (y).
- Allocate testis cancer event ($d = 0, 1$) to each.
- Analyze all $200,000,000$ records by a Poisson model.

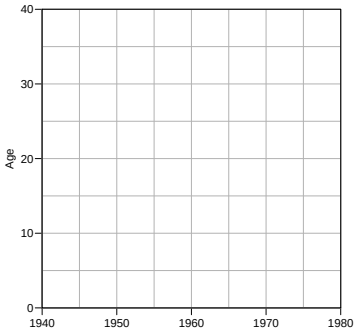
Realistic set-up

- Tabulate the follow-up time and events by age and period.
- 100 age-classes.
- 65 periods (single calendar years).
- 6500 aggregate records of (D, Y).
- Analyze by a Poisson model.

Practical set-up

- Tabulate only events (as obtained from the cancer registry) by age and period.
- 100 age-classes.
- 65 periods (single calendar years).
- 6500 aggregate records of D.
- Estimate the population follow-up based on census data from Statistics Denmark (Y_{pop}).
  . . . or get it from the human mortality database.
- If disease is common: tabulate follow-up **after** diagnosis (Y_{dis}), and subtract from population follow-up.
- Analyze (D, Y) by Poisson model.

Lexis diagram [1]

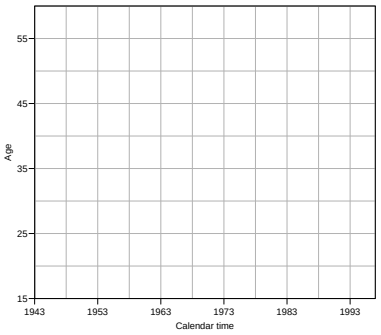


Disease registers record events.

Official statistics collect population data.

[1] Named after the German statistician and economist **William Lexis** (1837–1914), who devised this diagram in the book "Einleitung in die Theorie der Bevölkerungsstatistik" (Karl J. Trübner, Strassburg, 1875).

Lexis diagram



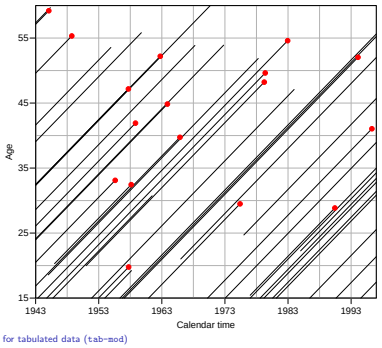
Registration of:

cases (D)

risk time,
person-years (Y)

in subsets of the Lexis diagram.

Lexis diagram



Registration of:

cases (D)

risk time,
person-years (Y)

in subsets of the Lexis diagram.

Rates available in each subset.

Register data

Classification of **cases** (D_{ap}) by age at diagnosis and date of diagnosis, and **population** (Y_{ap}) by age at risk and date at risk, in compartments of the Lexis diagram, e.g.:

```
> fCtable( xtabs( cbind(D,Y) ~ A + P, data=ts ), col.vars=3:2, w=8 )
```

	D				Y			
A \ P	1943	1948	1953	1958	1943	1948	1953	1958
15	2	3	4	1	773,812	744,217	794,123	972,853
20	7	7	17	8	813,022	744,706	721,810	770,859
25	28	23	26	35	790,501	781,827	722,968	698,612
30	28	43	49	51	799,293	774,542	769,298	711,596
35	36	42	39	44	769,356	782,893	760,213	760,452
40	24	32	46	53	694,073	754,322	768,471	749,912

In analysis format:

```
> ts
```

```
    A    P  D      Y
1  15 1943  2 773812
2  20 1943  7 813022
3  25 1943 28 790501
4  30 1943 28 799293
5  35 1943 36 769356
6  40 1943 24 694073
7  15 1948  3 744217
8  20 1948  7 744706
9  25 1948 23 781827
10 30 1948 43 774542
11 35 1948 42 782893
12 40 1948 32 754322
13 15 1953  4 794123
14 20 1953 17 721810
15 25 1953 26 722968
16 30 1953 49 769298
17 35 1953 39 760213
18 40 1953 46 768471
19 15 1958  1 972853
```


Tabulated data

Once data are in tabular form, models are restricted:

- Rates must be assumed constant in each cell of the table / subset of the Lexis diagram.
- With large cells (5×5 years) it is customary to put a separate parameter on each cell or on each levels of classifying factors.
- Output from the model will be rates and rate-ratios.
- Since we use multiplicative Poisson, usually the log rates and the log-RR are reported

Simple age-period model for the testis cancer rates:

```
> m0 <- glm( D ~ factor(A) + factor(P) + offset( log(Y/10^5) ),
+            family=poisson, data=ts )
> summary( m0 )
```

```
Call:
glm(formula = D ~ factor(A) + factor(P) + offset(log(Y/10^5)),
    family = poisson, data = ts)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5991  -0.6974   0.1284   0.6671   1.8904

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.4758     0.3267  -4.517 6.26e-06
factor(A)20    1.4539     0.3545   4.101 4.11e-05
factor(A)25    2.5321     0.3301   7.671 1.71e-14
factor(A)30    2.9327     0.3254   9.013  < 2e-16
factor(A)35    2.8613     0.3259   8.779  < 2e-16
factor(A)40    2.8521     0.3263   8.741  < 2e-16
factor(P)1948  0.1753     0.1211   1.447  0.14778
factor(P)1953  0.3822     0.1163   3.286  0.00102
```


`ci.exp()` from the `Epi` package extracts coefficients and computes confidence intervals:

```
> round( ci.exp( m0 ), 2 )
```

```
              exp(Est.) 2.5% 97.5%
(Intercept)        0.23 0.12  0.43
factor(A)20        4.28 2.14  8.57
factor(A)25       12.58 6.59 24.02
factor(A)30       18.78 9.92 35.53
factor(A)35       17.49 9.23 33.12
factor(A)40       17.32 9.14 32.84
factor(P)1948      1.19 0.94  1.51
factor(P)1953      1.47 1.17  1.84
factor(P)1958      1.59 1.27  2.00
```

. . . what do there parameters mean?

Subsets of parameter estimates accessed via a character string that is grep-ed to the names.

```
> round( ci.exp( m0, subset="P", pval=TRUE ), 3 )
```

```
              exp(Est.)  2.5% 97.5%     P
factor(P)1948     1.192 0.940 1.511 0.148
factor(P)1953     1.466 1.167 1.841 0.001
factor(P)1958     1.593 1.272 1.996 0.000
```

```
> round( ci.lin( m0, subset="P" ), 3 )
```

```
              Estimate StdErr     z     P   2.5% 97.5%
factor(P)1948    0.175  0.121 1.447 0.148 -0.062 0.413
factor(P)1953    0.382  0.116 3.286 0.001  0.154 0.610
factor(P)1958    0.466  0.115 4.052 0.000  0.241 0.691
```


Linear combinations of the parameters can be computed using the `ctr.mat` option:

```
> CM <- rbind( '1943 vs. 1953' = c( 0,-1, 0),
+              '1948 vs. 1953' = c( 1,-1, 0),
+               'Ref. (1953)' = c( 0, 0, 0),
+             '1958 vs. 1953' = c( 0,-1, 1) )
> round( ci.exp( m0, subset="P", ctr.mat=CM ), 3 )
```

```
              exp(Est.)  2.5% 97.5%
1943 vs. 1953     0.682 0.543 0.857
1948 vs. 1953     0.813 0.655 1.010
Ref. (1953)       1.000 1.000 1.000
1958 vs. 1953     1.087 0.887 1.332
```


Age-Period and Age-Cohort models

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019`

Register data — rates

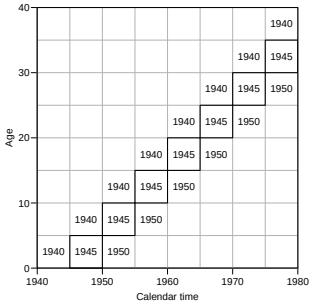
Rates in "tiles" of the Lexis diagram:

$$\lambda(a,p) = D_{ap}/Y_{ap}$$

Descriptive epidemiology based on disease registers:
How do the rates vary by age and time:

- Age-specific rates for a given period.
- Age-standardized rates as a function of calendar time. (Weighted averages of the age-specific rates).

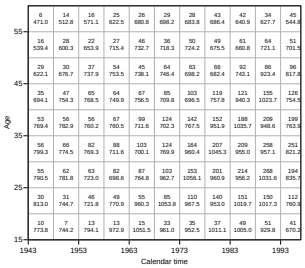
"Synthetic" cohorts



Events and risk time in cells along the diagonals are among persons with roughly same date of birth.

Successively overlapping 10-year periods.

Lexis diagram: data



Testis cancer cases in Denmark.

Male person-years in Denmark.

Data matrix: Testis cancer cases

Number of cases

	Date of diagnosis $(year - 1900)$								
Age	48–52	53–57	58–62	63–67	68–72	73–77	78–82	83–87	88–92
15–19	7	13	13	15	33	35	37	49	51
20–24	31	46	49	55	85	110	140	151	150
25–29	62	63	82	87	103	153	201	214	268
30–34	66	82	88	103	124	164	207	209	258
35–39	56	56	67	99	124	142	152	188	209
40–44	47	65	64	67	85	103	119	121	155
45–49	30	37	54	45	64	63	66	92	86
50–54	28	22	27	46	36	50	49	61	64
55–59	14	16	25	26	29	28	43	42	34

Data matrix: Male risk time

1000 person-years

	Date of diagnosis $(year - 1900)$								
Age	48–52	53–57	58–62	63–67	68–72	73–77	78–82	83–87	88–92
15–19	744.2	794.1	972.9	1051.5	961.0	952.5	1011.1	1005.0	929.8
20–24	744.7	721.8	770.9	960.3	1053.8	967.5	953.0	1019.7	1017.3
25–29	781.8	723.0	698.6	764.8	962.7	1056.1	960.9	956.2	1031.6
30–34	774.5	769.3	711.6	700.1	769.9	960.4	1045.3	955.0	957.1
35–39	782.9	760.2	760.5	711.6	702.3	767.5	951.9	1035.7	948.6
40–44	754.3	768.5	749.9	756.5	709.8	696.5	757.8	940.3	1023.7
45–49	676.7	737.9	753.5	738.1	746.4	698.2	682.4	743.1	923.4
50–54	600.3	653.9	715.4	732.7	718.3	724.2	675.5	660.8	721.1
55–59	512.8	571.1	622.5	680.8	698.2	683.8	686.4	640.9	627.7

Data matrix: Empirical rates

Rate per 1000,000 person-years

	Date of diagnosis $(year - 1900)$								
Age	48–52	53–57	58–62	63–67	68–72	73–77	78–82	83–87	88–92
15–19	9.4	16.4	13.4	14.3	34.3	36.7	36.6	48.8	54.8
20–24	41.6	63.7	63.6	57.3	80.7	113.7	146.9	148.1	147.4
25–29	79.3	87.1	117.4	113.8	107.0	144.9	209.2	223.8	259.8
30–34	85.2	106.6	123.7	147.1	161.1	170.8	198.0	218.8	269.6
35–39	71.5	73.7	88.1	139.1	176.6	185.0	159.7	181.5	220.3
40–44	62.3	84.6	85.3	88.6	119.8	147.9	157.0	128.7	151.4
45–49	44.3	50.1	71.7	61.0	85.7	90.2	96.7	123.8	93.1
50–54	46.6	33.6	37.7	62.8	50.1	69.0	72.5	92.3	88.7
55–59	27.3	28.0	40.2	38.2	41.5	40.9	62.6	65.5	54.2

The classical plots

Given a table of rates classified by age and period, we can do 4 "classical" plots:

- Rates versus age at diagnosis (period):
  — rates in the same age-class connected.
- Rates versus age at diagnosis:
  — rates in the same birth-cohort connected.
- Rates versus date of diagnosis:
  — rates in the same age-class connected.
- Rates versus date of date of birth:
  — rates in the same age-class connected.

These plots can be produced by the R-function `rateplot`.

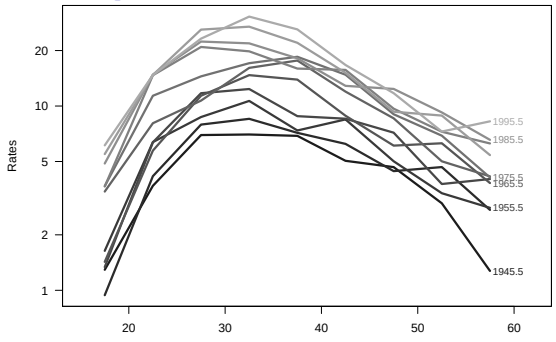

```
> library( Epi )
> data( testisDK )
> head( testisDK )

  A    P D        Y
1 0 1943 1 39649.50
2 1 1943 1 36942.83
3 2 1943 0 34588.33
4 3 1943 1 33267.00
5 4 1943 0 32614.00
6 5 1943 0 32020.33
```



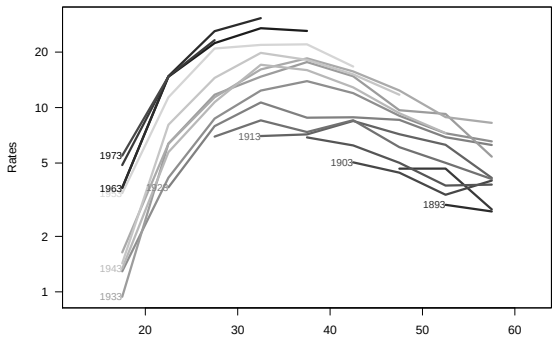
```
> ts <- transform( subset( testisDK, A>14 & A<60 ),
+                  A = floor( A      /5)*5     +2.5,
+                  P = floor((P-1943)/5)*5+1943+2.5 )
> ts$C <- ts$P - ts$A
> trate <- xtabs( D ~ A + P, data = ts ) /
+          xtabs( Y ~ A + P, data = ts ) * 100000
> #
> #
> #
> trate[1:5,1:6]

      P
A        1945.5    1950.5    1955.5    1960.5    1965.5    1970.5
  17.5 1.2923036 0.9405857 1.6370257 1.3362759 1.4264867 3.4340862
  22.5 3.6899378 4.1627194 6.3728682 6.3565492 5.7274822 8.0657826
  27.5 6.9576174 7.9301414 8.7140826 11.7375624 11.3753792 10.6996275
  32.5 7.0061961 8.5211703 10.6590661 12.3665762 14.7122260 16.1068525
  37.5 6.8888785 7.1529555 7.3663549 8.8105514 13.9126492 17.6571019
```

```
> par( mfrow=c(2,2) )
> rateplot( trate, col=gray(2:15/18), lwd=3, ann=TRUE )
> wh = c("ap","ac","pa","ca")
> for( ptp in wh ) {
+    pdf( paste("./AP-AC-",ptp,".pdf",sep=""), height=6, width=8 )
+    par( mar=c(3,3,1,1, mgp=c(3,1,0)/1.6, bty="n", las=1 ))
+    rateplot( trate, which=ptp,
+              col=gray(2:15/18), lwd=3, ann=TRUE, a.lim=c(15,60) )
+    dev.off()
+    }
```

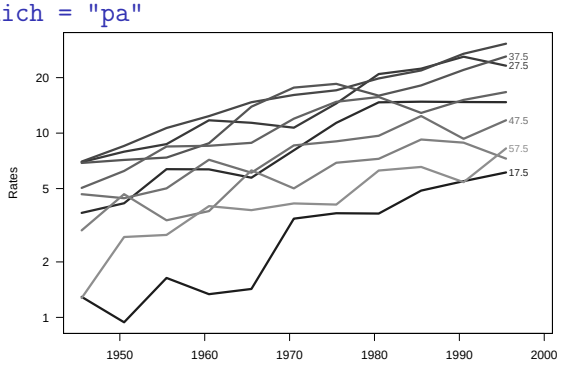

which = "ap"



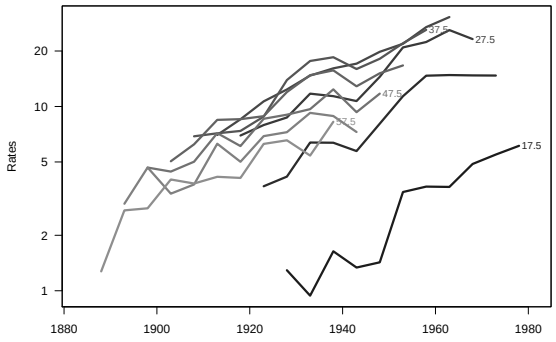
which = "ac"



which = "pa"



which = "ca"



Age-Period model

Rates are proportional between periods:

$$\lambda(a,p) = a_a \times b_p \qquad \text{or} \qquad \log[\lambda(a,p)] = \alpha_a + \beta_p$$

Choose p_0 as reference period, where $\beta_{p_0} = 0$

$$\log[\lambda(a,p_0)] = \alpha_a + \beta_{p_0} = \alpha_a$$

Fitting the A-P model in R I

Reference period is the 5th period (1970.5 $\sim$ 1968–72):

```
> ap <- glm( D ~ factor(A) - 1 + relevel( factor(P), "1970.5" ) +
+               offset( log(Y/10^5) ),
+            family=poisson, data=ts )
> summary( ap )

Call:
glm(formula = D ~ factor(A) - 1 + relevel(factor(P), "1970.5") +
    offset(log(Y/10^5)), family = poisson, data = ts)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-3.1850  -0.9465  -0.1683   0.5767   3.8610

Coefficients:
```


Fitting the A-P model in R II

```
                                        Estimate Std. Error z value Pr(>|z|)
factor(A)17.5                            1.06715    0.06791  15.715  < 2e-16
factor(A)22.5                            2.20732    0.04837  45.630  < 2e-16
factor(A)27.5                            2.65463    0.04465  59.449  < 2e-16
factor(A)32.5                            2.77057    0.04458  62.142  < 2e-16
factor(A)37.5                            2.63081    0.04606  57.122  < 2e-16
factor(A)42.5                            2.36224    0.04878  48.422  < 2e-16
factor(A)47.5                            2.01945    0.05341  37.811  < 2e-16
factor(A)52.5                            1.73119    0.05957  29.062  < 2e-16
factor(A)57.5                            1.45070    0.06748  21.498  < 2e-16
relevel(factor(P), "1970.5")1945.5 -0.75072    0.07011 -10.708  < 2e-16
relevel(factor(P), "1970.5")1950.5 -0.59740    0.06633  -9.006  < 2e-16
relevel(factor(P), "1970.5")1955.5 -0.43562    0.06299  -6.916 4.65e-12
relevel(factor(P), "1970.5")1960.5 -0.27952    0.05999  -4.659 3.18e-06
relevel(factor(P), "1970.5")1965.5 -0.16989    0.05751  -2.954  0.00313
relevel(factor(P), "1970.5")1975.5  0.16037    0.05143   3.118  0.00182
relevel(factor(P), "1970.5")1980.5  0.30022    0.04953   6.061 1.35e-09
relevel(factor(P), "1970.5")1985.5  0.37491    0.04853   7.726 1.11e-14
relevel(factor(P), "1970.5")1990.5  0.47047    0.04745   9.916  < 2e-16
relevel(factor(P), "1970.5")1995.5  0.54079    0.04862  11.123  < 2e-16
```

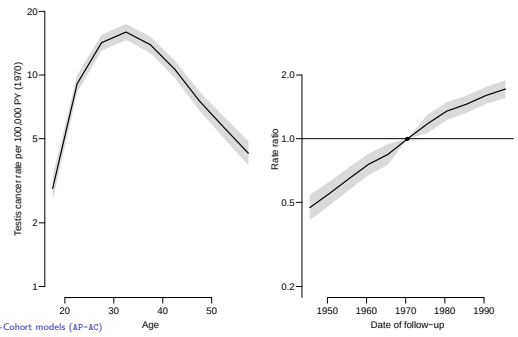
Fitting the A-P model in R III

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 29193.6  on 2430  degrees of freedom
Residual deviance:  2816.6  on 2411  degrees of freedom
AIC: 9005

Number of Fisher Scoring iterations: 5
```


Estimates with confidence intervals

```
> par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, bty="n", las=1 )
> matshade( seq(17.5,57.5,5), ci.exp(ap,subset="A"), plot=TRUE,
+           log="y", lwd=2, ylim=c(1,20), xlab="Age",
+           ylab="Testis cancer rate per 100,000 PY (1970)" )
> matshade( seq(1945.5,1995.5,5),
+           rbind( ci.exp(ap,subset="P")[1:5 ,], 1,
+                  ci.exp(ap,subset="P")[6:10,] ), plot=TRUE,
+           log="y", lwd=2, ylim=c(1,20)/5,
+           xlab="Date of follow-up", ylab="Rate ratio" )
> abline( h = 1)
> points( 1970.5, 1, pch=16 )
```


Estimates from Age-Period model



Age-cohort model

Rates are proportional between cohorts:

$$\lambda(a,c) = a_a \times c_c \qquad \text{or} \qquad \log[\lambda(a,p)] = \alpha_a + \gamma_c$$

Choose c_0 as reference cohort, where $\gamma_{c_0} = 0$

$$\log[\lambda(a,c_0)] = \alpha_a + \gamma_{c_0} = \alpha_a$$

Fitting the A-C model in R I

Reference period is the 1933 cohort:

```
> ac <- glm( D ~ factor(A) - 1 + relevel( factor(C), "1933" ) +
+                offset( log(Y/10^5) ),
+            family=poisson, data=ts )
> summary( ac )

Call:
glm(formula = D ~ factor(A) - 1 + relevel(factor(C), "1933") +
    offset(log(Y/10^5)), family = poisson, data = ts)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0796  -0.9538  -0.1620   0.5767   3.9525

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
factor(A)17.5              0.61513    0.07534   8.165 3.23e-16
```


Fitting the A-C model in R II

```
factor(A)22.5                  1.89965    0.05342  35.558  < 2e-16
factor(A)27.5                  2.46911    0.04842  50.990  < 2e-16
factor(A)32.5                  2.70635    0.04695  57.639  < 2e-16
factor(A)37.5                  2.71211    0.04758  57.006  < 2e-16
factor(A)42.5                  2.58676    0.04993  51.803  < 2e-16
factor(A)47.5                  2.36542    0.05459  43.327  < 2e-16
factor(A)52.5                  2.18192    0.06098  35.782  < 2e-16
factor(A)57.5                  2.01519    0.06939  29.041  < 2e-16
relevel(factor(C), "1933")1888 -1.77316   0.41400  -4.283 1.84e-05
relevel(factor(C), "1933")1893 -1.05641   0.19017  -5.555 2.77e-08
relevel(factor(C), "1933")1898 -0.79897   0.12600  -6.341 2.28e-10
relevel(factor(C), "1933")1903 -0.87599   0.10389  -8.432  < 2e-16
relevel(factor(C), "1933")1908 -0.76707   0.08352  -9.184  < 2e-16
relevel(factor(C), "1933")1913 -0.56290   0.07006  -8.035 9.36e-16
relevel(factor(C), "1933")1918 -0.56702   0.06683  -8.484  < 2e-16
relevel(factor(C), "1933")1923 -0.36836   0.06124  -6.015 1.79e-09
relevel(factor(C), "1933")1928 -0.18832   0.05903  -3.190 0.001421
relevel(factor(C), "1933")1938  0.08958   0.05439   1.647 0.099585
relevel(factor(C), "1933")1943 -0.03107   0.05443  -0.571 0.568091
```

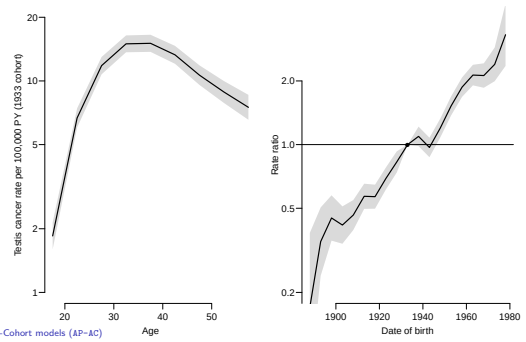

Fitting the A-C model in R III

```
relevel(factor(C), "1933")1948  0.18088    0.05256   3.441 0.000579
relevel(factor(C), "1933")1953  0.42239    0.05309   7.956 1.77e-15
relevel(factor(C), "1933")1958  0.62544    0.05421  11.537  < 2e-16
relevel(factor(C), "1933")1963  0.75687    0.05727  13.215  < 2e-16
relevel(factor(C), "1933")1968  0.75183    0.06799  11.057  < 2e-16
relevel(factor(C), "1933")1973  0.87343    0.09373   9.318  < 2e-16
relevel(factor(C), "1933")1978  1.19601    0.17340   6.898 5.29e-12

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 29193.6  on 2430  degrees of freedom
Residual deviance:  2767.8  on 2403  degrees of freedom
AIC: 8972.2

Number of Fisher Scoring iterations: 5
```


Estimates with confidence intervals

```
> par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, bty="n", las=1 )
> matshade( seq(17.5,57.5,5), ci.exp(ac,subset="A"), plot=TRUE,
+           log="y", lwd=2, ylim=c(1,20), xlab="Age",
+           ylab="Testis cancer rate per 100,000 PY (1933 cohort)" )
> matshade( seq(1888,1978,5),
+           rbind( ci.exp(ac,subset="C")[1:9 ,], 1,
+                  ci.exp(ac,subset="C")[10:18,] ), plot=TRUE,
+           log="y", lwd=2, ylim=c(1,20)/5,
+           xlab="Date of birth", ylab="Rate ratio" )
> abline( h = 1)
> points( 1933, 1, pch=16 )
```

Estimates from Age-Cohort model



Recap of Monday — rates

- Rate, intensity: $\lambda(t) = \mathrm{P}\{\text{ event in } (t, t+h) | \text{ alive at } t \}/h$
- Observe empirical rates (d, y) — possibly many per person.
- $\ell_{\mathsf{FU}} = d\log(\lambda) - \lambda y$, obs: (d, y), rate par: λ
- $\ell_{\mathsf{Poisson}} = d\log(\lambda y) - \lambda y$, obs: d, mean par: $\mu = \lambda y$
- $\ell_{\mathsf{Poisson}} - \ell_{\mathsf{FU}} = d\log(y)$ does not involve λ — use either to find m.l.e. of λ
- Poisson model is for $\log(\mu) = \log(\lambda y) = \log(\lambda) + \log(y)$ hence `offset=log(Y)`
- Once rates are known, we can construct survival curves and derivatives of that.

Recap Monday — models

- Empirical rate (d_t, y_t) relates to a **time** t
- Many for the same person — different times
- Not independent, but likelihood is a product
- One parameter per interval ⇒ exchangeable times
- Use the quantitative nature of t: ⇒ smooth continuous effects of time
- Predicted rates: `ci.pred( model, newdata=nd )`
- RR is the difference between two predictions:
- RR by period:
- `ndx<-data.frame(P=1947:1980,A=47)`
- `ndr<-data.frame(P=1870,A=47)`
- `ci.exp( model, ctr.mat=list(ndx-ndr))`

Age-drift model

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019` Ad

Linear effect of period:

$$\log[\lambda(a,p)] = \alpha_a + \beta_p = \alpha_a + \beta(p - p_0)$$

that is, $\beta_p = \beta(p - p_0)$.

Linear effect of cohort:

$$\log[\lambda(a,p)] = \tilde{\alpha}_a + \gamma_c = \tilde{\alpha}_a + \gamma(c - c_0)$$

that is, $\gamma_c = \gamma(c - c_0)$

Age and linear effect of period:

```
> apd <- glm( D ~ factor( A ) - 1 + I(P-1970.5) +
+             offset( log( Y ) ),
+             family=poisson )
> summary( apd )

Call:
glm(formula = D ~ factor(A) - 1 + I(P - 1970.5) + offset(log(Y)), family = poisson

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.97593  -0.77091    0.02809    0.95914    2.93076

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
factor(A)17.5 -3.58065    0.06306  -56.79   <2e-16
...
factor(A)57.5 -3.17579    0.06256  -50.77   <2e-16
I(P - 1970.5)  0.02653    0.00100   26.52   <2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 89358.53  on 81  degrees of freedom
Residual deviance:   126.07  on 71  degrees of freedom
```


Age and linear effect of cohort:

```
> acd <- glm( D ~ factor( A ) - 1 + I(C-1933) +
+             offset( log( Y ) ),
+             family=poisson )
> summary( acd )

Call:
glm(formula = D ~ factor(A) - 1 + I(C - 1933) + offset(log(Y)), family = poisson)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.97593  -0.77091    0.02809    0.95914    2.93076

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
factor(A)17.5 -4.11117    0.06760  -60.82   <2e-16
...
factor(A)57.5 -2.64527    0.06423  -41.19   <2e-16
I(C - 1933)    0.02653    0.00100   26.52   <2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 89358.53  on 81  degrees of freedom
Residual deviance:   126.07  on 71  degrees of freedom
```

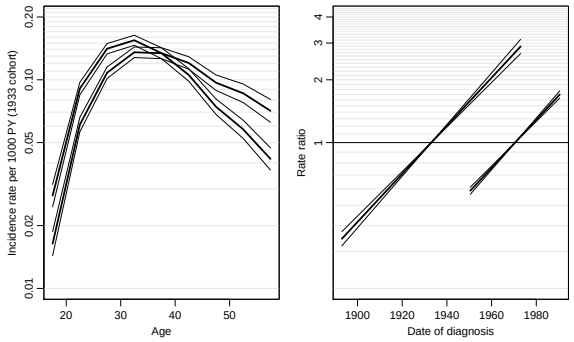

What goes on?

$$p = a + c \qquad p_0 = a_0 + c_0$$

$$\begin{aligned} \alpha_a + \beta(p - p_0) &= \alpha_a + \beta\big(a + c - (a_0 + c_0)\big) \\ &= \underbrace{\alpha_a + \beta(a - a_0)}_{\text{cohort age-effect}} + \beta(c - c_0) \end{aligned}$$

The two **models** are the same.
The **parametrization** is different.

The age-curve refers either
- to a period (cross-sectional rates) or
- to a cohort (longitudinal rates).



Which age-curve is period and which is cohort?

Age-Period-Cohort model

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

http://BendixCarstensen/APC/EDSD-2019 APC-cat

The age-period-cohort model

$$\log[\lambda(a,p)] = \alpha_a + \beta_p + \gamma_c$$

- Three effects:
  - a — Age (at diagnosis)
  - p — Period (of diagnosis)
  - c — Cohort (of birth)
- No assumptions about the **shape** of effects.
- Levels of A, P and C are assumed **exchangeable**
- *i.e.* no assumptions about the relationship between parameter estimates and the **scaled values** of A, P and C

Fitting the model in R I

```
> library( Epi )
> data( testisDK )
> tc <- transform( subset( testisDK, A>14 & A<60 & P<1993),
+                  A = floor( A      /5)*5     +2.5,
+                  P = floor((P-1943)/5)*5+1943+2.5 )
> tc <- aggregate( tc[,c("D","Y")], tc[,c("A","P")], FUN=sum )
> tc$C <- tc$P - tc$A
> str( tc )

'data.frame': 90 obs. of  5 variables:
 $ A: num  17.5 22.5 27.5 32.5 37.5 42.5 47.5 52.5 57.5 17.5 ...
 $ P: num  1946 1946 1946 1946 1946 ...
 $ D: num  10 30 55 56 53 35 29 16 6 7 ...
 $ Y: num  773812 813022 790500 799293 769356 ...
 $ C: num  1928 1923 1918 1913 1908 ...
```


Fitting the model in R II

```
> m.apc <- glm( D ~ factor(A) + factor(P) + factor(C) + offset(log(Y)),
+               family = poisson, data = tc )
> summary( m.apc )

Call:
glm(formula = D ~ factor(A) + factor(P) + factor(C) + offset(log(Y)),
    family = poisson, data = tc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5406  -0.5534   0.0000   0.4934   1.2966

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -11.39890    0.23316 -48.889  < 2e-16
factor(A)22.5    1.19668    0.07789  15.364  < 2e-16
factor(A)27.5    1.63551    0.08627  18.957  < 2e-16
factor(A)32.5    1.71939    0.10223  16.819  < 2e-16
factor(A)37.5    1.57062    0.12205  12.869  < 2e-16
```


Fitting the model in R III

```
factor(A)42.5    1.29418    0.14416   8.977  < 2e-16
factor(A)47.5    0.87209    0.16828   5.182 2.19e-07
factor(A)52.5    0.51257    0.19309   2.655  0.00794
factor(A)57.5    0.12801    0.21109   0.606  0.54424
factor(P)1950.5  0.20286    0.08247   2.460  0.01390
factor(P)1955.5  0.42044    0.09081   4.630 3.66e-06
factor(P)1960.5  0.64099    0.10548   6.077 1.23e-09
factor(P)1965.5  0.82135    0.12407   6.620 3.60e-11
factor(P)1970.5  1.06435    0.14444   7.369 1.72e-13
factor(P)1975.5  1.27796    0.16653   7.674 1.67e-14
factor(P)1980.5  1.43441    0.18961   7.565 3.88e-14
factor(P)1985.5  1.50578    0.21339   7.056 1.71e-12
factor(P)1990.5  1.58798    0.23562   6.740 1.59e-11
factor(C)1893    0.50556    0.42894   1.179  0.23855
factor(C)1898    0.56443    0.38398   1.470  0.14158
factor(C)1903    0.28430    0.35557   0.800  0.42397
factor(C)1908    0.20683    0.32836   0.630  0.52877
factor(C)1913    0.22302    0.30344   0.735  0.46236
factor(C)1918    0.02713    0.28150   0.096  0.92322
```


Fitting the model in R IV

```
factor(C)1923    0.03280    0.25971   0.126  0.89950
factor(C)1928    0.02155    0.23945   0.090  0.92830
factor(C)1933    0.02518    0.21988   0.115  0.90881
factor(C)1938   -0.07240    0.20268  -0.357  0.72094
factor(C)1943   -0.35284    0.18706  -1.886  0.05927
factor(C)1948   -0.30472    0.17308  -1.761  0.07831
factor(C)1953   -0.17916    0.16258  -1.102  0.27047
factor(C)1958   -0.11739    0.15585  -0.753  0.45133
factor(C)1963   -0.10882    0.15410  -0.706  0.48008
factor(C)1968   -0.16807    0.16235  -1.035  0.30053
factor(C)1973         NA         NA      NA       NA

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2761.230  on 89  degrees of freedom
Residual deviance:   38.783  on 56  degrees of freedom
AIC: 637.64

Number of Fisher Scoring iterations: 4
```


Fitting the model in R V

No. of parameters

A has $9(A)$ levels
P has $10(P)$ levels
C=P-A has $18(C = A + P - 1)$ levels
Age-drift model has $A + 1 = 10$ parameters
Age-period model has $A + P - 1 = 18$ parameters
Age-cohort model has $A + C - 1 = 26$ parameters
Age-period-cohort model has $A + P + C - 3 = 34$ parameters:

```
> length( coef(m.apc) ) ; sum( !is.na(coef(m.apc)) )
[1] 35
[1] 34
```

The missing parameter is because of the **identifiability problem**.

Test for effects

```
> tc.acp <- apc.fit( tc, model="factor", ref.c=1943, scale=10^5 )

Latest version: TRUErefs: 1975.5 1943
[1] "ML of APC-model Poisson with log(Y) offset : ( ACP ):\n"

Analysis of deviance for Age-Period-Cohort model

                  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
Age                      81    1114.65
Age-drift                80     131.77   1   982.88 < 2.2e-16
Age-Cohort               64      70.20  16    61.57  2.84e-07
Age-Period-Cohort        56      38.78   8    31.42 0.0001183
Age-Period               72     122.23 -16   -83.45  3.95e-11
Age-drift                80     131.77  -8    -9.54 0.2989863
```


How to choose a parametrization

- Standard approach: Put extremes of periods or cohorts to 0, and choose a reference for the other.
- Clayton & Schifflers: only 2nd order differences are invariants:

$$\alpha_{i-1} - 2\alpha_i + \alpha_{i+1}$$

  Implemented in `Epi` via the contrast type `contr.2nd` (later).
- Holford: Extract linear effects by regression:

$$\begin{array}{rcccccccc} \lambda(a,p) & = & \hat\alpha_a & + & = & \tilde\alpha_a & + & \hat\mu_a & + & \hat\delta_a a & + \\ & & \hat\beta_p & + & & \tilde\beta_p & + & \hat\mu_p & + & \hat\delta_p p & + \\ & & \hat\gamma_c & & & \tilde\gamma_c & + & \hat\mu_c & + & \hat\delta_c c & \end{array}$$

Assumptions

Assumptions are needed to do this, *e.g.*:

- Age is the major time scale
- Cohort is the secondary time scale (the major secular trend)
- c_0 is the reference cohort
- Period is the residual time scale: 0 on average, 0 slope
- ...constraining first and last period parameter to 0 is a crude way of obtaining this.

Relocating effects between A, P and C

Period effect, 0 on average, slope is 0: a regression of β_p on p:

$$g(p) = \tilde\beta_p = \beta_p - \hat\mu_p - \hat\delta_p p$$

Cohort effect, absorbing all time-trend ($\delta_p p = \delta_p(a + c)$) and risk relative to c_0:

$$h(c) = \gamma_c - \gamma_{c_0} + \hat\delta_p(c - c_0)$$

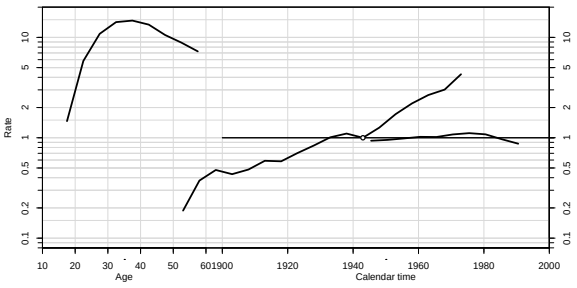
The rest is the age-effect:

$$f(a) = \alpha_a + \hat\mu_p + \hat\delta_p a + \hat\delta_p c_0 + \gamma_{c_0}$$

How it all adds up:

$$\begin{array}{rcccccccc} \lambda(a,p) & = & \hat\alpha_a + \hat\beta_p + \hat\gamma_c & & & & & & \\ & = & \hat\alpha_a + \gamma_{c_0} & + & \hat\mu_p & + & \hat\delta_p(a + c_0) & + \\ & & \hat\beta_p & - & \hat\mu_p & - & \hat\delta_p(a + c) & + \\ & & \hat\gamma_c - \gamma_{c_0} & & & + & \hat\delta_p(c - c_0) & \end{array}$$

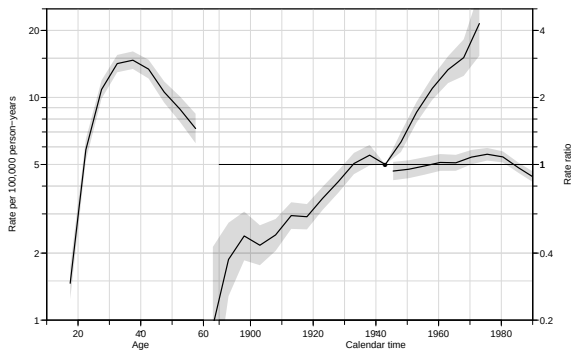
Only the regression on period is needed! (For this model...)




```
> plot( tc.acp )

cp.offset    RR.fac
    1835         1
```


Customize the frame for nicer plot of parameter estimates:

```
> par( mar=c(3,4,0.1,4), mgp=c(3,1,0)/1.6, las=1 )
> apc.frame( a.lab=c(2,4,6)*10,
+            a.tic=1:6*10,
+           cp.lab=1900+0:4*20,
+           cp.tic=1890+0:10*10,
+            r.lab=c(c(1,2,5),c(1,2)*10),
+            r.tic=c(1:10,15,20,25),
+           rr.ref=5 )
>    matshade( tc.acp$Age[,1], tc.acp$Age[,-1], lwd=2 )
> pc.matshade( tc.acp$Per[,1], tc.acp$Per[,-1], lwd=2 )
> pc.matshade( tc.acp$Coh[,1], tc.acp$Coh[,-1], lwd=2 )
> pc.points( 1943, 1, pch=16 )
```



A simple practical approach

- First fit the age-cohort model, with cohort c_0 as reference and get estimates $\hat{\alpha}_a$ and $\hat{\gamma}_c$:

$$\log[\lambda(a,p)] = \hat{\alpha}_a + \hat{\gamma}_c$$

- Then consider the full APC-model with age and cohort effects constrained to be as estimated from the AC-model:

$$\log[\lambda(a,p)] = \hat{\alpha}_a + \hat{\gamma}_c + \beta_p$$

- The residual period effect can be estimated if we note that for the number of cases we have:

$$\log(\text{expected cases}) = \log[\hat{\lambda}(a,p)\,Y] = \underbrace{\hat{\alpha}_a + \hat{\gamma}_c + \log(Y)}_{\text{"known"}} + \beta_p$$

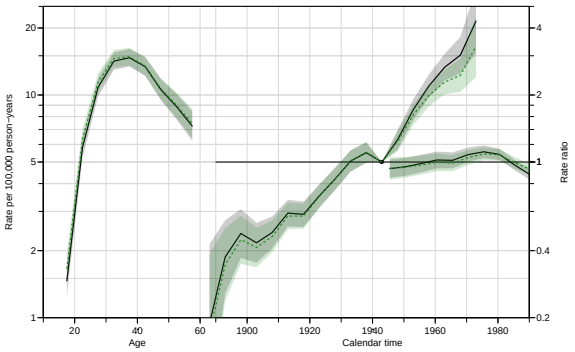
- This is analogous to the expression for a Poisson model in general,
- . . . but now is the offset not just $\log(Y)$ but $\hat{\alpha}_a + \hat{\gamma}_c + \log(Y)$, the log of the fitted values from the age-cohort model.
- β_ps are estimated in a Poisson model with this as offset.
- Advantage: We get the standard errors for free.

Customize the frame for nicer plot of parameter estimates:

```
> par( mar=c(3,4,0.1,4), mgp=c(3,1,0)/1.6, las=1 )
> apc.frame( a.lab=c(2,4,6)*10,
+            a.tic=1:6*10,
+           cp.lab=1900+0:4*20,
+           cp.tic=1890+0:10*10,
+            r.lab=c(c(1,2,5),c(1,2)*10),
+            r.tic=c(1:10,15,20,25),
+           rr.ref=5 )
>    matshade( tc.acp$Age[,1], tc.acp$Age[,-1], lwd=2, alpha=0.2 )
> pc.matshade( tc.acp$Per[,1], tc.acp$Per[,-1], lwd=2, alpha=0.2 )
> pc.matshade( tc.acp$Coh[,1], tc.acp$Coh[,-1], lwd=2, alpha=0.2 )
> pc.points( 1943, 1, pch=16 )
> #
> # The stepwise conditioning:
> tc.ac.p <- apc.fit( tc, model="factor", parm="AC-P", ref.c=1943, scale=10^5 )
```



```
Latest version: TRUErefs: 1975.5 1943
[1] "Sequential modelling Poisson with log(Y) offset : ( AC-P ):\n"

Analysis of deviance for Age-Period-Cohort model

                  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
Age                      81    1114.65
Age-drift                80     131.77   1   982.88 < 2.2e-16
Age-Cohort               64      70.20  16    61.57  2.84e-07
Age-Period-Cohort        56      38.78   8    31.42 0.0001183
Age-Period               72     122.23 -16   -83.45  3.95e-11
Age-drift                80     131.77  -8    -9.54 0.2989863

> #
>    matshade( tc.ac.p$Age[,1], tc.ac.p$Age[,-1], lwd=2, alpha=0.2, lty='22', lenc
> pc.matshade( tc.ac.p$Per[,1], tc.ac.p$Per[,-1], lwd=2, alpha=0.2, lty='22', lenc
> pc.matshade( tc.ac.p$Coh[,1], tc.ac.p$Coh[,-1], lwd=2, alpha=0.2, lty='22', lenc
```

Age at entry
Age-Duration-Diagnosis

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

http://BendixCarstensen/APC/EDSD-2019

Age at entry (diagnosis) as covariate

t: time since entry (duration)
e: age at entry
$a = e + t$: current age (age at follow-up)

Duration as basic time-scale; linear effect of age at entry:

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

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

Non-linear effects of time-scales

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

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

Three quantities can be arbitrarily moved between the three functions:

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

because $t - a + e = 0$.

This is the age-period-cohort modeling problem again.

"Controlling for age"

— is not a well defined statement:

- Mostly it means that age **at entry** is included in the model.
- But ideally one would check whether there were non-linear effects of age at entry and current age.
- This would require modeling of multiple timescales.
- Which is best accomplished by splitting follow up and using Poisson models, with time scales as covariates.

Tabulation in the Lexis diagram

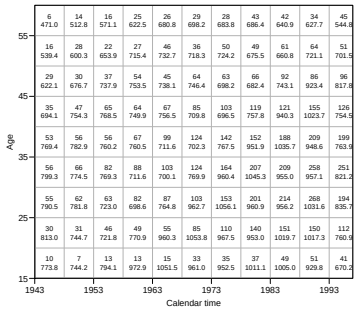
Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019` `Lexis-tab`

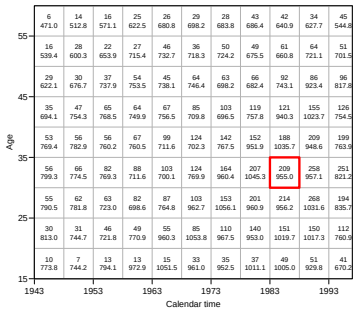
Tabulation of register data



Testis cancer cases in Denmark.

Male person-years in Denmark.

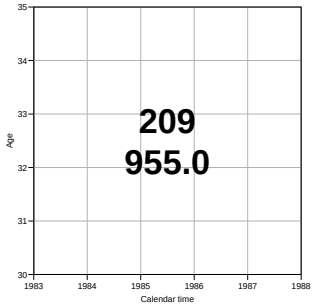
Tabulation of register data



Testis cancer cases in Denmark.

Male person-years in Denmark.

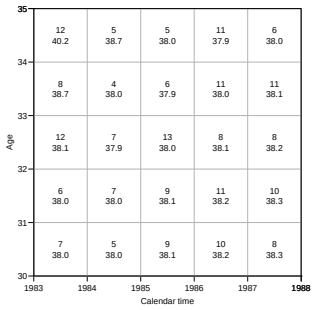
Tabulation of register data



Testis cancer cases in Denmark.

Male person-years in Denmark.

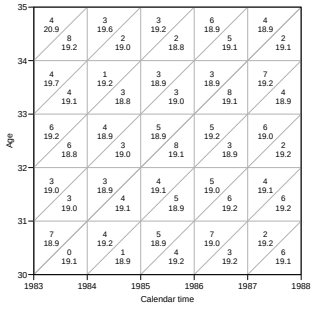
Tabulation of register data



Testis cancer cases in Denmark.

Male person-years in Denmark.

Tabulation of register data



Testis cancer cases in Denmark.

Male person-years in Denmark.

Subdivision by year of birth (cohort).

Major sets in the Lexis diagram

A-sets: Classification by age and period. (□)

B-sets: Classification by age and cohort. (▱)

C-sets: Classification by cohort and period. (▯)

The mean age, period and cohort for these sets is just the mean of the tabulation interval.

The mean of the third variable is found by using $a = p - c$.

Analysis of rates from a complete observation in a Lexis diagram need not be restricted to these classical sets classified by two factors.

We may classify cases and risk time by all three factors **Lexis triangles**:

Upper triangles: Classification by age and period, earliest born cohort. (◸)

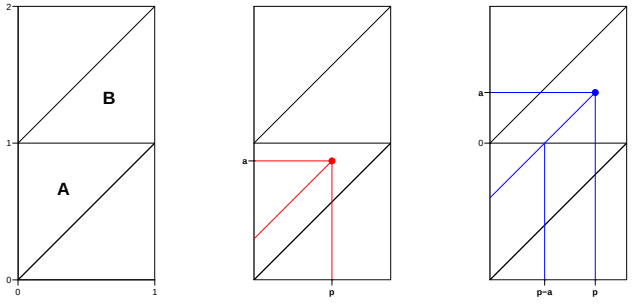
Lower triangles: Classification by age and period, latest born cohort. (◿)

Mean a, p and c during FU in triangles

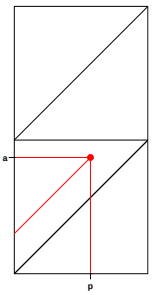
Modeling requires that each set (=observation in the dataset) be assigned a value of age, period and cohort. So for each triangle we need:

- mean age at risk.
- mean date at risk.
- mean cohort at risk.

Means in upper (A) and lower (B) triangles:



Upper triangles (◸), A:

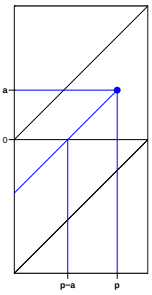


$$\mathrm{E}_{\mathbf{A}}(a) = \int_{p=0}^{p=1}\int_{a=p}^{a=1} a \times 2 \,\mathrm{d}a \,\mathrm{d}p = \int_{p=0}^{p=1} 1 - p^2 \,\mathrm{d}p = \frac{2}{3}$$

$$\mathrm{E}_{\mathbf{A}}(p) = \int_{a=0}^{a=1}\int_{p=0}^{p=a} p \times 2 \,\mathrm{d}p \,\mathrm{d}a = \int_{a=0}^{a=1} a^2 \,\mathrm{d}p = \frac{1}{3}$$

$$\mathrm{E}_{\mathbf{A}}(c) = \frac{1}{3} - \frac{2}{3} = -\frac{1}{3}$$

Lower triangles (◿), B:

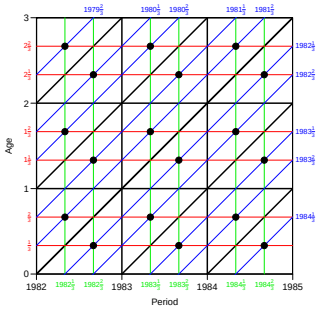


$$\mathrm{E}_{\mathbf{B}}(a) = \int_{p=0}^{p=1}\int_{a=0}^{a=p} a \times 2 \,\mathrm{d}a \,\mathrm{d}p = \int_{p=0}^{p=1} p^2 \,\mathrm{d}p = \frac{1}{3}$$

$$\mathrm{E}_{\mathbf{B}}(p) = \int_{a=0}^{a=1}\int_{p=a}^{p=1} p \times 2 \,\mathrm{d}p \,\mathrm{d}a = \int_{a=0}^{a=1} 1 - a^2 \,\mathrm{d}p = \frac{2}{3}$$

$$\mathrm{E}_{\mathbf{B}}(c) = \frac{2}{3} - \frac{1}{3} = \frac{1}{3}$$

Tabulation by age, period and cohort



Gives triangular sets with differing mean age, period and cohort:

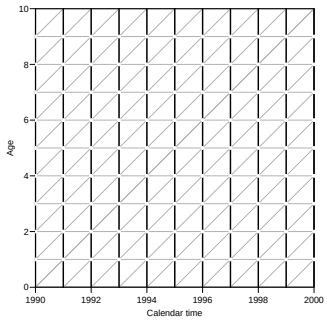
These correct midpoints for age, period and cohort must be used in modeling.

From population figures to risk time

Population figures in the form of size of the population at certain date are available from most statistical bureaus.

This corresponds to population sizes along the vertical lines in the diagram.

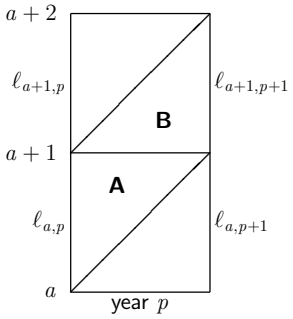
We want risk time figures for the population in the squares and triangles in the diagram.



Prevalent population figures

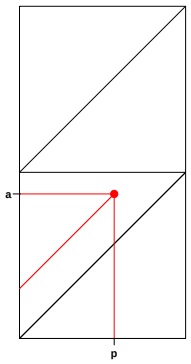
$\ell_{a,p}$ is the number of persons in age class a alive at the beginning of period (=year) p.

The aim is to compute person-years for the triangles **A** and **B**, respectively.



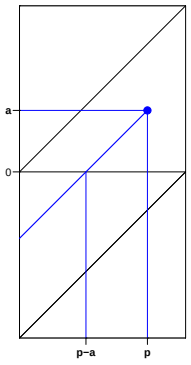
The area of the triangle is $1/2$, so the uniform measure over the triangle has density 2. Therefore a person dying in age a at date p in **A** contributes p risk time in **A**, so the average will be:

$$\int_{p=0}^{p=1}\int_{a=p}^{a=1} 2p \; \mathrm{d}a \; \mathrm{d}p$$
$$= \int_{p=0}^{p=1} 2p - 2p^2 \; \mathrm{d}p$$
$$= \left[p^2 - \frac{2p^3}{3}\right]_{p=0}^{p=1} = \frac{1}{3}$$



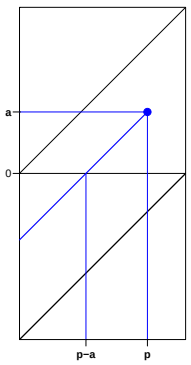
A person dying in age a at date p in **B** contributes $p - a$ risk time in **A**, so the average will be (again using the density 2 of the uniform measure):

$$\int_{p=0}^{p=1}\int_{a=0}^{a=p} 2(p-a) \; \mathrm{d}a \; \mathrm{d}p$$
$$= \int_{p=0}^{p=1} \left[2pa - a^2\right]_{a=0}^{a=p} \; \mathrm{d}p$$
$$= \int_{p=0}^{p=1} p^2 \; \mathrm{d}p = \frac{1}{3}$$



A person dying in age a at date p in **B** contributes a risk time in **B**, so the average will be:

$$\int_{p=0}^{p=1}\int_{a=0}^{a=p} 2a \; \mathrm{d}a \; \mathrm{d}p$$
$$= \int_{p=0}^{p=1} p^2 \; \mathrm{d}p = \frac{1}{3}$$



Mean contributions to risk time in A and B:

	A:	**B**:
Survivors:	$\ell_{a+1,p+1} \times \frac{1}{2}$y	$\ell_{a+1,p+1} \times \frac{1}{2}$y
Dead in **A**:	$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}$y	
Dead in **B**:	$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}$y	$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}$y
$\sum$	$(\frac{1}{3}\ell_{a,p} + \frac{1}{6}\ell_{a+1,p+1}) \times 1$y	$(\frac{1}{6}\ell_{a,p} + \frac{1}{3}\ell_{a+1,p+1}) \times 1$y

The number of deaths in **A** and **B** is $\ell_{a,p} - \ell_{a+1,p+1}$, and we assume that half occur in **A** and half in **B**.

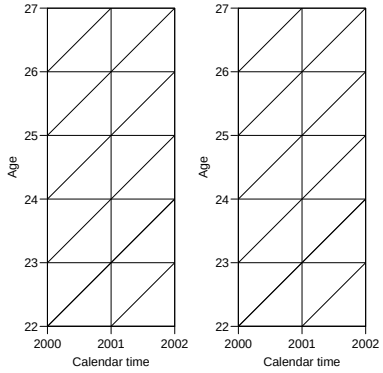
Population as of 1. January from Statistics Denmark:

	Men			Women		
	2000	2001	2002	2000	2001	2002
Age						
22	33435	33540	32272	32637	32802	31709
23	35357	33579	33742	34163	32853	33156
24	38199	35400	33674	37803	34353	33070
25	37958	38257	35499	37318	37955	34526
26	38194	38048	38341	37292	37371	38119
27	39891	38221	38082	39273	37403	37525

Exercise:

Fill in the risk time figures in as many triangles as possible from the previous table for men and women, respectively.

Look at the `N2Y` function in `Epi`.



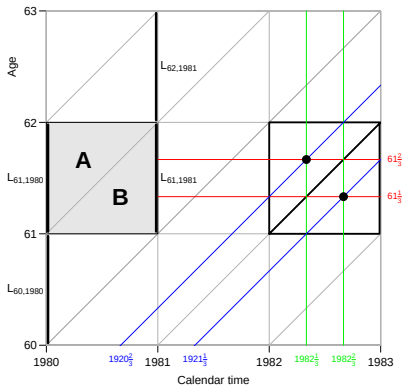
Summary:

Population risk time:

A: $(\frac{1}{3}\ell_{a,p} + \frac{1}{6}\ell_{a+1,p+1}) \times 1$y

B: $(\frac{1}{6}\ell_{a-1,p} + \frac{1}{3}\ell_{a,p+1}) \times 1$y

Mean age, period and cohort: $\frac{1}{3}$ into the interval.



APC-model for triangular data

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019`

Model for triangular data

- One parameter per distinct value on each timescale.
- Example: 3 age-classes and 3 periods:
  - 6 age parameters
  - 6 period parameters
  - 10 cohort parameters
- Model:

$$\lambda_{ap} = \alpha_a + \beta_p + \gamma_c$$

Problem: Disconnected design!

Log-likelihood contribution from one triangle:

$$D_{ap}\log(\lambda_{ap}) - \lambda_{ap}\,Y_{ap} = D_{ap}\log(\alpha_a + \beta_p + \gamma_c) - (\alpha_a + \beta_p + \gamma_c)\,Y_{ap}$$

The log-likelihood can be separated:

$$\sum_{a,p\,\in\,\bigtriangledown} D_{ap}\log(\lambda_{ap}) - \lambda_{ap}\,Y_{ap} + \sum_{a,p\,\in\,\bigtriangleup} D_{ap}\log(\lambda_{ap}) - \lambda_{ap}\,Y_{ap}$$

No common parameters between terms
— we have two separate models:
One for upper triangles, one for lower.

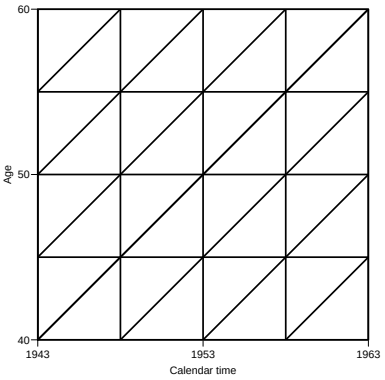
Illustration by lung cancer data

```
> library( Epi )
> data( lungDK )
> lungDK[1:10,]
   A5   P5   C5 up       Ax       Px       Cx  D        Y
1  40 1943 1898  1 43.33333 1944.667 1901.333 52 336233.8
2  40 1943 1903  0 41.66667 1946.333 1904.667 28 357812.7
3  40 1948 1903  1 43.33333 1949.667 1906.333 51 363783.7
4  40 1948 1908  0 41.66667 1951.333 1909.667 30 390985.8
5  40 1953 1908  1 43.33333 1954.667 1911.333 50 391925.3
6  40 1953 1913  0 41.66667 1956.333 1914.667 23 377515.3
7  40 1958 1913  1 43.33333 1959.667 1916.333 56 365575.5
8  40 1958 1918  0 41.66667 1961.333 1919.667 43 383689.0
9  40 1963 1918  1 43.33333 1964.667 1921.333 44 385878.5
10 40 1963 1923  0 41.66667 1966.333 1924.667 38 371361.5
```


Fill in the number of cases (`D`) and person-years (`Y`) from previous slide.

Indicate birth cohorts on the axes for upper and lower triangles.

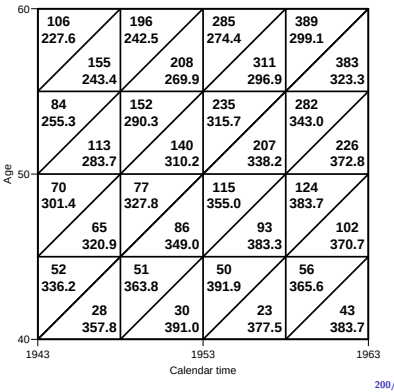
Mark mean date of birth for these.



Fill in the number of cases (`D`) and person-years (`Y`) from previous slide.

Indicate birth cohorts on the axes for upper and lower triangles.

Mark mean date of birth for these.



APC-model with "synthetic" cohorts

```
> mc <- glm( D ~ factor(A5) - 1 +
+                factor(P5-A5) +
+                factor(P5) + offset( log( Y ) ),
+            family=poisson )
> summary( mc )

...

    Null deviance: 1.0037e+08  on 220  degrees of freedom
Residual deviance: 8.8866e+02  on 182  degrees of freedom
```

No. parameters: $220 - 182 = 38$.

$$A = 10, \quad P = 11, \quad C = 20 \quad \Rightarrow \quad A + P + C - 3 = 38$$

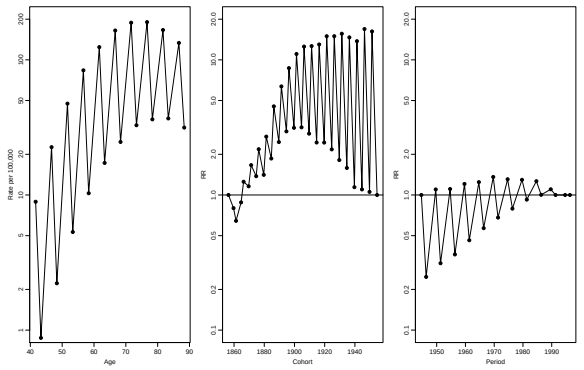
APC-model with "correct" cohorts

```
> mx <- glm( D ~ factor(Ax) - 1 +
+                factor(Cx) +
+                factor(Px) + offset( log( Y ) ),
+            family=poisson )
> summary( mx )
...
    Null deviance: 1.0037e+08  on 220  degrees of freedom
Residual deviance: 2.8473e+02  on 144  degrees of freedom
```

No. parameters: $220 - 144 = 76 \quad (= 38 \times 2)$.

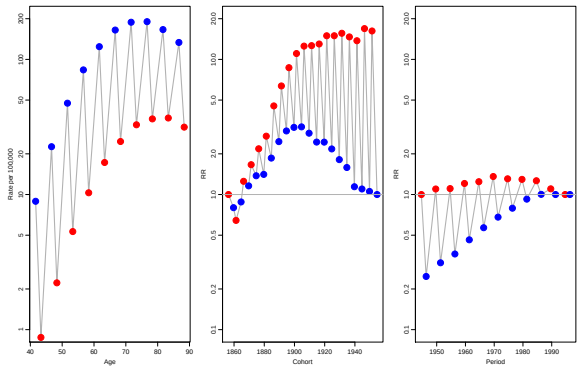
$$A = 20, \quad P = 22, \quad C = 40 \quad \Rightarrow \quad A + P + C - 3 = 79 \neq 76!$$

We have fitted two age-period-cohort models separately to upper and lower triangles.



Modeling for Lexis triangles

- Modeling by factors not possible
- Two separate models that cannot be fitted together
- We are not using the **quantitative** values of age, period and cohort.
- **Solution**: parametric models using the quantitative nature of a, p and $c = p - a$.
- . . . so we need to handle smooth parametric functions.



Non-linear effects

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019` `crv-mod`

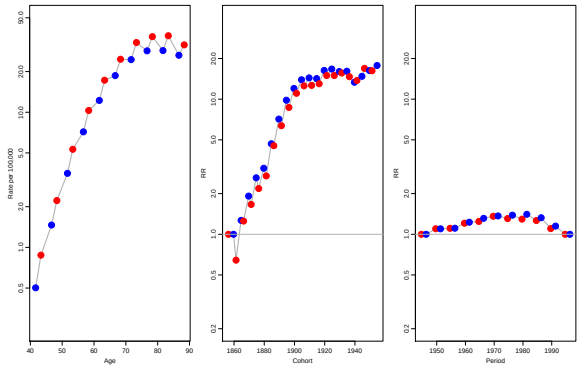
Now, explicitly fit models for upper and lower triangles:

```
> mx.u <- glm( D ~ factor(Ax) - 1 +
+                  factor(Cx) +
+                  factor(Px) + offset( log( Y/10^5 ) ), family=poisson,
+                  data=lungDK[lungDK$up==1,] )
> mx.l <- glm( D ~ factor(Ax) - 1 +
+                  factor(Cx) +
+                  factor(Px) + offset( log( Y/10^5 ) ), family=poisson,
+                  data=lungDK[lungDK$up==0,] )
> mx$deviance
[1] 284.7269
> mx.l$deviance
[1] 134.4566
> mx.u$deviance
[1] 150.2703
> mx.l$deviance+mx.u$deviance
[1] 284.7269
```


Testis cancer

Testis cancer in Denmark:

```
> library( Epi )
> data( testisDK )
>  str( testisDK )

'data.frame': 4860 obs. of  4 variables:
 $ A: num  0 1 2 3 4 5 6 7 8 9 ...
 $ P: num  1943 1943 1943 1943 1943 ...
 $ D: num  1 1 0 1 0 0 0 0 0 0 ...
 $ Y: num  39650 36943 34588 33267 32614 ...

> head( testisDK )

  A    P D        Y
1 0 1943 1 39649.50
2 1 1943 1 36942.83
3 2 1943 0 34588.33
4 3 1943 1 33267.00
5 4 1943 0 32614.00
6 5 1943 0 32020.33
```

Cases, PY and rates

```
> print(
+ stat.table( list( A = floor(A/10)*10,
+                     P = floor(P/10)*10),
+               list( D = sum(D),
+                     Y = sum(Y/1000),
+                  rate = ratio(D,Y,10^6) ),
+               margins = TRUE, data = testisDK ), digits=c(sum=0,ratio=2) )
```

	P						
A	1940	1950	1960	1970	1980	1990	Total
0	10	7	16	18	9	10	70
	2605	4037	3885	3821	3071	2166	19584
	3.84	1.73	4.12	4.71	2.93	4.62	3.57
10	13	27	37	72	97	75	321
	2136	3505	4004	3906	3847	2261	19659
	6.09	7.70	9.24	18.43	25.21	33.17	16.33
20	124	221	280	535	724	557	2441

Linear effects in glm

How do rates depend on age?

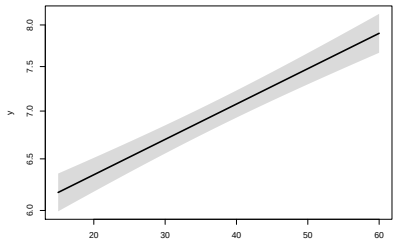
```
> ml <- glm( D ~ A, offset=log(Y), family=poisson, data=testisDK )
> round( ci.lin( ml ), 4 )

            Estimate StdErr         z P    2.5%   97.5%
(Intercept)  -9.7755 0.0207 -472.3164 0 -9.8160 -9.7349
A             0.0055 0.0005   11.3926 0  0.0045  0.0064

> round( ci.exp( ml ), 4 )

            exp(Est.)   2.5%  97.5%
(Intercept)    0.0001 0.0001 0.0001
A              1.0055 1.0046 1.0064
```

Linear increase of log-rates by age

Linear effects in glm

```
> nd <- data.frame( A=15:60, Y=10^5 )
> pr <- ci.pred( ml, newdata=nd )
> head( pr )

  Estimate     2.5%    97.5%
1 6.170105 5.991630 6.353896
2 6.204034 6.028525 6.384652
3 6.238149 6.065547 6.415662
4 6.272452 6.102689 6.446937
5 6.306943 6.139944 6.478485
6 6.341624 6.177301 6.510319

> matplot( nd$A, pr, type="l", lty=1, lwd=c(3,1,1), col="black", log="y" )
```


Linear effects in glm



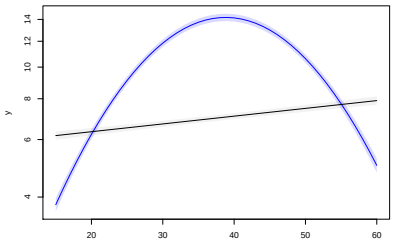

```
> nd <- data.frame( A=15:60, Y=10^5 )
> pr <- ci.pred( ml, newdata=nd )
> matshade( nd$A, pr, plot=TRUE, lwd=3, log="y" )
```


Quadratic effects in glm

How do rates depend on age?

```
> mq <- glm( D ~ A + I(A^2),
+            offset=log(Y), family=poisson, data=testisDK )
> round( ci.lin( mq ), 4 )

            Estimate StdErr         z P     2.5%    97.5%
(Intercept) -12.3656 0.0596 -207.3611 0 -12.4825 -12.2487
A             0.1806 0.0033   54.8290 0   0.1741   0.1871
I(A^2)       -0.0023 0.0000  -53.7006 0  -0.0024  -0.0022

> round( ci.exp( mq ), 4 )

            exp(Est.)   2.5%  97.5%
(Intercept)    0.0000 0.0000 0.0000
A              1.1979 1.1902 1.2057
I(A^2)         0.9977 0.9976 0.9978
```


Quadratic effect in glm

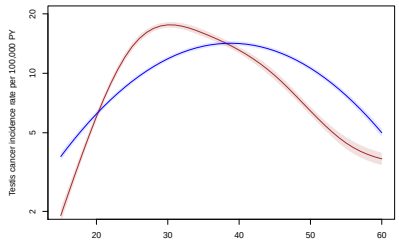



```
> matshade( nd$A, cbind( ci.pred(mq,nd), ci.pred(ml,nd) ), plot=TRUE,
+           log="y", col=c("blue","black"), alpha=c(15,5)/100 )
```


Spline effects in glm

```
> library( splines )
> ms <- glm( D ~ Ns(A,knots=seq(15,65,10)),
+            offset=log(Y), family=poisson, data=testisDK )
> round( ci.exp( ms ), 3 )

                                exp(Est.)   2.5%  97.5%
(Intercept)                         0.000  0.000  0.000
Ns(A, knots = seq(15, 65, 10))1     8.548  7.650  9.551
Ns(A, knots = seq(15, 65, 10))2     5.706  4.998  6.514
Ns(A, knots = seq(15, 65, 10))3     1.002  0.890  1.128
Ns(A, knots = seq(15, 65, 10))4    14.402 11.896 17.436
Ns(A, knots = seq(15, 65, 10))5     0.466  0.429  0.505

> matplot( nd$A, ci.pred( ms, nd ),
+          log="y", xlab="Age", ylab="Testis cancer incidence rate per 100,000 PY
+          type="l", lty=1, lwd=c(3,1,1), col="black", ylim=c(2,20) )
```


Spline effects in glm




```
> matshade( nd$A, cbind(ci.pred(ms,nd),ci.pred(mq,nd)), plot=TRUE,
+           log="y", xlab="Age", ylab="Testis cancer incidence rate per 100,000 PY
+           col=c("brown","blue"), alpha=c(15,10)/100, ylim=c(2,20) )
```


Adding a linear period effect

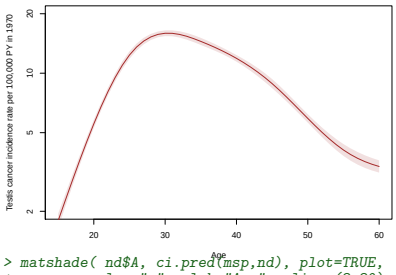
```
> msp <- glm( D ~ Ns(A,knots=seq(15,65,10)) + P,
+             offset=log(Y), family=poisson, data=testisDK )
> nd <- data.frame( A=15:60, Y=10^5, P=1970 )
```

A multiplicative model:

$$\lambda(a,p) = f(a) \times g(p), \quad g(p_{\text{ref}}) = 1$$

$f(a)$: Rate at p_{ref}
$g(p)$: Rate ratio relative to p_{ref}

Adding a linear period effect




```
> matshade( nd$A, ci.pred(msp,nd), plot=TRUE,
+           log="y", xlab="Age", ylim=c(2,20), col="brown", alpha=0.15,
+           ylab="Testis cancer incidence rate per 100,000 PY in 1970" )
```


The period effect

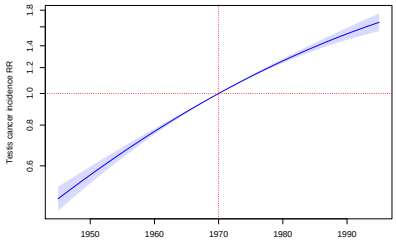
```
> nd.p <- data.frame( P=1945:1995 )
> nd.r <- data.frame( P=1970 )
> str( nd.p )

'data.frame': 51 obs. of  1 variable:
 $ P: int  1945 1946 1947 1948 1949 1950 1951 1952 1953 1954 ...

> str( nd.r )

'data.frame': 1 obs. of  1 variable:
 $ P: num 1970

> RR <- ci.exp( msp, ctr.mat=list(nd.p,nd.r), xvars="A" )
> matshade( nd.p$P, RR, plot=TRUE,
+           log="y", xlab="Date", ylab="Testis cancer incidence RR",
+           type="l", lty=1, lwd=c(3,1,1), col="black" )
> abline( v=1970, h=1, col="red" )
```


A quadratic period effect

```
> mspq <- glm( D ~ Ns(A,knots=seq(15,65,10)) + P + I(P^2),
+              offset=log(Y), family=poisson, data=testisDK )
> round( ci.exp( mspq ), 4 )
```

	exp(Est.)	2.5%	97.5%
(Intercept)	0.0000	0.0000	0.0000
Ns(A, knots = seq(15, 65, 10))1	8.3560	7.4783	9.3366
Ns(A, knots = seq(15, 65, 10))2	5.5133	4.8290	6.2945
Ns(A, knots = seq(15, 65, 10))3	1.0060	0.8935	1.1326
Ns(A, knots = seq(15, 65, 10))4	13.4388	11.1008	16.2691
Ns(A, knots = seq(15, 65, 10))5	0.4582	0.4223	0.4971
P	2.1893	1.4566	3.2906
I(P^2)	0.9998	0.9997	0.9999

A quadratic period effect



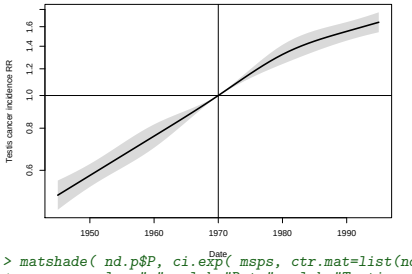

```
> matshade( nd.p$P, ci.exp( mspq, ctr.mat=list(nd.p,nd.r), xvars="A" ), plot=TRUE,
+           log="y", xlab="Date", ylab="Testis cancer incidence RR", col="blue" )
> abline( h=1, v=1970, col="red", lty="13" )
```


A spline period effect

```
> msps <- glm( D ~ Ns(A,knots=seq(15,65,10)) +
+                  Ns(P,knots=seq(1950,1990,10)),
+                  offset=log(Y), family=poisson, data=testisDK )
> round( ci.exp( msps ), 3 )
```

	exp(Est.)	2.5%	97.5%
(Intercept)	0.000	0.000	0.000
Ns(A, knots = seq(15, 65, 10))1	8.327	7.452	9.305
Ns(A, knots = seq(15, 65, 10))2	5.528	4.842	6.312
Ns(A, knots = seq(15, 65, 10))3	1.007	0.894	1.133
Ns(A, knots = seq(15, 65, 10))4	13.447	11.107	16.279
Ns(A, knots = seq(15, 65, 10))5	0.458	0.422	0.497
Ns(P, knots = seq(1950, 1990, 10))1	1.711	1.526	1.918
Ns(P, knots = seq(1950, 1990, 10))2	2.190	2.028	2.364
Ns(P, knots = seq(1950, 1990, 10))3	3.222	2.835	3.661
Ns(P, knots = seq(1950, 1990, 10))4	2.299	2.149	2.459

Period effect

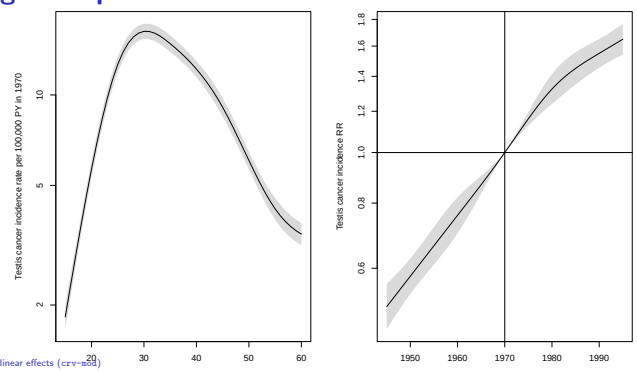



```
> matshade( nd.p$P, ci.exp( msps, ctr.mat=list(nd.p,nd.r), xvars="A" ), plot=TRUE,
+           log="y", xlab="Date", ylab="Testis cancer incidence RR", lwd=3 )
> abline( h=1, v=1970 )
```


Period effect

```
> par( mfrow=c(1,2) )
> matshade( nd$A, ci.pred( msps, nd ), plot=TRUE,
+           log="y", xlab="Age", col="black",
+           ylab="Testis cancer incidence rate per 100,000 PY in 1970" )
> matshade( nd.p$P, ci.exp( msps, ctr.mat=list(nd.p,nd.r), xvars="A" ), plot=TRUE,
+           log="y", xlab="Date", ylab="Testis cancer incidence RR",
+           col="black" )
> abline( h=1, v=1970 )
```


Age and period effect



Age and period effect with `ci.exp`

- In rate models there is always one term with the **rate** dimension.
  Usually **age**
- But it must refer to specific **reference** values for **all other** variables (in this case only `P`).
- For the "other" variables, report the RR **relative** to the reference point.
- Only parameters relevant for the variable (`P`) actually used in the calculation.
- We are computing the difference between two predictions.
- . . . as well as the confidence intervals for it.

APC-model: Parametrization

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019` `APC-par`

What's the problem?

- One parameter is assigned to each distinct value of the timescales, the **scale** of the variables is not used.
- The solution is to "tie together" the points on the scales together with smooth functions of the **mean** age, period and cohort with three functions:

$$\lambda_{ap} = f(a) + g(p) + h(c)$$

- The practical problem is how to choose a reasonable parametrization of these functions, and how to get estimates.

The identifiability problem still exists:

$$c = p - a \quad \Leftrightarrow \quad p - a - c = 0$$

$$\begin{aligned}
\lambda_{ap} &= f(a) + g(p) + h(c) \\
&= f(a) + g(p) + h(c) + \gamma(p - a - c) \\
&= f(a) - \mu_a - \gamma a + \\
&\quad g(p) + \mu_a + \mu_c + \gamma p + \\
&\quad h(c) - \mu_c - \gamma c
\end{aligned}$$

A decision on parametrization is needed.
. . . it must be **external to the model**.

Smooth functions

$$\log\big(\lambda(a,p)\big) = f(a) + g(p) + h(c)$$

Possible choices for non-linear parametric functions describing the effect of the three **quantitative** variables:

- Polynomials / fractional polynomials.
- Linear / quadratic / cubic splines.
- Natural splines.

All of these contain the linear effect as special case.

Parametrization of effects

There are still three "free" parameters:

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

Any set of 3 numbers, μ_a, μ_c and γ will produce effects with the same sum:

$$\tilde{f}(a) + \tilde{g}(p) + \tilde{h}(c) = f(a) + g(p) + h(c)$$

The problem is to choose μ_a, μ_c and γ according to some criterion for the functions.

Parametrization principle

1. The age-function should be interpretable as log age-specific rates in a cohort c_0 after adjustment for the period effect.
2. The cohort function is 0 at a reference cohort c_0, interpretable as log-RR relative to cohort c_0.
3. The period function is 0 on average with 0 slope, interpretable as log-RR relative to the age-cohort prediction. (residual log-RR).

This will yield cohort age-effects a.k.a. **longitudinal** age effects.

Biologically interpretable:
— what happens during the lifespan of a cohort?

Period-major parametrization

- Alternatively, the period function could be constrained to be 0 at a reference date, p_0.
- Then, age-effects at $a_0 = p_0 - c_0$ would equal the fitted rate for period p_0 (and cohort c_0), and the period effects would be residual log-RRs relative to p_0.
- Gives period or **cross-sectional** age-effects
- Bureaucratically interpretable:
  — what was seen at a particular date?

Implementation:

1. Obtain any set of parameters $f(a)$, $g(p)$, $h(c)$.
2. Extract the trend from the period effect (find μ and β):
$$\tilde{g}(p) = \hat{g}(p) - (\mu + \beta p)$$
3. Decide on a reference cohort c_0.
4. Use the functions:
$$\begin{aligned}\tilde{f}(a) &= \hat{f}(a) + \mu + \beta a + \hat{h}(c_0) + \beta c_0 \\ \tilde{g}(p) &= \hat{g}(p) - \mu - \beta p \\ \tilde{h}(c) &= \hat{h}(c) \qquad + \beta c - \hat{h}(c_0) - \beta c_0\end{aligned}$$

"Extract the trend"

- **Not** a well-defined concept:
  - Regress $\hat{g}(p)$ on p for all units in the dataset.
  - Regress $\hat{g}(p)$ on p for all different values of p.
  - Weighted regression — what weights?
- How do we get the standard errors?
- Matrix-algebra!
- Projections!
- Weighted inner product...

Parametic function

Suppose that $g(p)$ is parametrized using the design matrix $\mathbf{M}$, with the estimated parameters π.

Example: 2nd degree polynomial:

$$\mathbf{M} = \begin{bmatrix} 1 & p_1 & p_1^2 \\ 1 & p_2 & p_2^2 \\ \vdots & \vdots & \vdots \\ 1 & p_n & p_n^2 \end{bmatrix} \qquad \pi = \begin{bmatrix} \pi_0 \\ \pi_1 \\ \pi_2 \end{bmatrix} \qquad g(p) = \mathbf{M}\pi$$

`nrow(M)` is the no. of observations in the dataset,

`ncol(M)` is the no. of parameters

Extract the trend from g:

Vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if the inner product is 0

$$\mathbf{x} \perp \mathbf{y} \quad \Leftrightarrow \quad \langle \mathbf{x} | \mathbf{y} \rangle = \sum_i x_i y_i = 0$$

- $\langle \tilde{g}(p) | 1 \rangle = 0, \langle \tilde{g}(p) | p \rangle = 0$, i.e. $\tilde{g}$ is **orthogonal** to $[1{:}p]$.
- Suppose $\tilde{g}(p) = \tilde{\mathbf{M}}\pi$, then for **any** parameter vector π:
$$\langle \tilde{\mathbf{M}}\pi | 1 \rangle = 0, \quad \langle \tilde{\mathbf{M}}\pi | p \rangle = 0 \quad \Longrightarrow \quad \tilde{\mathbf{M}} \perp [1{:}p]$$
- Thus we just need to be able to produce $\tilde{\mathbf{M}}$ from $\mathbf{M}$:
Projection on the orthogonal complement of span$([1{:}p])$.
- **But:** orthogonality requires an inner product!

Practical parametization

1. Set up model matrices for age, period and cohort, M_a, M_p and M_c. Intercept in all three.
2. Extract the linear trend from M_p and M_c, by projecting their columns onto the orthogonal complement of $[1{:}p]$ and $[1{:}c]$, respectively
3. Center the cohort effect around c_0:
Take a row from $\tilde{M}_c$ corresponding to c_0, replicate to dimension as $\tilde{M}_c$, and subtract it from $\tilde{M}_c$ to form $\tilde{M}_{c_0}$.

4. Use:
M_a for the age-effects,
$\tilde{M}_p$ for the period effects and
$[c - c_0{:}\tilde{M}_{c_0}]$ for the cohort effects.
5. Value of $\hat{f}(a)$ is $M_a\hat{\beta}_a$, similarly for the other two effects.
Variance is found by $M_a'\hat{\Sigma}_a M_a$, where $\hat{\Sigma}_a$ is the variance-covariance matrix of $\hat{\beta}_a$.

Information about a parameter in the data

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

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

so $I(\hat{\theta}) = \mathrm{e}^{\hat{\theta}} Y = \hat{\lambda} Y = D.$

Information about sqare root of rate $\sigma = \sqrt{\lambda}$:

$$l(\sigma | D, Y) = D\log(\sigma^2) - \sigma^2 Y, \quad l'_\sigma = (D/\sigma^2)2\sigma - 2\sigma Y = 2D/\sigma - 2\sigma Y$$
$$l''_\sigma = -2D/\sigma^2 - 2Y$$

so $I(\hat{\sigma}) = -2D/\hat{\sigma}^2 - 2Y = -4Y$

Information in the data and inner product

- Inner products:
$$\langle \mathbf{m}_j | \mathbf{m}_k \rangle = \sum_i m_{ij} m_{ik} \qquad \langle \mathbf{m}_j | \mathbf{m}_k \rangle = \sum_i m_{ij} w_i m_{ik}$$
- Weights could be chosen as:
  - $w_i = Y_i$, *i.e.* proportional to the information content for $\sigma = \sqrt{\lambda}$, `dr.extr = Y` (the default)
  - $w_i = D_i$, *i.e.* proportional to the information content for $\theta = \log(\lambda)$, `dr.extr` $\in$ `c(D,T)`
  - $w_i = Y_i^2/D_i$, *i.e.* proportional to the information content for λ, `dr.extr` $\in$ `c(L,R)`
  - $w_i = 1$, the "usual" inner product — implicitly used in most of the literature — any other (character) value for `dr.extr`.

How to? I

Implemented in `apc.fit` in the `Epi` package:

```
> library( Epi )
> library( splines )
> data( lungDK )



> mw <- apc.fit( A = lungDK$Ax,
+                P = lungDK$Px,
+                D = lungDK$D,
+                Y = lungDK$Y/1000,
+           ref.c = 1900,
+            npar = 8,
+            parm = "ACP",
+         dr.extr = "y", print.AOV=FALSE ) # drift extraction - choice of inner p
```


How to? II

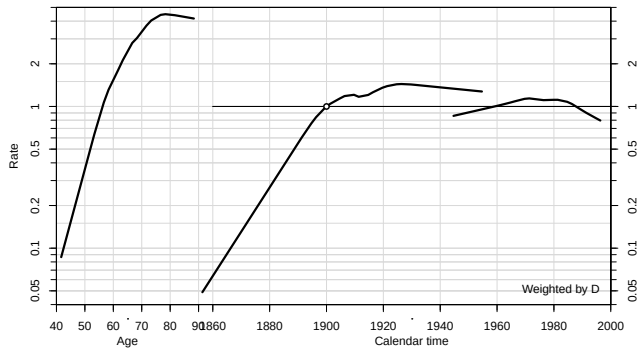
```
Latest version: FALSErefs: 1979.667 1900
NOTE: npar is specified as:
A P C
8 8 8

> mw$Ref

 Per  Coh
  NA 1900

> cbind( mw$Age[1:4,1:2], mw$Per[1:4,1:2], mw$Coh[1:4,1:2] )

          Age       Rate      Per      P-RR      Coh       C-RR
[1,] 41.66667 0.08648277 1944.667 0.8584085 1856.333 0.04890071
[2,] 43.33333 0.11507077 1946.333 0.8736385 1859.667 0.06214798
[3,] 46.66667 0.20372101 1949.667 0.9049139 1861.333 0.07006207
[4,] 48.33333 0.27106362 1951.333 0.9209689 1864.667 0.08904198

> plot( mw )
```


How to? III

```
cp.offset    RR.fac
     1765         1

> mw$Drift

                exp(Est.)     2.5%    97.5%
APC (Y-weights)  1.020305 1.019450 1.021161
A-d              1.023487 1.022971 1.024003
```


Consult the help page for: `apc.fit` to see options for weights in inner product, type of function, variants of parametrization etc.

`apc.plot`, `apc.lines` and `apc.frame` to see how to plot the results.



Other models I

```
> lungDK$A <- lungDK$Ax
> lungDK$P <- lungDK$Px
> ml <- apc.fit( data = lungDK,
+                npar = 8,
+              ref.c = 1900,
+           dr.extr = "y" )

Latest version: TRUErefs: 1979.667 1900
NOTE: npar is specified as:
A P C
8 8 8
[1] "ML of APC-model Poisson with log(Y) offset : ( ACP ):\n"

Analysis of deviance for Age-Period-Cohort model

                Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
Age                   212    15468.6
```


Other models II

```
Age-drift             211     6858.9  1   8609.7 < 2.2e-16
Age-Cohort            205     1034.7  6   5824.1 < 2.2e-16
Age-Period-Cohort     199      423.2  6    611.6 < 2.2e-16
Age-Period            205     3082.6 -6  -2659.4 < 2.2e-16
Age-drift             211     6858.9 -6  -3776.3 < 2.2e-16

> ##
> my <- apc.fit( A = lungDK$Ax,
+                P = lungDK$Px,
+                D = lungDK$D,
+                Y = lungDK$Y/10^5,
+             npar = 8,
+           ref.c = 1900,
+          dr.extr = "y" ) # person-yeras, weight Y
```


Other models III

```
Latest version: FALSErefs: 1979.667 1900
NOTE: npar is specified as:
A P C
8 8 8
[1] "ML of APC-model Poisson with log(Y) offset : ( ACP ):\n"

Analysis of deviance for Age-Period-Cohort model

                Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
Age                   212    15468.6
Age-drift             211     6858.9  1   8609.7 < 2.2e-16
Age-Cohort            205     1034.7  6   5824.1 < 2.2e-16
Age-Period-Cohort     199      423.2  6    611.6 < 2.2e-16
Age-Period            205     3082.6 -6  -2659.4 < 2.2e-16
Age-drift             211     6858.9 -6  -3776.3 < 2.2e-16
```

Other models IV

```
> ##
> m1 <- apc.fit( A = lungDK$Ax,
+                P = lungDK$Px,
+                D = lungDK$D,
+                Y = lungDK$Y/10^5,
+            npar = 8,
+           ref.c = 1900,
+         dr.extr = "1" ) # usual inner product

Latest version: FALSErefs: 1979.667 1900
NOTE: npar is specified as:
A P C
8 8 8
[1] "ML of APC-model Poisson with log(Y) offset : ( ACP ):\n"

Analysis of deviance for Age-Period-Cohort model

                  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
Age                     212    15468.6
```


Other models V

```
Age-drift               211     6858.9  1   8609.7 < 2.2e-16
Age-Cohort              205     1034.7  6   5824.1 < 2.2e-16
Age-Period-Cohort       199      423.2  6    611.6 < 2.2e-16
Age-Period              205     3082.6 -6  -2659.4 < 2.2e-16
Age-drift               211     6858.9 -6  -3776.3 < 2.2e-16

> ##
> dr <- cbind( mw$Drift, m1$Drift, my$Drift, m1$Drift )
> rownames(dr) <- c("APC extract","Age-Drift")
> colnames(dr)[0:3*3+1] <- c("D-wt","Y^2/D-wt","Y-wt","1-wt")
> round( dr, 2 )

            D-wt 2.5% 97.5% Y^2/D-wt 2.5% 97.5% Y-wt 2.5% 97.5% 1-wt 2.5% 97.5%
APC extract 1.02 1.02  1.02     1.02 1.02  1.02 1.02 1.02  1.02 1.03 1.03  1.03
Age-Drift   1.02 1.02  1.02     1.02 1.02  1.02 1.02 1.02  1.02 1.02 1.02  1.02

> # % change per year
> round( (dr-1)*100, 1 )
```


Other models VI

```
            D-wt 2.5% 97.5% Y^2/D-wt 2.5% 97.5% Y-wt 2.5% 97.5% 1-wt 2.5% 97.5%
APC extract  2.0  1.9   2.1      2.0  1.9   2.1  2.0  1.9   2.1  3.3  3.2   3.4
Age-Drift    2.3  2.3   2.4      2.3  2.3   2.4  2.3  2.3   2.4  2.3  2.3   2.4
```

Substantial differences between the estimated drifts.

Parametrization of the APC model is arbitrary

- Separation of the three effects relies on arbitrary principles, *e.g.*:
  - Age is the primary effect
  - Cohort the secondary, reference c_0
  - Period is the residual
  - Inner product for trend extraction
- There is no magical fix that allows you to escape this, it comes from modelling a, p and $p - a$
- Any fix has some (hidden) assumption(s)
- . . . but the **fitted values** are the same

Lee-Carter model

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

http://BendixCarstensen/APC/EDSD-2019 LeeCarter

Lee-Carter model for (mortality) rates

Lee & Carter, JASA, 1992:

$$\log(\lambda_{x,t}) = a_x + b_x \times k_t$$

x is age; t is calendar time

- Formulated originally using as step-functions with one parameter per age/period.
- Implicitly assumes a data lay out by age and period: A, B or C-sets, but **not** Lexis triangles
- Using Lexis triangles with categorical set-up would just produce separate models for upper and lower triangles.

Lee-Carter model in continuous time

For **any** set of subsets of a Lexis diagram:

$$\log\big(\lambda(a,t)\big) = f(a) + b(a) \times k(t)$$

- $f(a)$, $b(a)$ smooth functions of age, a is **quantitative**
- $k(t)$ smooth function of period, t is **quantitative**
- Relative **scaling** of $b(a)$ and $k(t)$ cannot be determined
- $k(t)$ only determined up to an **affine** transformation:

$$\begin{aligned} f(a) + b(a)k(t) &= f(a) + \big(b(a)/n\big)\big(m + k(t) \times n\big) \\ &\quad - \big(b(a)/n\big) \times m \\ &= \tilde{f}(a) + \tilde{b}(a)\tilde{k}(t) \end{aligned}$$

Lee-Carter model in continuous time

$$\log\big(\lambda(a,t)\big) = f(a) + b(a) \times k(t)$$

- Lee-Carter model is an extension of the age-period model; if $b(a) = 1$ it **is** the age-period model.
- The extension is an age×period interaction, but not a traditional one:

$$\log\big(\lambda(a,t)\big) = f(a) + b(a) \times k(t) = f(a) + k(t) + \big(b(a) - 1\big) \times k(t)$$

- Main effect and interaction component of t are constrained to be identical.

Main effect and interaction term

Main effect and interaction component of t are constrained to be identical.

None of these are Lee-Carter models:

```
> glm( D ~ Ns(A,kn=a1.kn) + Ns(A,kn=a2.kn,i=T):Ns(P,kn=p.kn), ... )
> glm( D ~ Ns(A,kn=a1.kn) + Ns(A,kn=a2.kn,i=T)*Ns(P,kn=p.kn), ... )
> glm( D ~ Ns(A,kn=a1.kn) + Ns(P,kn=p.kn) + Ns(A,kn=a2.kn,i=T):Ns(P,kn=p.kn), ...
```


Lee-Carter model interpretation

$$\log\big(\lambda(a,p)\big) = f(a) + b(a) \times k(p)$$

- Constraints:
- $f(a)$ is the basic age-specific mortality
- $k(p)$ is the rate-ratio (RR) as a function of p:
  - relative to a p_{ref} where $k(p_{\text{ref}}) = 1$
  - for persons aged a_{ref} where $b(a_{\text{ref}}) = 1$
- $b(a)$ is an age-specific multiplier for the RR $k(p)$
- Choose p_{ref} and a_{ref} *a priori*.

Danish lung cancer data I

```
> lung <- read.table( "../data/apc-Lung.txt", header=T )
> head( lung )

  sex A    P    C D       Y
1   1 0 1943 1942 0 19546.2
2   1 0 1943 1943 0 20796.5
3   1 0 1944 1943 0 20681.3
4   1 0 1944 1944 0 22478.5
5   1 0 1945 1944 0 22369.2
6   1 0 1945 1945 0 23885.0

> # Only A by P classification - and only men over 40
> ltab <- xtabs( cbind(D,Y) ~ A + P, data=subset(lung,sex==1) )
> str( ltab )
```


Danish lung cancer data II

```
 xtabs [1:90, 1:61, 1:2] 0 0 0 0 0 0 0 0 0 0 ...
 - attr(*, "dimnames")=List of 3
  ..$ A: chr [1:90] "0" "1" "2" "3" ...
  ..$ P: chr [1:61] "1943" "1944" "1945" "1946" ...
  ..$  : chr [1:2] "D" "Y"
 - attr(*, "call")= language xtabs(formula = cbind(D, Y) ~ A + P, data = subset(lu
```

Lee-Carter modeling in **R**-packages:

- `demography` (`lca`)
- `ilc` (`lca.rh`)
- `Epi` (`LCa.fit`).

Lee-Carter with `demography` I

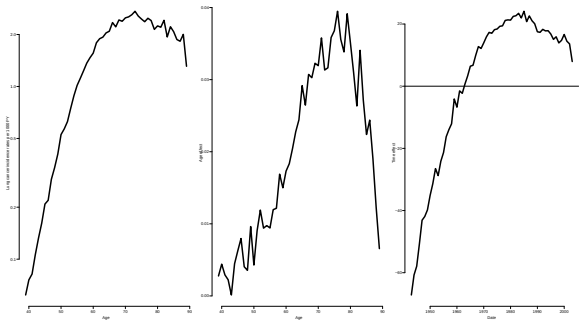
```
> library(demography)
> lcM <- demogdata( data = as.matrix(ltab[40:90,,"D"]/ltab[40:90,,"Y"]),
+                   pop = as.matrix(ltab[40:90,,"Y"]),
+                  ages = as.numeric(dimnames(ltab)[[1]][40:90]),
+                 years = as.numeric(dimnames(ltab)[[2]]),
+                  type = "Lung cancer incidence",
+                 label = "Denmark",
+                  name = "Male" )
```

`lca` estimation function checks the `type` argument, so we make a work-around, `mrt`:

Lee-Carter with `demography` II

```
> mrt <- function(x) { x$type <- "mortality" ; x }
> dmg.lcM <- lca( mrt(lcM), interpolate=TRUE )
> par( mfcol=c(1,3) )
> matplot( dmg.lcM$age, exp(dmg.lcM$ax)*1000,
+          log="y", ylab="Lung cancer incidence rates per 1000 PY",
+          xlab="Age", type="l", lty=1, lwd=4 )
> matplot( dmg.lcM$age, dmg.lcM$bx,
+          ylab="Age effect",
+          xlab="Age", type="l", lty=1, lwd=4 )
> matplot( dmg.lcM$year, dmg.lcM$kt,
+          ylab="Time effect",
+          xlab="Date", type="l", lty=1, lwd=4 )
> abline(h=0)
```


Lee-Carter with `demography`

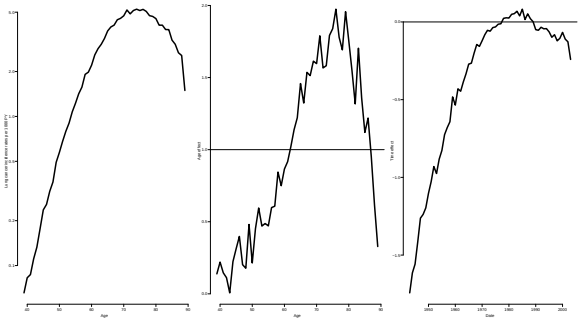


Lee-Carter re-scaled I

$$log\big(\hat{\lambda}(a,p)\big) = [f(a) + b(a) \times 20] + [b(a) \times 50] \times [(k(t) - 20)/50]$$

```
> par( mfcol=c(1,3) )
> matplot( dmg.lcM$age, exp(dmg.lcM$ax+dmg.lcM$bx*20)*1000,
+          log="y", ylab="Lung cancer incidence rates per 1000 PY",
+          xlab="Age", type="l", lty=1, lwd=4 )
> matplot( dmg.lcM$age, dmg.lcM$bx*50,
+          ylab="Age effect",
+          xlab="Age", type="l", lty=1, lwd=4 )
> abline(h=1)
> matplot( dmg.lcM$year, (dmg.lcM$kt-20)/50,
+          ylab="Time effect",
+          xlab="Date", type="l", lty=1, lwd=4 )
> abline(h=0)
```


Lee-Carter with demography rescaled



Lee-Carter with ilc

- The lca.rh function fits the model using maximum likeliood (proportional scaling)
- Fits the more general model and submodels of it:

$$\log\big(\lambda(a,p)\big) = f(a) + b(a) \times k(p) + c(a)m(p-a)$$

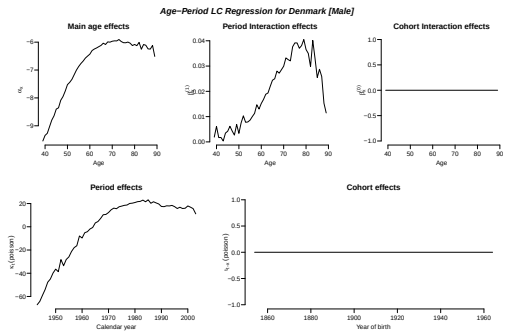
- Age interaction with between age and both period and/or cohort (=period-age)
- It is also an extension of the APC-model; if If $b(a) = 1$ and $c(a) = 1$ it's the APC-model.

$\Rightarrow$ suffers from the same identifiability problem

Lee-Carter with ilc I

```
> library( ilc )
> ilc.lcM <- lca.rh( mrt(lcM), model="lc", interpolate=TRUE, verbose=FALSE )

Original sample: Mortality data for Denmark
    Series: Male
    Years: 1943 - 2003
    Ages:  39 - 89
Applied sample: Mortality data for Denmark (Corrected: interpolate)
    Series: Male
    Years: 1943 - 2003
    Ages:  39 - 89

  Fitting model: [ LC = a(x)+b1(x)*k(t) ]
- with poisson error structure and with deaths as weights -

 Iterations finished in: 34 steps

> plot( ilc.lcM )
```


Lee-Carter with ilc



Lee-Carter with Epi

- LCa.fit fits the Lee-Carter model using natural splines for the **quantitative** effects of age and time.
- Normalizes effects to a reference age and period.
- The algoritm alternately fits a main age and period effects and the age-interaction effect.

$$\log\big(\lambda(a,p)\big) = f(a) + b(a) \times k(p) + c(a) \times m(p-a)$$
$$\log\big(\lambda(a,p)\big) = f(a) + b(a) \times k(p) + c(a) \times m(p-a)$$

Lee-Carter with Epi I

```
> library( Epi )
> Mlc <- subset( lung, sex==1 & A>39 )
> LCa.Mlc <- LCa.fit( Mlc, a.ref=60, p.ref=1980 )

LCa.fit convergence in  8  iterations, deviance: 8548.443 on 6084 d.f.

> LCa.Mlc

APa: Lee-Carter model with natural splines:
  log(Rate) = ax(Age) + pi(Age)kp(Per)
with 6, 5 and 5 parameters respectively.
Deviance: 8548.443 on 6084 d.f.

> plot( LCa.Mlc, rnam="Lung cancer incidence per 1000 PY" )
```


Lee-Carter with Epi

Lee-Carter and the APC-model

- Lee-Carter model is an interaction extension of the Age-Period model
- . . . or an interaction extension of the Age-Cohort model
- Age-Period-Cohort model is:
  - interaction extension
  - the smallest **union** of Age-Period and Age-Cohort
- Extended Lee-Carter (from the `ilc` package)

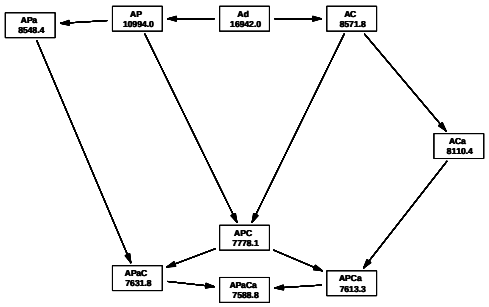
$$\log\big(\lambda(a,p)\big) = f(a) + b(a) \times k(p) + c(a)m(p-a)$$

is the union of all of these.

Lee-Carter and the APC-model

```
> system.time( allmod <- apc.LCa( Mlc, keep.models=TRUE ) )
> str( allmod )
> save( allmod, file='allmod.Rda' )


> load( file='allmod.Rda' )
> show.apc.LCa( allmod, top="Ad" )


> show.apc.LCa( allmod, top="AP" )


> show.apc.LCa( allmod, top="AC" )
```


Lee-Carter models and APC models



Lee-Carter models and APC models

- The classical Lee-Carter model is an extension of the Age-Period model with an interaction
- The Age-Period-Cohort model is an extension of the Age-Period model with an interaction
- Replacing period with cohort gives another type of Lee-Carter model
- The logical step is to consider all 9 models that comes from cross-classification of how the interaction term $b(a)$
  - Linear effect ($b(a) = 0$)
  - Non-linear effect ($b(a) = 1$)
  - Multiplicative interaction with age ($b(a)$ unconstrained)

Lee-Carter models and APC models

		$b_c(a)$		
		0	1	free
$b_p(a)$	0	Age	Age+Coh	LCa(C) AC, `ac`, `ACa`
	1	Age+Per	Age+Per+Coh H_0, `h0`	Age+Per+LCa(C) H_1, `h1`, `APCa`
	free	LCa(P) LC, `lc`, `APa`	Age+Coh+LCa(P) H_2, `h2`, `APaC`	Age+LCa(P)+LCa(C) M, `m`, `APaCa`

Model: `ilc: lca.rh( model= )` `Epi: LCa.fit( model= )`

APC-models for several datasets

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019`

APC2

Two APC-models

- APC-models for two sets of rates (men/women, say)

$$\log\big(\lambda_i(a,p)\big) = f_i(a) + g_i(p) + h_i(p-a), \quad i = 1,2$$

- Rate-ratio also an APC-model:

$$\begin{aligned}\log\big(\mathrm{RR}(a,p)\big) &= \log\big(\lambda_1(a,p)\big) - \log\big(\lambda_2(a,p)\big)\\ &= \big(f_1(a) - f_2(a)\big) + \big(g_1(p) - g_2(p)\big)\\ &\quad + \big(h_1(p-a) - h_2(p-a)\big)\\ &= f_{\mathrm{RR}}(a) + g_{\mathrm{RR}}(p) + h_{\mathrm{RR}}(p-a)\end{aligned}$$

- Model the two sets of rates separately and reporte the ratio effects as any other APC-model.
- Note; not all constraints carry over to RR

Two sets of data I

Example: Testis cancer in Denmark, Seminoma and non-Seminoma cases.

```
> th <- read.table( "../data/testis-hist.txt", header=TRUE )
> str( th )

'data.frame': 29160 obs. of  9 variables:
 $ a    : int  0 0 0 0 0 0 1 1 1 1 ...
 $ p    : int  1943 1943 1943 1943 1943 1943 1943 1943 1943 1943 ...
 $ c    : int  1942 1942 1942 1943 1943 1943 1941 1941 1941 1942 ...
 $ y    : num  18853 18853 18853 20796 20796 ...
 $ age  : num  0.667 0.667 0.667 0.333 0.333 ...
 $ diag : num  1943 1943 1943 1944 1944 ...
 $ birth: num  1943 1943 1943 1943 1943 ...
 $ hist : int  1 2 3 1 2 3 1 2 3 1 ...
 $ d    : int  0 1 0 0 0 0 0 0 0 0 ...
```

Two sets of data II

```
> head( th )

  a    p    c        y       age     diag    birth hist d
1 0 1943 1942 18853.0 0.6666667 1943.333 1942.667    1 0
2 0 1943 1942 18853.0 0.6666667 1943.333 1942.667    2 1
3 0 1943 1942 18853.0 0.6666667 1943.333 1942.667    3 0
4 0 1943 1943 20796.5 0.3333333 1943.667 1943.333    1 0
5 0 1943 1943 20796.5 0.3333333 1943.667 1943.333    2 0
6 0 1943 1943 20796.5 0.3333333 1943.667 1943.333    3 0

> th <- transform( th,
+                  hist = factor( hist, labels=c("Sem","nS","Oth") ),
+                     A = age,
+                     P = diag,
+                     D = d,
+                     Y = y/10^4 )[,c("A","P","D","Y","hist")]
> th <- subset( th, A>15 & A<65 & hist!="Oth" )
> th$hist <- factor( th$hist )
```



```
> library( Epi )
> stat.table( list( Histology = hist ),
+             list( D = sum(D),
+                   Y = sum(Y) ),
+             margins = TRUE,
+                data = th )

 ----------------------------
 Histology         D        Y
 ----------------------------
 Sem         4461.00  8435.49
 nS          3494.00  8435.49

 Total       7955.00 16870.99
 ----------------------------
```

First step is separate analyses for each subtype (Sem, nS, resp.)

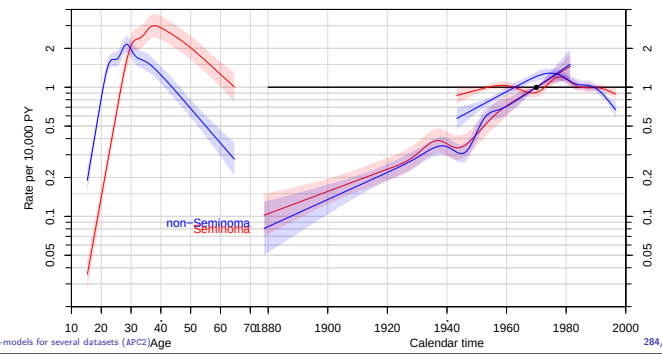

```
> apc.Sem <- apc.fit( subset( th, hist=="Sem" ),
+                     parm = "ACP",
+                     ref.c = 1970,
+                     npar = c(A=8,P=8,C=8) )

[1] "ML of APC-model Poisson with log(Y) offset : ( ACP ):\n"

Analysis of deviance for Age-Period-Cohort model

                   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
Age                     5392     5677.5
Age-drift               5391     5074.1  1   603.33 < 2.2e-16
Age-Cohort              5385     5038.7  6    35.47 3.495e-06
Age-Period-Cohort       5379     5014.7  6    24.01 0.0005201
Age-Period              5385     5061.5 -6   -46.80 2.049e-08
Age-drift               5391     5074.1 -6   -12.68 0.0484745
```

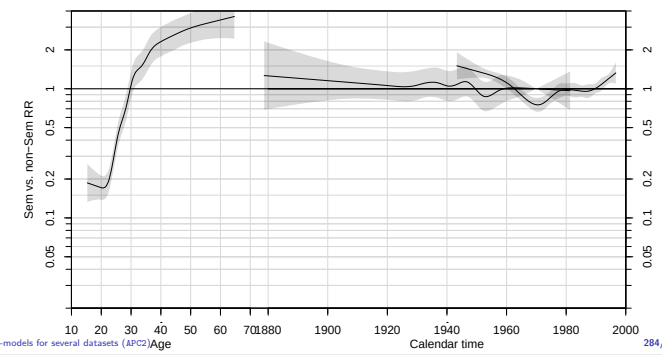


```
> apc.nS  <- apc.fit( subset( th, hist=="nS" ),
+                     parm = "ACP",
+                     ref.c = 1970,
+                     npar = c(A=8,P=8,C=8) )

[1] "ML of APC-model Poisson with log(Y) offset : ( ACP ):\n"

Analysis of deviance for Age-Period-Cohort model

                   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
Age                     5392     5202.5
Age-drift               5391     4501.5  1   701.08 < 2.2e-16
Age-Cohort              5385     4459.7  6    41.77 2.046e-07
Age-Period-Cohort       5379     4375.2  6    84.53 4.133e-16
Age-Period              5385     4427.6 -6   -52.43 1.531e-09
Age-drift               5391     4501.5 -6   -73.87 6.563e-14

> round( (cbind( apc.Sem$Drift,
+                apc.nS$Drift ) -1)*100, 1 )
```



```
                  exp(Est.) 2.5% 97.5% exp(Est.) 2.5% 97.5%
APC (D-weights)         2.5  2.3   2.7       3.1  2.8   3.3
A-d                     2.5  2.3   2.7       3.1  2.8   3.3


> plot( apc.Sem, "Sem vs. non-Sem RR", col="transparent" )

cp.offset    RR.fac
     1804         1

>    matshade( apc.nS$Age[,1], ci.ratio(apc.Sem$Age[,-1],apc.nS$Age[,-1]), col=1 )
> pc.matshade( apc.nS$Per[,1], ci.ratio(apc.Sem$Per[,-1],apc.nS$Per[,-1]), col=1 )
> pc.matshade( apc.nS$Coh[,1], ci.ratio(apc.Sem$Coh[,-1],apc.nS$Coh[,-1]), col=1 )
> abline( h=1 )
```



Analysis of two rates: Formal tests I

Separate models with the **same** parametrization:

```
> ( Akn <- (apc.Sem$Knots$Age+apc.nS$Knots$Age)/2 )

[1] 22.66667 26.50000 29.50000 32.33333 35.16667 38.83333 43.83333 52.66667

> ( Pkn <- (apc.Sem$Knots$Per+apc.nS$Knots$Per)/2 )

[1] 1952.417 1964.000 1972.333 1978.167 1983.000 1987.500 1991.500 1995.000

> ( Ckn <- (apc.Sem$Knots$Coh+apc.nS$Knots$Coh)/2 )

[1] 1913.500 1926.000 1934.833 1942.000 1947.833 1953.333 1958.958 1966.000

> apc.sem <- apc.fit( subset(th,hist=="Sem"), npar=list(A=Akn,P=Pkn,C=Ckn), pr=F )

No reference period given:
 Reference period for age-effects is chosen as
 the median date of birth for persons  with event:  1939.667 .
```

Analysis of two rates: Formal tests II

```
> apc.ns  <- apc.fit( subset(th,hist=="nS" ), npar=list(A=Akn,P=Pkn,C=Ckn), pr=F )

No reference period given:
 Reference period for age-effects is chosen as
 the median date of birth for persons  with event:  1949.667 .
```

Joint model, parametrize interactions separately:

Analysis of two rates: Formal tests III

```
> Ma <- with( th, Ns(   A, knots=Akn, intercept=TRUE ) )
> Mp <- with( th, Ns( P  , knots=Pkn ) )
> Mc <- with( th, Ns( P-A, knots=Ckn ) )
> # extract the linear trend
> Mp <- detrend( Mp, th$P     , weight=th$D )
> Mc <- detrend( Mc, th$P-th$A, weight=th$D )
> m.apc <- glm( D ~ -1 + Ma:hist + Mp:hist + Mc:hist +
+                  P:hist + # note seperate slopes extracted
+                                offset( log(Y)),
+              family=poisson, data=th )
> m.apc$deviance

[1] 9410.446

> # Same as the sum from separate modeles
> apc.ns$Model$deviance + apc.sem$Model$deviance

[1] 9410.446
```


Analysis of two rates: Formal tests IV

Tests for equality of non-linear part of shapes

```
> m.ap  <- update( m.apc, . ~ . - Mc:hist + Mc )
> m.ac  <- update( m.apc, . ~ . - Mp:hist + Mp )
> m.a   <- update( m.ap , . ~ . - Mp:hist + Mp )
> m.d   <- update( m.ap , . ~ . - Mp:hist )
> m.0   <- update( m.ap , . ~ . -  P:hist + P )
> AOV <- anova( m.a, m.ac, m.apc, m.ap, m.a, m.d, m.0, test="Chisq")
> rownames( AOV ) <- c("","cohRR","perRR|coh","cohRR|per","perRR","drift","Smdrift
> AOV
```


Analysis of two rates: Formal tests V

```
Analysis of Deviance Table

Model 1: D ~ Mc + Mp + Ma:hist + hist:P + offset(log(Y)) - 1
Model 2: D ~ Mp + Ma:hist + hist:Mc + hist:P + offset(log(Y)) - 1
Model 3: D ~ -1 + Ma:hist + Mp:hist + Mc:hist + P:hist + offset(log(Y))
Model 4: D ~ Mc + Ma:hist + hist:Mp + hist:P + offset(log(Y)) - 1
Model 5: D ~ Mc + Mp + Ma:hist + hist:P + offset(log(Y)) - 1
Model 6: D ~ Mc + Ma:hist + hist:P + offset(log(Y)) - 1
Model 7: D ~ Mc + P + Ma:hist + hist:Mp + offset(log(Y)) - 1
          Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
              10770     9467.4
cohRR         10764     9447.3  6   20.094  0.002665
perRR|coh     10758     9410.4  6   36.886 1.854e-06
cohRR|per     10764     9421.6 -6  -11.196  0.082496
perRR         10770     9467.4 -6  -45.783 3.270e-08
drift         10776     9538.2 -6  -70.807 2.793e-13
Smdrift       10765     9425.6 11  112.612 < 2.2e-16
```


Several datasets I

- Separate models for each
- Rate-ratios between two sets of fitted rates also follow an APC model
- Constraints does not necessarily carry over to RRs
- Test for equality of effects: non-linear and linear
- Take care not to violate the **principle of marginality**: — do not test linear terms when non-linear terms are in the model.

APC-model: Interactions

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

http://BendixCarstensen/APC/EDSD-2019 APC-int

Analysis of DM-rates: Age×sex interaction I

- 10 centres
- 2 sexes
- Age: 0–15
- Period 1989–1999

- Is the sex-effect the same between all centres?
- How is timetrend by birth cohort?

Analysis of DM-rates: Age×sex interaction II

```
> library( Epi )
> library( splines )
> # load( file="c:/Bendix/Artikler/A_P_C/IDDM/Eurodiab/data/tri.Rdata" )
> load( file = "~/teach/APC/examples/EuroDiab/tri.Rdata" )
> str(dm)

'data.frame': 5940 obs. of  8 variables:
 $ sex: Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ cen: Factor w/ 10 levels "Z2: Czech","A1: Austria",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ per: num  1989 1990 1991 1992 1993 ...
 $ D  : num  1 0 0 0 0 0 0 0 0 1 ...
 $ A  : num  0.333 0.333 0.333 0.333 0.333 ...
 $ P  : num  1990 1991 1992 1993 1994 ...
 $ C  : num  1989 1990 1991 1992 1993 ...
 $ Y  : num  21970 22740 22886 23026 22323 ...
```

Analysis of DM-rates: Age×sex interaction III

```
> dm <- dm[dm$cen=="D1: Denmark",]
> attach( dm )
> # Define knots and points of prediction
> n.A <- 5
> n.C <- 8
> n.P <- 5
> c0  <- 1985
> attach( dm, warn.conflicts=FALSE )
> A.kn <- quantile( rep( A, D ), probs=(1:n.A-0.5)/n.A )
> P.kn <- quantile( rep( P, D ), probs=(1:n.P-0.5)/n.P )
> C.kn <- quantile( rep( C, D ), probs=(1:n.C-0.5)/n.C )
> A.pt <- sort( A[match( unique(A), A )] )
> P.pt <- sort( P[match( unique(P), P )] )
> C.pt <- sort( C[match( unique(C), C )] )
> # Age-cohort model with age-sex interaction
> # The model matrices for the ML fit
> # - note that intercept is in age term, and drift is added to the cohort term:
> Ma <-                    Ns( A, kn=A.kn, intercept=T )
> Mc <- cbind( C-c0, detrend( Ns( C, kn=C.kn ), C, weight=D ) )
```


Analysis of DM-rates: Age×sex interaction IV

```
> Mp <-             detrend( Ns( P, kn=P.kn ), P, weight=D )
> # The prediction matrices - corresponding to ordered unique values of A, P and C
> Pa <- Ma[match(A.pt,A),,drop=F]
> Pp <- Mp[match(P.pt,P),,drop=F]
> Pc <- Mc[match(C.pt,C),,drop=F]
> # Fit the apc model using the cohort major parametrization
> apcs <- glm( D ~ Ma:sex - 1 + Mc + Mp +
+                  offset( log (Y/10^5) ),
+                  family=poisson, epsilon = 1e-10,
+                  data=dm )
> ci.exp( apcs )
```


Analysis of DM-rates: Age×sex interaction V

```
             exp(Est.)       2.5%      97.5%
Mc           1.0053157  0.9719640  1.0398118
Mc1          0.6496197  0.3305926  1.2765132
Mc2          1.2576228  0.6868926  2.3025652
Mc3          0.5366885  0.2787860  1.0331743
Mc4          0.9207689  0.4877809  1.7381069
Mc5          0.6898805  0.3999550  1.1899714
Mc6          1.1005438  0.5817089  2.0821352
Mp1          0.5735223  0.3489977  0.9424928
Mp2          1.0534148  0.6090201  1.8220792
Mp3          0.9412582  0.4032633  2.1969937
Ma1:sexF    11.9104421  6.7605869 20.9831831
Ma2:sexF    22.0985163 11.9531639 40.8548253
Ma3:sexF    16.5201055  9.6623215 28.2451673
Ma4:sexF   360.8119685 225.4568974 577.4286708
Ma5:sexF     2.5694234  1.5219041  4.3379452
Ma1:sexM    17.0238730  9.9414867 29.1518021
Ma2:sexM    13.4664178  7.0861312 25.5914549
Ma3:sexM    14.4664367  8.6164003 24.2883087
```


Analysis of DM-rates: Age×sex interaction VI

```
Ma4:sexM 531.9214375 343.2221445 824.3652694
Ma5:sexM   3.1485499   1.9406858   5.1081770

> # Average trend (D-projection)
> round( ( ci.exp( apcs, subset=1 ) - 1 ) *100, 1 )

   exp(Est.) 2.5% 97.5%
Mc       0.5 -2.8     4

> ci.exp( apcs, subset="sexF" )

           exp(Est.)       2.5%      97.5%
Ma1:sexF   11.910442   6.760587  20.983183
Ma2:sexF   22.098516  11.953164  40.854825
Ma3:sexF   16.520106   9.662321  28.245167
Ma4:sexF  360.811968 225.456897 577.428671
Ma5:sexF    2.569423   1.521904   4.337945

> cbind( A.pt, ci.exp( apcs, subset="sexF", ctr.mat=Pa ) )
```


Analysis of DM-rates: Age×sex interaction VII

```
          A.pt exp(Est.)      2.5%    97.5%
 [1,] 0.3333333  4.943285  2.363023 10.34102
 [2,] 0.6666667  5.309563  2.676029 10.53481
 [3,] 1.3333333  6.125551  3.416160 10.98379
 [4,] 1.6666667  6.579431  3.847562 11.25100
 [5,] 2.3333333  7.590575  4.833655 11.91993
 [6,] 2.6666667  8.153008  5.380890 12.35326
 [7,] 3.3333333  9.401089  6.531373 13.53168
 [8,] 3.6666667 10.085197  7.103019 14.31943
 [9,] 4.3333333 11.561158  8.190634 16.31869
[10,] 4.6666667 12.344483  8.712446 17.49064
[11,] 5.3333333 13.969938  9.777715 19.95959
[12,] 5.6666667 14.794673 10.355375 21.13708
[13,] 6.3333333 16.412682 11.674678 23.07354
[14,] 6.6666667 17.179232 12.425801 23.75107
[15,] 7.3333333 18.578132 13.958075 24.72741
[16,] 7.6666667 19.228123 14.579373 25.35917
[17,] 8.3333333 20.513353 15.309646 27.48579
[18,] 8.6666667 21.190703 15.538953 28.89808
```


Analysis of DM-rates: Age×sex interaction VIII

```
[19,]  9.3333333 22.742587 16.317839 31.69692
[20,]  9.6666667 23.679333 17.100960 32.78827
[21,] 10.3333333 25.893547 19.499950 34.38346
[22,] 10.6666667 26.999519 20.607727 35.37382
[23,] 11.3333333 28.605296 21.348779 38.32832
[24,] 11.6666667 28.831963 21.013988 39.55851
[25,] 12.3333333 27.526786 19.701501 38.46022
[26,] 12.6666667 25.941507 18.827598 35.74337
[27,] 13.3333333 21.900696 16.035816 29.91058
[28,] 13.6666667 19.869417 14.038380 28.12246
[29,] 14.3333333 16.320075 10.026866 26.56312
[30,] 14.6666667 14.790766  8.323640 26.28258
```

```
> # Extract the effects
> F.inc <- ci.exp( apcs, subset="sexF", ctr.mat=Pa)
> M.inc <- ci.exp( apcs, subset="sexM", ctr.mat=Pa)
> MF.RR <- ci.exp( apcs, subset=c("sexM","sexF"), ctr.mat=cbind(Pa,-Pa))
>  c.RR <- ci.exp( apcs, subset="Mc", ctr.mat=Pc)
>  p.RR <- ci.exp( apcs, subset="Mp", ctr.mat=Pp)
```

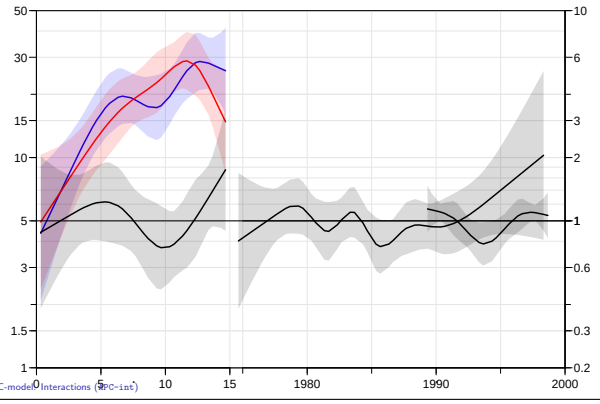

Analysis of DM-rates: Age×sex interaction IX

The the frame for the effects

```
> par( mar=c(4,4,1,4), mgp=c(3,1,0)/1.6, las=1 )
> apc.frame( a.lab=c(0,5,10,15),
+            a.tic=c(0,5,10,15),
+            r.lab=c(c(1,1.5,3,5),c(1,1.5,3,5)*10),
+            r.tic=c(c(1,1.5,2,5),c(1,1.5,2,5)*10),
+           cp.lab=seq(1980,2000,10),
+           cp.tic=seq(1975,2000,5),
+           rr.ref=5,
+              gap=1,
+         col.grid=gray(0.9),
+            a.txt="",
+           cp.txt="",
+            r.txt="",
+           rr.txt="" )
> ###
> ### Draw the estimates
> ###
```


Analysis of DM-rates: Age×sex interaction X

```
>    matshade( A.pt, M.inc, lwd=2, col="blue" )
>    matshade( A.pt, F.inc, lwd=2, col="red" )
>    matshade( A.pt, MF.RR*5, lwd=2 ) ; abline( h=5 )
> pc.matshade( C.pt, c.RR, lwd=2 )
> pc.matshade( P.pt, p.RR, lwd=2 )
```



Analysis of DM-rates: Age×sex interaction I

A bit more intuitive, independent of parametrization:

```
> apcS <- glm( D ~ Ns(A,knots=A.kn,intercept=TRUE):sex +
+                  Ns(P,knots=P.kn) + Ns(C,knots=C.kn) +
+                  offset( log (Y/10^5) ),
+                  family=poisson, epsilon = 1e-10,
+                  data=dm )
> apcS$deviance

[1] 633.5838

> apcs$deviance

[1] 633.5838
```

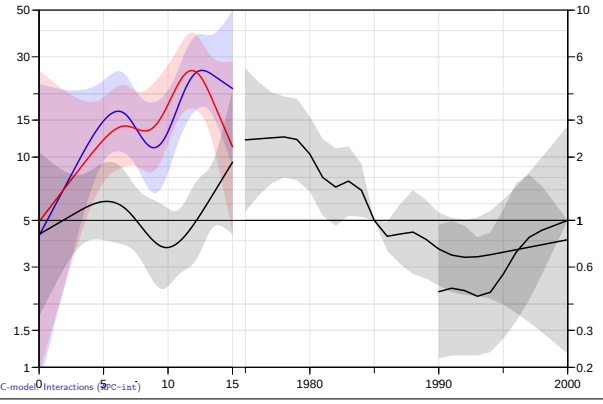

Analysis of DM-rates: Age×sex interaction II

```
> # rates for the 1985 birth cohort and the RR
> a.pt <- seq(0,15,0.1)
> ndaM <- data.frame( A=a.pt, P=1985+a.pt, C=1985, Y=10^5, sex="M" )
> ndaF <- data.frame( A=a.pt, P=1985+a.pt, C=1985, Y=10^5, sex="F" )
> a.pM <- ci.pred( apcS, ndaM )
> a.pF <- ci.pred( apcS, ndaF )
> a.RR <- ci.exp ( apcS, list(ndaM,ndaF) )
> # Cohort RRs relative to C=1985
> ndc <- data.frame( A=10, P=2000, C=1975:2000, Y=10^5 )
> ndr <- data.frame( A=10, P=2000, C=1985     , Y=10^5 )
> c.RR <- ci.exp( apcS, list(ndc,ndr) )
> # Period RRs relative to P=2000
> ndp <- data.frame( A=10, P=1990:2000, C=1985, Y=10^5 )
> ndr <- data.frame( A=10, P=2000,      C=1985, Y=10^5 )
> p.RR <- ci.exp( apcS, list(ndp,ndr) )
> # plt( paste( "DM-DK" ), width=11 )
> par( mar=c(4,4,1,4), mgp=c(3,1,0)/1.6, las=1 )
> #
> # The the frame for the effects
```

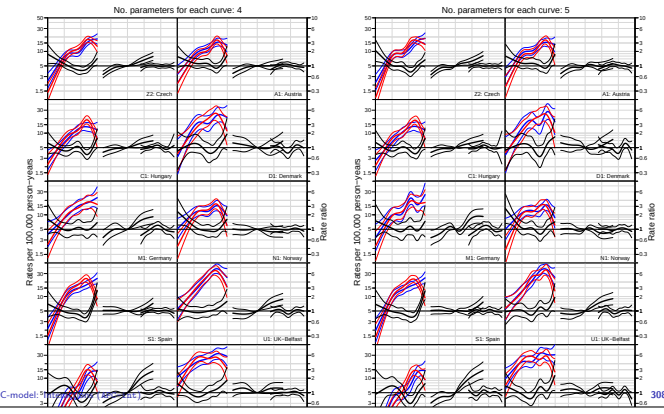

Analysis of DM-rates: Age×sex interaction III

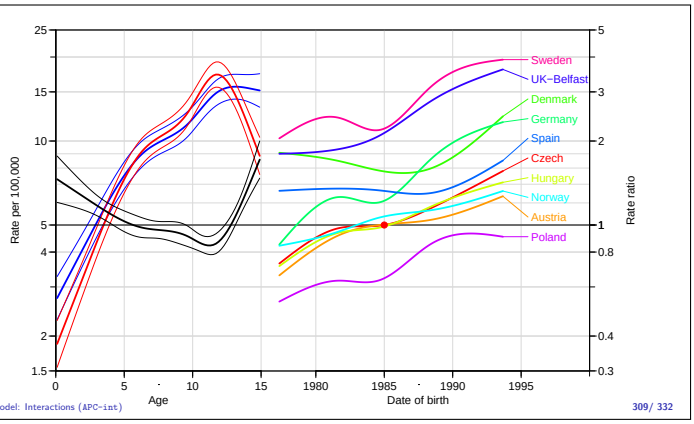
```
> apc.frame( a.lab=c(0,5,10,15),
+            a.tic=c(0,5,10,15),
+            r.lab=c(c(1,1.5,3,5),c(1,1.5,3,5)*10),
+            r.tic=c(c(1,1.5,2,5),c(1,1.5,2,5)*10),
+           cp.lab=seq(1980,2000,10),
+           cp.tic=seq(1975,2000,5),
+           rr.ref=5,
+              gap=1,
+         col.grid=gray(0.9),
+            a.txt="",
+           cp.txt="",
+            r.txt="",
+           rr.txt="" )
> # Draw the estimates
>    matshade( a.pt, a.pM, lwd=2, col="blue" )
>    matshade( a.pt, a.pF, lwd=2, col="red" )
>    matshade( a.pt, a.RR*5, lwd=2 ) ; abline( h=5 )
> pc.matshade( 1975:2000, c.RR, lwd=2 )
> pc.matshade( 1990:2000, p.RR, lwd=2 )
```

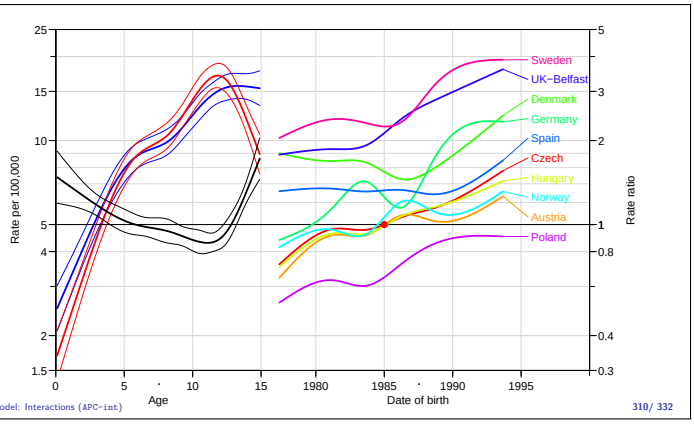

Analysis of DM-rates: Age×sex interaction IV



- . . . but these are not the estimates we really want as before.
- The detrended estimates are not avilable from the fitted values, because the parametrization they rely on is a function of **data**.
- Of course the parameters can be extracted but it requires a construction of the model matrices as we did first
- How is shown in the section "Reparametrizations" in the notes on "Introductory linear algebra with R".







Predicting future rates

Bendix Carstensen

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019` `predict`

Prediction of future rates

Model:

$$\log\big(\lambda(a,p)\big) = f(a) + g(p) + h(c)$$

- Why not just extend the estimated functions into the future?
- Natural splines lend themselves easily to this [**?**]
- The parametrization curse — the model as stated is not uniquely parametrized.
- Predictions from the model must be invariant under reparametrization.

Identifiability

Predictions based in the three functions $(f(a)$, $g(p)$ and $h(c)$ must give the same prediction also for the reparametrized version:

$$\begin{aligned}\log(\lambda(a,p)) &= \tilde{f}(a) + \tilde{g}(p) + \tilde{h}(c)\\ &= \big(f(a) - \gamma a\big) + \\ &\quad \big(g(p) + \gamma p\big) + \\ &\quad \big(h(c) - \gamma c\big)\end{aligned}$$

A prediction based on the parametrization $\big(f(a), g(p), h(c)\big)$ must give the same predictions as one based on $\big(\tilde{f}(a), \tilde{g}(p), \tilde{h}(c)\big)$

Parametrization invariance

- Prediction of the future course of g and h must preserve addition of a linear term in the argument:

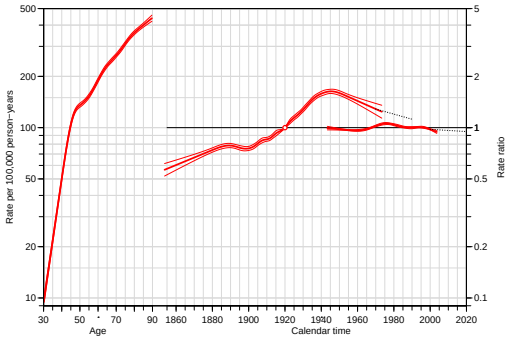
$$\begin{aligned}\text{pred}\big(g(p) + \gamma p\big) &= \text{pred}\big(g(p)\big) + \gamma p\\ \text{pred}\big(h(c) - \gamma c\big) &= \text{pred}\big(h(c)\big) - \gamma c\end{aligned}$$

- If this is met, the predictions made will not depend on the parametrizatioin chosen.
- If one of the conditions does not hold, the prediction wil depend on the parametrization chosen.
- Any linear combination of (known) function values of $g(p)$ and $h(c)$ will work.

Identifiability

- Any linear combination of function values of $g(p)$ and $h(c)$ will work.
- Coefficients in the linear combinations used for g and h must be the same; otherwise the prediction will depend on the specific parametrization.
- What works best in reality is difficult to say:
  depends on the subject matter.

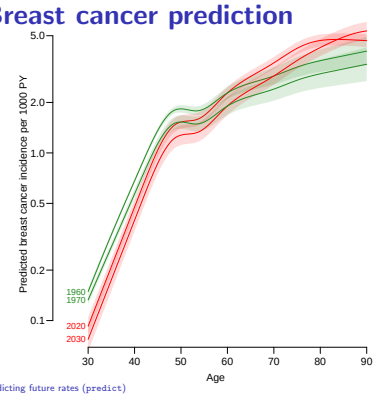
Example: Breast cancer in Denmark



Practicalities

- Long term predictions notoriously unstable.
- Decreasing slopes are possible, the requirement is that at any future point changes in the parametrization should cancel out in the predictions.

Breast cancer prediction



Predicted age-specific breast cancer rates at 2020 & 2030,

in the 1960 and 1970 cohorts.

APC-model for quantitative outcomes

- Model:

$$\text{BMI}_i = f\big(a(i)\big)+g\big(p(i)\big)+h\big(p(i)-a(i)\big)+e_i, \quad e_i \sim \mathcal{N}(o,\sigma^2)$$

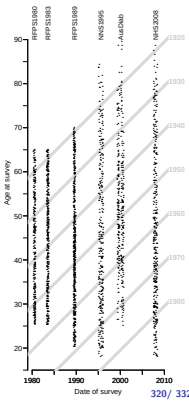
- But the identification problem is still the same:

$$c(i) = p(i) - a(i), \quad \forall i$$

- But the same machinery applies with extraction of the effects
- — and plotting of predictions of
  - E(BMI)
  - quantiles of BMI

APC-model for quantitative outcomes

- Australian surveys
- 40,000+ person suveyed at different times
- Date of birth, data of survey, sex and BMI known.
- How does BMI evolve **in the population**?
- Linear model (E(BMI))
- Quantile regression (median, quantile)
- — the latter is not a model



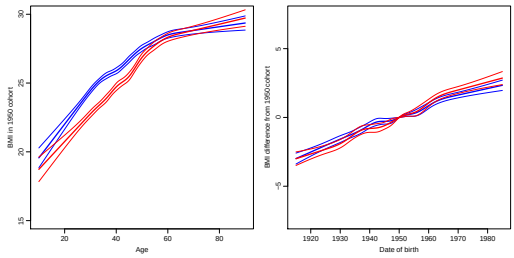
Continuous outcomes

Bendix Carstensen

Statistical Analysis in the Lexis Diagram:

Age-Period-Cohort models
— and some cousins
European Doctoral School of Demography, Odense,April 2019

`http://BendixCarstensen/APC/EDSD-2019` cont



APC-model for quantitative outcomes

- The classical model is:

$$\log(\lambda(a,p)) = f(a) + g(p) + h(p-a)$$

- In principle it would be possible to use an identity-link model:

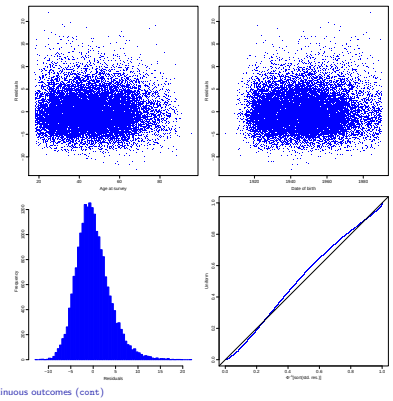
$$\lambda(a,p) = f(a) + g(p) + h(p-a)$$

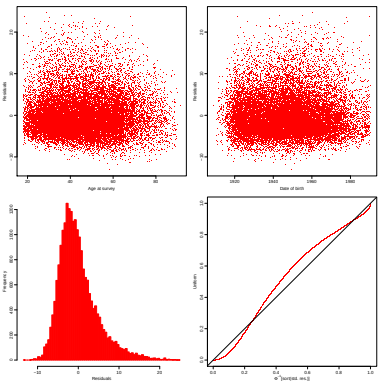
- . . . or use APC-modelling for **measurement** data such as BMI, measured at different times and ages:

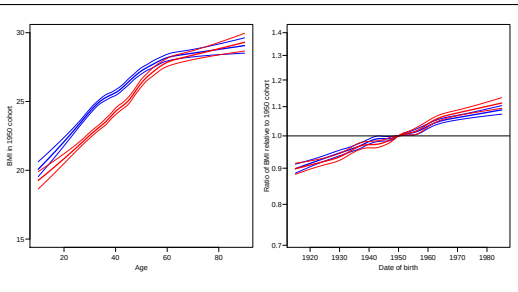
$$\text{BMI}_{ap} = f(a) + g(p) + h(p-a) + e_{ap}, \quad e_i \sim \mathcal{N}(o,\sigma^2)$$

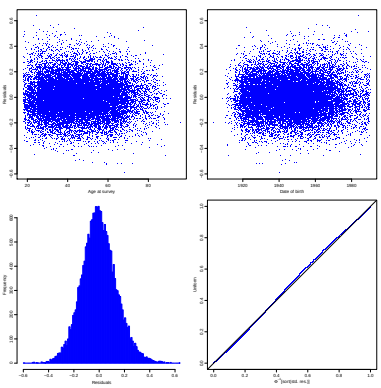
- . . . or more precisely:

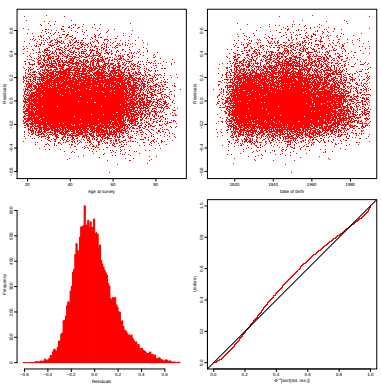
$$\text{BMI}_i = f\big(a(i)\big)+g\big(p(i)\big)+h\big(p(i)-a(i)\big)+e_i, \quad e_i \sim \mathcal{N}(o,\sigma^2)$$

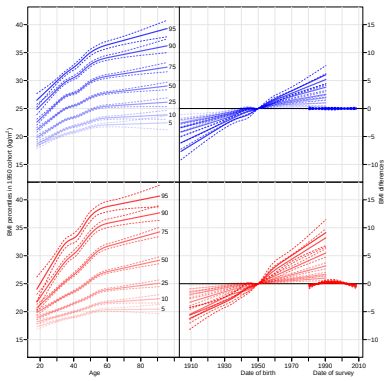


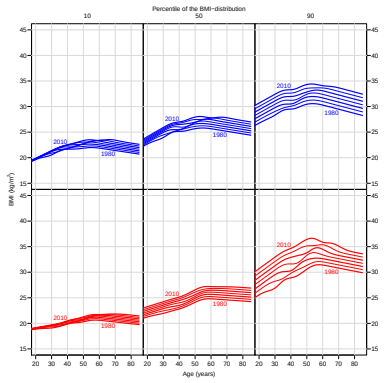


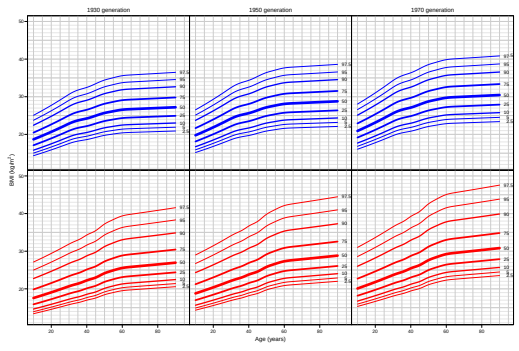

BMI in 1950 cohort
Age
Ratio of BMI relative to 1950 cohort
Date of birth

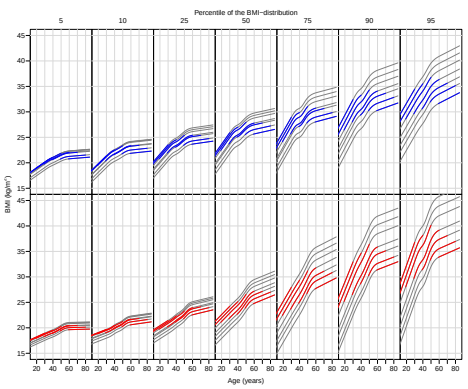


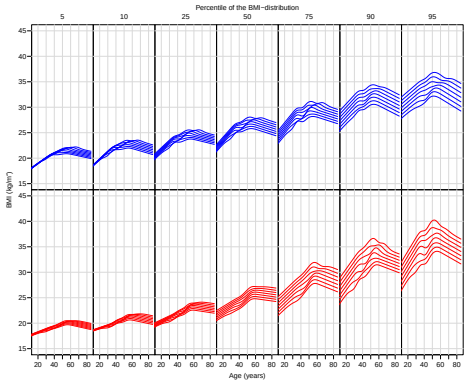



BMI percentiles in 1950 cohort (kg/m²)
Age
Date of birth
Date of survey
BMI differences


Percentile of the BMI−distribution
10
50
90
BMI (kg/m²)
Age (years)


1930 generation
1950 generation
1970 generation
BMI (kg/m²)
Age (years)


Percentile of the BMI−distribution
5
10
25
50
75
90
95
BMI (kg/m²)
Age (years)


Percentile of the BMI−distribution
5
10
25
50
75
90
95
BMI (kg/m²)
Age (years)

References