

Statistical Analysis in the Lexis Diagram: Age-Period-Cohort models

Bendix Carstensen Steno Diabetes Center, Gentofte, Denmark
<http://staff.pubhealth.ku.dk/~bxc/>

Max Planck Institut for Demographic Research, Rostock
March 2009

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

Introduction

Monday 23rd, morning

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

Welcome

- ▶ Purpose of the course:
 - ▶ Knowledge about APC-models
 - ▶ Technical knowledge of handling them
 - ▶ insight in the basic survival concepts
- ▶ Remedies of the course:
 - ▶ Lectures with handouts (BxC)
 - ▶ Practicals with suggested solutions (BxC + EG)

Scope of the course

- ▶ Rates as observed in populations — disease registers for example.
- ▶ Understanding of survival analysis (statistical analysis of rates) is needed — that is the content of much of the first day.
- ▶ Besides the concepts the practical understanding of the actual computations (in R) are emphasized.
- ▶ There is a section in the practicals:
“Probability concepts in follow-up studies”

About the lectures

- ▶ Please interrupt:
Most likely I did a mistake or left out a crucial argument.
- ▶ The handpout are not perfect — please comment on them, prospective students would benefit from it.
- ▶ There is a time-schedule in the practicals.
I might need revision as we go.

About the practicals

- ▶ You should use you preferred **R**-enviroment.
- ▶ Epi-package for **R** is needed.
- ▶ Data are all on the net; but we also have them on USB-sticks.
- ▶ 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.
- ▶ (An opportunity to learn R-weave?)
- ▶ A minor bug in `apc.fit`, `apc.plot` and `apc.frame` is fixed in Epi 1.0.11

Rates and Survival

Monday 23rd, morning

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

Survival data

Persons enter the study at some date.

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

Observation:

Actual time span to death (“event”)

or

Some time alive (“at least this long”)

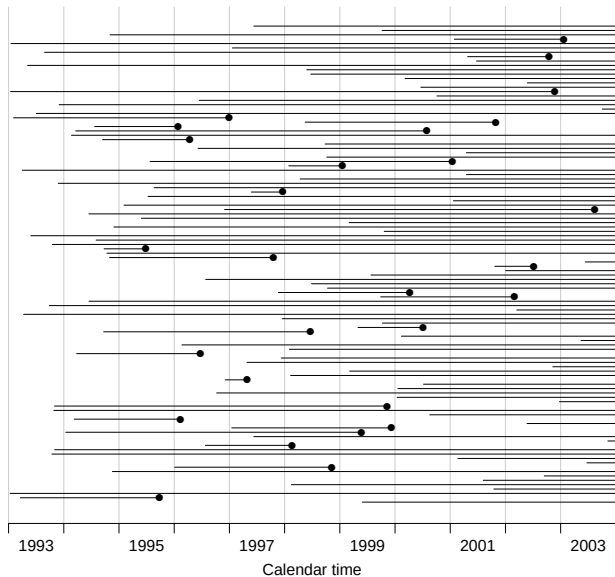
Examples of time-to-event measurements

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

Each line a
person

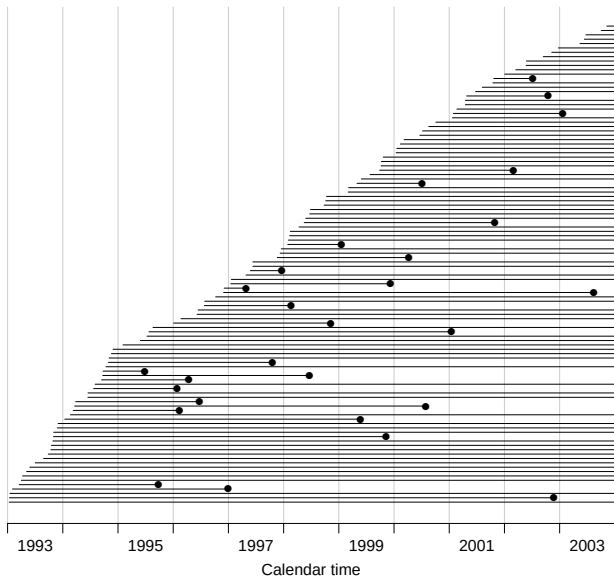
Each blob a
death

Study ended
at 31 Dec.
2003

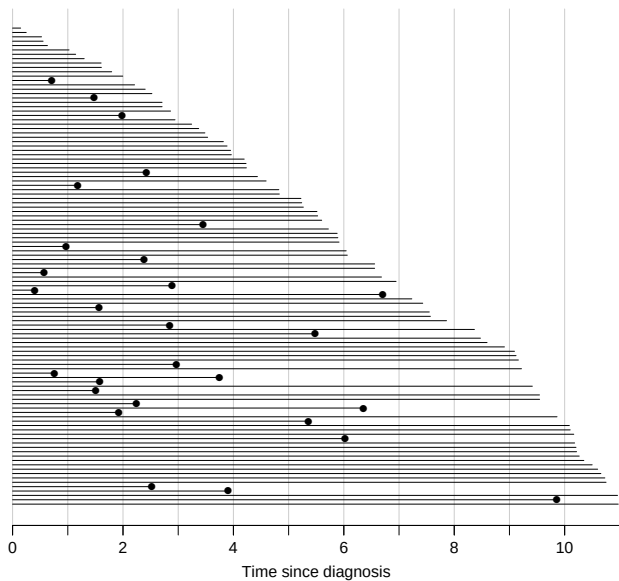


Ordered by
date of entry

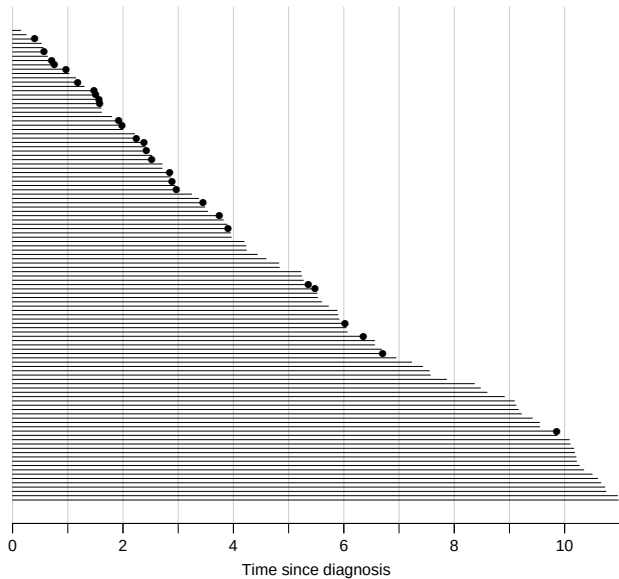
Most likely
the order in
your
database.



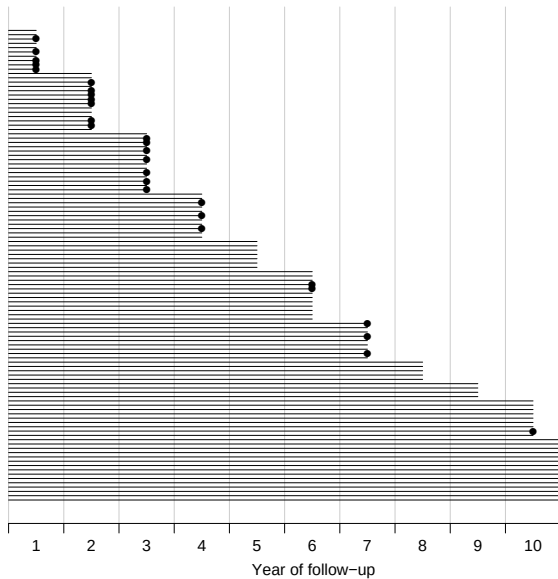
Timescale
changed to
“Time since
diagnosis”.



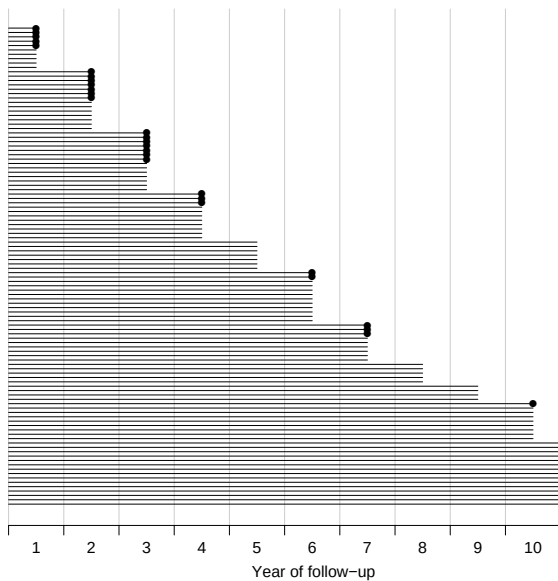
Patients
ordered by
survival
time.



Survival
times
grouped into
bands of
survival.



Patients
ordered by
survival
status within
each band.



Survival after Cervix cancer

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

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

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

Life-table estimator.

Survival function

Persons enter at time 0:

- Date of birth

- Date of randomization

- Date of diagnosis.

How long do they survive, survival time T

— a stochastic variable.

Distribution is characterized by the survival function:

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

Intensity or rate

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

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

Theoretical counterpart of a **rate**.

Relationships

$$-\frac{d\log S(t)}{dt} = \lambda(t)$$

$$\Updownarrow$$

$$S(t) = \exp\left(-\int_0^t \lambda(u) du\right) = \exp(-\Lambda(t))$$

$\Lambda(t) = \int_0^t \lambda(s) ds$ is called the *integrated intensity*.

Not an intensity — it is dimensionless.

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

Rate and survival

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

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

Note: A cumulative measure requires an origin!

Observed survival and rate

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

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

— sometimes conditional on $T > t_0$
(left truncated).

- Epidemiological studies:
Observation of (components of) a rate:

$$D/Y$$

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

Empirical rates for individuals

At the *individual* level we introduce the

empirical rate: (d, y) ,

— no. of events ($d \in \{0, 1\}$) during y risk time.

Each person contributes several obs. of (d, y) .

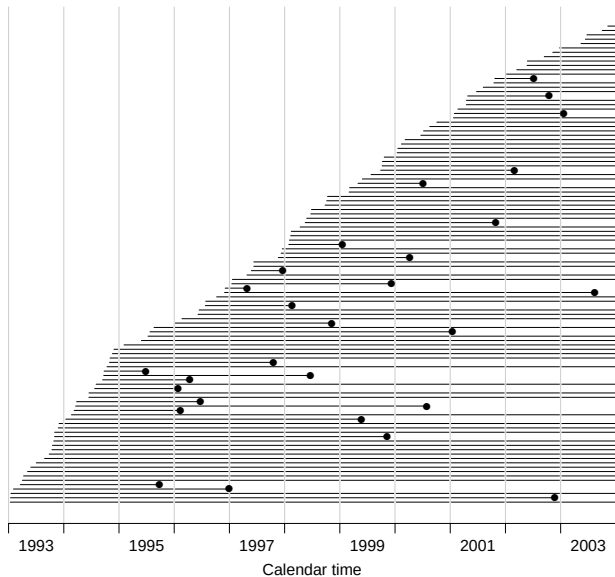
Empirical rates are **responses** in survival analysis.

The timescale is a **covariate** — varies across empirical rates from one individual:

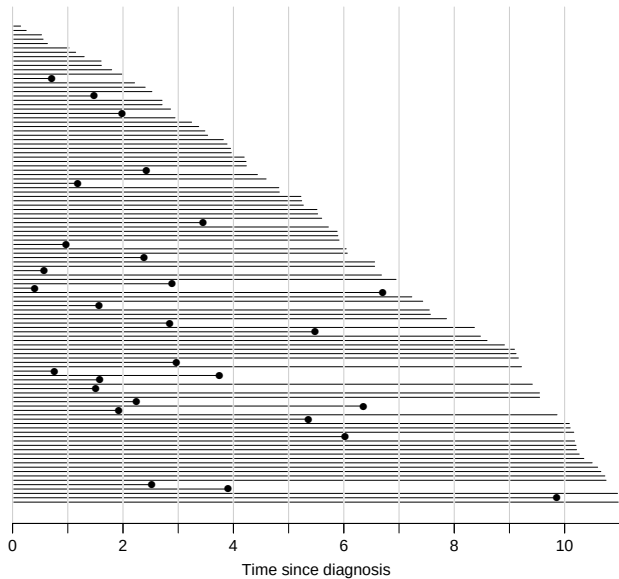
Age, calendar time, time since diagnosis.

Don't confuse timescale with y — difference between two points on **any** timescale we may choose.

Empirical
rates by
calendar
time.



Empirical
rates by
time since
diagnosis.



Two timescales

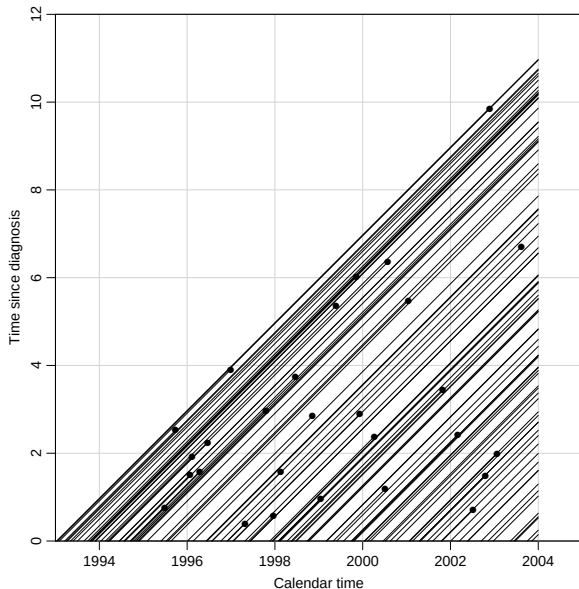
Note that we actually have two timescales:

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

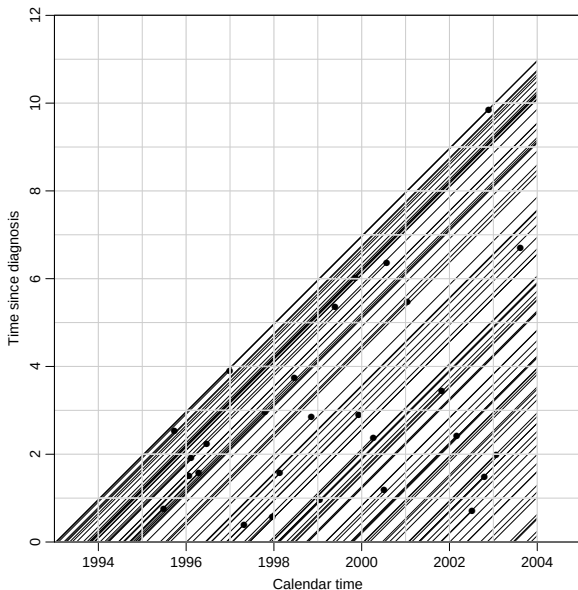
These can be shown simultaneously in a Lexis diagram.

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

A Lexis
diagram!



Empirical
rates by
calendar
time *and*
time since
diagnosis



Likelihood for rates

Monday 23rd, morning

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

Likelihood from one person

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

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

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

Each term refers to one empirical rate (d, y)

— $y = t_i - t_{i-1}$ and mostly $d = 0$.

Likelihood for an empirical rate

Model: the rate is constant in the interval we are looking at. The interval should sufficiently small for this assumption to be reasonable.

If $\pi = 1 - e^{-\lambda y}$ is the death probability:

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

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

Log-likelihood:

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

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

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

Likelihood

Probability of the data and the parameter:

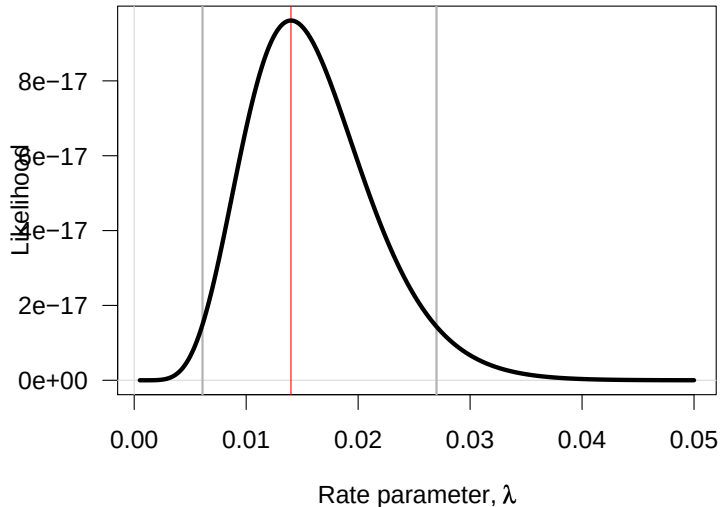
Assuming the rate (intensity) is constant, λ , the probability of observing 7 deaths in the course of 500 person-years:

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

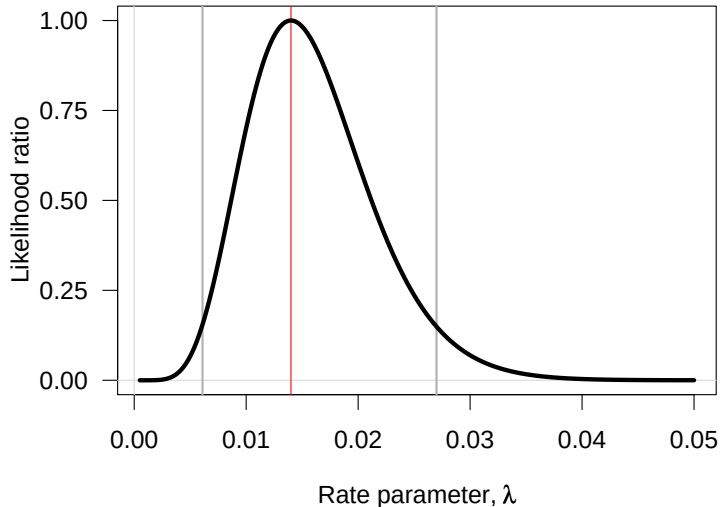
Best guess of λ is where this function is as large as possible.

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

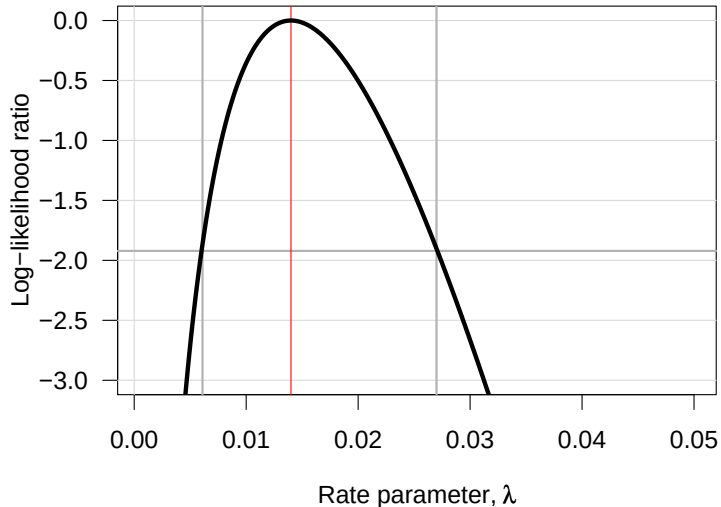
Likelihood function



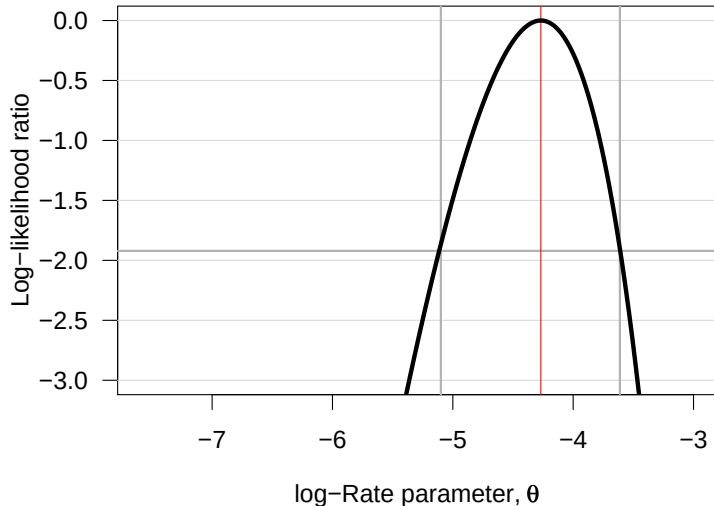
Likelihood-ratio function



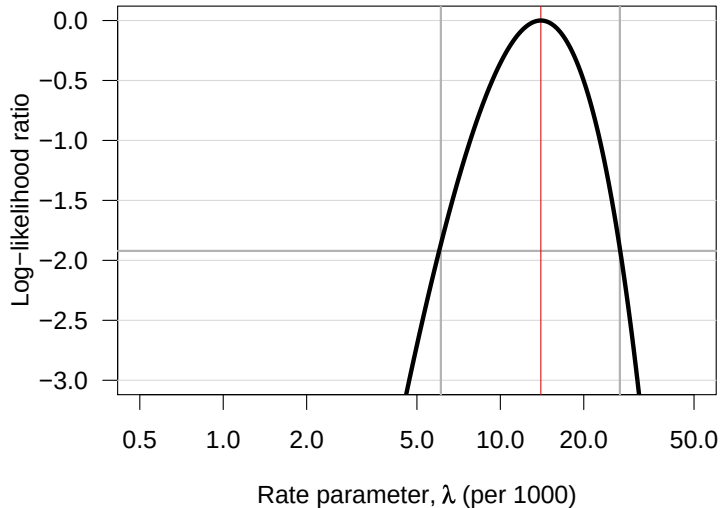
Log-likelihood ratio



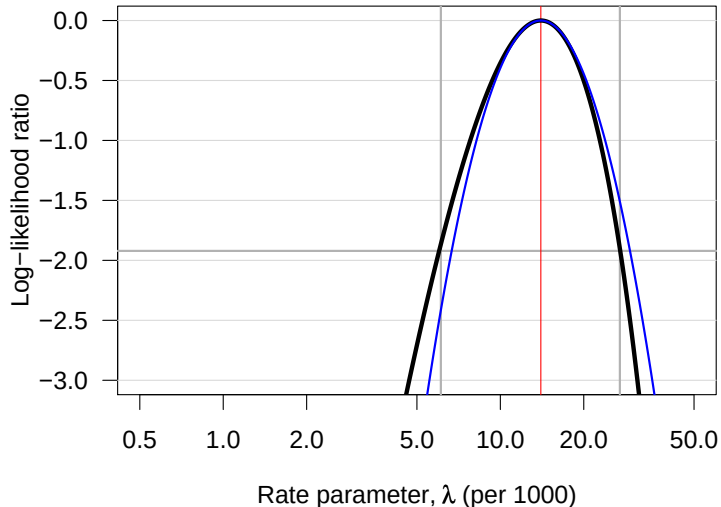
Log-likelihood ratio, $\theta = \log(\lambda)$



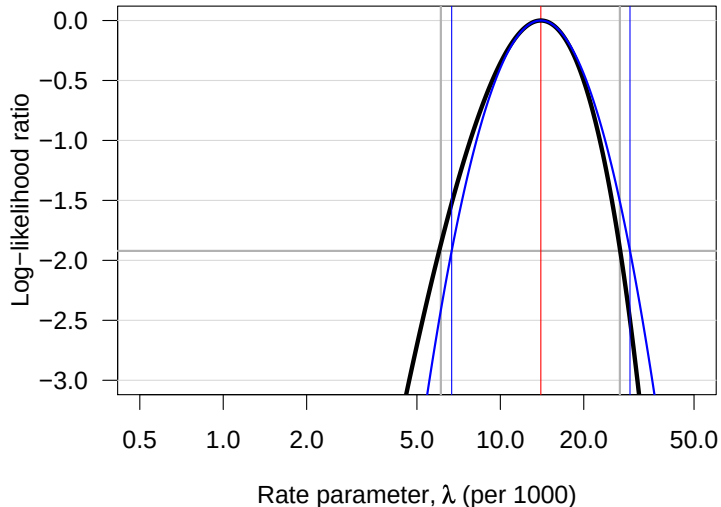
Log-likelihood ratio



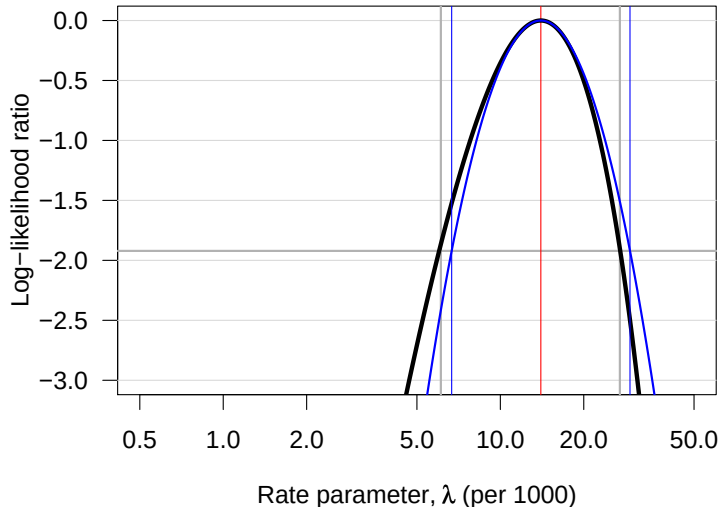
Log-likelihood ratio



Log-likelihood ratio



Log-likelihood ratio



$$\hat{\lambda} = 7/500 = 14 \quad \hat{\lambda} \div \exp(1.96/\sqrt{7}) = (6.7, 29.4)$$

Poisson likelihood

The contributions from **one** individual:

$$d_t \log(\lambda(t)) - \lambda(t) y_t, \quad t = 1, \dots, T$$

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

log-mean $\log(\lambda(t)) + \log(y_t)$

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

- ▶ d is the response variable.
- ▶ $\log(y)$ is the offset variable.

Likelihood for follow-up of many subjects

Adding empirical rates over the follow-up of persons:

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

- ▶ Persons are assumed independent
- ▶ Contribution from the same person are *conditionally* independent, hence give separate contributions to the log-likelihood.

The log-likelihood is maximal for:

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

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

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

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

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

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

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:

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

Exercise

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

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

Exercise – solution

The rate is computed as:

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

The confidence interval is computed as:

$$\hat{\lambda} \div_{\text{erf}} = 20.1 \div_{\text{exp}}(1.96/\sqrt{D}) = (12.5, 32.4)$$

per 1000 person-years.

Ratio of two rates

If we have observations two rates λ_1 and λ_0 , based on (D_1, Y_1) and (D_0, Y_0) the variance of the difference of the ratio of the rates, RR, is:

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

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

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

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.

Exercise – solution

The rate-ratio is computed as:

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

The 95% confidence interval is computed as:

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

Lifetables

Monday 23rd, morning

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

The life table method

The simplest analysis is by the “life-table method”:

interval	alive	dead	cens.	
i	n_i	d_i	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	$8/(59 - 1/2) = 0.137$

$$p_i = \text{P}\{\text{death in interval } i\} = 1 - d_i/(n_i - l_i/2)$$

$$S(t) = (1 - p_1) \times \cdots \times (1 - p_t)$$

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 $(n_i - d_i/2 - l_i/2) \times \ell_i$:

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

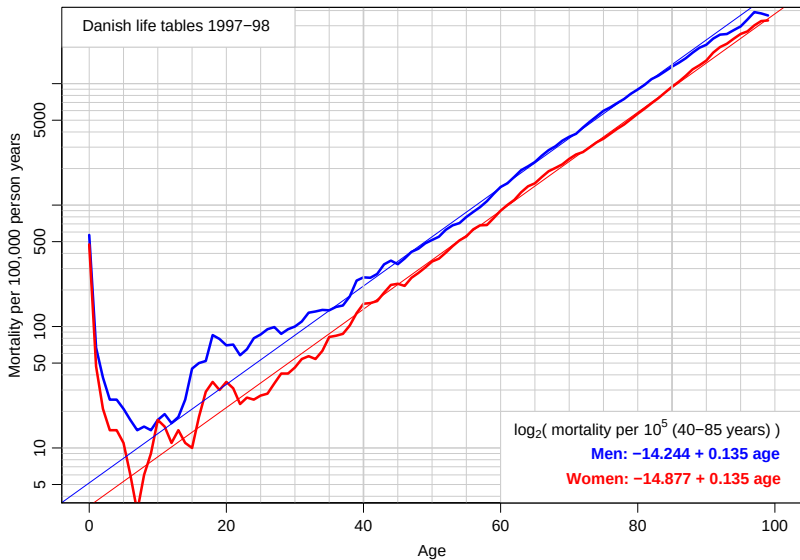
and hence the death probability:

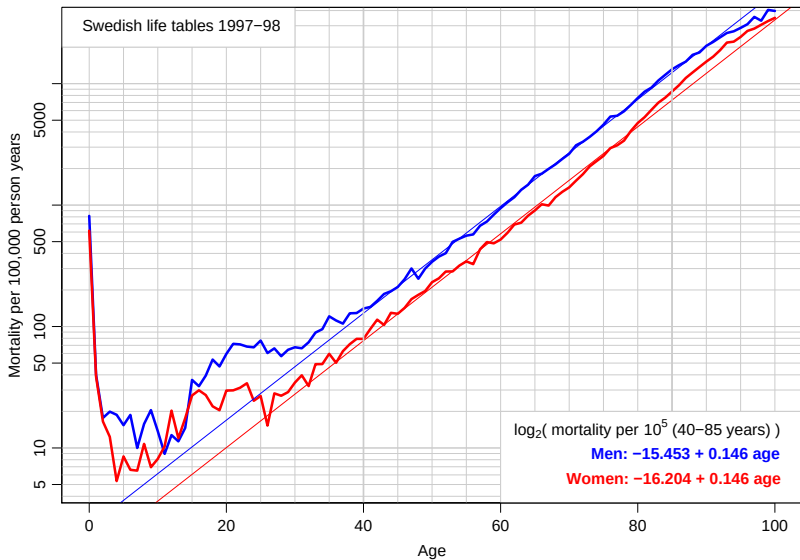
$$p_i = 1 - \exp(-\lambda_i \ell_i) = 1 - \exp\left(-\frac{d_i}{(n_i - d_i/2 - l_i/2)}\right)$$

The modified life-table estimator.

Population life table, DK 1997–98

a	Men			Women		
	$S(a)$	$\lambda(a)$	$E[\ell_{\text{res}}(a)]$	$S(a)$	$\lambda(a)$	$E[\ell_{\text{res}}(a)]$
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

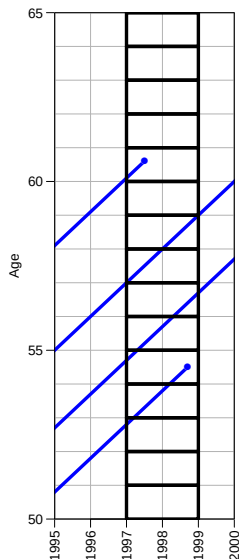
Based on the previous slides answer the following for both Danish and Swedish lifetables:

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

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

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:

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

— 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:

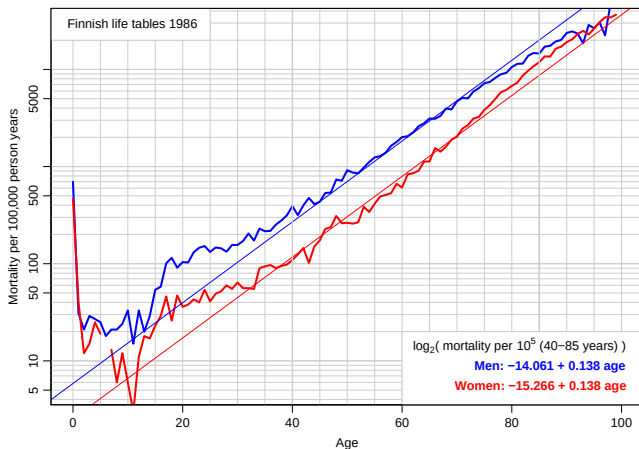
D: Deaths (events).

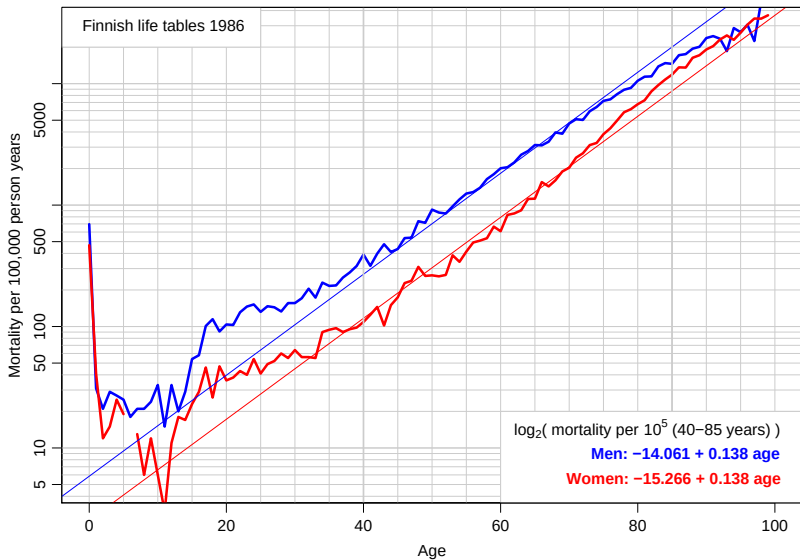
Y: Person-years (risk time).

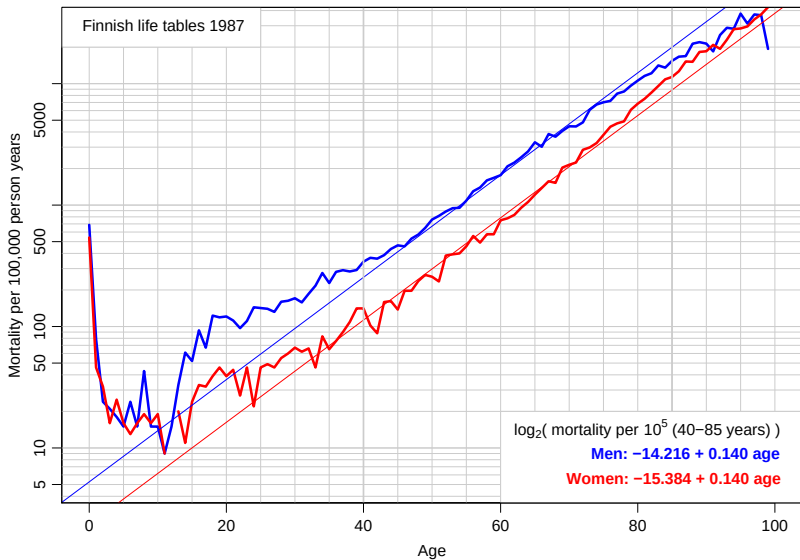
The classical lifetable analysis compiles these for prespecified intervals of age, and computes age-specific mortality **rates**.

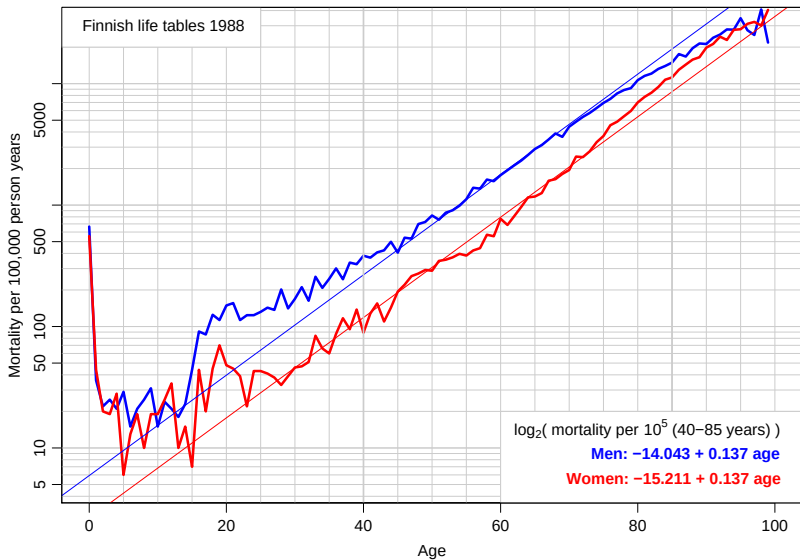
Data are collected cross-sectionally, but interpreted longitudinally.

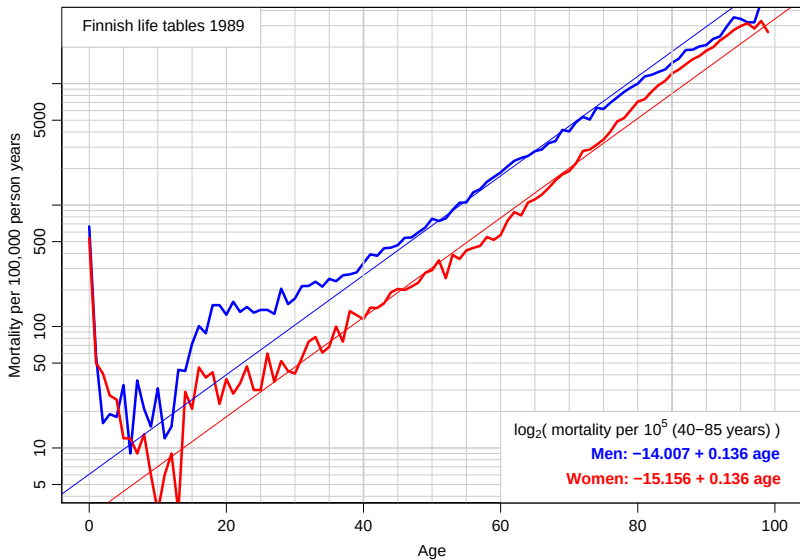
Rates vary over time:

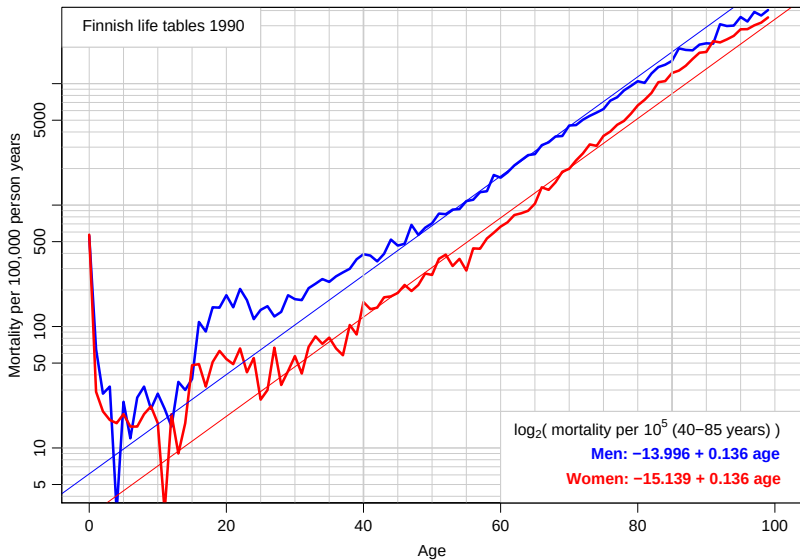


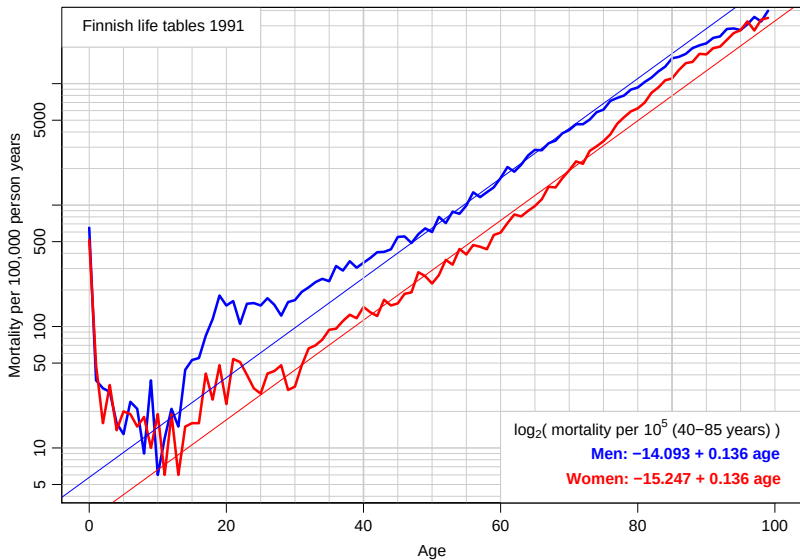


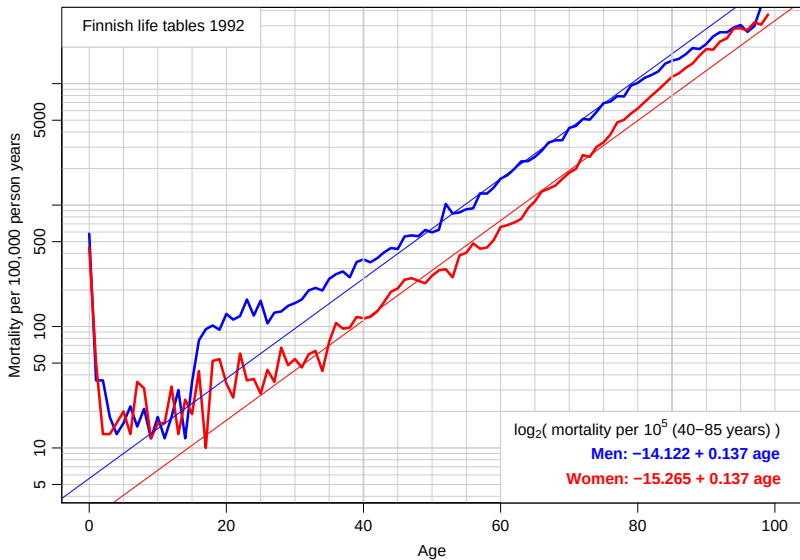


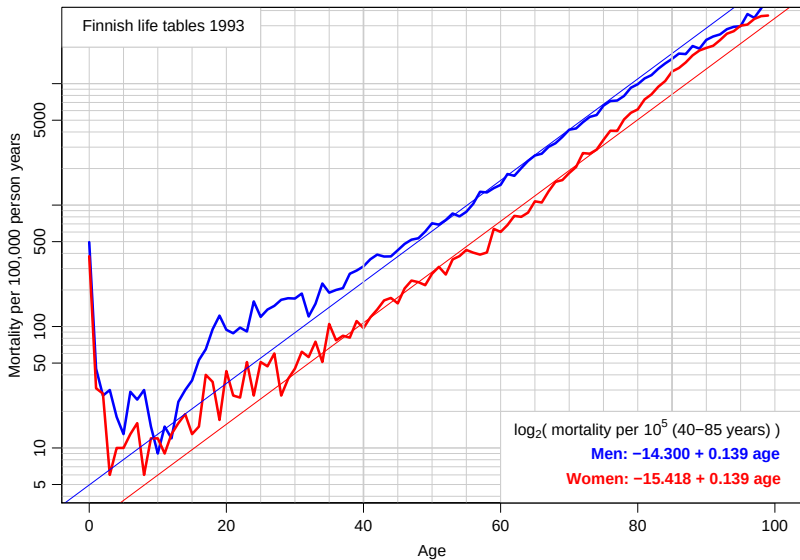


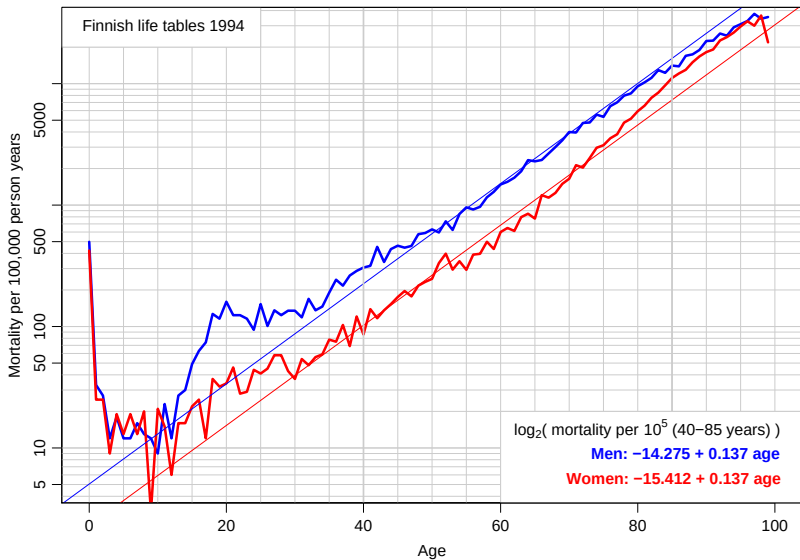


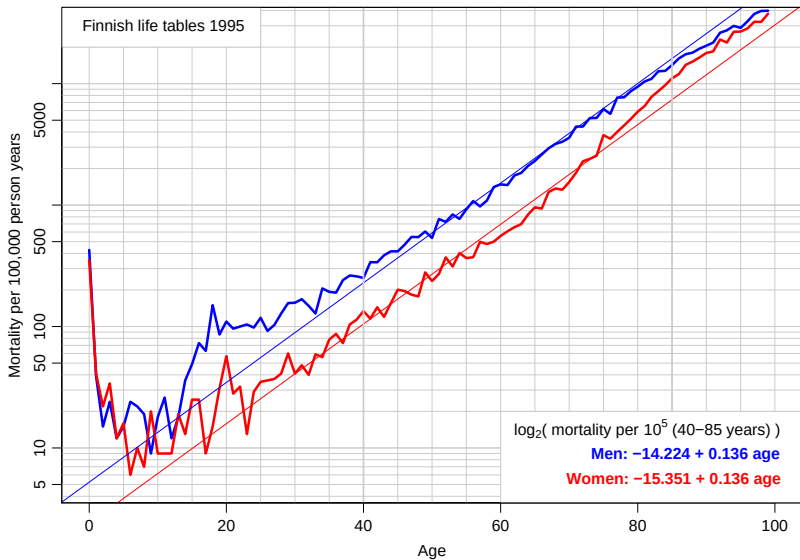


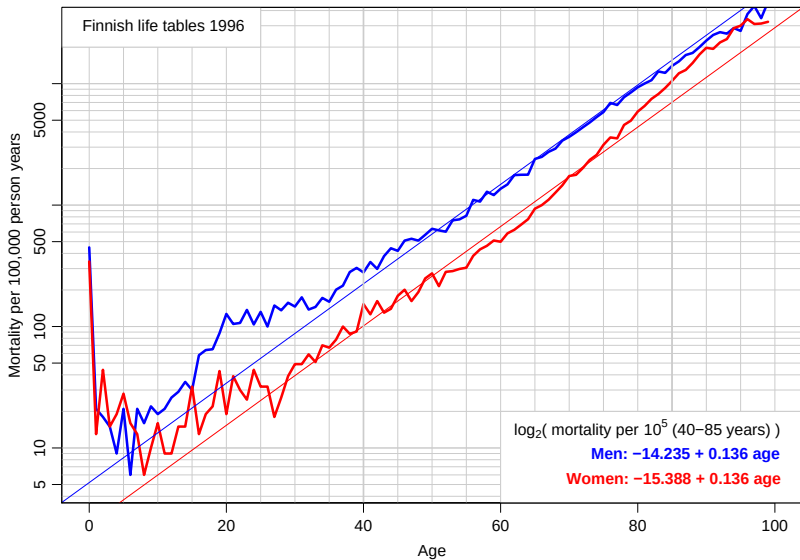


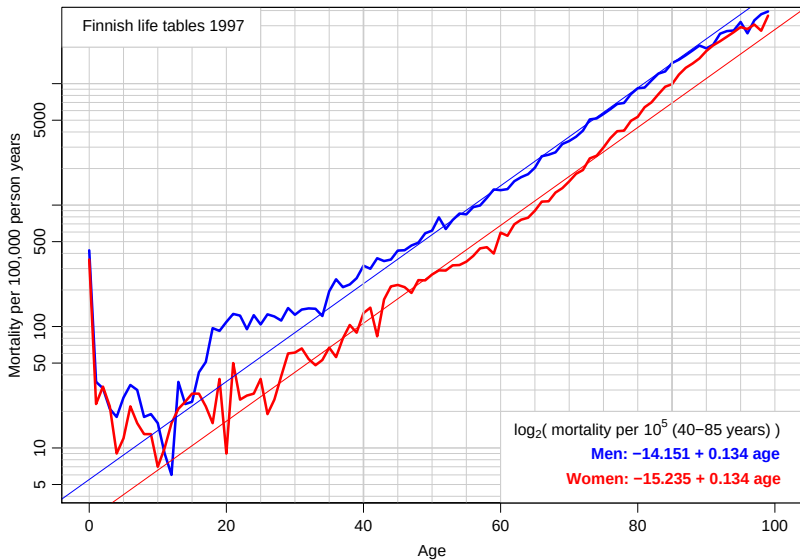


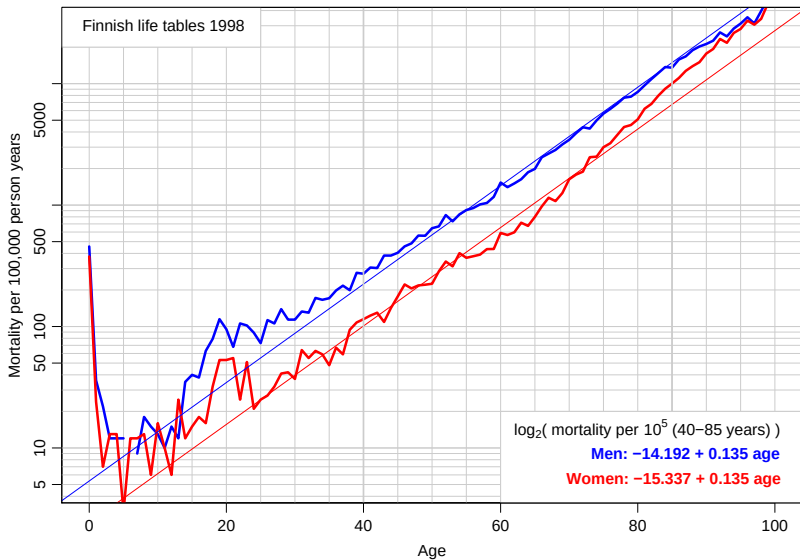


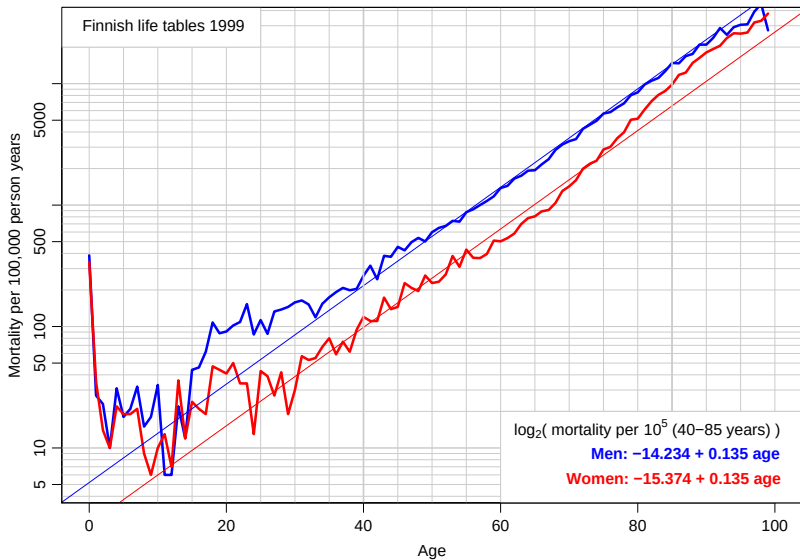


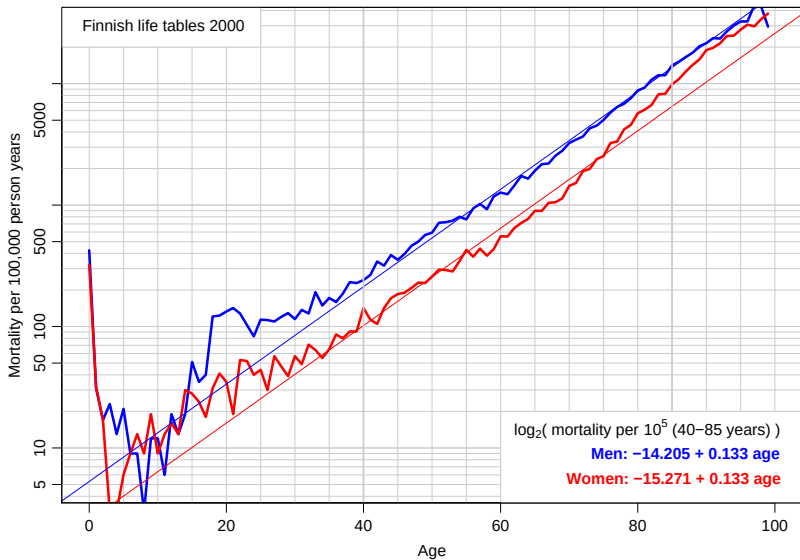


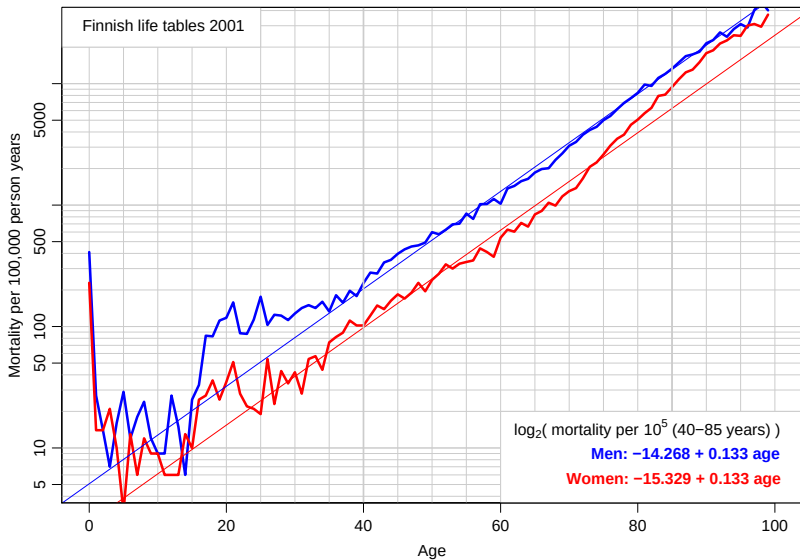


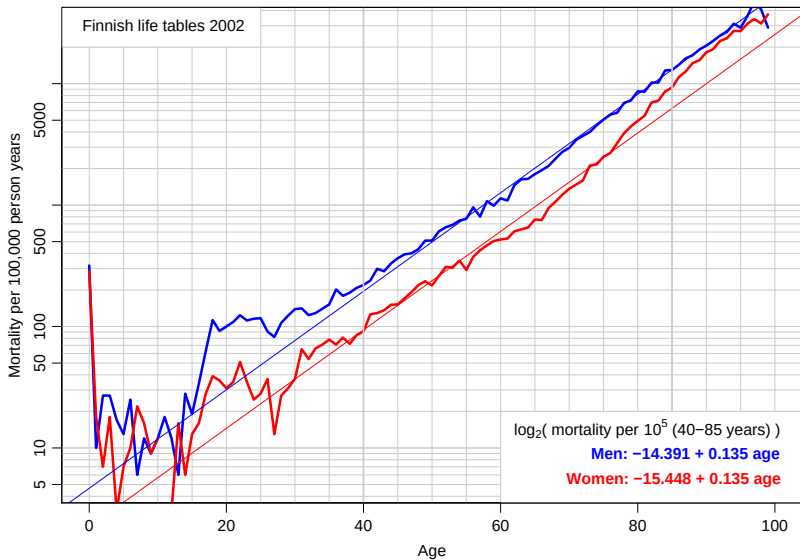


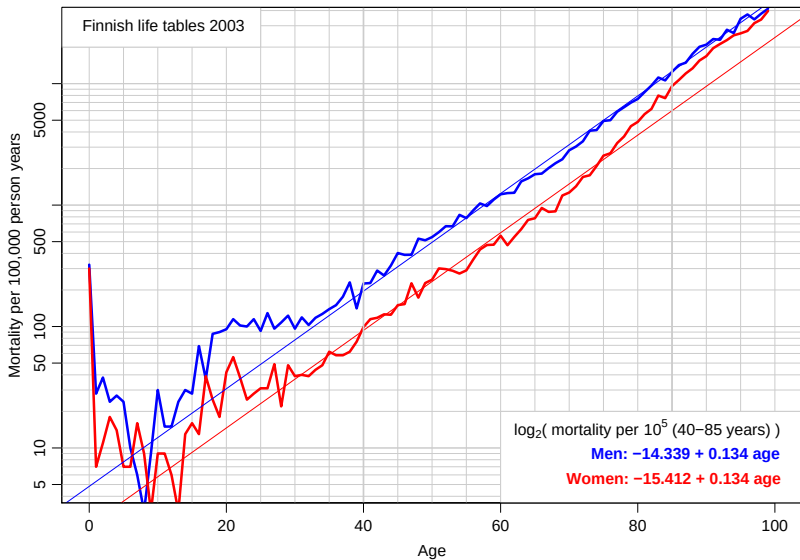












Who needs the Cox-model anyway?

Monday 23rd, afternoon

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

Empirical rates

At the individual level we introduce the

empirical rate: (d, y) ,

— the number of events ($d \in \{0, 1\}$) during y risk time.

Each individual contributes several observations.

Empirical rates are **responses** in survival analysis.

The timescale is a **covariate** — varies across empirical rates from one individual.

Likelihood

The likelihood from several empirical rates from one individual is a product of conditional probabilities. Hence, the log-likelihood contribution from one individual is a sum of terms.

The log-likelihood from one empirical rate (d, y) , assuming the event rate λ is constant is:

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

so the contribution from one individual is as the contribution from several independent Poisson observations.

The proportional hazards model

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

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

The covariate t has a special status:

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

Cox-likelihood

The partial likelihood for the regression parameters:

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

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

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

The Cox-likelihood as profile likelihood

Regression parameters describing the effect of covariates (other than the chosen underlying time scale).

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

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

Suppose the time scale has been divided into small intervals with at most one death in each.

Assume w.l.o.g. the y s in the empirical rates all are 1.

Log-likelihood contributions that contain information on a specific time-scale parameter α_t will be from:

- ▶ the (only) empirical rate $(d, y) = (1, 1)$ from the person that died at time t .
- ▶ all other empirical rates $(d, y) = (0, 1)$ from those who were at risk at time t .

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

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

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

The derivative w.r.t. α_t is:

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

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

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

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

What the Cox-model really is

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

- ▶ dividing time as finely as possible,
- ▶ modelling one covariate, the time-scale, with one parameter per distinct value,
- ▶ profiling these parameters out by maximizing the profile likelihood

Subsequently, one may recover the effect of the timescale by smoothing an estimate of the cumulative sum of these.

Sensible modelling

Replace the α_t s by a parametric function $f(t)$ with a limited number of parameters, for example:

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

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

Note that the intervals need not be derived from the death times.

Just small enough to make the constant rate assumption reasonable.

Splitting the dataset

The Poisson approach needs a dataset of empirical rates with small values of y .

Larger than the original: each individual contributes many empirical rates. From each empirical rate we get:

- ▶ Poisson-response d
- ▶ Risk time y
- ▶ Covariate value for the timescale (time since entry, current age, current date, ...)
- ▶ other covariates

Example: Mayo Clinic lung cancer

```
time status age sex
1 306      2  74  1
2 455      2  68  1
```

```
> Lx <- Lexis( exit=list( tfd=time), exit.status=(status==2), da
NOTE: entry is assumed to be 0 on the tfd timescale.
```

```
> tab(Lx,scale=365.25)
```

```
States:
```

```
  #records:
```

```
  To
```

From	FALSE	TRUE	Sum	#events:	#risk time:	Rate	(95
FALSE	63	165	228	165	190.5352	0.8659815	0.743432

```
> dx <- splitLexis( Lx, "tfd", breaks=c(0,unique(Lx$time)) )
```

```
> tab( dx, scale=365.25 )
```

```
States:
```

```
  #records:
```

```
  To
```

From	FALSE	TRUE	Sum	#events:	#risk time:	Rate	(
FALSE	19857	165	20022	165	190.5352	0.8659815	0.7434

The baseline hazard and survival functions

Using a parametric function to model the baseline hazard gives the possibility to plot this with confidence intervals for a given set of covariate values, x_0

The survival function in a multiplicative Poisson model has the form:

$$S(t) = \exp\left(-\sum_{\tau < t} \exp(g(\tau) + x_0' \gamma)\right)$$

This is just a non-linear function of the parameters in the model, g and γ . So the variance can be computed using the δ -method.

δ -method for survival function

1. Select timepoints t_i (fairly close).
2. Get estimates of log-rates $f(t_i) = g(t_i) + x'_0\gamma$ for these points:

$$\hat{f}(t_i) = \mathbf{B} \hat{\beta}$$

where β is the total parameter vector in the model.

3. Variance-covariance matrix of $\hat{\beta}$: $\hat{\Sigma}$.
4. Variance-covariance of $\hat{f}(t_i)$: $\mathbf{B}\hat{\Sigma}\mathbf{B}'$.
5. Transformation to the rates is the coordinate-wise exponential function, with derivative $\text{diag}[\exp(\hat{f}(t_i))]$

6. Variance-covariance matrix of the rates at the points t_i :

$$\text{diag}(e^{\hat{f}(t_i)}) \mathbf{B} \hat{\Sigma} \mathbf{B}' \text{diag}(e^{\hat{f}(t_i)})'$$

7. Transformation to cumulative hazard (ℓ is interval length):

$$\ell \times \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} e^{\hat{f}(t_1)} \\ e^{\hat{f}(t_2)} \\ e^{\hat{f}(t_3)} \\ e^{\hat{f}(t_4)} \end{bmatrix} = \mathbf{L} \begin{bmatrix} e^{\hat{f}(t_1)} \\ e^{\hat{f}(t_2)} \\ e^{\hat{f}(t_3)} \\ e^{\hat{f}(t_4)} \end{bmatrix}$$

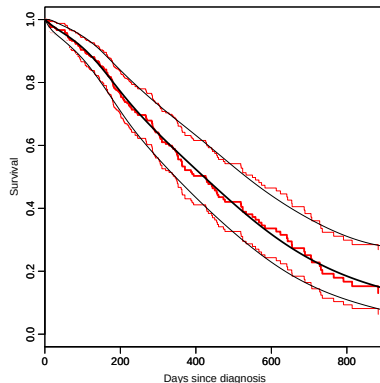
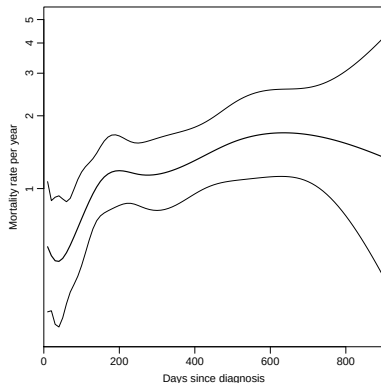
8. Variance-covariance matrix for the cumulative hazard is:

$$\mathbf{L} \text{diag}(e^{\hat{f}(t_i)}) \mathbf{B} \hat{\Sigma} \mathbf{B}' \text{diag}(e^{\hat{f}(t_i)})' \mathbf{L}'$$

This is all implemented in the `ci.cum()` function in `Epi`.

Mayo clinic lung cancer data

Smoothing by natural splines with 7 parameters;
knots at 0, 25, 75, 150, 250, 500, 1000 days



Computational tools for time-splitting

R: A function `splitLexis`, written by Martyn Plummer, in the package `Epi` available at CRAN.

Stata: The function `stsplit` (part of standard Stata).

Descendant of `stlexis` written by Michael Hills & David Clayton.

SAS: A macro `%Lexis`, available at <http://www.biostat.ku.dk/~bxc/Lexis>.

Conclusion

- ▶ 1 — 1 correspondence between life-tables and classical survival analysis.
- ▶ Cox-model (and the Kaplan-Meier estimator) is the lifetable taken ad absurdum, estimating one probability per interval defined by events/censorings.
- ▶ The natural modification is to use the modified life table estimator as basis for Poisson modelling.

Follow-up data

Monday 23rd, afternoon

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

Follow-up and rates

- ▶ Follow-up studies:
 - ▶ D — events, deaths
 - ▶ Y — person-years
 - ▶ $\lambda = D/Y$ rates
- ▶ Rates differ between persons.
- ▶ Rates differ **within** persons:
 - ▶ Along age
 - ▶ Along calendar time
- ▶ Multiple timescales.

Representation of follow-up data

In a cohort study we have records of:

Events and **Risk time**.

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

- ▶ Date of entry — date variable.
- ▶ Date of exit — date variable
- ▶ Status at exit — indicator-variable (0/1)

Specific for each *type* of outcome.

Aim of dividing time into bands:

Put D — events
 Y — risk time in intervals on the timescale:

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?

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

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

Splitting the follow up

	subj. 1	subj. 2	subj. 3
Age at E ntry:	13.06	18.44	4.54
Age at e X it:	44.95	41.14	11.12
S tatus at exit:	Dead	Alive	Dead

<i>Y</i>	31.89	22.70	6.58
<i>D</i>	1	0	1

Age	subj. 1		subj. 2		subj. 3		Σ	
	<i>Y</i>	<i>D</i>	<i>Y</i>	<i>D</i>	<i>Y</i>	<i>D</i>	<i>Y</i>	<i>D</i>
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
Σ	31.89	1	22.70	0	6.58	1	60.17	2

Splitting the follow-up

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

- but what if we want to keep track of calendar time too?

Timescales

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

Representation of follow-up on several timescales

- ▶ The time followed is the same on all timescales.
- ▶ Only use 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.

Follow-up data in Epi: Lexis objects

A follow-up study:

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

Timescales of interest:

- ▶ Age
- ▶ Calendar time
- ▶ Time since injection

Definition of Lexis object

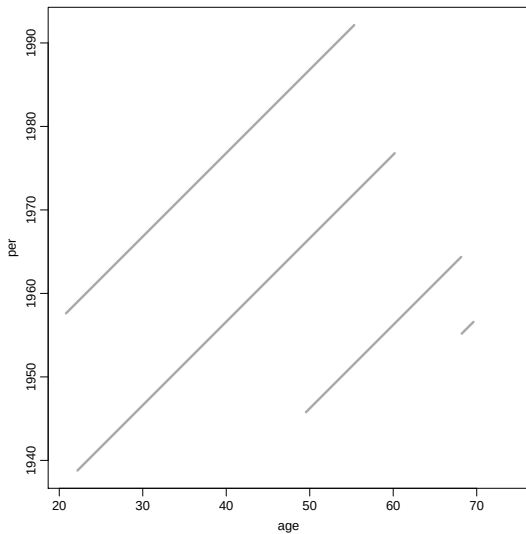
```
> thL <- Lexis( entry = list( age=injecdat-birthdat,  
+                             per=injecdat,  
+                             tfi=0 ),  
+               exit = list( per=exitdat ),  
+               exit.status = (exitstat==1)*1,  
+               data = th )
```

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

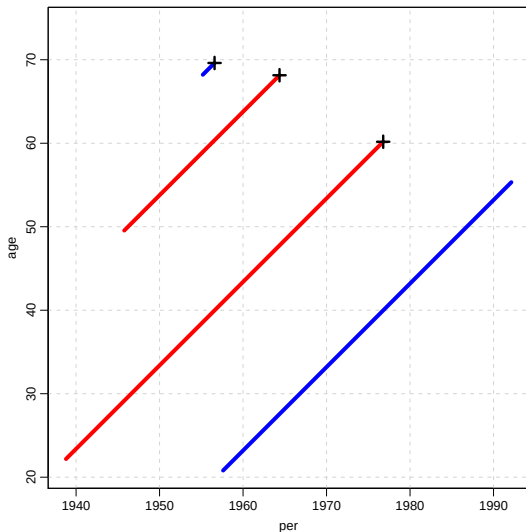
The looks of a Lexis object

```
> round( thL[,c(1:8,14,15)], 2 )
```

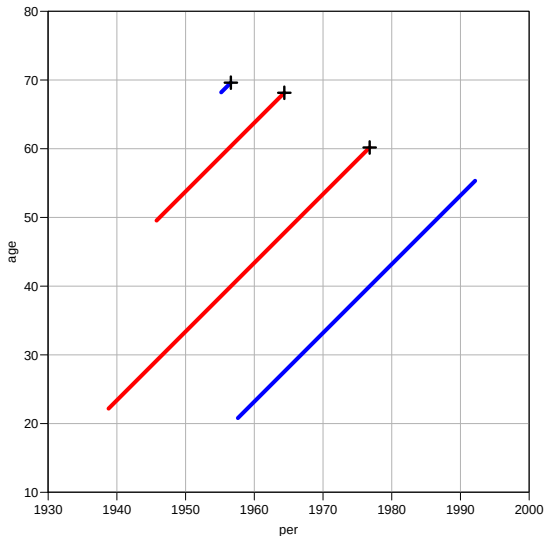
	age	per	tfi	lex.dur	lex.Cst	lex.Xst	lex.id	ic
1	22.18	1938.79	0	38.00	0	1	1	1
2	49.55	1945.77	0	18.60	0	1	2	640
3	68.21	1955.18	0	1.40	0	1	3	3425
4	20.80	1957.61	0	34.52	0	0	4	4017



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

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



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

Splitting follow-up time

```
> spl1 <- splitLexis( thL, "age", breaks=seq(0,100,20) )  
> round( spl1, 2 )
```

	lex.id	age	per	tfi	lex.dur	lex.Cst	lex.Xst	id	sex	bi
1	1	22.18	1938.79	0.00	17.82	0	0	1	2	1
2	1	40.00	1956.61	17.82	20.00	0	0	1	2	1
3	1	60.00	1976.61	37.82	0.18	0	1	1	2	1
4	2	49.55	1945.77	0.00	10.45	0	0	640	2	1
5	2	60.00	1956.23	10.45	8.14	0	1	640	2	1
6	3	68.21	1955.18	0.00	1.40	0	1	3425	1	1
7	4	20.80	1957.61	0.00	19.20	0	0	4017	2	1
8	4	40.00	1976.81	19.20	15.33	0	0	4017	2	1

Split on a second timescale

```
> # Split further on tfi:  
> spl2 <- splitLexis( spl1, "tfi", breaks=c(0,1,5,20,100) )  
> round( spl2, 2 )
```

	lex.id	age	per	tfi	lex.dur	lex.Cst	lex.Xst	id	sex	b
1	1	22.18	1938.79	0.00	1.00	0	0	1	2	
2	1	23.18	1939.79	1.00	4.00	0	0	1	2	
3	1	27.18	1943.79	5.00	12.82	0	0	1	2	
4	1	40.00	1956.61	17.82	2.18	0	0	1	2	
5	1	42.18	1958.79	20.00	17.82	0	0	1	2	
6	1	60.00	1976.61	37.82	0.18	0	1	1	2	
7	2	49.55	1945.77	0.00	1.00	0	0	640	2	
8	2	50.55	1946.77	1.00	4.00	0	0	640	2	
9	2	54.55	1950.77	5.00	5.45	0	0	640	2	
10	2	60.00	1956.23	10.45	8.14	0	1	640	2	
11	3	68.21	1955.18	0.00	1.00	0	0	3425	1	
12	3	69.21	1956.18	1.00	0.40	0	1	3425	1	
13	4	20.80	1957.61	0.00	1.00	0	0	4017	2	
14	4	21.80	1958.61	1.00	4.00	0	0	4017	2	
15	4	25.80	1962.61	5.00	14.20	0	0	4017	2	
16	4	40.00	1976.61	10.00	0.00	0	0	4017	2	

The Poisson likelihood for time-split data

One record per person-**interval** (i, t) :

$$D \ln(\lambda) - \lambda Y = \sum_{i,t} (d_{it} \ln(\lambda) - \lambda y_{it})$$

Assuming that the death indicator ($d_i \in \{0, 1\}$) is Poisson, with log-offset y_i will give the same result.

The model assume that rates are constant.

But the split data allows relaxing this to models that assume different rates for different (d_{it}, y_{it}) .

Where are the (d_{it}, y_{it}) in the split data?

The Poisson likelihood for time-split data

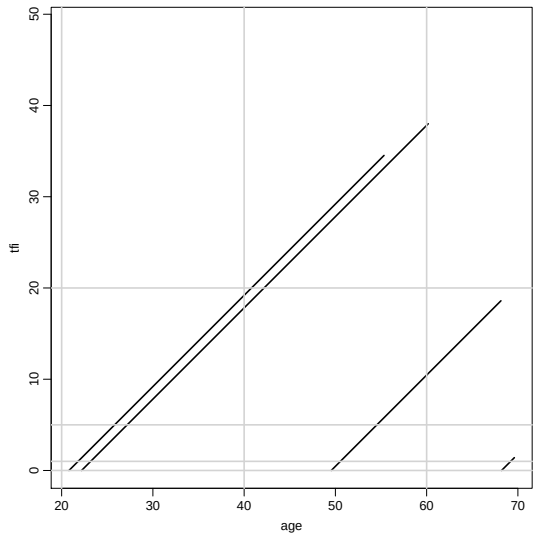
If $d \sim \text{Poisson}(\lambda y)$, i.e. with mean (λy) then the log-likelihood is

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

If we assume a multiplicative model for the rates, i.e. an additive model for the log-rates, we can use a Poisson model which is multiplicative in the mean, μ , i.e. linear in $\log(\mu)$:

$$\log(\mu) = \log(\lambda y) = \log(\lambda) + \log(y)$$

Regression model must include $\log(y)$ as covariate with coefficient fixed to 1 — an offset-variable.



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

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

```
> round( spl2, 2 )
```

	lex.id	age	per	tfi	lex.dur	lex.Cst	lex.Xst	id	sex	b
1	1	22.18	1938.79	0.00	1.00	0	0	1	2	
2	1	23.18	1939.79	1.00	4.00	0	0	1	2	
3	1	27.18	1943.79	5.00	12.82	0	0	1	2	
4	1	40.00	1956.61	17.82	2.18	0	0	1	2	
5	1	42.18	1958.79	20.00	17.82	0	0	1	2	
6	1	60.00	1976.61	37.82	0.18	0	1	1	2	
7	2	49.55	1945.77	0.00	1.00	0	0	640	2	
8	2	50.55	1946.77	1.00	4.00	0	0	640	2	
9	2	54.55	1950.77	5.00	5.45	0	0	640	2	
10	2	60.00	1956.23	10.45	8.14	0	1	640	2	
11	3	68.21	1955.18	0.00	1.00	0	0	3425	1	
12	3	69.21	1956.18	1.00	0.40	0	1	3425	1	
13	4	20.80	1957.61	0.00	1.00	0	0	4017	2	
14	4	21.80	1958.61	1.00	4.00	0	0	4017	2	
15	4	25.80	1962.61	5.00	14.20	0	0	4017	2	
16	4	40.00	1976.81	19.20	0.80	0	0	4017	2	
17	4	40.80	1977.61	20.00	14.52	0	0	4017	2	

Analysis of results

- ▶ d_i — events in the variable: `lex.Xst`.
- ▶ y_i — risk time: `lex.dur` (duration).
Enters in the model via $\log(y)$ as offset.
- ▶ 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.

Poisson model for split data

- ▶ Each interval contribute λY to the log-likelihood.
- ▶ All intervals with the same set of covariate values (age, exposure, . . .) have the same λ .
- ▶ The log-likelihood contribution from these is $\lambda \sum Y$ — the same as from aggregated data.
- ▶ The event intervals contribute each $D \log \lambda$.
- ▶ The log-likelihood contribution from those with the same lambda is $\sum D \log \lambda$ — the same as from aggregated data.
- ▶ The log-likelihood is the same for split data and aggregated data — no need to tabulate first.

Age at entry

Monday 23rd, afternoon

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

Age at entry as covariate

t : time since entry

e : age at entry

$a = e + t$: current age

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

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

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

Non-linear effects of time-scales

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

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

Three quantities can be arbitrarily moved between the three functions:

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

because $t - a + e = 0$.

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

“Controlling for age”

— is not a well defined statement.

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

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

This would require modelling of multiple timescales.

Which is best accomplished by splitting time.

Models for tabulated data

Monday 23rd, afternoon

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

Conceptual set-up

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

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

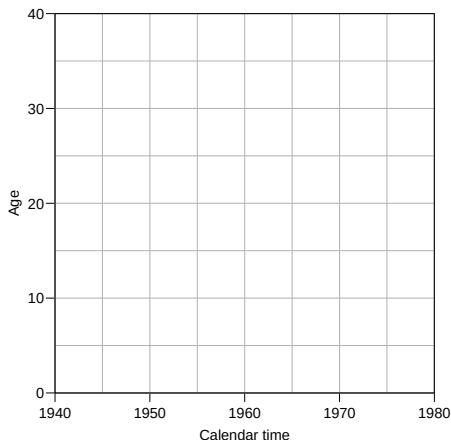
Realistic set-up

- ▶ 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.
Or get it from the human mortality database.
- ▶ Analyse by Poisson model.

Lexis diagram ¹

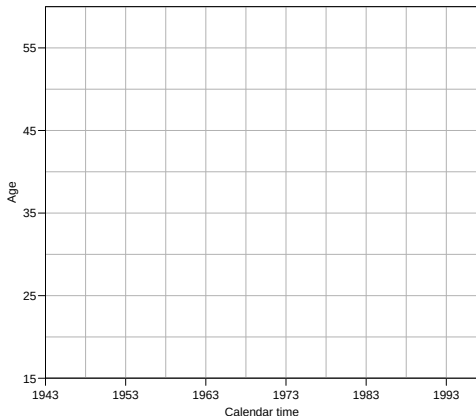


Disease registers
record events.

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



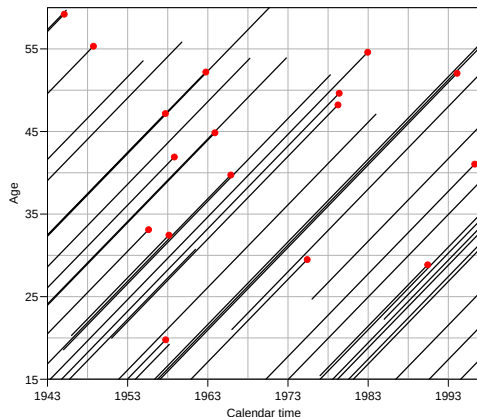
Registration of:

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.:

Age	Seminoma cases				Person-years			
	1943	1948	1953	1958	1943	1948	1953	1958
15	2	3	4	1	773812	744217	794123	972853
20	7	7	17	8	813022	744706	721810	770859
25	28	23	26	35	790501	781827	722968	698612
30	28	43	49	51	799293	774542	769298	711596
35	36	42	39	44	769356	782893	760213	760452
40	24	32	46	53	694073	754322	768471	749912

Reshape data to analysis form:

	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
1	15	1948	3	744217
2	20	1948	7	744706
3	25	1948	23	781827
4	30	1948	43	774542
5	35	1948	42	782893
6	40	1948	32	754322
1	15	1953	4	794123
2	20	1953	17	721810
3	25	1953	26	722968
4	30	1953	49	769298
5	35	1953	39	760213
6	40	1953	46	768471
1	15	1958	1	972853
2	20	1958	8	770859
3	25	1958	35	698612

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 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 model for the testiscancer 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.lin()` from the `Epi` package extracts coefficients and computes confidence intervals:

```
> round( ci.lin( m0 ), 3 )
```

	Estimate	StdErr	z	P	2.5%	97.5%
(Intercept)	-1.476	0.327	-4.517	0.000	-2.116	-0.836
factor(A)20	1.454	0.354	4.101	0.000	0.759	2.149
factor(A)25	2.532	0.330	7.671	0.000	1.885	3.179
factor(A)30	2.933	0.325	9.013	0.000	2.295	3.570
factor(A)35	2.861	0.326	8.779	0.000	2.223	3.500
factor(A)40	2.852	0.326	8.741	0.000	2.213	3.492
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

Subsets of parameter estimates accessed via a character string that is greped to the names.

```
> 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

Rates / rate-ratios are computed on the fly by
Exp=TRUE:

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

	Estimate	StdErr	z	P	exp(Est.)	2.5%	97.5%
factor(P)1948	0.175	0.121	1.447	0.148	1.192	0.940	1.511
factor(P)1953	0.382	0.116	3.286	0.001	1.466	1.167	1.841
factor(P)1958	0.466	0.115	4.052	0.000	1.593	1.272	1.996

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

```
> CM <- rbind( c( 0,-1, 0),  
+              c( 1,-1, 0),  
+              c( 0, 0, 0),  
+              c( 0,-1, 1) )  
> round( ci.lin( m0, subset="P", ctr.mat=CM, Exp=TRUE ), 3 )
```

	Estimate	StdErr	z	P	exp(Est.)	2.5%	97.5%
[1,]	-0.382	0.116	-3.286	0.001	0.682	0.543	0.857
[2,]	-0.207	0.110	-1.874	0.061	0.813	0.655	1.010
[3,]	0.000	0.000	NaN	NaN	1.000	1.000	1.000
[4,]	0.084	0.104	0.808	0.419	1.087	0.887	1.332

Age-Period and Age-Cohort models

Tuesday 24th, morning

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

Register data - rates

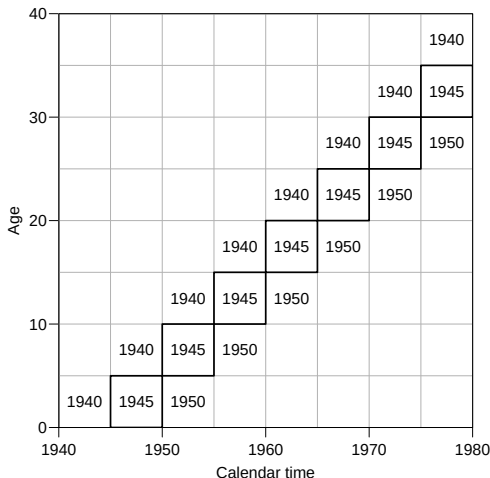
Rates in “tiles” of the Lexis diagram:

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

Descriptive epidemiology based on disease registers:
How do the rates vary across 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).

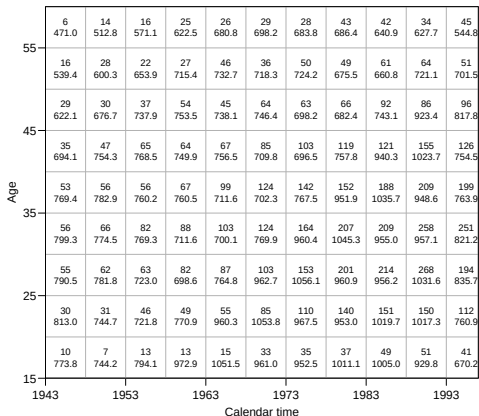
Synthetic cohorts



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

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

Data matrix: Male risk time

1000 person-years

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

Data matrix: Empirical rates

Rate per 1000,000 person-years

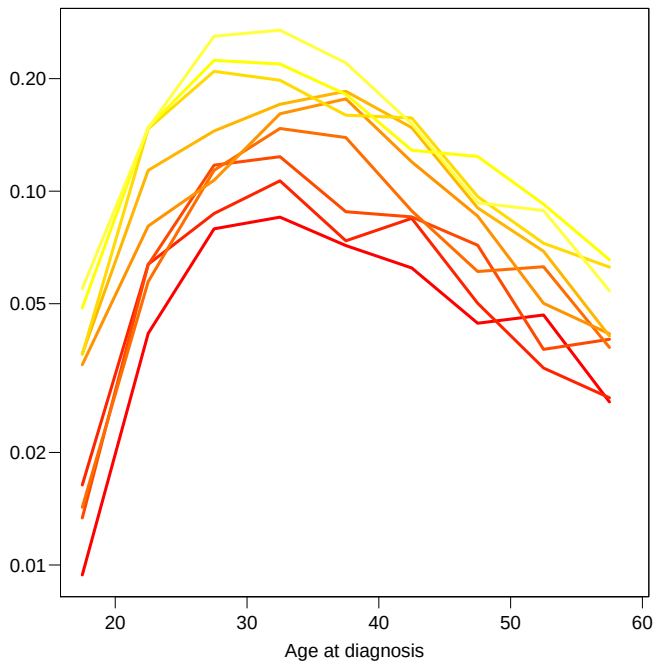
Age	Date of diagnosis (<i>year</i> – 1900)								8
	48–52	53–57	58–62	63–67	68–72	73–77	78–82	83–87	
15–19	9.4	16.4	13.4	14.3	34.3	36.7	36.6	48.8	
20–24	41.6	63.7	63.6	57.3	80.7	113.7	146.9	148.1	
25–29	79.3	87.1	117.4	113.8	107.0	144.9	209.2	223.8	
30–34	85.2	106.6	123.7	147.1	161.1	170.8	198.0	218.8	
35–39	71.5	73.7	88.1	139.1	176.6	185.0	159.7	181.5	
40–44	62.3	84.6	85.3	88.6	119.8	147.9	157.0	128.7	
45–49	44.3	50.1	71.7	61.0	85.7	90.2	96.7	123.8	
50–54	46.6	33.6	37.7	62.8	50.1	69.0	72.5	92.3	
55–59	27.3	28.0	40.2	38.2	41.5	40.9	62.6	65.5	

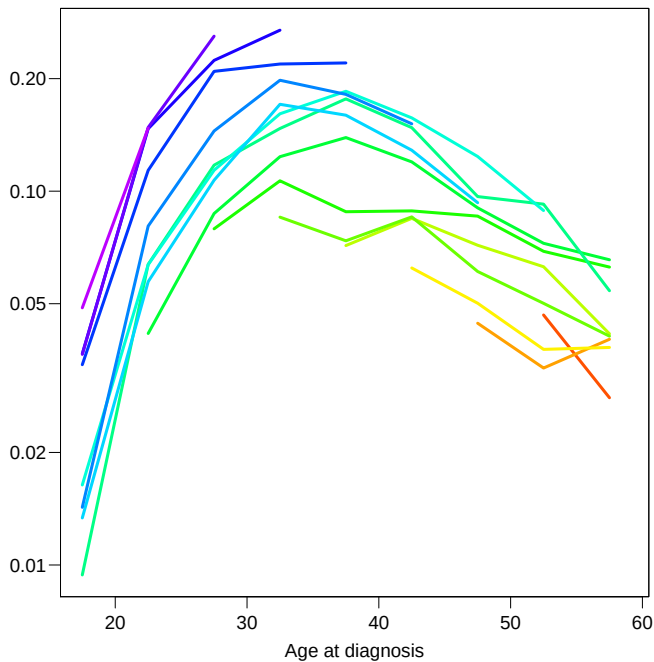
The classical plots

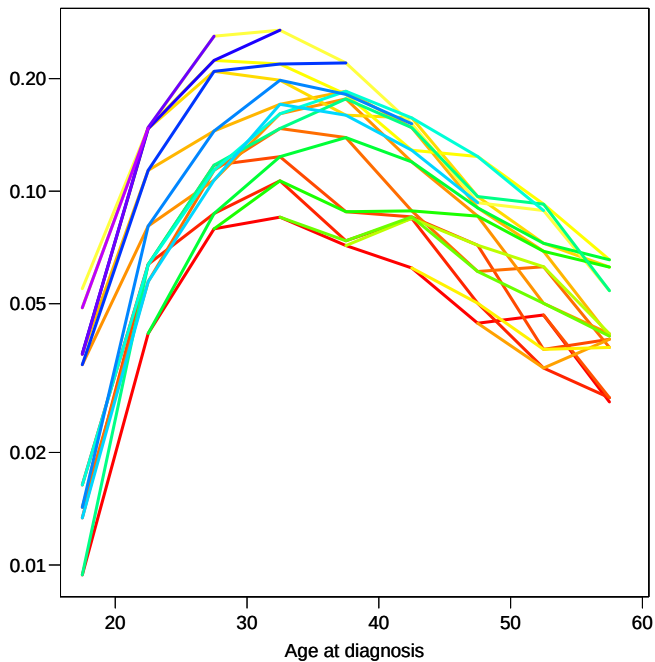
Given a table of rates classified by age and period, we can do 4 “classical” plots:

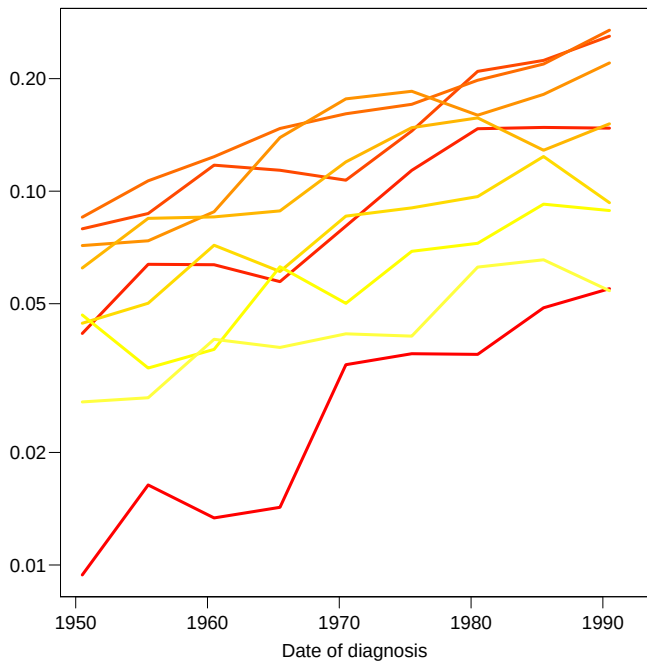
- ▶ Rates versus age at diagnosis (period):
— rates in the same ageclass connected.
- ▶ Rates versus age at diagnosis:
— rates in the same birth-cohort connected.
- ▶ Rates versus date of diagnosis:
— rates in the same ageclass connected.
- ▶ Rates versus date of date of birth:
— rates in the same ageclass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

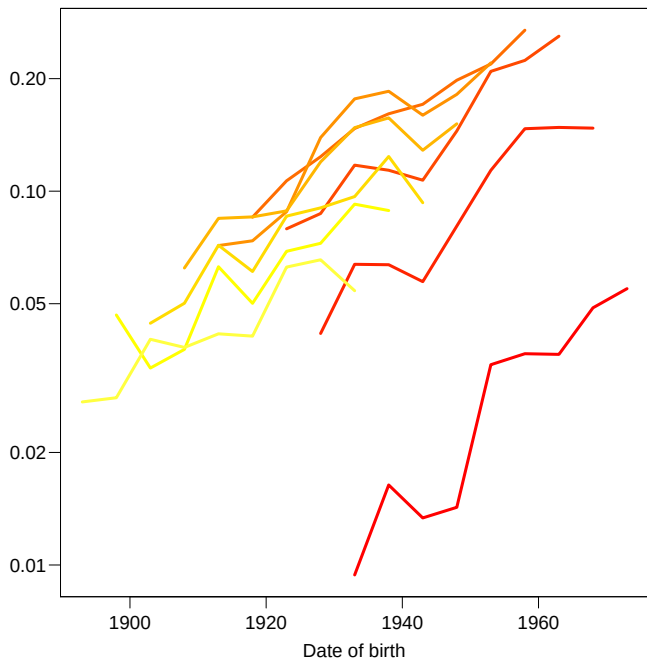
These plots can be produced by the R-function `rateplot`.











Age-period model

Rates are proportional between periods:

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

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 model in R

Reference period is the 5th period (1970.5 ~ 1968-72):

```
> ap <- glm( D ~ factor( A ) - 1 + relevel( factor( P ), 5 ) +  
+           offset( log( Y ) ),  
+           family=poisson )  
> summary( ap )
```

Call:

```
glm(formula = D ~ factor(A) - 1 + relevel(factor(P), 5) + offset
```

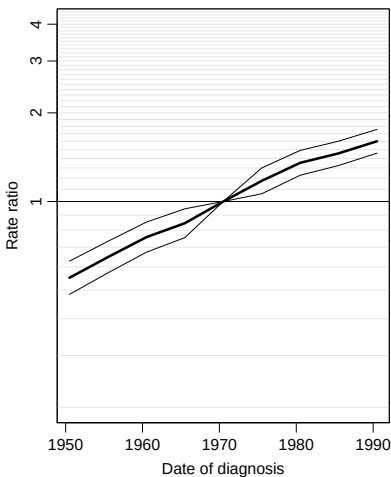
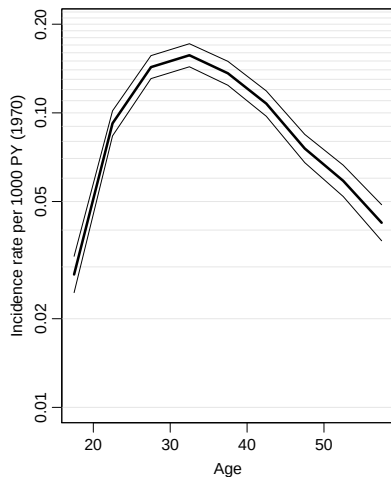
Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.0925	-0.8784	0.1148	0.9790	2.7653

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
factor(A)17.5	-3.56605	0.07249	-49.194	< 2e-16
factor(A)22.5	-2.38447	0.04992	-47.766	< 2e-16
factor(A)27.5	-1.94496	0.04583	-42.442	< 2e-16
factor(A)32.5	-1.85214	0.04597	-40.294	< 2e-16
factor(A)37.5	-1.99308	0.04770	-41.787	< 2e-16
factor(A)42.5	-2.23017	0.05057	-44.104	< 2e-16
factor(A)47.5	-2.58125	0.05631	-45.839	< 2e-16

Graph of estimates with confidence intervals



Age-cohort model

Rates are proportional between cohorts:

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

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

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

Fit the model in R

Reference period is the 9th cohort (1933 ~ 1928-38):

```
> ac <- glm( D ~ factor( A ) - 1 + relevel( factor( C ), 9 ) +  
+           offset( log( Y ) ),  
+           family=poisson )  
> summary( ac )
```

Call:

```
glm(formula = D ~ factor(A) - 1 + relevel(factor(C), 9) + offset
```

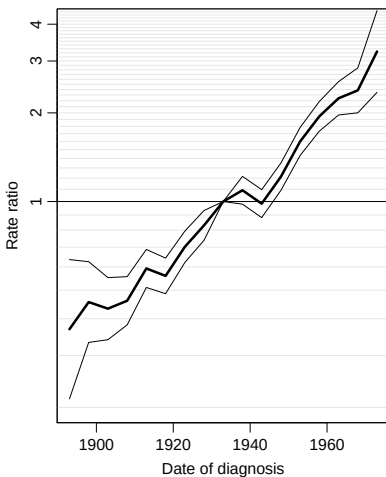
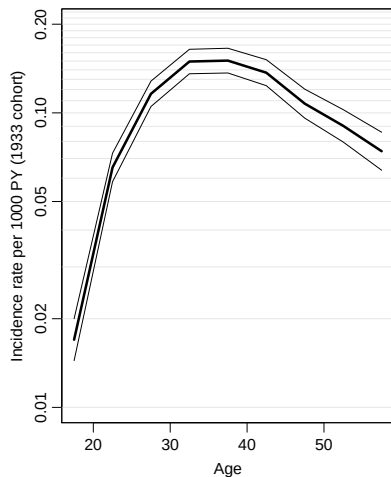
Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.92700	-0.72364	-0.02422	0.59623	1.87770

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
factor(A)17.5	-4.07597	0.08360	-48.753	< 2e-16
factor(A)22.5	-2.72942	0.05683	-48.031	< 2e-16
factor(A)27.5	-2.15347	0.05066	-42.505	< 2e-16
factor(A)32.5	-1.90118	0.04878	-38.976	< 2e-16
factor(A)37.5	-1.89404	0.04934	-38.387	< 2e-16
factor(A)42.5	-1.98846	0.05178	-38.399	< 2e-16
factor(A)47.5	-2.23047	0.05775	-38.626	< 2e-16

Graph of estimates with confidence intervals



Age-drift model

Tuesday 24th, morning

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

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))
```

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)),
```

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?

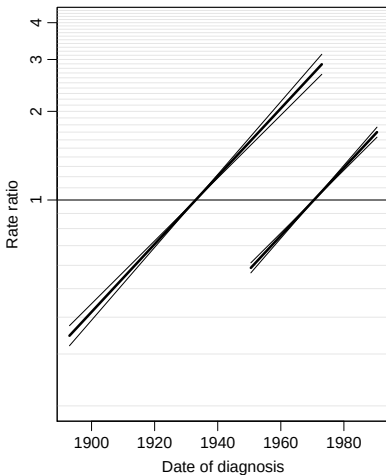
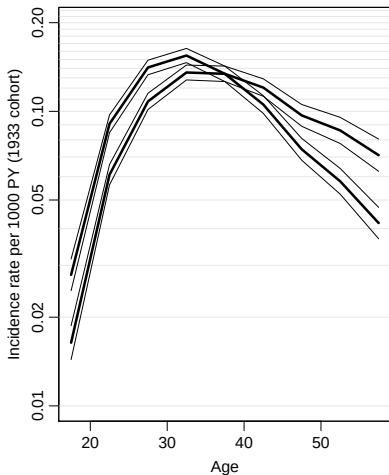
$$\begin{aligned}\alpha_a + \beta(p - p_0) &= \alpha_a + \beta(a + c - (a_0 + c_0)) \\ &= \underbrace{\alpha_a + \beta(a - a_0)}_{\text{cohort age-effect}} + \beta(c - c_0)\end{aligned}$$

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

Tuesday 24th, afternoon

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

The age-period-cohort model

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

- ▶ Three effects:
 - ▶ Age (at diagnosis)
 - ▶ Period (of diagnosis)
 - ▶ Cohort (of birth)
- ▶ Modelled on the same *scale*.
- ▶ No assumptions about the *shape* of effects.

Fitting the model in R

```
> c1933.p <- glm( D ~ factor( A ) - 1 +  
+                 relevel( factor( C ), "1933" ) +  
+                 factor( P ) + offset( log( Y ) ), family=poisson)  
> summary( c1933.p )  
Coefficients: (1 not defined because of singularities)  
              Estimate Std. Error z value Pr(>|z|)      
factor(A)17.5      -4.27754    0.10479  -40.819  < 2e-16 ***  
...  
factor(A)57.5       -2.75892    0.08380  -32.922  < 2e-16 ***  
relevel(factor(C), "1933")1893 -0.84187    0.28009   -3.006  0.0025 **  
...  
relevel(factor(C), "1933")1928 -0.17922    0.05965   -3.005  0.0025 **  
relevel(factor(C), "1933")1938  0.07540    0.05592    1.348  0.1767 .  
...  
relevel(factor(C), "1933")1973  1.37438    0.17490    7.858  3.90e-16 ***  
factor(P)1955.5      0.04793    0.07022    0.683  0.4915 .  
...  
factor(P)1985.5      0.09276    0.04091    2.267  0.0235 *  
factor(P)1990.5      NA          NA          NA          NA  
  
(Dispersion parameter for poisson family taken to be 1)  
Null deviance: 89358.526 on 81 degrees of freedom  
Residual deviance: 35.459 on 49 degrees of freedom
```

No. of parameters

A has 9 levels

P has 9 levels

C has 17 levels

Age-drift model has $A + 1 = 10$ parameters

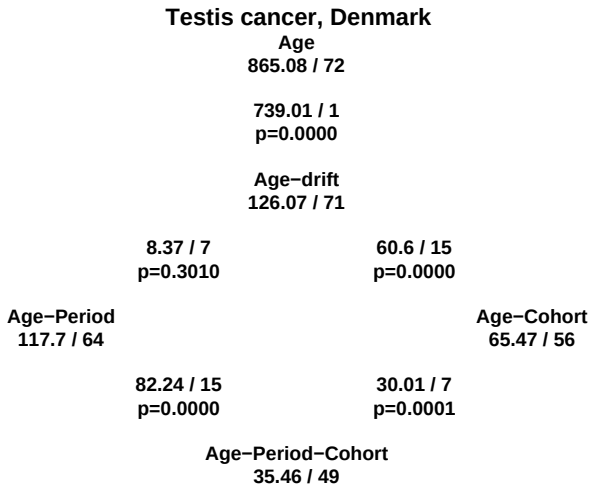
Age-period model has $A + P - 1 = 17$ parameters

Age-cohort model has $A + C - 1 = 25$ parameters

Age-period-cohort model has $A + P + C - 3 = 32$ parameters

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

Relationship of models



Test for effects

Model	Deviance	d.f.	p-value
Age - drift	126.07	71	
Δ	60.60	15	0.000
Age - cohort	65.47	56	
Δ	30.01	7	0.000
Age - period - cohort	35.46	49	
Δ	82.24	15	0.000
Age - period	117.70	64	
Δ	8.37	7	0.301
Age - drift	126.07	71	

How to choose a parametrization

Standard programs: Put extremes of periods or cohorts to 0, and choose a reference for the other.

Holford: Extract linear effects by regression:

$$\begin{aligned}\lambda(a, p) &= \hat{\alpha}_a + \\ &\quad \hat{\beta}_p + \\ &\quad \hat{\gamma}_c \\ \\ &= \tilde{\alpha}_a + \hat{\mu}_a + \hat{\delta}_a a + \\ &\quad \tilde{\beta}_p + \hat{\mu}_p + \hat{\delta}_p p + \\ &\quad \tilde{\gamma}_c + \hat{\mu}_c + \hat{\delta}_c c\end{aligned}$$

Putting it together again

Assumptions are needed, 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.

Period effect, on average 0, slope is 0:

$$g(p) = \tilde{\beta}_p = \beta_p - \hat{\mu}_p - \hat{\delta}_p p$$

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

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

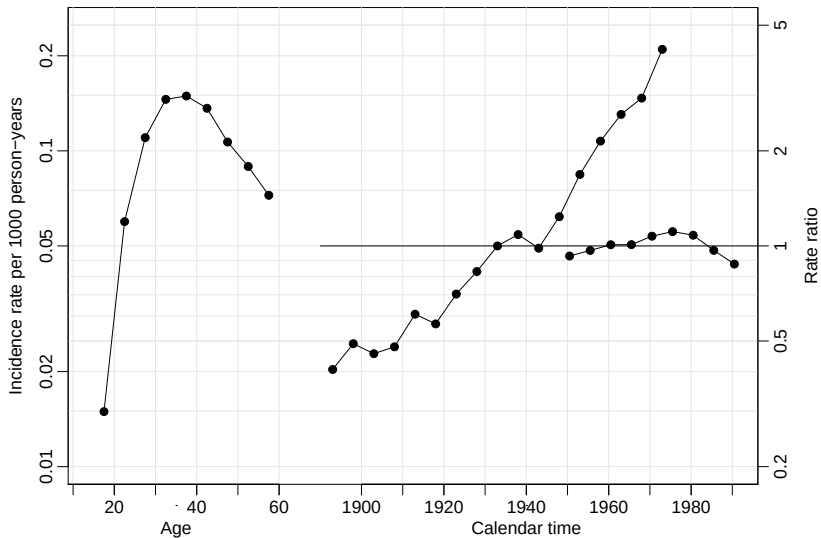
The rest is the age-effect:

$$f(a) = \alpha_a + \hat{\mu}_p + \hat{\delta}_p a + \hat{\delta}_p c_0 + \gamma_{c_0}$$

How it adds up:

$$\begin{aligned}\lambda(a, p) &= \hat{\alpha}_a + \hat{\beta}_p + \hat{\gamma}_c \\ &= \hat{\alpha}_a + \gamma_{c_0} + \hat{\mu}_p + \hat{\delta}_p(a + c_0) + \\ &\quad \hat{\beta}_p - \hat{\mu}_p - \hat{\delta}_p(a + c) + \\ &\quad \hat{\gamma}_c - \gamma_{c_0} + \hat{\delta}_p(c - c_0)\end{aligned}$$

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



A simple practical approach

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

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

Now consider the full APC-model with age and cohort effects as estimated:

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

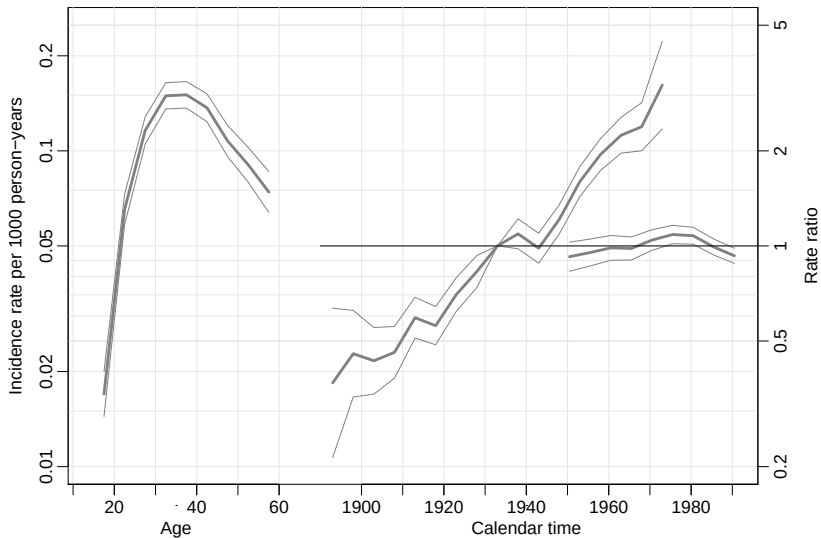
The residual period effect can be estimated if we note that for the number of cases we have:

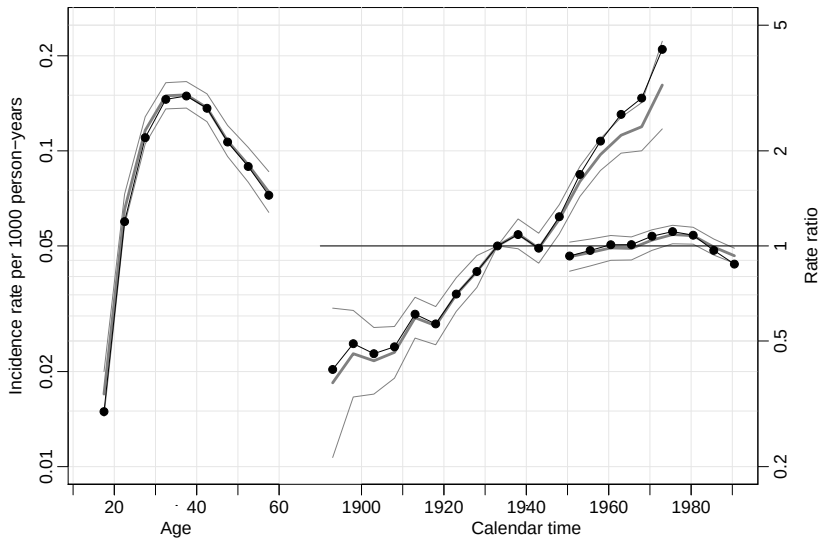
$$\log(\text{expected cases}) = \log[\lambda(a, p)Y] = \underbrace{\hat{\alpha}_a + \hat{\gamma}_c + \log(Y)}_{\text{"known"}}$$

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_p s are estimated in a Poisson model with this as offset.

Advantage: We get the standard errors for free..





Tabulation in the Lexis diagram

Tuesday 24th, morning

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

Tabulation of register data

55	6 471.0	14 512.8	16 571.1	25 622.5	26 680.8	29 698.2	28 683.8	43 686.4	42 640.9	34 627.7	45 544.8
	16 539.4	28 600.3	22 653.9	27 715.4	46 732.7	36 718.3	50 724.2	49 675.5	61 660.8	64 721.1	51 701.5
45	29 622.1	30 676.7	37 737.9	54 753.5	45 738.1	64 746.4	63 698.2	66 682.4	92 743.1	86 923.4	96 817.8
	35 694.1	47 754.3	65 768.5	64 749.9	67 756.5	85 709.8	103 696.5	119 757.8	121 940.3	155 1023.7	126 754.5
35	53 769.4	56 782.9	56 760.2	67 760.5	99 711.6	124 702.3	142 767.5	152 951.9	188 1035.7	209 948.6	199 763.9
	56 799.3	66 774.5	82 769.3	88 711.6	103 700.1	124 769.9	164 960.4	207 1045.3	209 955.0	258 957.1	251 821.2
25	55 790.5	62 781.8	63 723.0	82 698.6	87 764.8	103 962.7	153 1056.1	201 960.9	214 956.2	268 1031.6	194 835.7
	30 813.0	31 744.7	46 721.8	49 770.9	55 960.3	85 1053.8	110 967.5	140 953.0	151 1019.7	150 1017.3	112 760.9
15	10 773.8	7 744.2	13 794.1	13 972.9	15 1051.5	33 961.0	35 952.5	37 1011.1	49 1005.0	51 929.8	41 670.2
	1943	1953	1963	1973	1983	1993					
	Calendar time										

Testis cancer cases in Denmark.

Male person-years in Denmark.

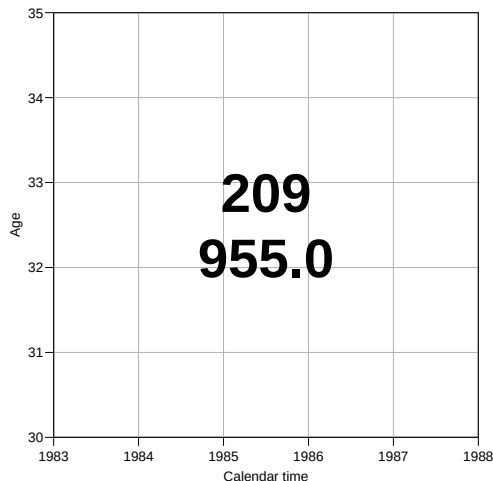
Tabulation of register data

55	6 471.0	14 512.8	16 571.1	25 622.5	26 680.8	29 698.2	28 683.8	43 686.4	42 640.9	34 627.7	45 544.8
	16 539.4	28 600.3	22 653.9	27 715.4	46 732.7	36 718.3	50 724.2	49 675.5	61 660.8	64 721.1	51 701.5
45	29 622.1	30 676.7	37 737.9	54 753.5	45 738.1	64 746.4	63 698.2	66 682.4	92 743.1	86 923.4	96 817.8
	35 694.1	47 754.3	65 768.5	64 749.9	67 756.5	85 709.8	103 696.5	119 757.8	121 940.3	155 1023.7	126 754.5
35	53 769.4	56 782.9	56 760.2	67 760.5	99 711.6	124 702.3	142 767.5	152 951.9	188 1035.7	209 948.6	199 763.9
	56 799.3	66 774.5	82 769.3	88 711.6	103 700.1	124 769.9	164 960.4	207 1045.3	209 955.0	258 957.1	251 821.2
25	55 790.5	62 781.8	63 723.0	82 698.6	87 764.8	103 962.7	153 1056.1	201 960.9	214 956.2	268 1031.6	194 835.7
	30 813.0	31 744.7	46 721.8	49 770.9	55 960.3	85 1053.8	110 967.5	140 953.0	151 1019.7	150 1017.3	112 760.9
15	10 773.8	7 744.2	13 794.1	13 972.9	15 1051.5	33 961.0	35 952.5	37 1011.1	49 1005.0	51 929.8	41 670.2
	1943	1953	1963	1973	1983	1993					
Calendar time											

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

35	12 40.2	5 38.7	5 38.0	11 37.9	6 38.0	
34	8 38.7	4 38.0	6 37.9	11 38.0	11 38.1	
33	12 38.1	7 37.9	13 38.0	8 38.1	8 38.2	
32	6 38.0	7 38.0	9 38.1	11 38.2	10 38.3	
31	7 38.0	5 38.0	9 38.1	10 38.2	8 38.3	
30						
	1983	1984	1985	1986	1987	1988

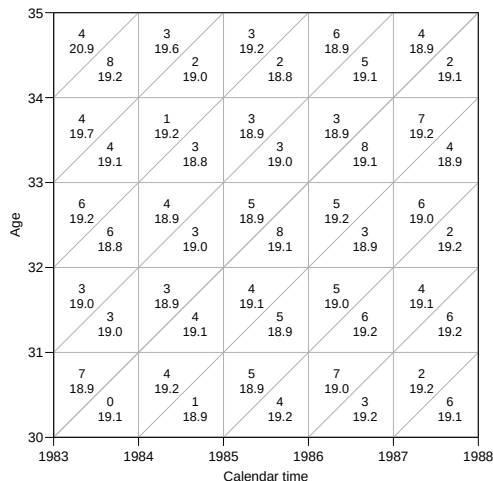
Age

Calendar time

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. ($\square$)

B-sets: Classification by age and cohort. ($\parallel$)

C-sets: Classification by cohort and period. ($\backslash$)

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:

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

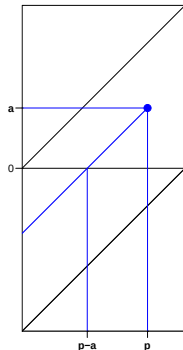
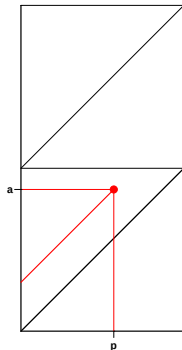
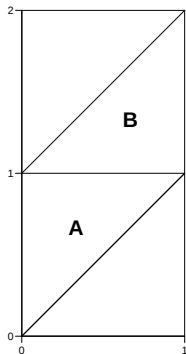
Lower triangles: Classification by age and cohort, last born cohort. ($\triangle$)

Mean time in triangles

Modelling 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**:

$$E_{\mathbf{A}}(a) = \int_{p=0}^{p=1} \int_{a=p}^{a=1} a \times 2 \, da \, dp = \int_{p=0}^{p=1} 1 - p^2 \, dp = \frac{2}{3}$$

$$E_{\mathbf{A}}(p) = \int_{a=0}^{a=1} \int_{p=0}^{p=a} p \times 2 \, dp \, da = \int_{a=0}^{a=1} a^2 \, dp = \frac{1}{3}$$

$$E_{\mathbf{A}}(c) = \frac{1}{3} - \frac{2}{3} = -\frac{1}{3}$$

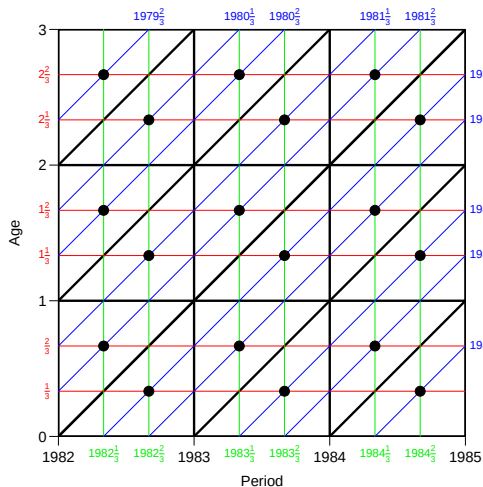
Lower triangles ($\triangle$), **B**:

$$E_{\mathbf{B}}(a) = \int_{p=0}^{p=1} \int_{a=0}^{a=p} a \times 2 \, da \, dp = \int_{p=0}^{p=1} p^2 \, dp = \frac{1}{3}$$

$$E_{\mathbf{B}}(p) = \int_{a=0}^{a=1} \int_{p=a}^{p=1} p \times 2 \, dp \, da = \int_{a=0}^{a=1} 1 - a^2 \, dp = \frac{2}{3}$$

$$E_{\mathbf{B}}(c) = \frac{2}{3} - \frac{1}{3} = \frac{1}{3}$$

Tabulation by age, period and cohort



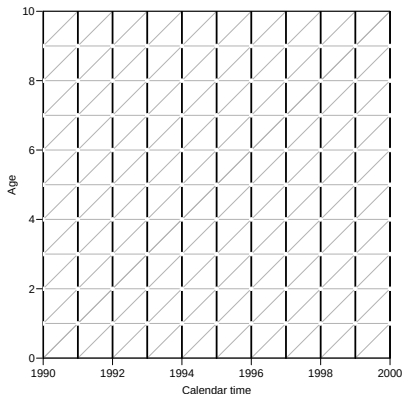
Gives triangular sets with differing mean age, period and cohort:

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

Population figures

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

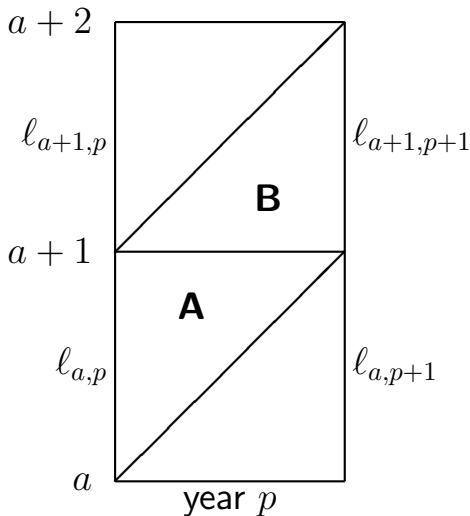
This corresponds to population sizes along the vertical lines indicated in the diagram. We want risk time figures for the population in the squares and triangles in the diagram.



Prevalent population figures

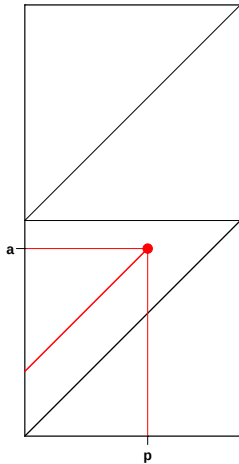
$\ell_{a,p}$ is the number of persons in age class a alive at the beginning of period (=year) p .

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



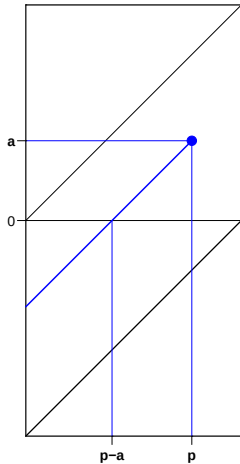
A person dying in age a at date p in **A** contributes p risk time, so the average will be:

$$\begin{aligned} & \int_{p=0}^{p=1} \int_{a=p}^{a=1} 2p \, da \, dp \\ &= \int_{p=0}^{p=1} 2p(1-p) \, dp \\ &= \left[p^2 - \frac{2}{3}p^3 \right]_{p=0}^{p=1} = \frac{1}{3} \end{aligned}$$



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

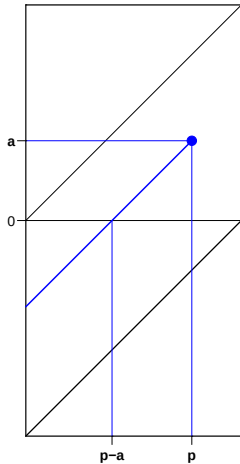
$$\begin{aligned} & \int_{p=0}^{p=1} \int_{a=0}^{a=p} 2(p - a) \, da \, dp \\ &= \int_{p=0}^{p=1} [2pa - a^2]_{a=0}^{a=p} \, dp \\ &= \int_{p=0}^{p=1} p^2 \, dp = \frac{1}{3} \end{aligned}$$



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

$$\int_{p=0}^{p=1} \int_{a=0}^{a=p} 2a \, da \, dp$$

$$= \int_{p=0}^{p=1} p^2 \, dp = \frac{1}{3}$$



Contributions to risk time in A and B:

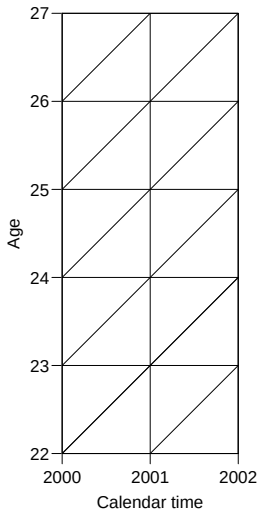
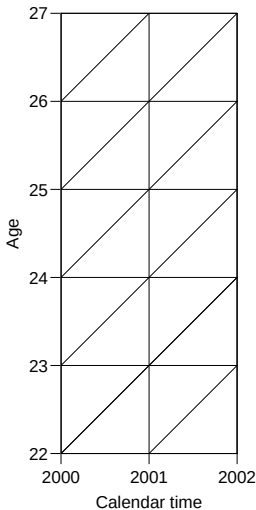
	A:	B:
Survivors:	$\ell_{a+1,p+1} \times \frac{1}{2}y$	$\ell_{a+1,p+1} \times \frac{1}{2}y$
Dead in A :	$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}y$	
Dead in B :	$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}y$	$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}y$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

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.



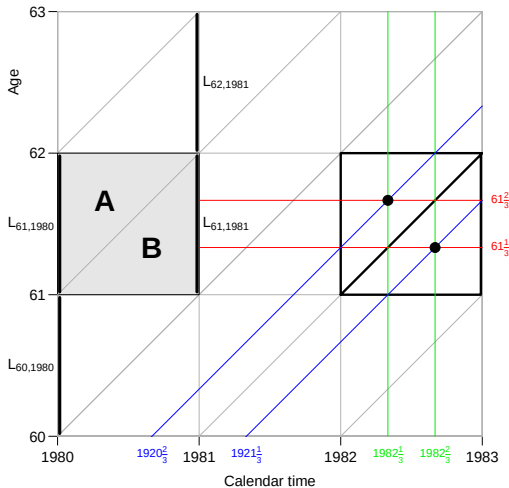
Summary:

Population
risk time:

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

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

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



APC-model for triangular data

Wednesday 25th, morning

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

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:

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

The log-likelihood can be separated:

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

No common parameters between terms — we have two separate models:

One for upper triangles, one for lower.

Illustration by lung cancer data

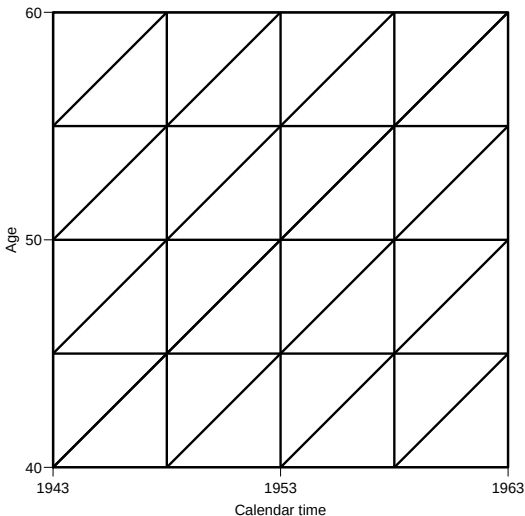
```
> library( Epi )
> data( lungDK )
> lungDK[1:10,]
```

	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.

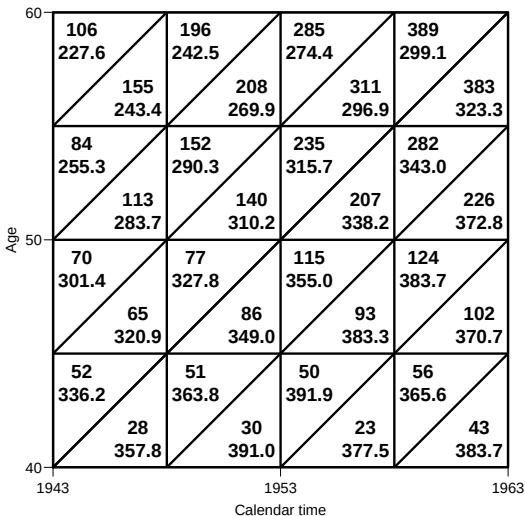
Mark mean date of birth for these.



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: $220 - 182 = 38$.

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

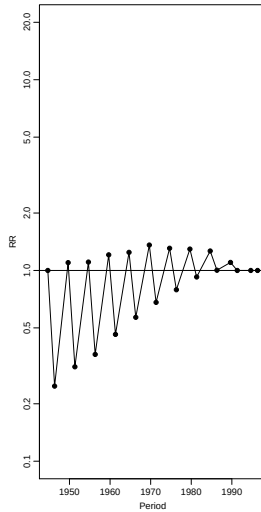
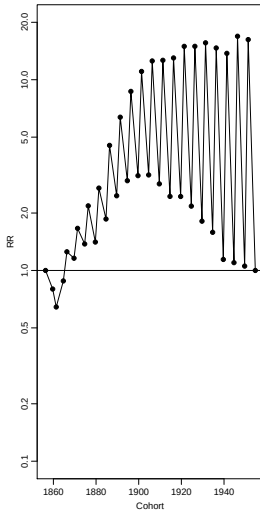
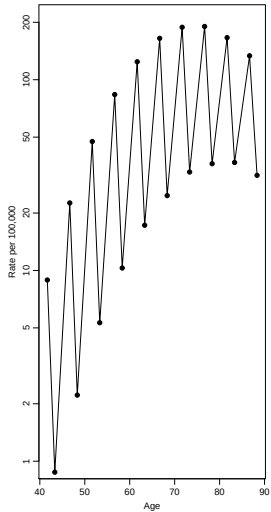
APC-model with “correct” cohorts

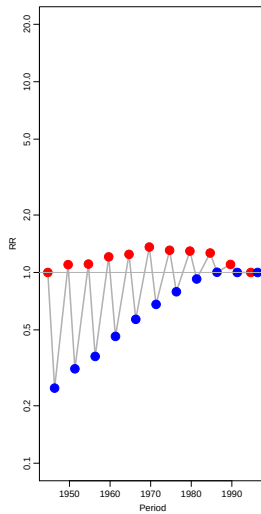
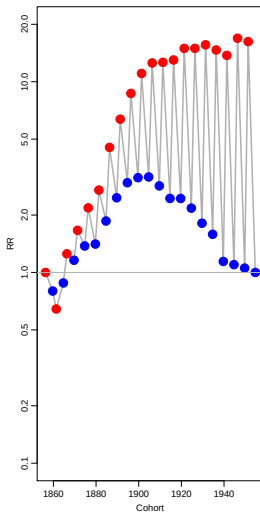
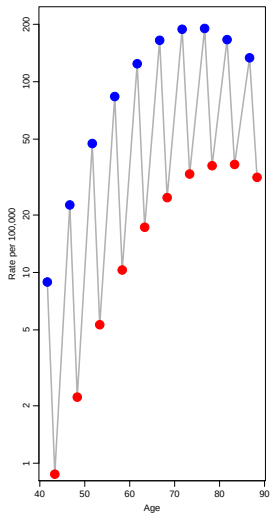
```
> mx <- glm( D ~ factor(Ax) - 1 +  
+           factor(Cx) +  
+           factor(Px) + offset( log( Y ) ),  
+           family=poisson )  
> summary( mx )  
...  
Null deviance: 1.0037e+08 on 220 degrees of freedom  
Residual deviance: 2.8473e+02 on 144 degrees of freedom
```

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

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

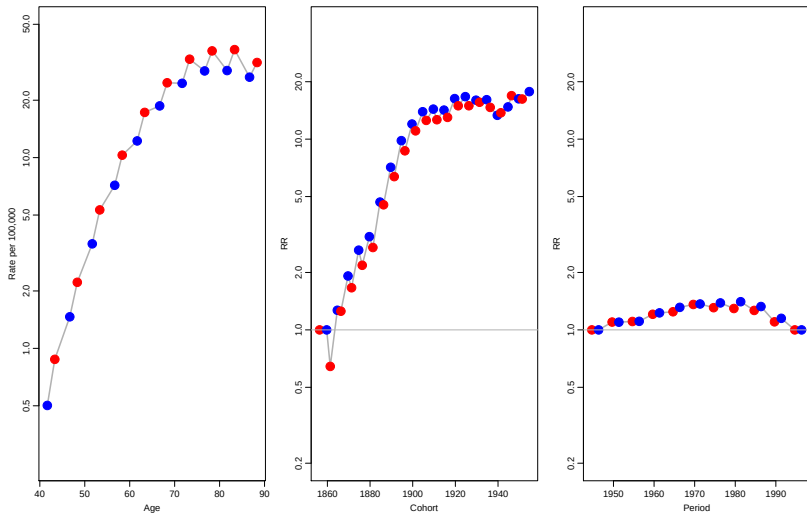
We have fitted two age-period-cohort models separately to upper and lower triangles.





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



APC-model: Parametrization

Wednesday 25th, afternoon

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

What's the problem?

One parameter is assigned to each distinct value of the timescales, the ordering 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

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

Possible choices for parametric functions describing the effect of the three continuous 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:

$$\check{f}(a) = f(a) - \mu_a - \gamma a$$

$$\check{g}(p) = g(p) + \mu_a + \mu_c + \gamma p$$

$$\check{h}(c) = h(c) - \mu_c - \gamma c$$

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 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).

Longitudinal or cohort age-effects.

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

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 .

Cross-sectional or period age-effects?

Bureaucratically interpretable — what's is 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:

$$\tilde{g}(p) = \hat{g}(p) - (\mu + \beta p)$$

3. Use the functions:

$$\begin{aligned}\tilde{f}(a) &= \hat{f}(a) + \mu + \beta a + \hat{h}(c_0) + \beta c_0 \\ \tilde{g}(p) &= \hat{g}(p) - \mu - \beta p \\ \tilde{h}(c) &= \hat{h}(c) + \beta c - \hat{h}(c_0) - \beta c_0\end{aligned}$$

These functions fulfill the criteria.

“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?

How do we get the standard errors?

Matrix-algebra! Projections!

Parametric function

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

Example: 2nd order polynomial:

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

$\text{nrow}(\mathbf{M})$ is the number of observations in the dataset.

Extract the trend from g :

$$\langle \tilde{g}(p) | 1 \rangle = 0, \quad \langle \tilde{g}(p) | p \rangle = 0$$

i.e. $\tilde{g}$ is *orthogonal* to $[1|p]$.

Suppose $\tilde{g}(p) = \tilde{\mathbf{M}}\pi$, then for any parameter vector π :

$$\langle \tilde{\mathbf{M}}\pi | 1 \rangle = 0, \quad \langle \tilde{\mathbf{M}}\pi | p \rangle = 0 \quad \implies \quad \tilde{\mathbf{M}} \perp [1|p]$$

Thus we just need to be able to produce $\tilde{\mathbf{M}}$ from $\mathbf{M}$: Projection on the orthogonal space of $\text{span}([1|p])$.

(Orthogonality requires an inner product!)

Practical parametrization

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 $[1|c]$.
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 in the data and inner product

Log-lik for an observation (D, Y) , log-rate θ :

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

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

Two relevant inner products:

$$\langle \mathbf{m}_j | \mathbf{m}_k \rangle = \sum_i m_{ij} m_{ik} \qquad \langle \mathbf{m}_j | \mathbf{m}_k \rangle = \sum_i m_{ij} w_i m_{ik}$$

the weights could be chosen as $w_i = D_i$, i.e.
proportional to the information content in the units
of the dataset.

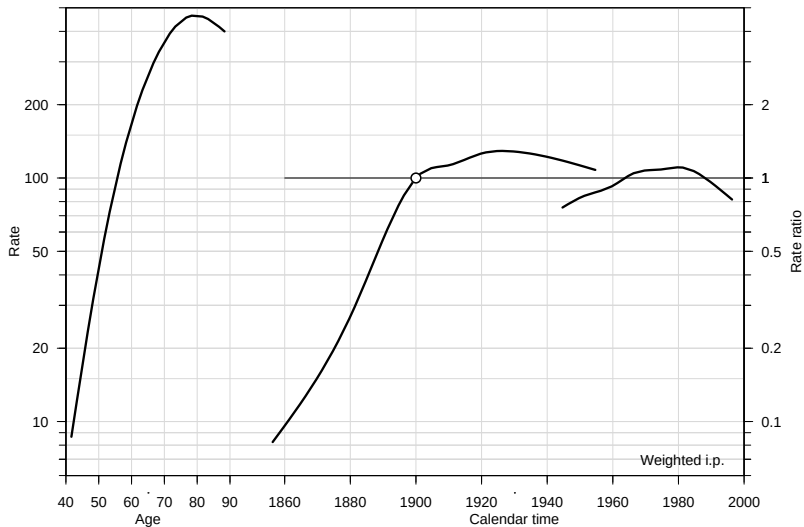
How to?

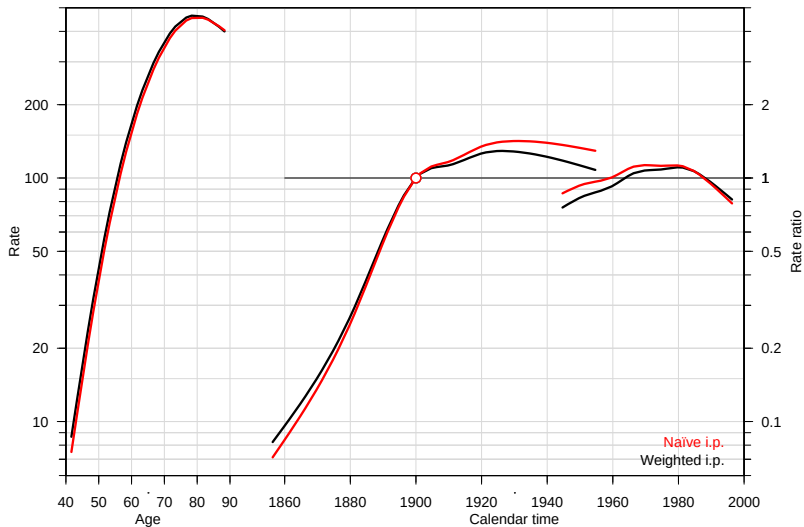
Implemented in `apc.fit`:

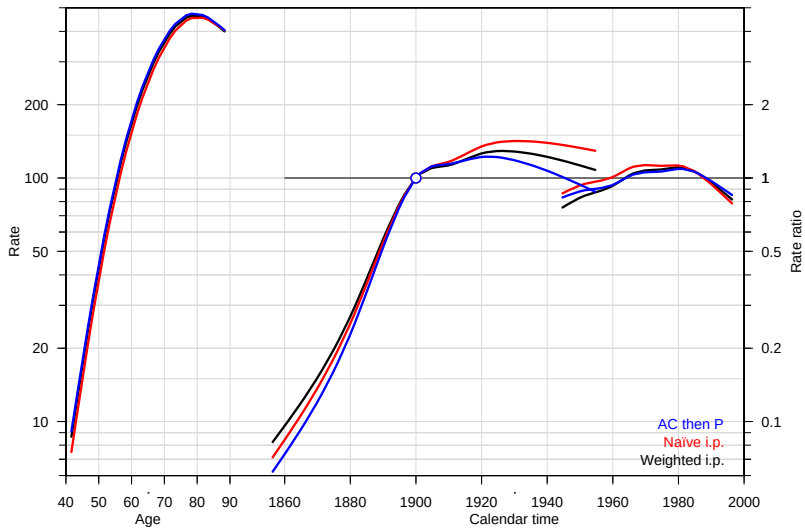
```
m1 <- apc.fit( A=lungDK$Ax,  
              P=lungDK$Px,  
              D=lungDK$D,  
              Y=lungDK$Y/10^5,  
              ref.c=1900 )  
apc.plot( m1 )
```

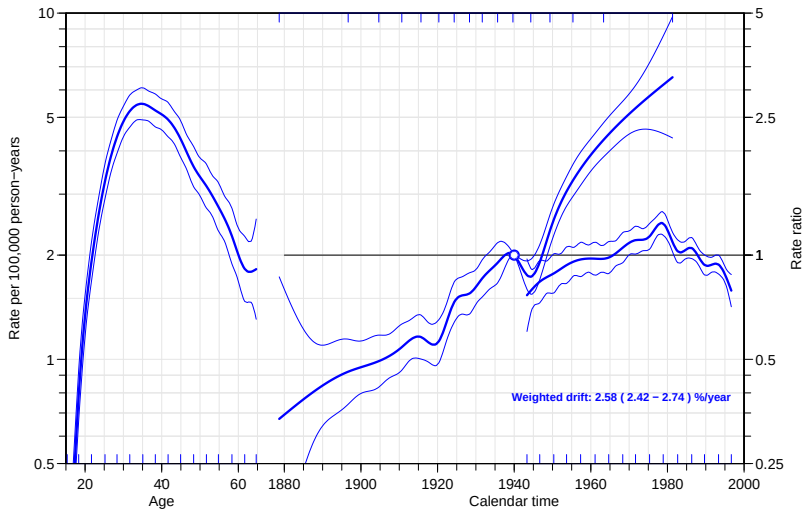
Consult the help page for:

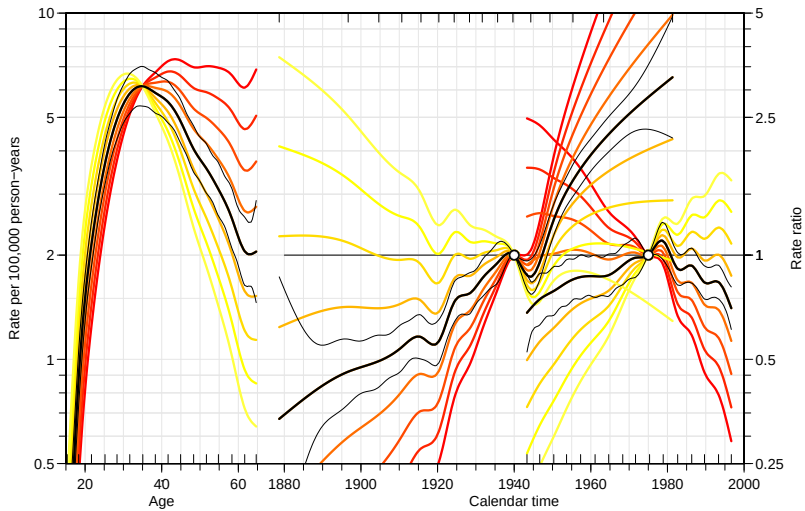
`apc.fit` to see options for weights in inner product,
type of function, variants of parametrization etc.
`apc.plot`, `apc.lines` and `apc.frame` to see how
to plot the results.











APC-models for several datasets

Wednesday 25th, afternoon

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

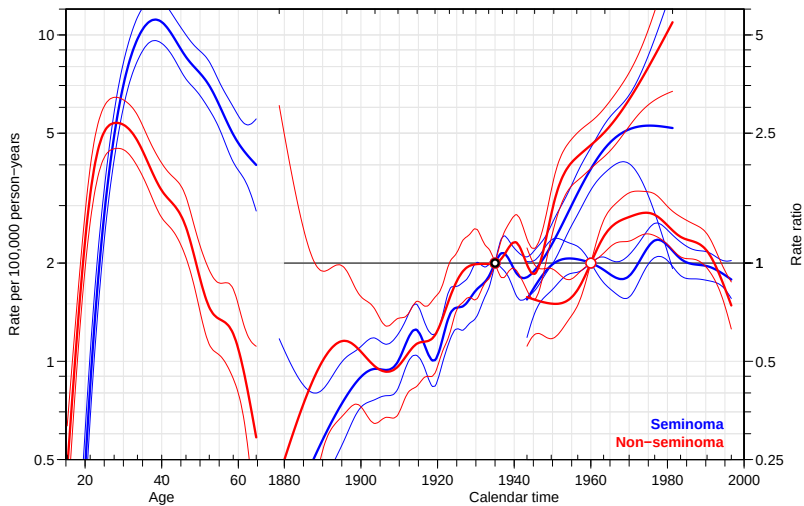
Two sets of data

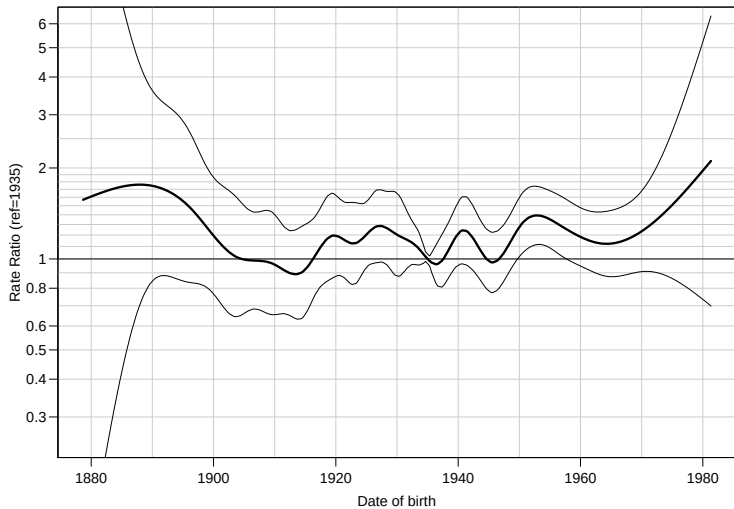
Example: Testis cancer in Denmark, Seminoma and non-Seminoma cases.

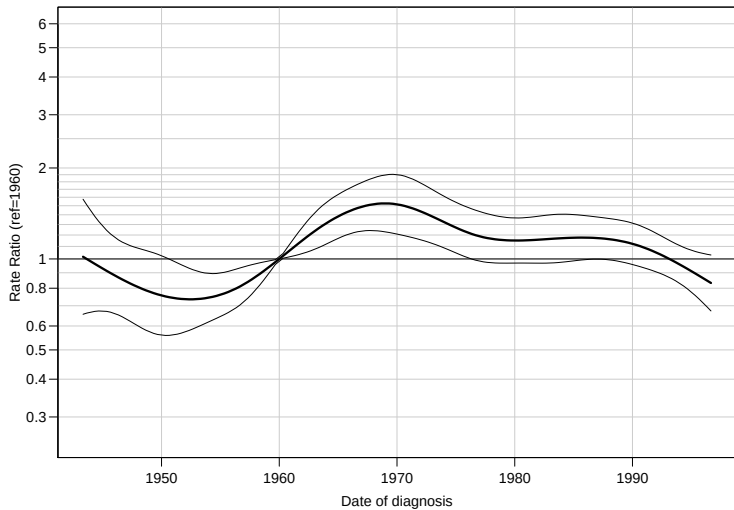
```
> stat.table( list( Histology=hist ),  
+             list( D=sum(d), Y=sum(y/10^6) ),  
+             margins = TRUE )
```

Histology	D	Y
1	4708.00	127.53
2	3632.00	127.53
3	466.00	127.53
Total	8806.00	382.58

First step is separate analyses for each subtype.







Analysis of two rates: Formal tests I

```
> Ma <- ns( A, df=15, intercept=TRUE )
> Mp <- ns( P, df=15 )
> Mc <- ns( P-A, df=20 )
> Mp <- detrend( Mp, P, weight=D )
> Mc <- detrend( Mc, P-A, weight=D )
>
> m.apc <- glm( D ~ -1 + Ma:type + Mp:type + Mc:type + offset( 1
> m.ap <- update( m.apc, . ~ . - Mc:type + Mc )
> m.ac <- update( m.apc, . ~ . - Mp:type + Mp )
> m.a <- update( m.ap , . ~ . - Mp:type + Mp )
>
> anova( m.a, m.ac, m.apc, m.ap, m.a, test="Chisq")
Analysis of Deviance Table

Model 1: D ~ Mc + Mp + Ma:type + offset(log(Y)) - 1
Model 2: D ~ Mp + Ma:type + type:Mc + offset(log(Y)) - 1
Model 3: D ~ -1 + Ma:type + Mp:type + Mc:type + offset(log(Y))
Model 4: D ~ Mc + Ma:type + type:Mp + offset(log(Y)) - 1
Model 5: D ~ Mc + Mp + Ma:type + offset(log(Y)) - 1
      Resid. Df Resid. Dev      Df Deviance P(>|Chi|)
```

Analysis of two rates: Formal tests II

1	10737	10553.7			
2	10718	10367.9	19	185.7	2.278e-29
3	10704	10199.6	14	168.3	1.513e-28
4	10723	10508.6	-19	-309.0	2.832e-54
5	10737	10553.7	-14	-45.0	4.042e-05

APC-model: Interactions

Friday 27th, morning

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

Analysis of DM-rates: Age $\times$ sex interaction

- ▶ 10 centres
 - ▶ 2 sexes
 - ▶ Age: 0-15
 - ▶ Period 1989–1999
-
- ▶ Is the sex-effect the same between all centres?
 - ▶ How are the timetrends.

Analysis of DM-rates: Age×sex interaction

II

```
library( Epi )
library( splines )
load( file="c:/Bendix/Artikler/A_P_C/IDDM/Eurodiab/data/tri.Rdat" )
dm <- dm[dm$cen=="D1: Denmark",]

# Define knots and points of prediction
n.A <- 5
n.C <- 8
n.P <- 5
pA <- seq(1/(3*n.A),1-1/(3*n.A),,n.A )
pC <- seq(1/(3*n.C),1-1/(3*n.C),,n.P )
pP <- seq(1/(3*n.P),1-1/(3*n.P),,n.C )
c0 <- 1985
attach( dm, warn.conflicts=FALSE )
A.kn <- quantile( rep( A, D ), probs=pA[-c(1,n.A)] )
A.ok <- quantile( rep( A, D ), probs=pA[ c(1,n.A)] )
A.pt <- sort( A[match( unique(A), A )] )
C.kn <- quantile( rep( C, D ), probs=pC[-c(1,n.C)] )
C.ok <- quantile( rep( C, D ), probs=pC[ c(1,n.C)] )
C.pt <- sort( C[match( unique(C), C )] )
```

Analysis of DM-rates: Age×sex interaction

III

```
P.kn <- quantile( rep( P, D ), probs=pP[-c(1,n.P)] )
P.ok <- quantile( rep( P, D ), probs=pP[ c(1,n.P)] )
P.pt <- sort( P[match( unique(P), P )] )

# Age-cohort model with age-sex interaction
# The model matrices for the ML fit
Ma <- ns( A, kn=A.kn, Bo=A.ok, intercept=T )
Mc <- cbind( C-c0, detrend( ns( C, kn=C.kn, Bo=C.ok ), C, weight
Mp <- detrend( ns( P, kn=P.kn, Bo=P.ok ), P, weight
# The prediction matrices
Pa <- Ma[match(A.pt,A),,drop=F]
Pc <- Mc[match(C.pt,C),,drop=F]
Pp <- Mp[match(P.pt,P),,drop=F]

# Fit the apc model by ML
apcs <- glm( D ~ Ma:sex - 1 + Mc + Mp +
             offset( log (Y/10^5) ),
             family=poisson,
             data=dm )

summary( apcs )
```

Analysis of DM-rates: Age×sex interaction

IV

```
ci.lin( apcs )
ci.lin( apcs, subset="sexF", Exp=T)
ci.lin( apcs, subset="sexF", ctr.mat=Pa, Exp=T)

# Extract the effects
F.inc <- ci.lin( apcs, subset="sexF", ctr.mat=Pa, Exp=T)[,5:7]
M.inc <- ci.lin( apcs, subset="sexM", ctr.mat=Pa, Exp=T)[,5:7]
MF.RR <- ci.lin( apcs, subset=c("sexM","sexF"), ctr.mat=cbind(Pa
  c.RR <- ci.lin( apcs, subset="Mc", ctr.mat=Pc, Exp=T)[,5:7]
  p.RR <- ci.lin( apcs, subset="Mp", ctr.mat=Pp, Exp=T)[,5:7]

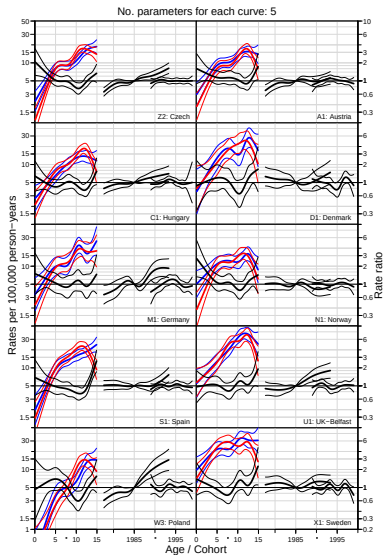
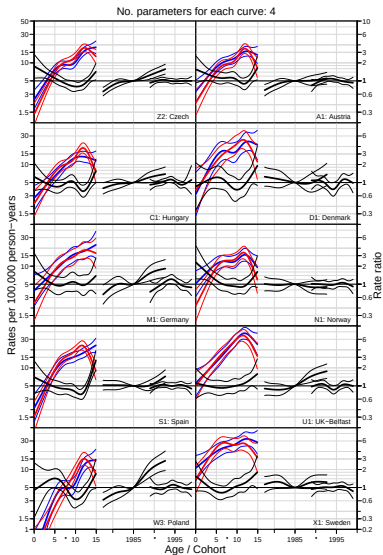
# plt( paste( "DM-DK" ), width=11 )
par( mar=c(4,4,1,4), mgp=c(3,1,0)/1.6, las=1 )
# The the frame for the effects
fr <- apc.frame( a.lab=c(0,5,10,15),
                 a.tic=c(0,5,10,15),
                 r.lab=c(c(1,1.5,3,5),c(1,1.5,3,5)*10),
                 r.tic=c(c(1,1.5,2,5),c(1,1.5,2,5)*10),
                 cp.lab=seq(1980,2000,10),
                 cp.tic=seq(1975,2000,5),
```

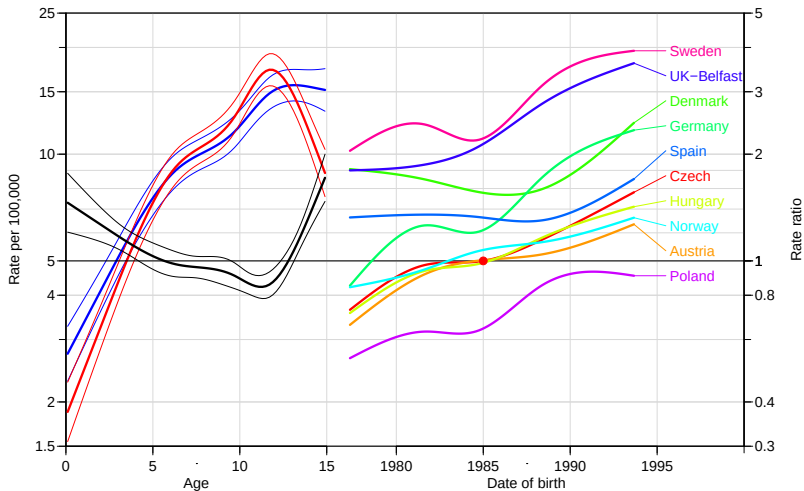
Analysis of DM-rates: Age×sex interaction

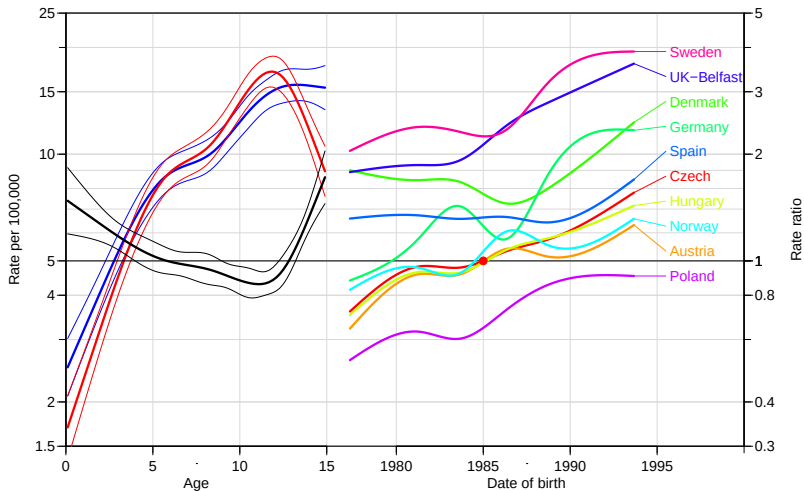
V

```
rr.ref=5,
gap=1,
col.grid=gray(0.9),
a.txt="",
cp.txt="",
r.txt="",
rr.txt="" )

# Draw the estimates
matlines( A.pt, M.inc, lwd=c(3,1,1), lty=1, col="blue" )
matlines( A.pt, F.inc, lwd=c(3,1,1), lty=1, col="red" )
matlines( C.pt - fr[1], c.RR * fr[2],
          lwd=c(3,1,1), lty=1, col="black" )
matlines( P.pt - fr[1], p.RR * fr[2],
          lwd=c(3,1,1), lty=1, col="black" )
matlines( A.pt, MF.RR * fr[2],
          lwd=c(3,1,1), lty=1, col=gray(0.6) )
abline(h=fr[2])
```





Predicting future rates

Friday 27th, morning

Bendix Carstensen

Age-Period-Cohort models

March 2009

Max Planck Institut for Demographic Research, Rostock

www.biostat.ku.dk/~bxc/APC/MPIDR-2009

Prediction of future rates

Model:

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

- ▶ Why not just extend the estimated functions into the future?
- ▶ The parametrization curse — the model as stated is not uniquely parametrized.
- ▶ Prediction 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 version:

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

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

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

If this is met, the predictions made will not depend on the parametrization chosen.

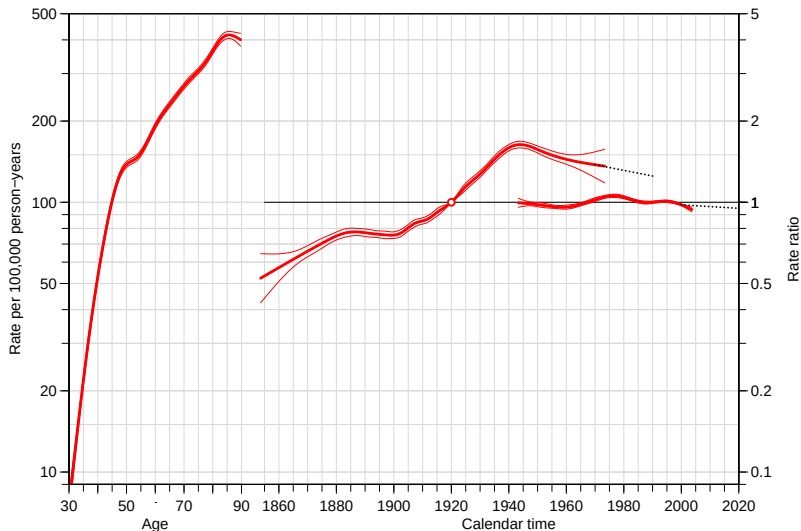
If one of the conditions does not hold, the prediction will depend on the parametrization chosen.

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.

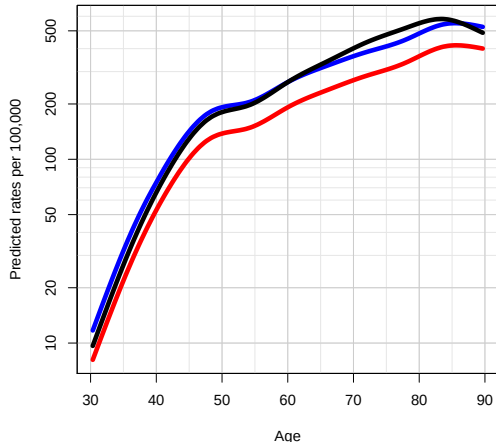
Example: Breast cancer in Denmark



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.

Bresat cancer prediction



Predicted age-specific breast cancer rates at 2020 (black),

in the 1950 cohort (blue),

and the estimated age-effects (red).