

Statistical Analysis in the Lexis Diagram:

Age-Period-Cohort models

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

Max Planck Institut for Demographic Research, Rostock

May 2016

<http://BendixCarstensen/APC/MPIDR-2016>

1/ 327

About the practicals

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

Introduction (intro)

5/ 327

Introduction

Statistical Analysis in the Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

intro

Rates and Survival

Statistical Analysis in the Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

surv-rate

Welcome

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

Introduction (intro)

2/ 327

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")

Rates and Survival (surv-rate)

6/ 327

Scope of the course

- ▶ Rates as observed in populations — disease registers for example.
- ▶ Understanding of survival analysis (statistical analysis of rates) — this is the content of much of the first day.
- ▶ Besides concepts, practical understanding of the actual computations (in **R**) are emphasized.
- ▶ There is a section in the practicals: "Basic concepts in analysis of rates and survival" — read it.

Introduction (intro)

3/ 327

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 from jail release to re-offending

Rates and Survival (surv-rate)

7/ 327

About the lectures

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

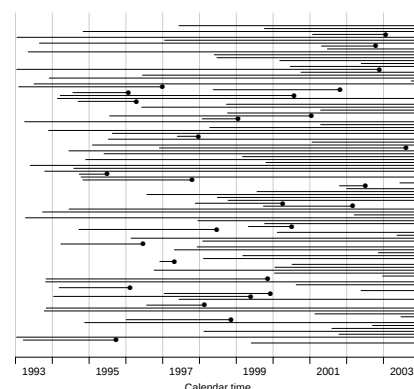
Introduction (intro)

4/ 327

Each line a person

Each blob a death

Study ended at 31 Dec. 2003

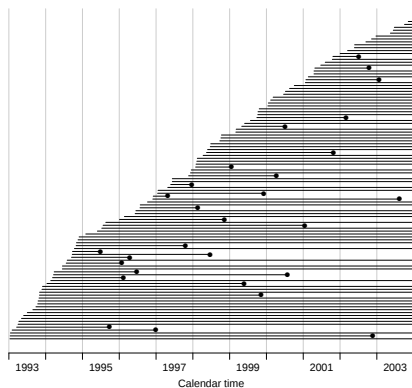


Rates and Survival (surv-rate)

8/ 327

Ordered by date of entry

Most likely the order in your database.



Rates and Survival (surv-rate)

9/ 327

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

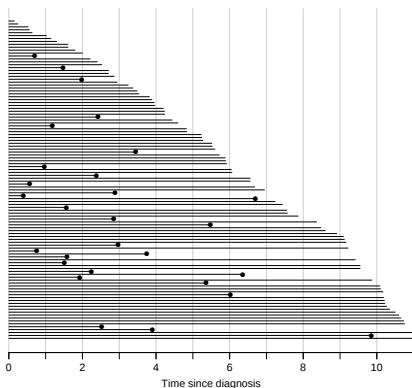
Estimated risk in year 1 for Stage I women is $5/107.5 = 0.0465$

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

Rates and Survival (surv-rate)

14/ 327

Timescale changed to "Time since diagnosis".



Rates and Survival (surv-rate)

10/ 327

Survival function

Persons enter at time 0:

Date of birth
Date of randomization
Date of diagnosis.

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

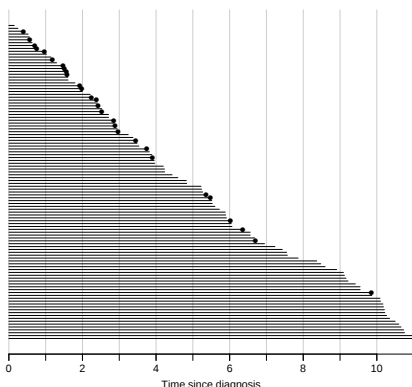
Distribution is characterized by the survival function:

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

Rates and Survival (surv-rate)

15/ 327

Patients ordered by survival time.



Rates and Survival (surv-rate)

11/ 327

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.

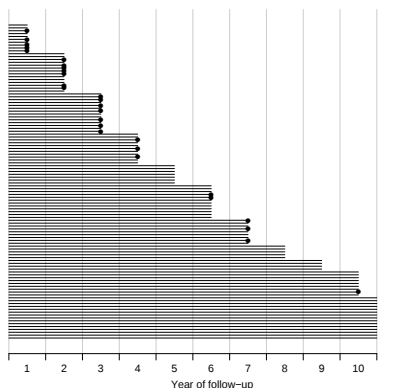
Characterizes the survival distribution as does f or F .

Theoretical counterpart of a **rate**.

Rates and Survival (surv-rate)

16/ 327

Survival times grouped into bands of survival.



Rates and Survival (surv-rate)

12/ 327

Relationships

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

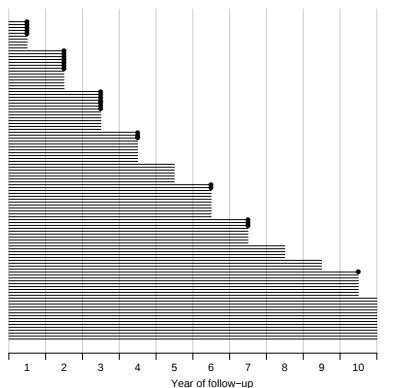
$\Lambda(t) = \int_0^t \lambda(s) ds$ is called the **integrated intensity** or **cumulative hazard**.

$\Lambda(t)$ is **not** an intensity — it is dimensionless.

Rates and Survival (surv-rate)

17/ 327

Patients ordered by survival status within each band.



Rates and Survival (surv-rate)

13/ 327

Rate and survival

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

- Survival is a **cumulative** measure
- A rate is an **instantaneous** measure.
- **Note:** A cumulative measure requires an origin!

Rates and Survival (surv-rate)

18/ 327

Observed survival and rate

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

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

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

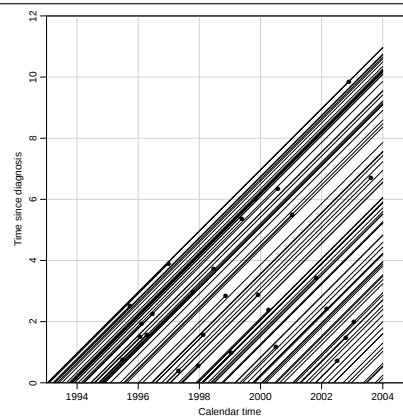
- Epidemiological studies:
Observation of (components of) a rate:

$$D, \quad Y, \quad D/Y$$

D : no. events, Y no of person-years.

Follow-up by
calendar time
and
time since
diagnosis:

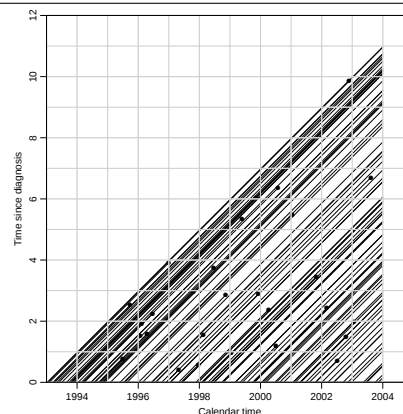
A Lexis
diagram!



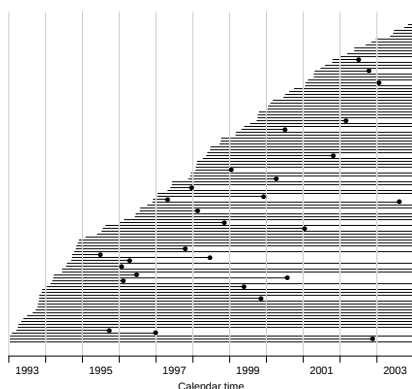
Empirical rates for individuals

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

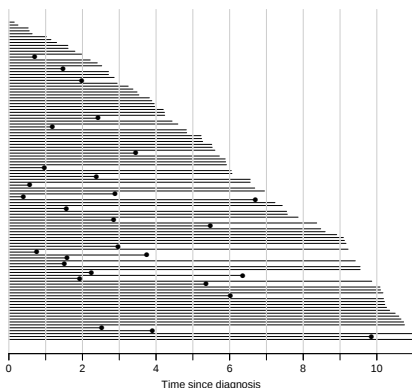
Empirical rates
by
calendar time
and
time since
diagnosis



Empirical rates by
calendar time.



Empirical rates by
time since diagnosis.



Likelihood for rates

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

likelihood

Likelihood contribution from one person

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

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

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

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

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

Two timescales

Note that we actually have two timescales:

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

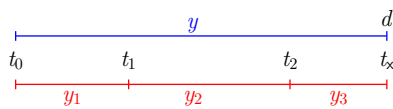
These can be shown simultaneously in a Lexis diagram.

Likelihood for an empirical rate

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

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

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



Probability

$P(d \text{ at } t_x | \text{entry } t_0)$

$$= P(\text{surv } t_0 \rightarrow t_1 | \text{entry } t_0) \\ \times P(\text{surv } t_1 \rightarrow t_2 | \text{entry } t_1) \\ \times P(d \text{ at } t_x | \text{entry } t_2)$$

log-Likelihood

$$d \log(\lambda) - \lambda y \\ = 0 \log(\lambda) - \lambda y_1 \\ + 0 \log(\lambda) - \lambda y_2 \\ + d \log(\lambda) - \lambda y_3$$

Likelihood for rates (likelihood)

28 / 327

Likelihood example

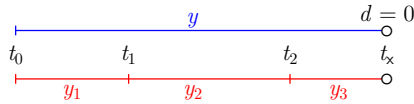
- Assuming the rate (intensity) is constant, λ ,
- the probability of observing 7 deaths in the course of 500 person-years:

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

- 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 for rates (likelihood)

33 / 327



Probability

$P(\text{surv } t_0 \rightarrow t_x | \text{entry } t_0)$

$$= P(\text{surv } t_0 \rightarrow t_1 | \text{entry } t_0) \\ \times P(\text{surv } t_1 \rightarrow t_2 | \text{entry } t_1) \\ \times P(\text{surv } t_2 \rightarrow t_x | \text{entry } t_2)$$

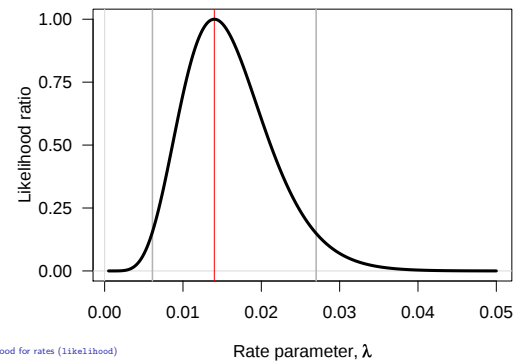
log-Likelihood

$$0 \log(\lambda) - \lambda y \\ = 0 \log(\lambda) - \lambda y_1 \\ + 0 \log(\lambda) - \lambda y_2 \\ + 0 \log(\lambda) - \lambda y_3$$

Likelihood for rates (likelihood)

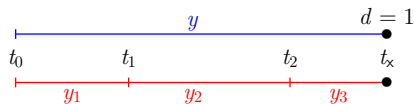
29 / 327

Likelihood-ratio function



Likelihood for rates (likelihood)

35 / 327



Probability

$P(\text{event at } t_x | \text{entry } t_0)$

$$= P(\text{surv } t_0 \rightarrow t_1 | \text{entry } t_0) \\ \times P(\text{surv } t_1 \rightarrow t_2 | \text{entry } t_1) \\ \times P(\text{event at } t_x | \text{entry } t_2)$$

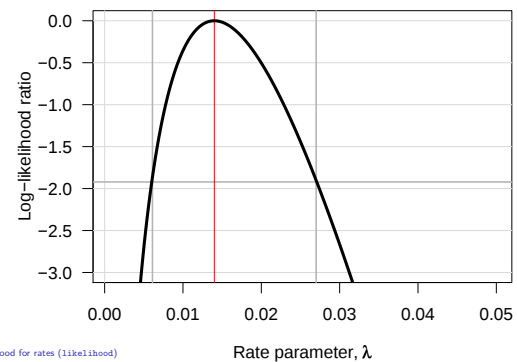
log-Likelihood

$$1 \log(\lambda) - \lambda y \\ = 0 \log(\lambda) - \lambda y_1 \\ + 0 \log(\lambda) - \lambda y_2 \\ + 1 \log(\lambda) - \lambda y_3$$

Likelihood for rates (likelihood)

30 / 327

Log-likelihood ratio



Likelihood for rates (likelihood)

36 / 327

Aim of dividing time into bands:

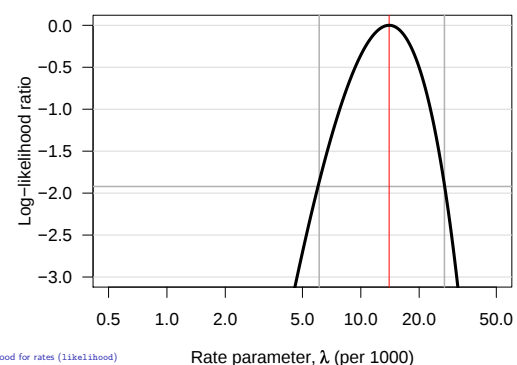
- Compute rates in different bands of:
 - age
 - calendar time
 - disease duration
 - ...
- Allow rates to vary along the timescale:

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

Likelihood for rates (likelihood)

31 / 327

Log-likelihood ratio



Likelihood for rates (likelihood)

38 / 327

Log-likelihood from more persons

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

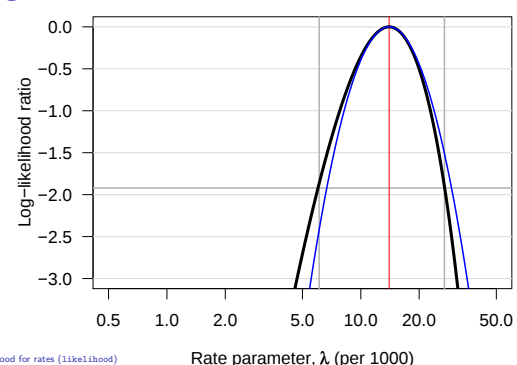
$$\begin{aligned} \sum_t \sum_p (d_{pt} \log(\lambda_t) - \lambda_t y_{pt}) &= \sum_t ((\sum_p d_{pt}) \log(\lambda_t) - \lambda_t (\sum_p y_{pt})) \\ &= \sum_t (D_t \log(\lambda_t) - \lambda_t Y_t) \end{aligned}$$

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

Likelihood for rates (likelihood)

32 / 327

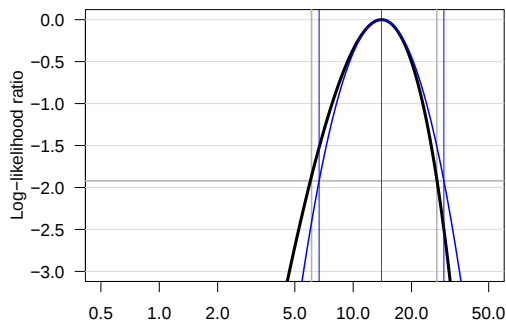
Log-likelihood ratio



Likelihood for rates (likelihood)

39 / 327

Log-likelihood ratio



Likelihood for rates (likelihood)

Rate parameter, λ (per 1000)

40/ 327

The log-likelihood is maximal for:

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

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

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

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

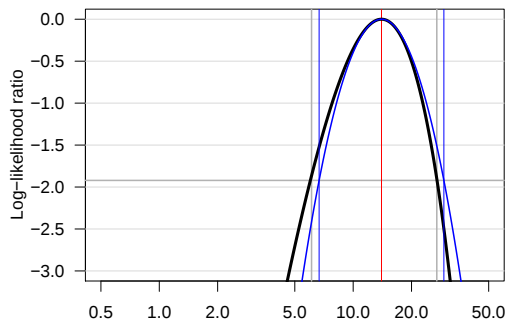
Standard error of log-rate: $1/\sqrt{D}$.

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

Likelihood for rates (likelihood)

45/ 327

Log-likelihood ratio



Likelihood for rates (likelihood)

Rate parameter, λ (per 1000)

41/ 327

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

$$\hat{\lambda} \pm \frac{1.96}{\sqrt{7}} = (6.7, 29.4)$$

The log-likelihood is maximal for:

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

Information about the rate itself, λ :

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

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

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

Likelihood for rates (likelihood)

46/ 327

Poisson likelihood

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

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

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

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

Extra term does not depend on the rate parameter λ .

Likelihood for rates (likelihood)

42/ 327

Poisson likelihood

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

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

- Terms are **not** independent,
 - but the log-likelihood is a **sum** of Poisson-like terms,
 - the **same** as a likelihood for **independent** Poisson variates, d_{pt}
 - with mean $\mu = \lambda_t y_{py} \Leftrightarrow \log \mu = \log(\lambda_t) + \log(y_{py})$
- ⇒ Analyse rates λ based on empirical rates (d, y) Poisson model with log-link applied to where:
- d is the response variable.
 - $\log(y)$ is the offset variable.

Likelihood for rates (likelihood)

43/ 327

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:

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

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

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

Likelihood for rates (likelihood)

47/ 327

Exercise

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

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

Likelihood for rates (likelihood)

48/ 327

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.

Likelihood for rates (likelihood)

44/ 327

Rates with glm

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

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

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

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

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

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

Likelihood for rates (likelihood)

50/ 327

Ratio of two rates

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

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

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

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

Likelihood for rates (likelihood)

51/ 327

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.

Likelihood for rates (likelihood)

52/ 327

Lifetables

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
May 2016

Max Planck Institut for Demographic Research, Rostock
<http://BendixCarstensen/APC/MPIDR-2016>

Lifetables

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	$8/(59 - 1/2) = 0.137$

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

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

Lifetables (lifetable)

54/ 327

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.

Lifetables (lifetable)

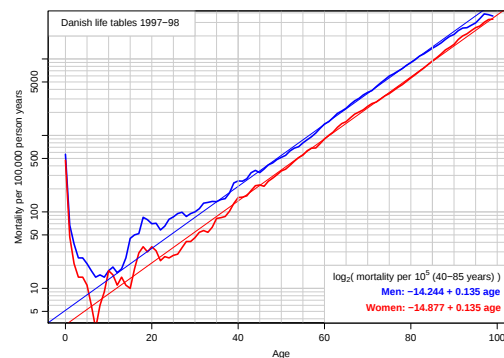
55/ 327

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

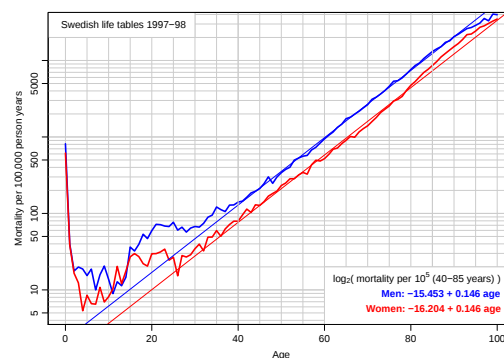
Lifetables (lifetable)

56/ 327



Lifetables (lifetable)

57/ 327



Lifetables (lifetable)

58/ 327

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?

Lifetables (lifetable)

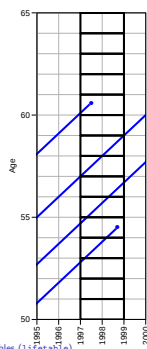
59/ 327

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	$(-15.453 + 16.204)/0.146 = 5.14 \text{ years}$	

Lifetables (lifetable)

60/ 327

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.

Lifetables (lifetable)

61 / 327

Who needs the Cox-model anyway?

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

WntOsa

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.

Lifetables (lifetable)

62 / 327

A look at the Cox model

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

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

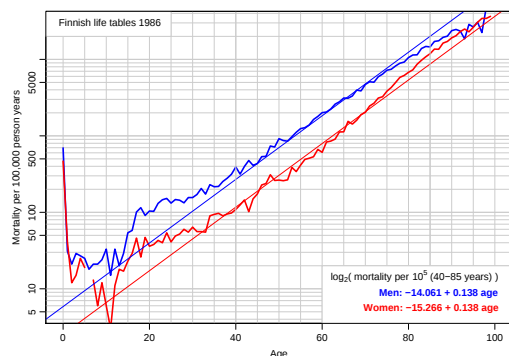
The covariate t has a special status:

- ▶ Computationally, because all individuals contribute to (some of) the range of t .
- ▶ ... the scale along which time is split (the risk sets)
- ▶ Conceptually it is less clear — t is but a covariate that varies within individual.
- ▶ Cox's approach profiles $\lambda_0(t)$ out.

Who needs the Cox-model anyway? (WntOsa)

64 / 327

Rates vary over time:



Lifetables (lifetable)

63 / 327

Cox-likelihood

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

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

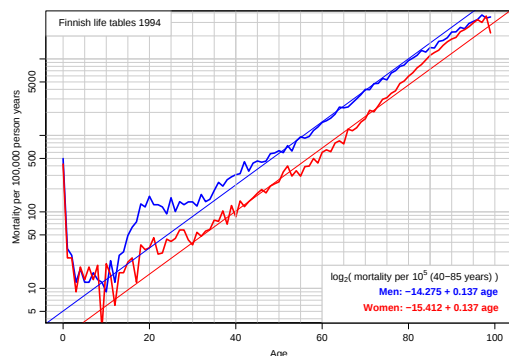
is also a **profile likelihood** in the model where observation time has been subdivided in small pieces (empirical rates) and each small piece provided with its own parameter:

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

Who needs the Cox-model anyway? (WntOsa)

65 / 327

Rates vary over time:



Lifetables (lifetable)

63 / 327

The Cox-likelihood as profile likelihood

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

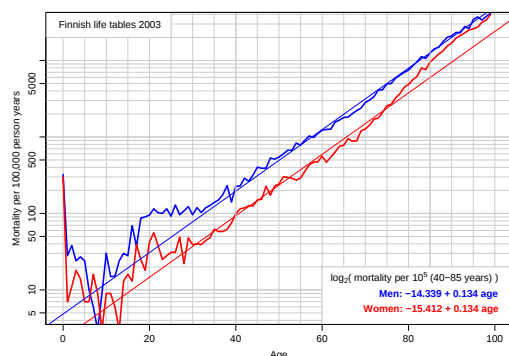
$$\log(\lambda(t, x_i)) = \log(\lambda_0(t)) + \beta_1 x_{1i} + \dots + \beta_p x_{pi} = \alpha_t + \eta_i$$

- ▶ Profile likelihood:
 - ▶ Derive estimates of α_t as function of data and β s — assuming constant rate between death times
 - ▶ Insert in likelihood, now only a function of data and β s
 - ▶ Turns out to be Cox's partial likelihood

Who needs the Cox-model anyway? (WntOsa)

66 / 327

Rates vary over time:



Lifetables (lifetable)

63 / 327

- ▶ 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 (1, 1) with the death at time t .
 - ▶ all other empirical rates (0, 1) from those who were at risk at time t .

Who needs the Cox-model anyway? (WntOsa)

67 / 327

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.

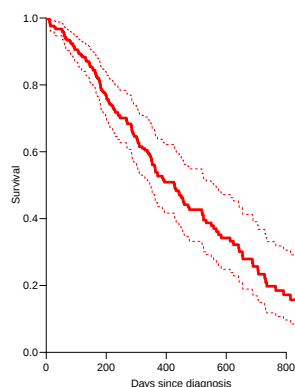
Splitting the dataset

- ▶ The Poisson approach needs a dataset of empirical rates (d, y) with suitably small values of y .
- ▶ — much larger than the original dataset
- ▶ — each individual contributes many empirical rates
- ▶ (one per risk-set contribution in Cox-modelling)
- ▶ 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
- ▶ Modelling is by standard glm Poisson

Example: Mayo Clinic lung cancer

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

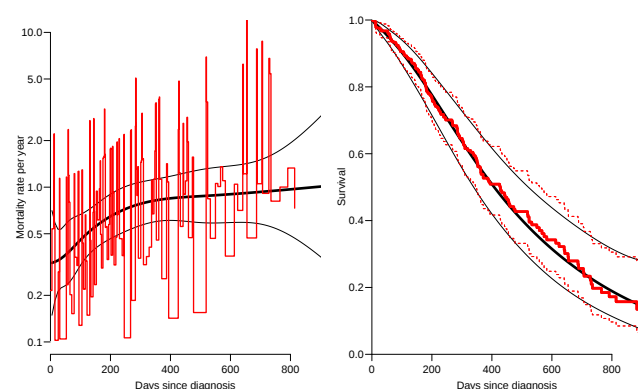
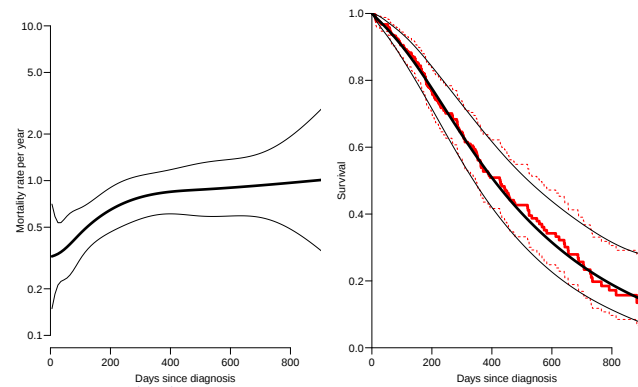
Mayo Clinic lung cancer 60 year old woman



Example: Mayo Clinic lung cancer I

```
> round( cmp, 5 )
```

	age	2.5%	97.5%	sex	2.5%	97.5%
Cox	1.01716	0.99894	1.03571	0.59896	0.43137	0.83165
Poisson-factor	1.01716	0.99894	1.03571	0.59896	0.43137	0.83165
Poisson-spline	1.01619	0.99803	1.03468	0.59983	0.43199	0.83287



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

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

What the Cox-model really is

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

- ▶ dividing time very finely and
- ▶ modeling one covariate, the time-scale, with one parameter per distinct value.
- ▶ ⇒ difficult to access the baseline hazard.
- ▶ ⇒ uninitiated tempted to show survival curves where irrelevant

Modeling in this world

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

Who needs the Cox-model anyway? (WatQaa)

77 / 327

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.

Who needs the Cox-model anyway? (WatQaa)

78 / 327

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 $\text{diag}[\exp(\hat{f}(t_i))]$

Who needs the Cox-model anyway? (WatQaa)

79 / 327

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):

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

Who needs the Cox-model anyway? (WatQaa)

80 / 327

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`.

Practical: Cox and Poisson modelling

Who needs the Cox-model anyway? (WatQaa)

81 / 327

(non)-Linear models: Estimates and predictions

Statistical Analysis in the Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

lin-mod

Linear models

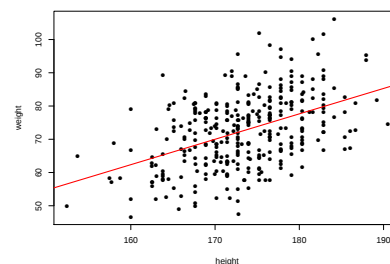
```
> library( Epi )
> data( diet )
> names( diet )

[1] "id"      "doe"      "dox"      "dob"      "y"      "fail"
[8] "month"   "energy"    "height"    "weight"    "fat"    "fibre"
[15] "chd"

> with( diet, plot( weight ~ height, pch=16 ) )
> abline( lm( weight ~ height, data=diet ), col="red", lwd=2 )
```

(non)-Linear models: Estimates and predictions (lin-mod)

82 / 327



```
> with( diet, plot( weight ~ height, pch=16 ) )
> abline( lm( weight ~ height, data=diet ), col="red", lwd=2 )
```

(non)-Linear models: Estimates and predictions (lin-mod)

83 / 327

Linear models, extracting estimates

```
> ml <- lm( weight ~ height, data=diet )
> summary( ml )

Call:
lm(formula = weight ~ height, data = diet)

Residuals:
    Min       1Q   Median       3Q      Max
-24.7361  -7.4553   0.1608   6.9384  27.8130

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -59.91601    14.31557  -4.185 3.66e-05
height       0.76421     0.08252   9.261 < 2e-16

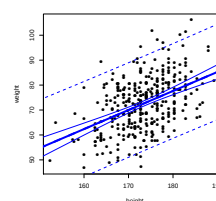
Residual standard error: 9.625 on 330 degrees of freedom
(5 observations deleted due to missingness)
Multiple R-squared:  0.2063, Adjusted R-squared:  0.2039
F-statistic: 85.76 on 1 and 330 DF, p-value: < 2.2e-16
```

```
> round( ci.lin( ml ), 4 )
```

(non)-Linear models: Estimates and predictions (lin-mod)

84 / 327

Linear models, prediction

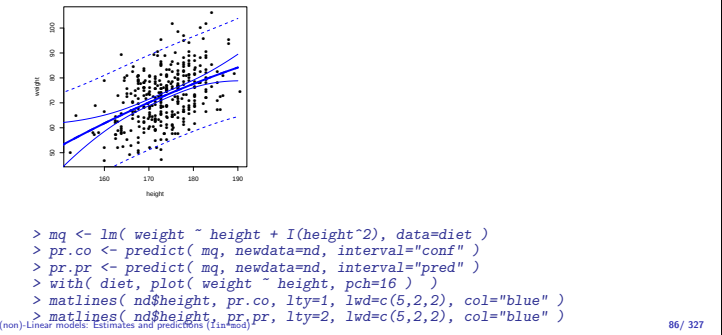


```
> ml <- lm( weight ~ height, data=diet )
> nd <- data.frame( height = 150:190 )
> pr.co <- predict( ml, newdata=nd, interval="conf" )
> pr.pr <- predict( ml, newdata=nd, interval="pred" )
> with( diet, plot( weight ~ height, pch=16 ) )
> matlines( nd$height, pr.co, lty=1, lwd=c(5,2,2), col="blue" )
> matlines( nd$height, pr.pr, lty=2, lwd=c(5,2,2), col="blue" )
```

(non)-Linear models: Estimates and predictions (lin-mod)

85 / 327

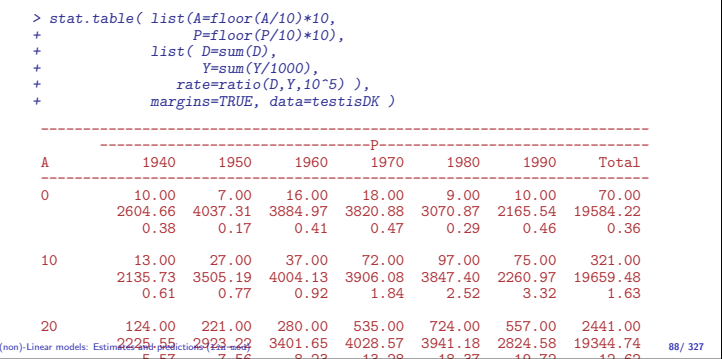
non-Linear models, prediction



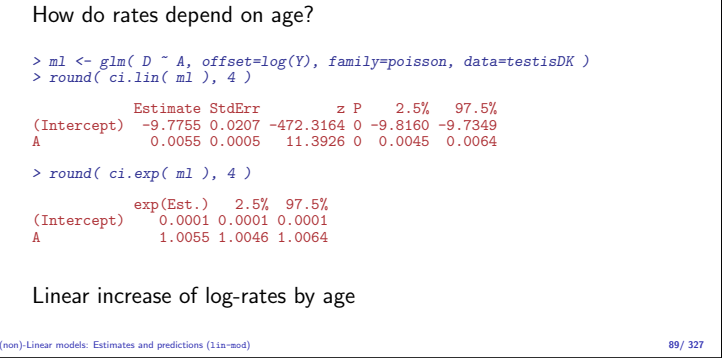
Testis cancer



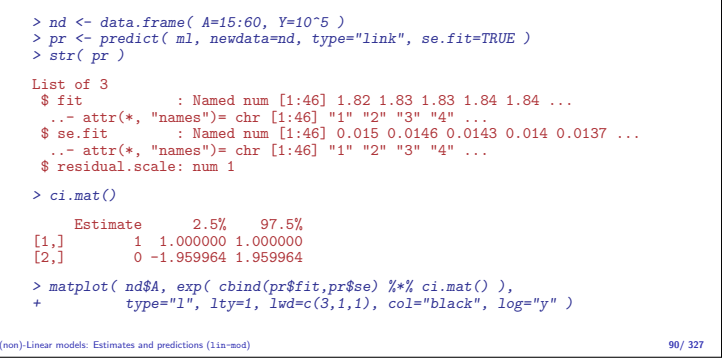
Cases, PY and rates



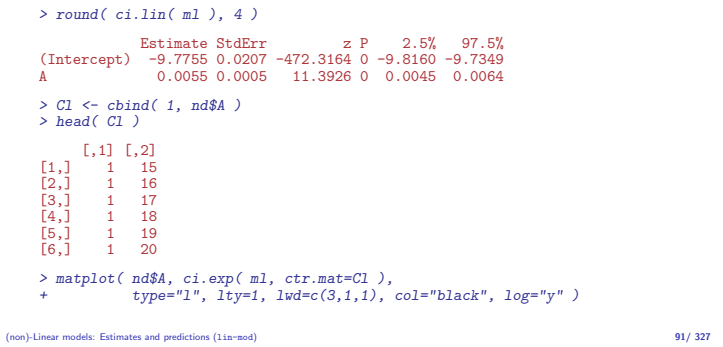
Linear effects in glm



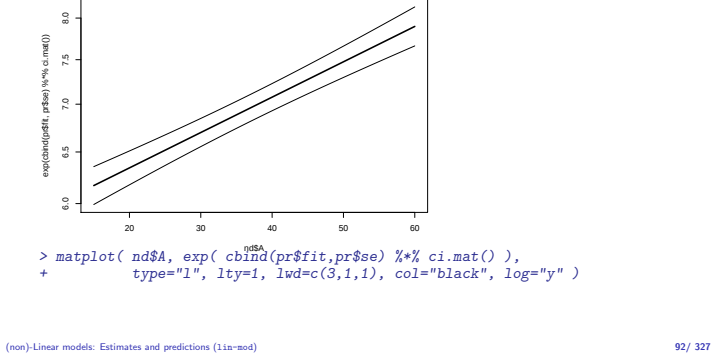
Linear effects in glm



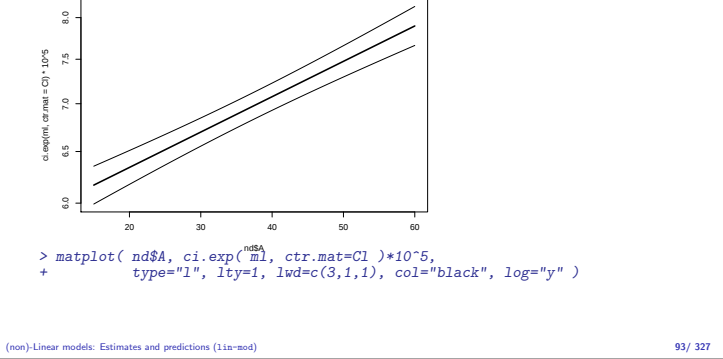
Linear effects in glm



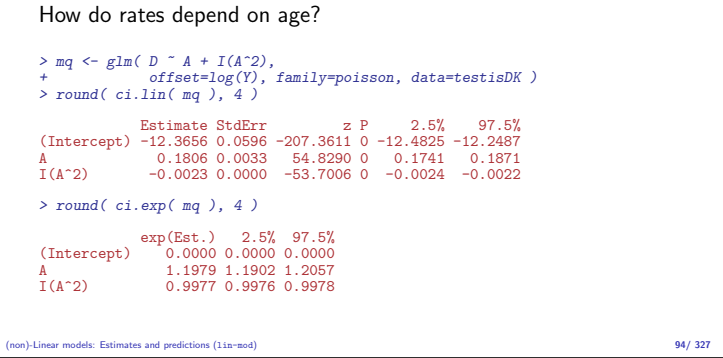
Linear effects in glm



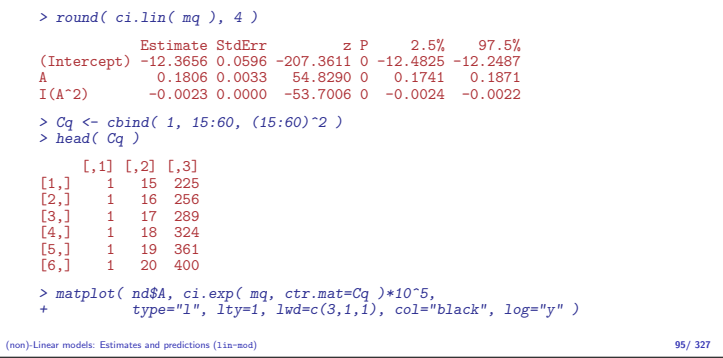
Linear effects in glm



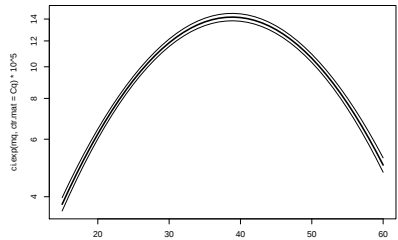
Quadratic effects in glm



Quadratic effect in glm

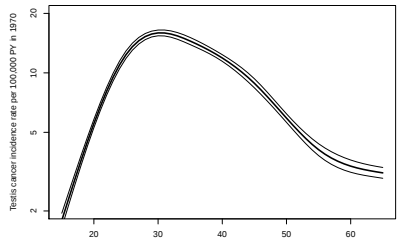


Quadratic effect in glm



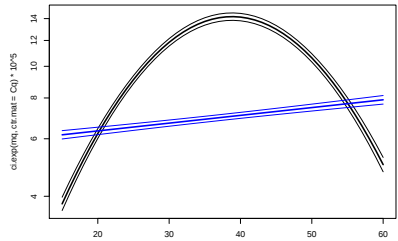
```
> matplot( nd$A, ci.exp( msp, ctr.mat=Cq ) * 10^5,
+         type="l", lty=1, lwd=c(3,1,1), col="black", log="y" )
```

Adding a linear period effect



```
> matplot( aa, ci.exp( msp, ctr.mat=Ca ) * 10^5,
+         log="y", xlab="Age",
+         ylab="Testis cancer incidence rate per 100,000 PY in 1970",
+         type="l", lty=1, lwd=c(3,1,1), col="black", ylim=c(2,20) )
```

Quadratic effect in glm



```
> matplot( nd$A, ci.exp( msp, ctr.mat=Cq ) * 10^5,
+         type="l", lty=1, lwd=c(3,1,1), col="black", log="y" )
+ matlines( nd$A, ci.exp( ml, ctr.mat=Cl ) * 10^5,
+         type="l", lty=1, lwd=c(3,1,1), col="blue" )
```

The period effect

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

	Estimate	StdErr	z	P	2.5%	97.5%
(Intercept)	-58.105	1.444	-40.229	0.000	-60.935	-55.274
Ns(A, knots = seq(15, 65, 10))1	2.120	0.057	37.444	0.000	2.009	2.231
Ns(A, knots = seq(15, 65, 10))2	1.700	0.068	25.157	0.000	1.567	1.832
Ns(A, knots = seq(15, 65, 10))3	0.007	0.060	0.110	0.913	-0.112	0.125
Ns(A, knots = seq(15, 65, 10))4	2.596	0.097	26.631	0.000	2.405	2.787
Ns(A, knots = seq(15, 65, 10))5	-0.780	0.042	-18.748	0.000	-0.861	-0.698
P	0.024	0.001	32.761	0.000	0.023	0.025

```
> pp <- 1945:1995
> Cp <- cbind( pp ) - 1970
> head( Cp )
```

	pp
[1,]	-25
[2,]	-24
[3,]	-23
[4,]	-22
[5,]	-21
[6,]	-20

Spline effects in glm

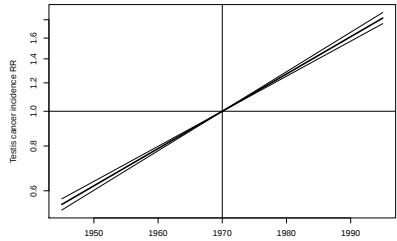
```
> library( splines )
> aa <- 15:65
> ms <- glm( D ~ Ns(A,knots=seq(15,65,10)),
+         offset=log(Y), family=poisson, data=testisDK )
> round( ci.exp( ms ), 3 )
```

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

```
> As <- Ns( aa, knots=seq(15,65,10) )
> head( As )
```

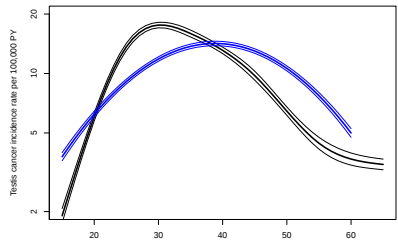
	1	2	3	4	5
[1,]	0.0000000000	0.00000000	0.00000000	0.00000000	0.00000000
[2,]	0.0001666667	0	-0.02527011	0.07581034	-0.05054022
[3,]	0.0013333333	0	-0.05003313	0.15009940	-0.10006626
[4,]	0.0045000000	0	-0.07378197	0.22134590	-0.14756393
[5,]	0.0100000000	0	-0.09600952	0.28802857	-0.19201905

Period effect



```
> matplot( pp, ci.exp( msp, subset="P", ctr.mat=Cp ),
+         log="y", xlab="Date", ylab="Testis cancer incidence RR",
+         type="l", lty=1, lwd=c(3,1,1), col="black" )
> abline( h=1, v=1970 )
```

Spline effects in glm



```
> matplot( aa, ci.exp( msp, ctr.mat=cbind(1,As) ) * 10^5,
+         log="y", xlab="Age", ylab="Testis cancer incidence rate per 100,000 PY",
+         type="l", lty=1, lwd=c(3,1,1), col="black", ylim=c(2,20) )
+ matlines( nd$A, ci.exp( mq, ctr.mat=Cq ) * 10^5,
+         type="l", lty=1, lwd=c(3,1,1), col="blue" )
```

A quadratic period effect

```
> mspq <- glm( D ~ Ns(A,knots=seq(15,65,10)) + P + I(P^2),
+         offset=log(Y), family=poisson, data=testisDK )
> round( ci.exp( mspq ), 3 )
```

	exp(Est.)	2.5%	97.5%
(Intercept)	0.000	0.000	0.000
Ns(A, knots = seq(15, 65, 10))1	8.356	7.478	9.337
Ns(A, knots = seq(15, 65, 10))2	5.513	4.829	6.295
Ns(A, knots = seq(15, 65, 10))3	1.006	0.894	1.133
Ns(A, knots = seq(15, 65, 10))4	13.439	11.101	16.269
Ns(A, knots = seq(15, 65, 10))5	0.458	0.422	0.497
P	2.189	1.457	3.291
I(P^2)	1.000	1.000	1.000

```
> pp <- 1945:1995
> Cp <- cbind( pp-1970, pp^2-1970^2 )
> head( Cp )
```

	[,1]	[,2]
[1,]	-25	-97875
[2,]	-24	-93984
[3,]	-23	-90091

Adding a linear period effect

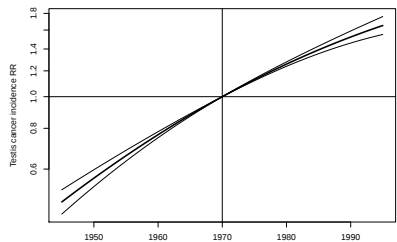
```
> msp <- glm( D ~ Ns(A,knots=seq(15,65,10)) + P,
+         offset=log(Y), family=poisson, data=testisDK )
> round( ci.lin( msp ), 3 )
```

	Estimate	StdErr	z	P	2.5%	97.5%
(Intercept)	-58.105	1.444	-40.229	0.000	-60.935	-55.274
Ns(A, knots = seq(15, 65, 10))1	2.120	0.057	37.444	0.000	2.009	2.231
Ns(A, knots = seq(15, 65, 10))2	1.700	0.068	25.157	0.000	1.567	1.832
Ns(A, knots = seq(15, 65, 10))3	0.007	0.060	0.110	0.913	-0.112	0.125
Ns(A, knots = seq(15, 65, 10))4	2.596	0.097	26.631	0.000	2.405	2.787
Ns(A, knots = seq(15, 65, 10))5	-0.780	0.042	-18.748	0.000	-0.861	-0.698
P	0.024	0.001	32.761	0.000	0.023	0.025

```
> Ca <- cbind( 1, Ns( aa, knots=seq(15,65,10) ), 1970 )
> head( Ca )
```

	1	2	3	4	5
[1,]	1	0.0000000000	0	0.00000000	0.00000000
[2,]	1	0.0001666667	0	-0.02527011	0.07581034
[3,]	1	0.0013333333	0	-0.05003313	0.15009940
[4,]	1	0.0045000000	0	-0.07378197	0.22134590
[5,]	1	0.0100000000	0	-0.09600952	0.28802857

A quadratic period effect



```
> matplot( pp, ci.exp( mspq, subset="P", ctr.mat=Cq ),
+         log="y", xlab="Date", ylab="Testis cancer incidence RR",
+         type="l", lty=1, lwd=c(3,1,1), col="black" )
> abline( h=1, v=1970 )
```

Age and period effect

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

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

(non-)Linear models: Estimates and predictions (11a-mod)

106/ 327

Age and period effect with ci.exp

- In rate models there is always one term with the **rate** dimension. Usually **age**
- But it must refer to a specific **reference** value for all **other** variables (P).
- All** parameters must be used in computing rates, at reference value.
- For the “other” variables, report the RR **relative** to the reference point.
- Only parameters relevant for the variable (P) used.
- Contrast matrix is a **difference** between prediction points and the reference point.

(non-)Linear models: Estimates and predictions (11a-mod)

111/ 327

A spline period effect

```
> pp <- 1945:1995
> Cs <- Ns( pp ,knots=seq(1950,1990,10))
> Cr <- Ns(rep(1970,length(pp)),knots=seq(1950,1990,10))
> head( Cs )
```

	1	2	3	4
[1,]	0.0	0.12677314	-0.38031941	0.25354628
[2,]	0.0	0.10141851	-0.30425553	0.20283702
[3,]	0.0	0.07606388	-0.22819165	0.15212777
[4,]	0.0	0.05070926	-0.15212777	0.10141851
[5,]	0.0	0.02535463	-0.07606388	0.05070926
[6,]	0.0	0.00000000	0.00000000	0.00000000

```
> head( Cr )
```

	1	2	3	4
[1,]	0.6666667	0.1125042	0.1624874	-0.1083249
[2,]	0.6666667	0.1125042	0.1624874	-0.1083249
[3,]	0.6666667	0.1125042	0.1624874	-0.1083249
[4,]	0.6666667	0.1125042	0.1624874	-0.1083249
[5,]	0.6666667	0.1125042	0.1624874	-0.1083249
[6,]	0.6666667	0.1125042	0.1624874	-0.1083249

(non-)Linear models: Estimates and predictions (11a-mod)

107/ 327

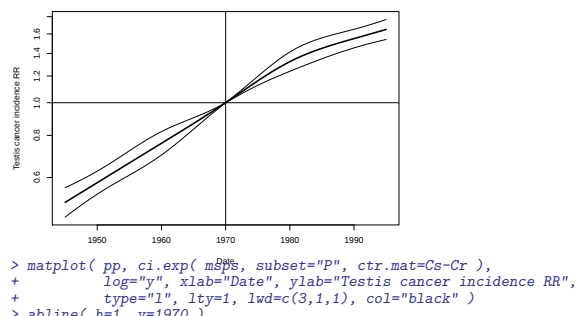
Recap of Monday — rates

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

(non-)Linear models: Estimates and predictions (11a-mod)

112/ 327

Period effect



(non-)Linear models: Estimates and predictions (11a-mod)

108/ 327

Recap Monday — models

- Empirical rate (d_t, y_t) relates to a **time** t
- Many for the same person — different times
- Not independent, but likelihood is a product
- One parameter per interval $\Rightarrow$ exchangeable times
- Use scaling of t : $\Rightarrow$ smooth continuous effects of time
- ... technically complicated:
- Construct `CA <- Ns(a.pt,knots=a.kn)`
- `ci.exp( model, ctr.mat=CA )`
- RR by period: `CP <- Ns(p.pt,knots=p.kn)`
and: `CR <- Ns(rep(p.ref,nrow(CP)),knots=p.kn)`
- `ci.exp( model, ctr.mat=CP-CR)`
- ... actually: `CP <- Ns(p.pt,knots=p.kn,ref=p.ref)`

(non-)Linear models: Estimates and predictions (11a-mod)

113/ 327

Period effect

```
> par( mfrow=c(1,2) )
> Cap <- cbind( 1, Ns( aa ,knots=seq(15,65,10)),
+             Ns(rep(1970,length(aa)),knots=seq(1950,1990,10)) )
> matplot( aa, ci.exp( msp, ctr.mat=Cap )*10^5,
+         log="y", xlab="Age",
+         ylab="Testis cancer incidence rate per 100,000 PY in 1970",
+         type="l", lty=1, lwd=c(3,1,1), col="black" )
> matplot( pp, ci.exp( msp, subset="P", ctr.mat=Cs-Cr ),
+         log="y", xlab="Date", ylab="Testis cancer incidence RR",
+         type="l", lty=1, lwd=c(3,1,1), col="black" )
> abline( h=1, v=1970 )
```

(non-)Linear models: Estimates and predictions (11a-mod)

109/ 327

Follow-up data

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models

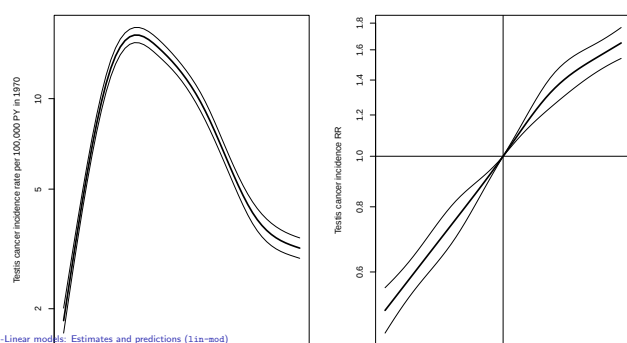
May 2016

Max Planck Institut für Demographische Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

FU-rep-Lexis

Age and period effect



(non-)Linear models: Estimates and predictions (11a-mod)

110/ 327

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.

Follow-up data (FU-rep-Lexis)

114/ 327

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.

Follow-up data (FU-rep-Lexis)

115/ 327

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?

Follow-up data (FU-rep-Lexis)

120/ 327

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?

Follow-up data (FU-rep-Lexis)

116/ 327

Timescales

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

Follow-up data (FU-rep-Lexis)

121/ 327

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.

Follow-up data (FU-rep-Lexis)

117/ 327

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 (FU-rep-Lexis)

122/ 327

Splitting the follow up

	subj. 1	subj. 2	subj. 3
Age at Entry:	13.06	18.44	4.54
Age at eXit:	44.95	41.14	11.12
Status at exit:	Dead	Alive	Dead
Y	31.89	22.70	6.58
D	1	0	1

Follow-up data (FU-rep-Lexis)

118/ 327

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

Follow-up data (FU-rep-Lexis)

123/ 327

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.

Follow-up data (FU-rep-Lexis)

124/ 327

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

Follow-up data (FU-rep-Lexis)

119/ 327

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  id exitdat exits
1 22.18 1938.79  0  38.00      0      1      1  1 1976.79
2 49.55 1945.77  0  18.60      0      1      2 640 1964.37
3 68.21 1955.18  0   1.40      0      1      3 3425 1956.59
4 20.80 1957.61  0  34.52      0      0      4 4017 1992.14
```

Follow-up data (FU-rep-Lexis)

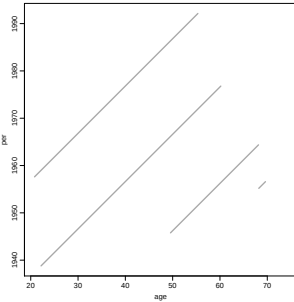
125/ 327

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 birthdat
1      1 22.18 1938.79  0.00    1.00      0      0      1  2 1916.61
2      1 23.18 1939.79  1.00    4.00      0      0      1  2 1916.61
3      1 27.18 1943.79  5.00   12.82      0      0      1  2 1916.61
4      1 40.00 1956.61 17.82    2.18      0      0      1  2 1916.61
5      1 42.18 1958.79 20.00   17.82      0      0      1  2 1916.61
6      1 60.00 1976.61 37.82    0.18      0      1      1  2 1916.61
7      2 49.55 1945.77  0.00    1.00      0      0     640  2 1896.23
8      2 50.55 1946.77  1.00    4.00      0      0     640  2 1896.23
9      2 54.55 1950.77  5.00    5.45      0      0     640  2 1896.23
10     2 60.00 1956.23 10.45    8.14      0      1     640  2 1896.23
11     3 68.21 1955.18  0.00    1.00      0      0    3425  1 1886.97
12     3 69.21 1956.18  1.00    0.40      0      1    3425  1 1886.97
13     4 20.80 1957.61  0.00    1.00      0      0    4017  2 1936.81
14     4 21.80 1958.61  1.00    4.00      0      0    4017  2 1936.81
15     4 25.80 1962.61  5.00   14.20      0      0    4017  2 1936.81
16     4 40.00 1976.81 19.20    0.80      0      0    4017  2 1936.81
```

Follow-up data (FU-rep-Lexis)

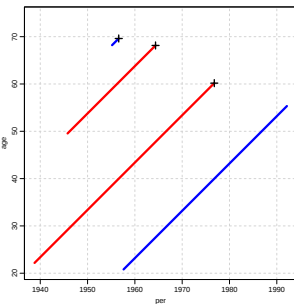
130/ 327



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

Follow-up data (FU-rep-Lexis)

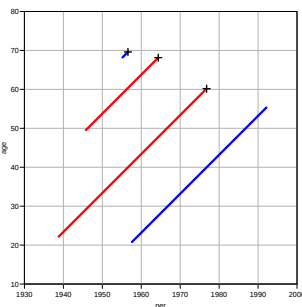
126/ 327



```
> 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 )
```

Follow-up data (FU-rep-Lexis)

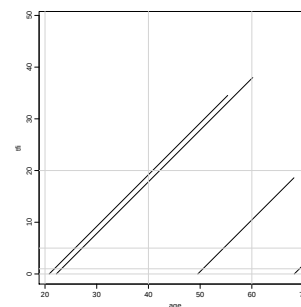
127/ 327



```
> 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 )
```

Follow-up data (FU-rep-Lexis)

128/ 327



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

Follow-up data (FU-rep-Lexis)

133/ 327

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 birthdat contrast
1      1 22.18 1938.79  0.00   17.82      0      0      1  2 1916.61      1
2      1 40.00 1956.61 17.82   20.00      0      0      1  2 1916.61      1
3      1 60.00 1976.61 37.82    0.18      0      1      1  2 1916.61      1
4      2 49.55 1945.77  0.00   10.45      0      0     640  2 1896.23      1
5      2 60.00 1956.23 10.45    8.14      0      1     640  2 1896.23      1
6      3 68.21 1955.18  0.00    1.40      0      1    3425  1 1886.97      2
7      4 20.80 1957.61  0.00   19.20      0      0    4017  2 1936.81      2
8      4 40.00 1976.81 19.20   15.33      0      0    4017  2 1936.81      2
```

Follow-up data (FU-rep-Lexis)

129/ 327

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 birthdat
1      1 22.18 1938.79  0.00    1.00      0      0      1  2 1916.61
2      1 23.18 1939.79  1.00    4.00      0      0      1  2 1916.61
3      1 27.18 1943.79  5.00   12.82      0      0      1  2 1916.61
4      1 40.00 1956.61 17.82    2.18      0      0      1  2 1916.61
5      1 42.18 1958.79 20.00   17.82      0      0      1  2 1916.61
6      1 60.00 1976.61 37.82    0.18      0      1      1  2 1916.61
7      2 49.55 1945.77  0.00    1.00      0      0     640  2 1896.23
8      2 50.55 1946.77  1.00    4.00      0      0     640  2 1896.23
9      2 54.55 1950.77  5.00    5.45      0      0     640  2 1896.23
10     2 60.00 1956.23 10.45    8.14      0      1     640  2 1896.23
11     3 68.21 1955.18  0.00    1.00      0      0    3425  1 1886.97
12     3 69.21 1956.18  1.00    0.40      0      1    3425  1 1886.97
13     4 20.80 1957.61  0.00    1.00      0      0    4017  2 1936.81
14     4 21.80 1958.61  1.00    4.00      0      0    4017  2 1936.81
15     4 25.80 1962.61  5.00   14.20      0      0    4017  2 1936.81
16     4 40.00 1976.81 19.20    0.80      0      0    4017  2 1936.81
```

Follow-up data (FU-rep-Lexis)

134/ 327

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.

Follow-up data (FU-`rep-Lexis`)

135/ 327

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.

Models for tabulated data (`tab-mod`)

139/ 327

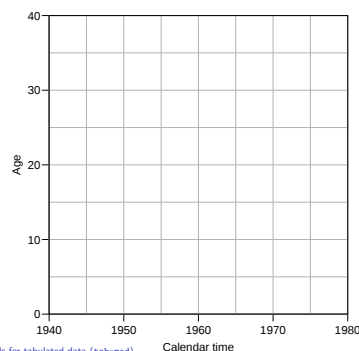
Poisson model for split data

- ▶ 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.

Follow-up data (FU-`rep-Lexis`)

136/ 327

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).

Models for tabulated data (`tab-mod`)

140/ 327

Models for tabulated data

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models

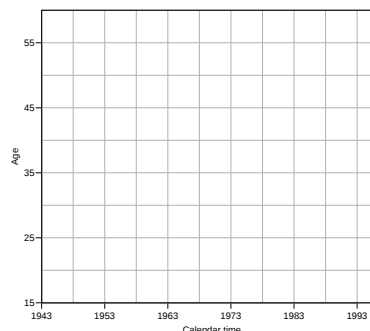
May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

`tab-mod`

Lexis diagram



Registration of:

cases (D)

risk time,
person-years (Y)

in subsets of the Lexis
diagram.

Models for tabulated data (`tab-mod`)

141/ 327

Conceptual set-up

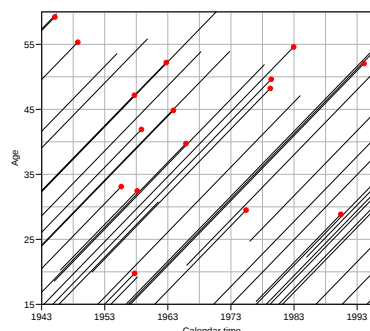
Follow-up of the entire (male) population from 1943–2006 w.r.t. occurrence of testiscancer:

- ▶ 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.

Models for tabulated data (`tab-mod`)

137/ 327

Lexis diagram



Registration of:

cases (D)

risk time,
person-years (Y)

in subsets of the Lexis
diagram.

Rates available in each
subset.

Models for tabulated data (`tab-mod`)

142/ 327

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).
- ▶ Analyse by a Poisson model.

Models for tabulated data (`tab-mod`)

138/ 327

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

Models for tabulated data (`tab-mod`)

143/ 327

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 773812
2 20 1958 7 813022
3 25 1958 28 790501
4 30 1958 28 799293
5 35 1958 36 769356
6 40 1958 24 694073
```

Models for tabulated data (tab-mod)

144/ 327

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
```

Models for tabulated data (tab-mod)

149/ 327

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

Models for tabulated data (tab-mod)

145/ 327

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
```

Models for tabulated data (tab-mod)

150/ 327

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
```

Models for tabulated data (tab-mod)

146/ 327

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
```

Models for tabulated data (tab-mod)

147/ 327

Subsets of parameter estimates accessed via a character string that is greged 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
```

Models for tabulated data (tab-mod)

148/ 327

Age-Period and Age-Cohort models

Statistical Analysis in the Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut für Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

AP-AC

Register data - rates

Rates in "tiles" of the Lexis diagram:

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

Descriptive epidemiology based on disease registers:

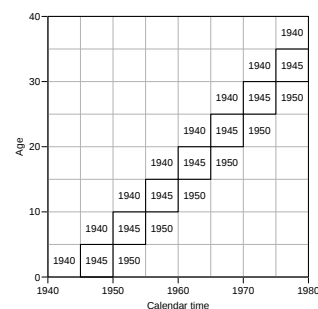
How do the rates vary across by age and time:

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

Age-Period and Age-Cohort models (AP-AC)

151/ 327

Synthetic cohorts



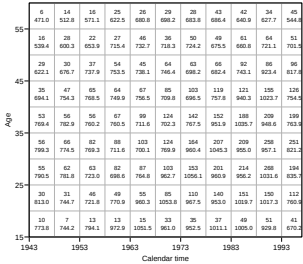
Events and risk time in cells along the diagonals are among persons with roughly same date of birth.

Successively overlapping 10-year periods.

Age-Period and Age-Cohort models (AP-AC)

152/ 327

Lexis diagram: data



Age-Period and Age-Cohort models (AP-AC)

Testis cancer cases in Denmark.

Male person-years in Denmark.

```
> library( Epi )
> load( file="../data/testisDK.Rda" )
> head( testisDK )

      A      P      D      Y
1 17.5 1950.5  7 744.2172
2 22.5 1950.5 31 744.7055
3 27.5 1950.5 62 781.8272
4 32.5 1950.5 66 774.5415
5 37.5 1950.5 56 782.8932
6 42.5 1950.5 47 754.3220

> xtabs( D ~ A + P, data = testisDK )
```

Age-Period and Age-Cohort models (AP-AC)

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	88–92
15–19	7	13	13	15	33	35	37	49	51
20–24	31	46	49	55	85	110	140	151	150
25–29	62	63	82	87	103	153	201	214	268
30–34	66	82	88	103	124	164	207	209	258
35–39	56	56	67	99	124	142	152	188	209
40–44	47	65	64	67	85	103	119	121	155
45–49	30	37	54	45	64	63	66	92	86
50–54	28	22	27	46	36	50	49	61	64
55–59	14	16	25	26	29	28	43	42	34

Age-Period and Age-Cohort models (AP-AC)

```
> wh = c("ap","ac","pa","ca")
> for( i in 1:4 ) {
+   pdf( paste("../graph/AP-AC-testisRate",i,".pdf",sep=""), height=6, width=6 )
+   par( mar=c(3,3,1,1, mgp=c(3,1,0)/1.6, bty="n", las=1 ))
+   rateplot( trate, wh[i], col=rainbow(15), lwd=3, ann=TRUE, a.lim=c(15,65) )
+   dev.off()
+ }
```

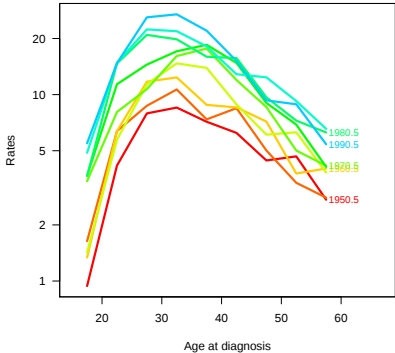
Age-Period and Age-Cohort models (AP-AC)

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	88–92
15–19	744.2	794.1	972.9	1051.5	961.0	952.5	1011.1	1005.0	929.8
20–24	744.7	721.8	770.9	960.3	1053.8	967.5	953.0	1019.7	1017.3
25–29	781.8	723.0	698.6	764.8	962.7	1056.1	960.9	956.2	1031.6
30–34	774.5	769.3	711.6	700.1	769.9	960.4	1045.3	955.0	957.1
35–39	782.9	760.2	760.5	711.6	702.3	767.5	951.9	1035.7	948.6
40–44	754.3	768.5	749.9	756.5	709.8	696.5	757.8	940.3	1023.7
45–49	676.7	737.9	753.5	738.1	746.4	698.2	682.4	743.1	923.4
50–54	600.3	653.9	715.4	732.7	718.3	724.2	675.5	660.8	721.1
55–59	512.8	571.1	622.5	680.8	698.2	683.8	686.4	640.9	627.7

Age-Period and Age-Cohort models (AP-AC)



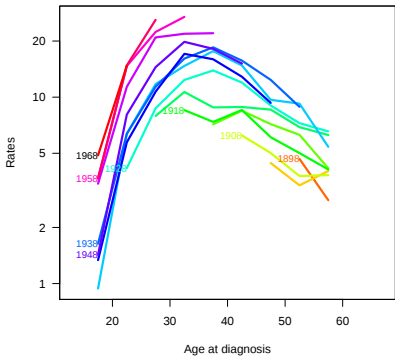
Age-Period and Age-Cohort models (AP-AC)

Data matrix: Empirical rates

Rate per 1000,000 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	88–92
15–19	9.4	16.4	13.4	14.3	34.3	36.7	36.6	48.8	54.8
20–24	41.6	63.7	63.6	57.3	80.7	113.7	146.9	148.1	147.4
25–29	79.3	87.1	117.4	113.8	107.0	144.9	209.2	223.8	259.8
30–34	85.2	106.6	123.7	147.1	161.1	170.8	198.0	218.8	269.6
35–39	71.5	73.7	88.1	139.1	176.6	185.0	159.7	181.5	220.3
40–44	62.3	84.6	85.3	88.6	119.8	147.9	157.0	128.7	151.4
45–49	44.3	50.1	71.7	61.0	85.7	90.2	96.7	123.8	93.1
50–54	46.6	33.6	37.7	62.8	50.1	69.0	72.5	92.3	88.7
55–59	27.3	28.0	40.2	38.2	41.5	40.9	62.6	65.5	54.2

Age-Period and Age-Cohort models (AP-AC)



Age-Period and Age-Cohort models (AP-AC)

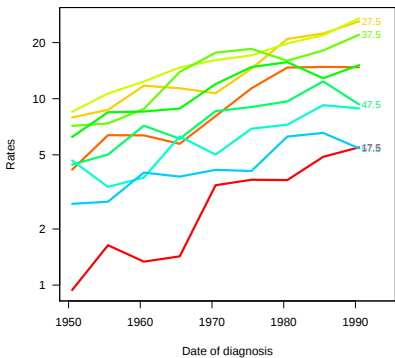
The classical plots

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

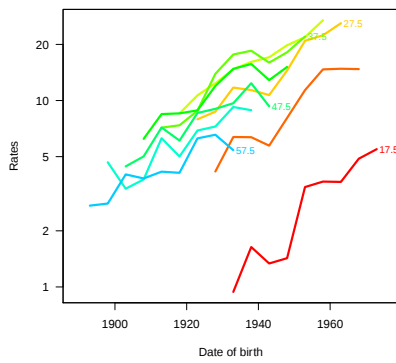
- ▶ Rates versus age at diagnosis (period):
 - rates in the same ageclass connected.
- ▶ Rates versus age at diagnosis:
 - rates in the same birth-cohort connected.
- ▶ Rates versus date of diagnosis:
 - rates in the same ageclass connected.
- ▶ Rates versus date of date of birth:
 - rates in the same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

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

Age-Period and Age-Cohort models (AP-AC)



Age-Period and Age-Cohort models (AP-AC)



Age-Period and Age-Cohort models (AP-AC)

162/ 327

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(log(Y)), family = 
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
```

Age-Period and Age-Cohort models (AP-AC)

167/ 327

Age-period model

Rates are proportional between periods:

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

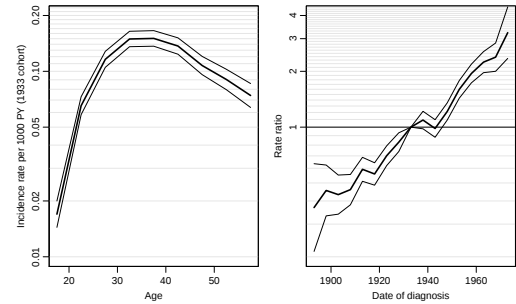
Choose p_0 as reference period, where $\beta_{p_0} = 0$

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

Age-Period and Age-Cohort models (AP-AC)

163/ 327

Graph of estimates with confidence intervals



Age-Period and Age-Cohort models (AP-AC)

168/ 327

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(log(Y)), family = 
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
```

Age-Period and Age-Cohort models (AP-AC)

164/ 327

Age-drift model

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models

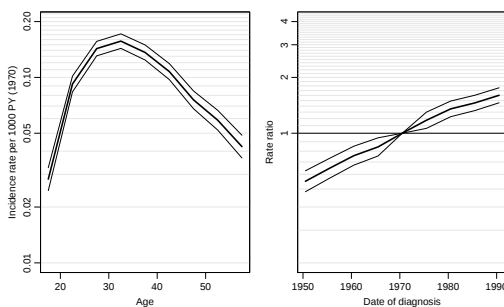
May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

Ad

Graph of estimates with confidence intervals



Age-Period and Age-Cohort models (AP-AC)

165/ 327

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-drift model (Ad)

169/ 327

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 $\gamma_{c_0} = 0$

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

Age-Period and Age-Cohort models (AP-AC)

166/ 327

Age and linear effect of period:

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

Call:
glm(formula = D ~ factor(A) - 1 + I(P - 1970.5) + offset(log(Y)), family = poisson
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.97593  -0.77091   0.02809   0.95914   2.93076

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

(Dispersion parameter for poisson family taken to be 1)

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

Age-drift model (Ad)

170/ 327

Age and linear effect of cohort:

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

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

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

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

(Dispersion parameter for poisson family taken to be 1)

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

Age-drift model (A&D)

171/ 327

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:

$$\tilde{f}(t) = f(t) - \mu_a - \mu_e + \gamma t$$

$$\tilde{g}(a) = g(a) + \mu_a - \gamma a$$

$$\tilde{h}(e) = h(e) + \mu_e + \gamma e$$

because $t - a + e = 0$.

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

Age at entry (Age-at-entry)

175/ 327

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).

Age-drift model (A&D)

172/ 327

“Controlling for age”

— is not a well defined statement.

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

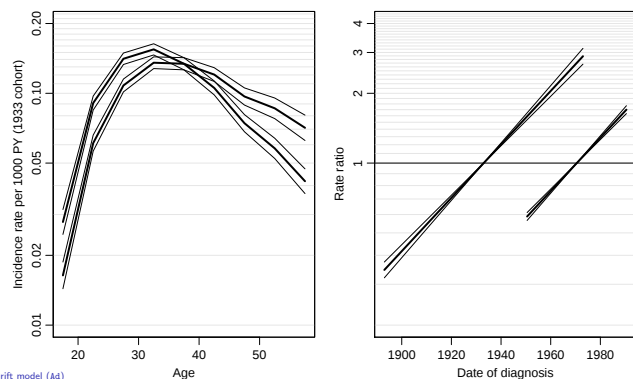
But ideally one would check whether there were non-linear effects of age at entry and current age.

This would require modelling of multiple timescales.

Which is best accomplished by splitting time.

Age at entry (Age-at-entry)

176/ 327



Age-drift model (A&D)

173/ 327

Age-Period-Cohort model

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

APC-cat

Age at entry

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

Age-at-entry

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.
- ▶ Levels of A, P and C are assumed **exchangeable**
- ▶ no assumptions about the relationship of parameter estimates and the **scaled values** of A, P and C

Age-Period-Cohort model (APC-cat)

177/ 327

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).

Age at entry (Age-at-entry)

174/ 327

Fitting the model in R I

```
> library( Epi )
> load( file = "../data/testisDK.Rda" )
> head( testisDK )

      A      P      D      Y
1 17.5 1950.5    7 744.2172
2 22.5 1950.5   31 744.7055
3 27.5 1950.5   62 781.8272
4 32.5 1950.5   66 774.5415
5 37.5 1950.5   56 782.8932
6 42.5 1950.5   47 754.3220

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

Age-Period-Cohort model (APC-cat)

178/ 327

Fitting the model in R II

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

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.55709	-0.56174	0.01096	0.51221	1.32770

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.01129	0.16094	-24.925	< 2e-16
factor(A)22.5	1.23961	0.07745	16.005	< 2e-16
factor(A)27.5	1.70594	0.08049	21.194	< 2e-16
factor(A)32.5	1.83935	0.08946	20.561	< 2e-16
factor(A)37.5	1.71786	0.10217	16.813	< 2e-16
factor(A)42.5	1.48259	0.11708	12.663	< 2e-16
factor(A)47.5	1.09057	0.13447	8.110	5.07e-16
factor(A)52.5	0.76631	0.15271	5.018	5.22e-07
factor(A)57.5	0.41050	0.16094	2.551	0.010751

Age-Period-Cohort model (APC-cat)179 / 327

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	

Age-Period-Cohort model (APC-cat)184 / 327

Fitting the model in R III

```
factor(P)1955.5 0.18645 0.07514 2.482 0.013082
factor(P)1960.5 0.37398 0.07949 4.705 2.54e-06
factor(P)1965.5 0.52062 0.08858 0.08858 5.877 4.17e-09
factor(P)1970.5 0.72806 0.10013 7.271 3.56e-13
factor(P)1975.5 0.90736 0.11422 7.944 1.96e-15
factor(P)1980.5 1.02698 0.12978 7.913 2.51e-15
factor(P)1985.5 1.06237 0.14641 7.256 3.98e-13
factor(P)1990.5 1.10813 0.16094 6.885 5.76e-12
factor(P - A)1898 0.04216 0.29749 0.142 0.887290
factor(P - A)1903 -0.17670 0.26768 -0.660 0.509173
factor(P - A)1908 -0.27238 0.24294 -1.121 0.262210
factor(P - A)1913 -0.18041 0.22226 -0.812 0.416942
factor(P - A)1918 -0.39714 0.20763 -1.913 0.055787
factor(P - A)1923 -0.32538 0.19267 -1.689 0.091249
factor(P - A)1928 -0.30696 0.18046 -1.701 0.088936
factor(P - A)1933 -0.26626 0.16917 -1.574 0.115521
factor(P - A)1938 -0.32937 0.16103 -2.045 0.040813
factor(P - A)1943 -0.57450 0.15417 -3.727 0.000194
factor(P - A)1948 -0.49088 0.14858 -3.304 0.000954
```

Age-Period-Cohort model (APC-cat)180 / 327

How to choose a parametrization

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

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

Implemented in Epi via the contrast type `contr.2nd` (later).

- Holford: Extract linear effects by regression:

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

Age-Period-Cohort model (APC-cat)185 / 327

Fitting the model in R IV

```
factor(P - A)1953 -0.32857 0.14601 -2.250 0.024430
factor(P - A)1958 -0.23140 0.14615 -1.583 0.113351
factor(P - A)1963 -0.18244 0.14978 -1.218 0.223200
factor(P - A)1968 -0.20961 0.16143 -1.298 0.194142
factor(P - A)1973 NA NA NA NA
```

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2463.197 on 80 degrees of freedom
Residual deviance: 35.459 on 49 degrees of freedom
AIC: 584.5

Number of Fisher Scoring iterations: 4

Age-Period-Cohort model (APC-cat)181 / 327

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.

Age-Period-Cohort model (APC-cat)186 / 327

No. of parameters

A has 9 levels
P has 9 levels
 $C = P - A$ 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:

```
> length( coef(m.apc) )
[1] 33
> sum( !is.na(coef(m.apc)) )
[1] 32
```

Age-Period-Cohort model (APC-cat)182 / 327

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:

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

Age-Period-Cohort model (APC-cat)187 / 327

Relationship of models

Testis cancer, Denmark

Age
865.08 / 72

739.01 / 1
p=0.0000

Age-drift
126.07 / 71

8.37 / 7
p=0.3010

60.6 / 15
p=0.0000

Age-Period
117.7 / 64

82.24 / 15
p=0.0000

30.01 / 7
p=0.0001

Age-Period-Cohort
35.46 / 49

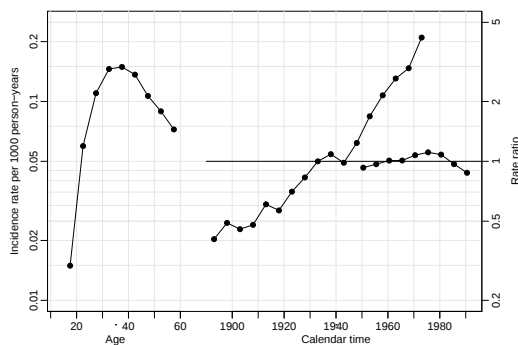
Age-Period-Cohort model (APC-cat)183 / 327

How it all adds up:

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

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

Age-Period-Cohort model (APC-cat)188 / 327



Age-Period-Cohort model (APC-cat)

189 / 327

Using contr.2nd I

```
> attach(testisDK)
> ( nA <- nlevels(factor(A)) )

[1] 9

> ( nP <- nlevels(factor(P)) )

[1] 9

> ( nC <- nlevels(factor(P-A)) )

[1] 17
```

Age-Period-Cohort model (APC-cat)

194 / 327

A simple practical approach

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

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

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

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

Age-Period-Cohort model (APC-cat)

190 / 327

Using contr.2nd II

```
> mp <- glm( D ~ factor(A) - 1 + I(P-1970) +
+           C( factor(P), contr.2nd, nP-2 ) +
+           C( factor(P-A), contr.2nd, nC-2 ),
+           offset = log(Y), family = poisson, data = testisDK )
> mc <- glm( D ~ factor(A) - 1 + I(P-1940) +
+           C( factor(P), contr.2nd, nP-2 ) +
+           C( factor(P-A), contr.2nd, nC-2 ),
+           offset = log(Y), family = poisson, data = testisDK )
> c( m.apc$deviance,
+   mp$deviance,
+   mc$deviance )

[1] 35.4587 35.4587 35.4587

> round( cbind( ci.exp(mp,subset="P"),
+               ci.exp(mc,subset="P") ), 4 )
```

Age-Period-Cohort model (APC-cat)

195 / 327

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

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

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

Age-Period-Cohort model (APC-cat)

191 / 327

Using contr.2nd III

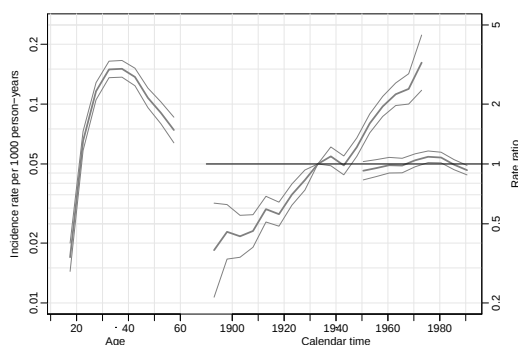
```
exp(Est.) 2.5% 97.5% exp(Est.) 2.5% 97.5%
C(factor(P), contr.2nd, nP - 2)1 1.0011 0.7860 1.2751 1.0011 0.7860 1.2751
C(factor(P), contr.2nd, nP - 2)2 0.9599 0.7680 1.1998 0.9599 0.7680 1.1998
C(factor(P), contr.2nd, nP - 2)3 1.0627 0.8651 1.3053 1.0627 0.8651 1.3053
C(factor(P), contr.2nd, nP - 2)4 0.9722 0.8080 1.1699 0.9722 0.8080 1.1699
C(factor(P), contr.2nd, nP - 2)5 0.9421 0.7977 1.1126 0.9421 0.7977 1.1126
C(factor(P), contr.2nd, nP - 2)6 0.9192 0.7893 1.0706 0.9192 0.7893 1.0706
C(factor(P), contr.2nd, nP - 2)7 1.0104 0.8750 1.1668 1.0104 0.8750 1.1668

> round( rbind( ci.exp(mp,subset="I"),
+               ci.exp(mc,subset="I") ), 4 )

exp(Est.) 2.5% 97.5%
I(P - 1970) 1.0468 0.926 1.1833
I(P - A - 1940) 1.0468 0.926 1.1833
```

Age-Period-Cohort model (APC-cat)

196 / 327



Age-Period-Cohort model (APC-cat)

192 / 327

Using contr.2nd IV

```
> matplot( sort(unique(testisDK$A)),
+           cbind(ci.exp(mp,subset="\\(A)",
+                       ci.exp(mc,subset="\\(A)")*100,
+                       log="y", xlab="Age", ylab="Incidence rate per 100,000 PY",
+                       type="l", lty=1, lwd=c(3,1,1), col=rep(c("red", "blue"), each=2) )
```

Age-Period-Cohort model (APC-cat)

197 / 327

Tabulation in the Lexis diagram

Statistical Analysis in the
Lexis Diagram:

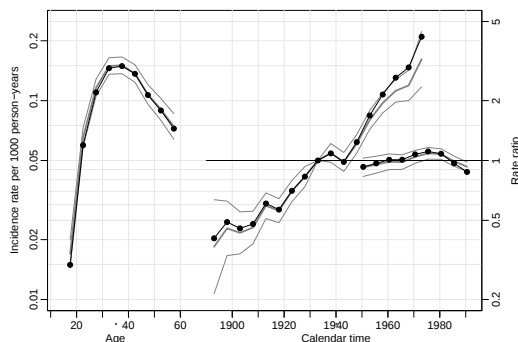
Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

Lexis-tab



Age-Period-Cohort model (APC-cat)

193 / 327

Tabulation of register data

	6	14	16	25	26	29	43	42	34	46
	471.0	522.8	571.1	622.5	680.8	689.2	686.4	660.9	627.7	544.8
55	16	38	22	27	46	36	50	49	61	64
	539.4	600.3	603.9	715.4	732.7	718.3	724.2	679.5	660.8	721.1
	29	30	37	54	45	64	63	66	92	86
	622.1	676.7	737.9	753.5	738.1	746.4	698.2	682.4	743.1	923.4
45	35	47	65	64	67	85	103	119	121	155
	694.1	754.3	768.5	749.9	756.5	709.8	696.5	757.8	940.3	1023.7
	53	56	56	67	99	124	142	152	188	209
	769.4	782.9	760.2	760.5	711.6	702.3	767.5	951.9	1035.7	948.6
35	56	66	82	88	103	124	164	207	209	258
	799.3	774.5	769.3	711.6	700.1	769.9	960.4	1045.3	955.0	957.1
	55	62	63	82	87	103	153	201	214	268
	795.5	781.8	723.0	698.6	764.8	962.7	1055.1	960.9	956.2	1031.6
25	30	31	46	49	55	85	110	140	151	150
	813.0	744.7	721.8	770.9	960.3	1053.8	967.5	953.0	1035.7	1017.3
	10	7	13	13	13	33	35	37	49	51
	773.8	744.2	794.1	972.9	1051.5	961.0	952.5	1011.1	1095.0	939.8
15										
	1943	1953	1963	1973	1983	1993				

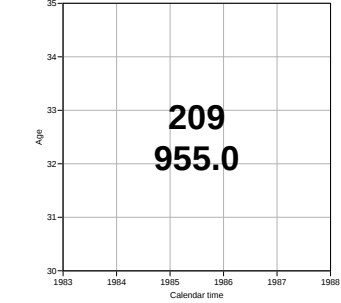
Tabulation in the Lexis diagram (Lexis-tab)

Tabulation of register data

	6	14	16	25	26	29	43	42	34	46
	471.0	522.8	571.1	622.5	680.8	689.2	686.4	660.9	627.7	544.8
55	16	38	22	27	46	36	50	49	61	64
	539.4	600.3	603.9	715.4	732.7	718.3	724.2	679.5	660.8	721.1
	29	30	37	54	45	64	63	66	92	86
	622.1	676.7	737.9	753.5	738.1	746.4	698.2	682.4	743.1	923.4
45	35	47	65	64	67	85	103	119	121	155
	694.1	754.3	768.5	749.9	756.5	709.8	696.5	757.8	940.3	1023.7
	53	56	56	67	99	124	142	152	188	209
	769.4	782.9	760.2	760.5	711.6	702.3	767.5	951.9	1035.7	948.6
35	56	66	82	88	103	124	164	207	209	258
	799.3	774.5	769.3	711.6	700.1	769.9	960.4	1045.3	955.0	957.1
	55	62	63	82	87	103	153	201	214	268
	795.5	781.8	723.0	698.6	764.8	962.7	1055.1	960.9	956.2	1031.6
25	30	31	46	49	55	85	110	140	151	150
	813.0	744.7	721.8	770.9	960.3	1053.8	967.5	953.0	1035.7	1017.3
	10	7	13	13	13	33	35	37	49	51
	773.8	744.2	794.1	972.9	1051.5	961.0	952.5	1011.1	1095.0	939.8
15										
	1943	1953	1963	1973	1983	1993				

Tabulation in the Lexis diagram (Lexis-tab)

Tabulation of register data



Tabulation in the Lexis diagram (Lexis-tab)

Tabulation of register data

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

Tabulation in the Lexis diagram (Lexis-tab)

Tabulation of register data

35	4	3	3	6	4
	20.9	19.6	19.2	18.9	18.9
34	4	1	3	3	7
	19.7	19.2	18.9	18.9	19.2
33	6	4	5	5	6
	19.2	18.9	18.9	19.2	19.0
32	3	3	4	5	4
	19.0	19.1	19.1	19.2	19.2
31	7	4	5	7	2
	18.9	19.2	18.9	19.0	19.2
30					
	1983	1984	1985	1986	1987

Tabulation in the Lexis diagram (Lexis-tab)

Testis cancer cases in Denmark.

Male person-years in Denmark.

Major sets in the Lexis diagram

A-sets: Classification by age and period. ($\square$)

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

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

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

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

Tabulation in the Lexis diagram (Lexis-tab)

Testis cancer cases in Denmark.

Male person-years in Denmark.

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 classifiy 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. ($\triangle$)

Tabulation in the Lexis diagram (Lexis-tab)

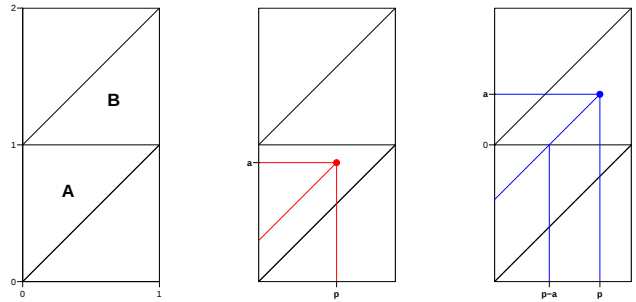
Mean time in triangles

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.

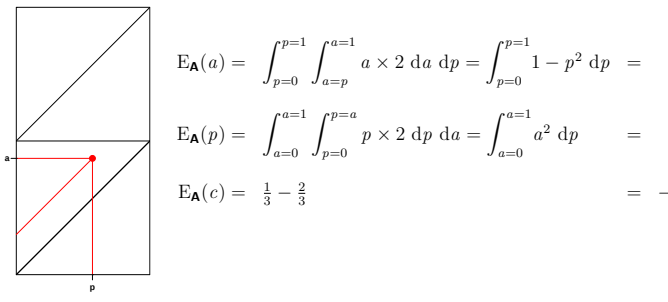
Tabulation in the Lexis diagram (Lexis-tab)

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



Tabulation in the Lexis diagram (Lexis-tab)

Upper triangles (∇), A:



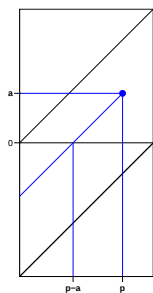
Tabulation in the Lexis diagram (Lexis-tab)

Testis cancer cases in Denmark.

Male person-years in Denmark.

Subdivision by year of birth (cohort).

Lower triangles ($\triangle$), B:



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

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

$$E_B(c) = \frac{2}{3} - \frac{1}{3} = \frac{1}{3}$$

Tabulation in the Lexis diagram (Lexis-tab)

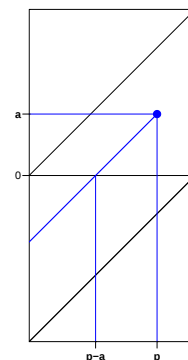
208/ 327

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

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

$$= \int_{p=0}^{p=1} [2pa - a^2]_{a=0}^{a=p} \, dp$$

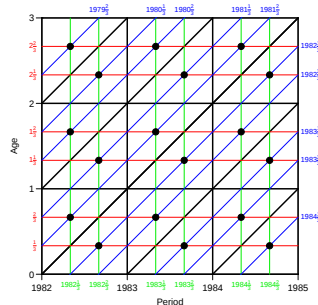
$$= \int_{p=0}^{p=1} p^2 \, dp = \frac{1}{3}$$



Tabulation in the Lexis diagram (Lexis-tab)

213/ 327

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.

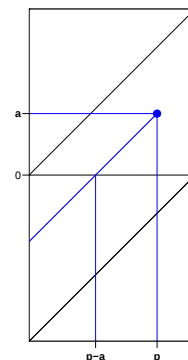
Tabulation in the Lexis diagram (Lexis-tab)

209/ 327

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

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

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



Tabulation in the Lexis diagram (Lexis-tab)

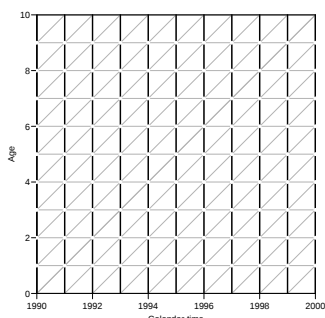
214/ 327

Population figures

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

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

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



Tabulation in the Lexis diagram (Lexis-tab)

210/ 327

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

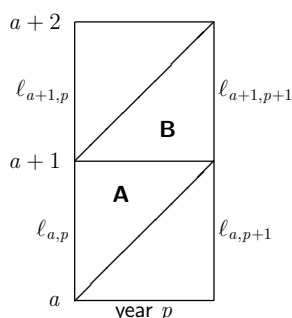
Tabulation in the Lexis diagram (Lexis-tab)

215/ 327

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.



Tabulation in the Lexis diagram (Lexis-tab)

211/ 327

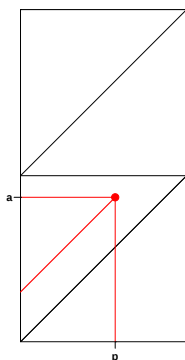
The area of the triangle is $1/2$, so the uniform measure over the triangle has density 2.

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

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

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

$$= \left[p^2 - \frac{2p^3}{3} \right]_{p=0}^{p=1} = \frac{1}{3}$$



Tabulation in the Lexis diagram (Lexis-tab)

212/ 327

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

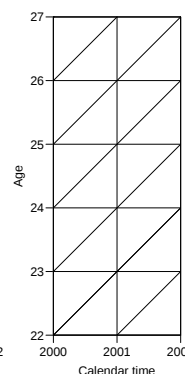
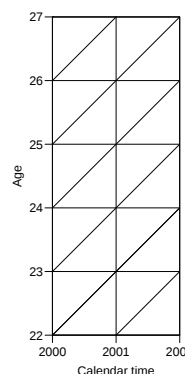
Tabulation in the Lexis diagram (Lexis-tab)

216/ 327

Exercise:

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

Look at the N2Y function in Epi.



Tabulation in the Lexis diagram (Lexis-tab)

217/ 327

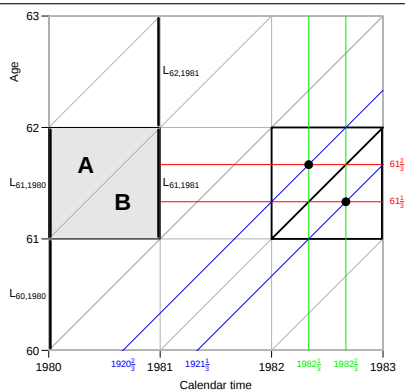
Summary:

Population risk time:

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

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

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



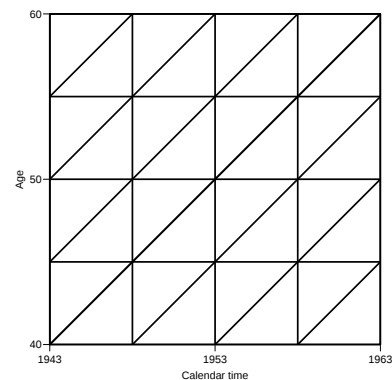
Tabulation in the Lexis diagram (Lexis-tab)

218/ 327

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 for triangular data (APC-tri)

222/ 327

APC-model for triangular data

Statistical Analysis in the Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

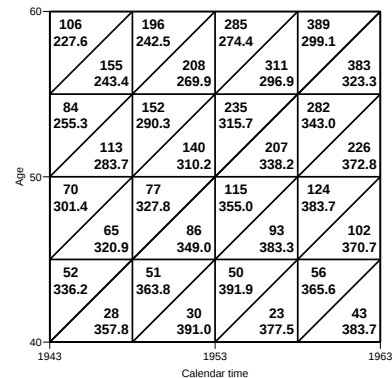
<http://BendixCarstensen/APC/MPIDR-2016>

APC-tri

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 for triangular data (APC-tri)

223/ 327

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:

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

APC-model for triangular data (APC-tri)

219/ 327

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: $220 - 182 = 38$.

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

APC-model for triangular data (APC-tri)

224/ 327

Problem: Disconnected design!

Log-likelihood contribution from one triangle:

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

The log-likelihood can be separated:

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

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

One for upper triangles, one for lower.

APC-model for triangular data (APC-tri)

220/ 327

APC-model with “correct” cohorts

```
> mx <- glm( D ~ factor(Ax) - 1 +
+           factor(Cx) +
+           factor(Px) + offset( log( Y ) ),
+           family=poisson )
> summary( mx )

...

Null deviance: 1.0037e+08 on 220 degrees of freedom
Residual deviance: 2.8473e+02 on 144 degrees of freedom
```

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

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

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

APC-model for triangular data (APC-tri)

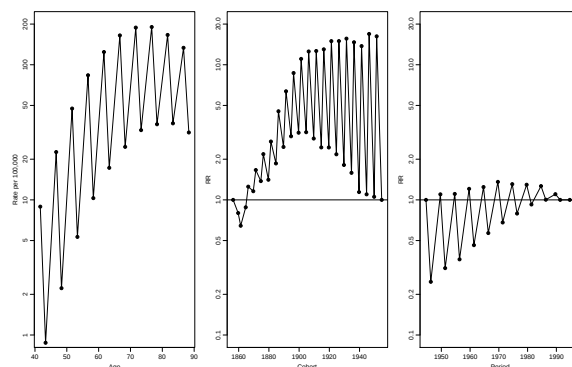
225/ 327

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
```

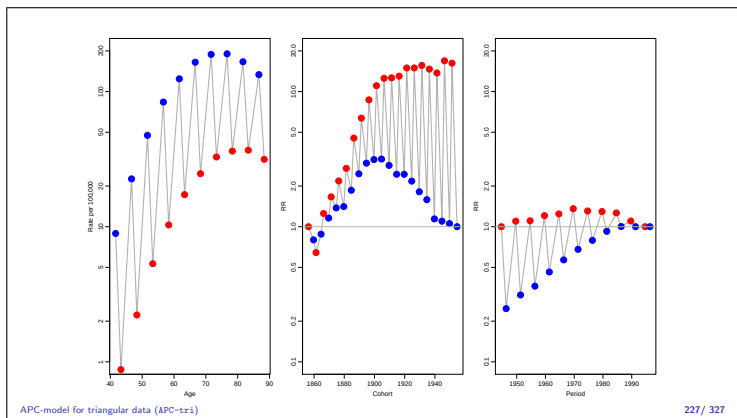
APC-model for triangular data (APC-tri)

221/ 327



APC-model for triangular data (APC-tri)

226/ 327



The identifiability problem still exists:

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

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

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

APC-model: Parametrization (APC-par)

231/ 327

Now, explicitly fit models for upper and lower triangles:

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

APC-model for triangular data (APC-tri)

228/ 327

Smooth functions

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

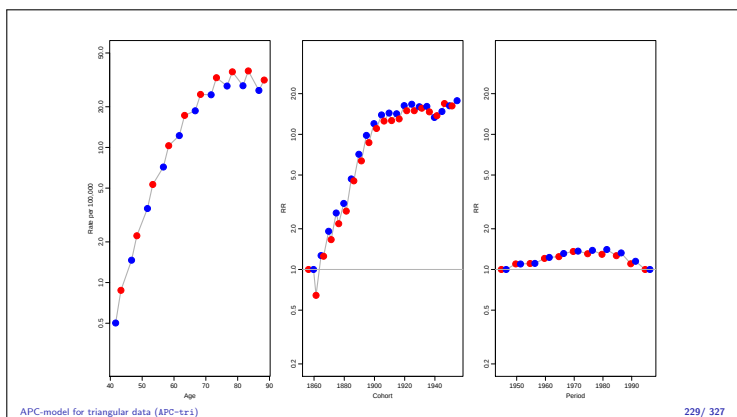
Possible choices for parametric functions describing the effect of the three continuous variables:

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

All of these contain the linear effect as special case.

APC-model: Parametrization (APC-par)

232/ 327



Parametrization of effects

There are still three “free” parameters:

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

Any set of 3 numbers, μ_a , μ_c and γ will produce effects with the same sum. Choose μ_a , μ_c and γ according to some criterion for the functions.

APC-model: Parametrization (APC-par)

233/ 327

APC-model: Parametrization

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

APC-par

Parametrization principle

- The age-function should be interpretable as log age-specific rates in cohort c_0 after adjustment for the period effect.
- The cohort function is 0 at a reference cohort c_0 , interpretable as log-RR relative to cohort c_0 .
- 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?

APC-model: Parametrization (APC-par)

234/ 327

What's the problem?

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

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

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

APC-model: Parametrization (APC-par)

230/ 327

Period-major parametrization

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

Might be wiser to look at predicted rates...

APC-model: Parametrization (APC-par)

235/ 327

Implementation:

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

These functions fulfill the criteria.

APC-model: Parametrization (APC-par)

236/ 327

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 $\hat{\beta}_a$.

APC-model: Parametrization (APC-par)

241/ 327

“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!

APC-model: Parametrization (APC-par)

237/ 327

Information in the data and inner product

Log-lik for an observation (D, Y) , with log-rate $\theta = \log(\lambda)$:

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

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

Log-lik for an observation (D, Y) , with rate λ :

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

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

APC-model: Parametrization (APC-par)

242/ 327

Parametric function

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

Example: 2nd order polynomial:

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

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

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

APC-model: Parametrization (APC-par)

238/ 327

Information in the data and inner product

- ▶ Two inner products:

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

- ▶ Weights could be chosen as:

- ▶ $w_i = D_i$, i.e. proportional to the information content for θ
- ▶ $w_i = Y_i^2/D_i$, i.e. proportional to the information content for λ

APC-model: Parametrization (APC-par)

243/ 327

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 $\tilde{g}(p) = \tilde{M}\pi$, then for **any** parameter vector π :
$$\langle \tilde{M}\pi | 1 \rangle = 0, \quad \langle \tilde{M}\pi | p \rangle = 0 \implies \tilde{M} \perp [1|p]$$
- ▶ Thus we just need to be able to produce $\tilde{M}$ from M : Projection on the orthogonal space of $\text{span}([1|p])$.
- ▶ **NOTE:** Orthogonality requires an inner product!

APC-model: Parametrization (APC-par)

239/ 327

How to? I

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

```
> library( Epi )
> sessionInfo()

R version 3.2.5 (2016-04-14)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.4 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C               LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8    LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8      LC_NAME=C                  LC_ADDRESS=C
[10] LC_TELEPHONE=C            LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] utils      datasets  graphics  grDevices  stats      methods   base
```

APC-model: Parametrization (APC-par)

244/ 327

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 $[1|c]$, respectively
3. Center the cohort effect around c_0 : Take a row from $\tilde{M}_c$ corresponding to c_0 , replicate to dimension as $\tilde{M}_c$, and subtract it from $\tilde{M}_c$ to form $\tilde{M}_{c_0}$.

APC-model: Parametrization (APC-par)

240/ 327

How to? II

other attached packages:

```
[1] Epi_2.3

loaded via a namespace (and not attached):
 [1] cmprsk_2.2-7      MASS_7.3-44       Matrix_1.2-1      plyr_1.8.3        parallel
 [6] survival_2.39-2   etm_0.6-2         Rcpp_0.11.6       splines_3.2.5     grid
[11] numDeriv_2014.2-1 lattice_0.20-31
```

```
> library( splines )
> data( lungDK )
> mw <- apc.fit( A=lungDK$Ax,
+               P=lungDK$Px,
+               D=lungDK$D,
+               Y=lungDK$Y/10^5, dr.extr="w", npar=8,
+               ref.c=1900 )
```

APC-model: Parametrization (APC-par)

245/ 327

How to? III

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

Analysis of deviance for Age-Period-Cohort model

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

> plot( mw )

cp.offset  RR.fac
   1765      100
```

```
+      Y=lungDK$Y/10^5, dr.extr="1", npar=8,
+      ref.c=1900 )

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

Analysis of deviance for Age-Period-Cohort model

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

> mw$Drift

exp(Est.)      2.5%      97.5%
APC (D-weights) 1.019662 1.019062 1.020263
A-d             1.023487 1.022971 1.024003

> ml$Drift
```

How to? IV

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.

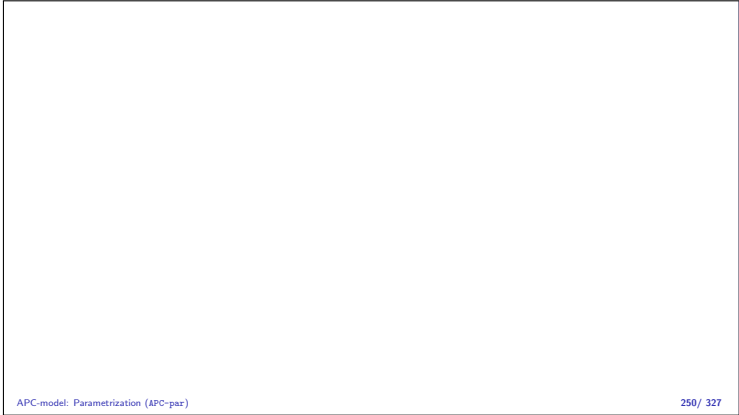
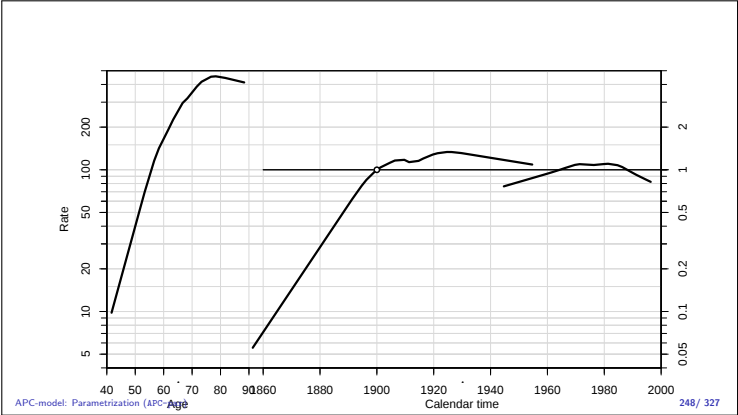
```
> cnr <-
+ function( xf, yf )
+ {
+   cn <- par()$usr
+   xf <- ifelse( xf>1, xf/100, xf )
+   yf <- ifelse( yf>1, yf/100, yf )
+   xx <- ( 1 - xf ) * cn[1] + xf * cn[2]
+   yy <- ( 1 - yf ) * cn[3] + yf * cn[4]
+   if ( par()$xlog ) xx <- 10^xx
+   if ( par()$ylog ) yy <- 10^yy
+   list( x=xx, y=yy )
+ }
```

```
exp(Est.)      2.5%      97.5%
APC (Y2/D-weights) 1.014869 1.013687 1.016053
A-d             1.023487 1.022971 1.024003

> m1$Drift

exp(Est.)      2.5%      97.5%
APC (1-weights) 1.033027 1.032174 1.033879
A-d             1.023487 1.022971 1.024003

> cnr <-
+ function( xf, yf )
+ {
+   cn <- par()$usr
+   xf <- ifelse( xf>1, xf/100, xf )
+   yf <- ifelse( yf>1, yf/100, yf )
+   xx <- ( 1 - xf ) * cn[1] + xf * cn[2]
+   yy <- ( 1 - yf ) * cn[3] + yf * cn[4]
+   if ( par()$xlog ) xx <- 10^xx
+   if ( par()$ylog ) yy <- 10^yy
+   list( x=xx, y=yy )
+ }
```



Other models I

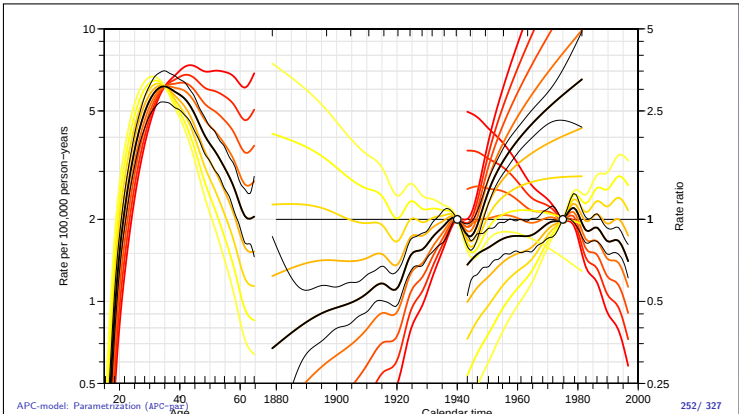
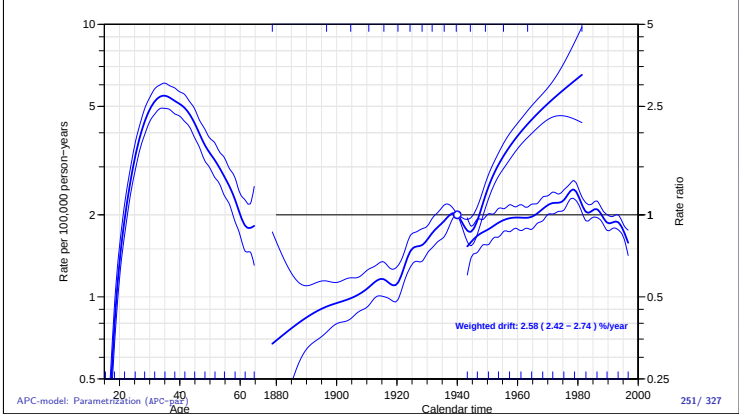
```
> m1 <- apc.fit( A=lungDK$Ax,
+               P=lungDK$Px,
+               D=lungDK$D,
+               Y=lungDK$Y/10^5, dr.extr="1", npar=8,
+               ref.c=1900 )

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

Analysis of deviance for Age-Period-Cohort model

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

> m1 <- apc.fit( A=lungDK$Ax,
+               P=lungDK$Px,
+               D=lungDK$D,
```



Lee-Carter model

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
May 2016

Max Planck Institut for Demographic Research, Rostock
<http://BendixCarstensen/APC/MPIDR-2016>

LeeCarter

Main effect and interaction the same

Main effect and interaction component of t are constrained to, $i=T$ be identical.

An interaction between two spline terms is **not** the same as the product of two terms:

```
> library( Epi )
> dfr <- data.frame( A=30:92, P=rep(1990:2010,3) )
> ( a.kn <- 4:8*10 ) ; ( p.kn <- c(1992+0:2*5) )

[1] 40 50 60 70 80
[1] 1992 1997 2002

> mA <- with( dfr, model.matrix( ~ Ns(A,k=a.kn,i=T) -1 ) )
> mP <- with( dfr, model.matrix( ~ Ns(P,k=p.kn) ) )
> mAP <- with( dfr, model.matrix( ~ Ns(A,k=a.kn,i=T):Ns(P,k=p.kn) -1 ) )
> map <- with( dfr, model.matrix( ~ Ns(A,k=a.kn,i=T)*Ns(P,k=p.kn) -1 ) )
> cbind( colnames(mA) )
```

Lee-Carter model (LeeCarter)

257/ 327

Lee-Carter model for (mortality) rates

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

x is age; t is calendar time

- Formulated originally using as step-functions with one parameter per age/period.
- Implicitly assumes a data lay out by age and period: A, B or C-sets, but **not** Lexis triangles.
- Relative **scaling** of b_x and k_t cannot be determined
- k_t only determined up to an **affine** transformation:

$$a_x + b_x(k_t + m) = (a_x + b_x m) = \tilde{a}_x + b_x k_t$$

Lee-Carter model (LeeCarter)

253/ 327

Lee-Carter model interpretation

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

- Constraints:
- $f(a)$ is the basic age-specific mortality
- $k(p)$ is the rate-ratio (RR) as a function of p :
 - relative to p_{ref} where $k(p_{\text{ref}}) = 1$
 - for persons aged a_{ref} where $b(a_{\text{ref}}) = 0$
- $b(a)$ is an age-specific multiplier for the RR
- Choose p_{ref} and a_{ref} *a priori*.

Lee-Carter model (LeeCarter)

258/ 327

Lee-Carter model in continuous time

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

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

$$f(a) + b(a)(k(t) + m) = (f(a) + b(a)m) = \tilde{f}(a) + b(a)k(t)$$

Lee-Carter model (LeeCarter)

254/ 327

Danish lung cancer data I

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

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

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

Lee-Carter model (LeeCarter)

259/ 327

Lee-Carter model in continuous time

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

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

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

- Main effect and interaction component of t are constrained to be identical.
- NOTE:** the time variable, t could be either period, p or cohort, $c = p - a$.

Lee-Carter model (LeeCarter)

255/ 327

Danish lung cancer data II

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

Lee-Carter model (LeeCarter)

260/ 327

Main effect and interaction the same

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

None of these are Lee-Carter models:

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

Lee-Carter model (LeeCarter)

256/ 327

Lee-Carter with demography I

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

Lee-Carter model (LeeCarter)

261/ 327

Lee-Carter with demography II

```
List of 7
$ year : num [1:61] 1943 1944 1945 1946 1947 ...
$ age : num [1:51] 39 40 41 42 43 44 45 46 47 48 ...
$ rate :List of 1
..$ Male: num [1:51, 1:61] 1.05e-04 7.10e-05 7.31e-05 3.73e-05 2.30e-04 ...
..- attr(*, "dimnames")=list of 2
.. ..$ : chr [1:51] "39" "40" "41" "42" ...
.. ..$ : chr [1:61] "1943" "1944" "1945" "1946" ...
$ pop :List of 1
..$ Male: num [1:51, 1:61] 28488 28152 27363 26791 26092 ...
..- attr(*, "dimnames")=list of 2
.. ..$ : chr [1:51] "39" "40" "41" "42" ...
.. ..$ : chr [1:61] "1943" "1944" "1945" "1946" ...
$ type : chr "Lung cancer incidence"
$ label : chr "Denmark"
$ lambda: num 1
- attr(*, "class")= chr "demogdata"
```

Lee-Carter with ilc

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

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

- ▶ Age interaction with betewwn age and both period and/or cohort (=period-age)
- ▶ Extension of APC-model:
 $b(a) = 1$ and $a(a) = 1 \Leftrightarrow$ APC model.

Lee-Carter with demography III

lca estimation function checks the type argument, so we make a workaround:

```
> mrt <- function(x) { x$type <- "mortality"; x }
> dmg.lcM <- lca( mrt(lcM), interpolate=TRUE )
> par( mfcol=c(2,2) )
> matplot( dmg.lcM$age, exp(dmg.lcM$ax)*1000,
+ log="y", ylab="Lung cancer incidence rates per 1000 PY",
+ xlab="Age", type="l", lty=1, lwd=4 )
> plot( NA, xlim=0:1, ylim=0:1, axes=FALSE, xlab="", ylab="" )
> matplot( dmg.lcM$age, dmg.lcM$bx,
+ ylab="Age effect",
+ xlab="Age", type="l", lty=1, lwd=4 )
> matplot( dmg.lcM$year, dmg.lcM$kt,
+ ylab="Time effect",
+ xlab="Date", type="l", lty=1, lwd=4 )
> abline(h=0)
```

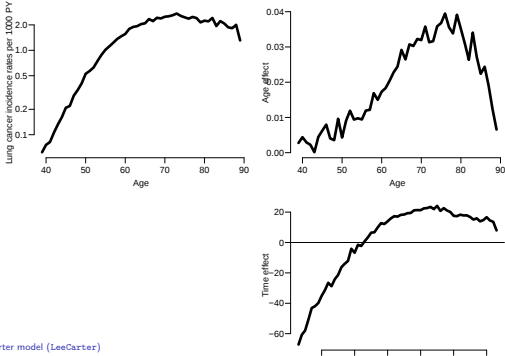
Lee-Carter with ilc I

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

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

Fitting model: [LC = a(x)+b1(x)*k(t)]
- with poisson error structure and with deaths as weights -
Note: 0 cells have 0/NA deaths and 0 have 0/NA exposure
out of a total of 3111 data cells.

Lee-Carter with demography



Lee-Carter with ilc II

Starting values are:

	per	per.c	age	age.c	bx1.c
1 1943	0	39	-9.687	0.02	
2 1944	0	40	-9.487	0.02	
3 1945	0	41	-9.408	0.02	
4 1946	0	42	-9.151	0.02	
5 1947	0	43	-8.929	0.02	
6 1948	0	44	-8.73	0.02	
7 1949	0	45	-8.475	0.02	
8 1950	0	46	-8.426	0.02	
9 1951	0	47	-8.145	0.02	
10 1952	0	48	-7.991	0.02	
11 1953	0	49	-7.808	0.02	
12 1954	0	50	-7.549	0.02	
13 1955	0	51	-7.473	0.02	
14 1956	0	52	-7.376	0.02	
15 1957	0	53	-7.199	0.02	
16 1958	0	54	-7.032	0.02	
17 1959	0	55	-6.893	0.02	

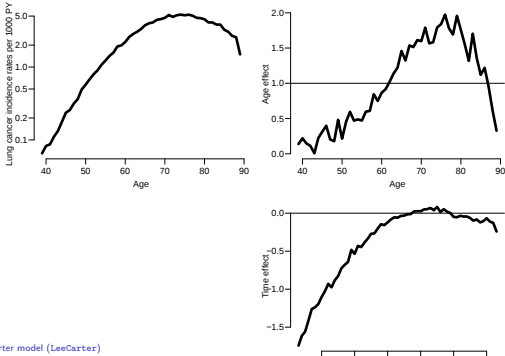
Lee-Carter re-scaled I

```
> par( mfcol=c(2,2) )
> matplot( dmg.lcM$age, exp(dmg.lcM$ax+dmg.lcM$bx*20)*1000,
+ log="y", ylab="Lung cancer incidence rates per 1000 PY",
+ xlab="Age", type="l", lty=1, lwd=4 )
> plot( NA, xlim=0:1, ylim=0:1, axes=FALSE, xlab="", ylab="" )
> matplot( dmg.lcM$age, dmg.lcM$bx*50,
+ ylab="Age effect",
+ xlab="Age", type="l", lty=1, lwd=4 )
> abline(h=1)
> matplot( dmg.lcM$year, (dmg.lcM$kt-20)/50,
+ ylab="Time effect",
+ xlab="Date", type="l", lty=1, lwd=4 )
> abline(h=0)
```

Lee-Carter with ilc III

18 1960	0	56	-6.798	0.02	
19 1961	0	57	-6.698	0.02	
20 1962	0	58	-6.596	0.02	
21 1963	0	59	-6.524	0.02	
22 1964	0	60	-6.463	0.02	
23 1965	0	61	-6.325	0.02	
24 1966	0	62	-6.271	0.02	
25 1967	0	63	-6.25	0.02	
26 1968	0	64	-6.194	0.02	
27 1969	0	65	-6.171	0.02	
28 1970	0	66	-6.056	0.02	
29 1971	0	67	-6.113	0.02	
30 1972	0	68	-6.021	0.02	
31 1973	0	69	-6.039	0.02	
32 1974	0	70	-5.993	0.02	
33 1975	0	71	-5.98	0.02	
34 1976	0	72	-5.951	0.02	
35 1977	0	73	-5.905	0.02	
36 1978	0	74	-5.969	0.02	

Lee-Carter with demography rescaled



Lee-Carter with ilc IV

37 1979	0	75	-6.008	0.02	
38 1980	0	76	-6.044	0.02	
39 1981	0	77	-5.998	0.02	
40 1982	0	78	-6.029	0.02	
41 1983	0	79	-6.146	0.02	
42 1984	0	80	-6.1	0.02	
43 1985	0	81	-6.118	0.02	
44 1986	0	82	-6.025	0.02	
45 1987	0	83	-6.247	0.02	
46 1988	0	84	-6.111	0.02	
47 1989	0	85	-6.177	0.02	
48 1990	0	86	-6.281	0.02	
49 1991	0	87	-6.305	0.02	
50 1992	0	88	-6.213	0.02	
51 1993	0	89	-6.638	0.02	
52 1994	0				
53 1995	0				
54 1996	0				
55 1997	0				

Lee-Carter with ilc V

```
56 1998      0
57 1999      0
58 2000      0
59 2001      0
60 2002      0
61 2003      0

Iterative fit:
#iter   Dev   non-conv
1      26123.55  0
2      9403.337  0
3      5219.715  0
4      4269.859  0
5      3982.804  0
6      3878.544  0
7      3836.703  0
8      3818.834  0
9      3810.839  0
10     3807.133  0
```

Lee-Carter with ilc X

```
48 1990 17.48203      86 -6.24171 0.02873
49 1991 17.31414      87 -6.24948 0.02586
50 1992 18.04416      88 -6.11791 0.01529
51 1993 17.91747      89 -6.51232 0.01146
52 1994 18.39041
53 1995 17.32639
54 1996 15.72621
55 1997 16.81425
56 1998 15.71813
57 1999 15.95432
58 2000 17.93764
59 2001 16.86795
60 2002 15.63661
61 2003 11.11935

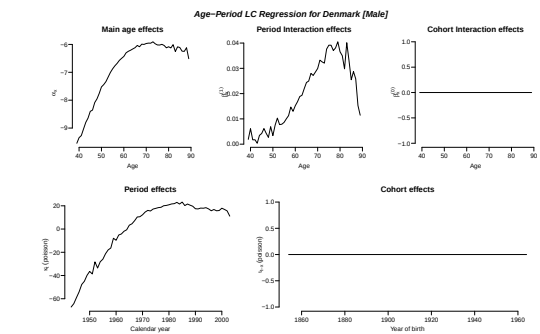
total sums are:
b0 b1 itx kt
0 1 0 0

> plot( ilc.lcM )
```

Lee-Carter with ilc VI

```
11 3805.368 0
12 3804.512 0
13 3804.089 0
14 3803.878 0
15 3803.772 0
16 3803.718 0
17 3803.69 0
18 3803.676 0
19 3803.669 0
20 3803.665 0
21 3803.663 0
22 3803.662 0
23 3803.661 0
24 3803.661 0
25 3803.661 0
26 3803.661 0
27 3803.66 0
28 3803.66 0
29 3803.66 0
```

Lee-Carter with ilc



Lee-Carter with ilc VII

```
30 3803.66 0
31 3803.66 0
32 3803.66 0
33 3803.66 0
34 3803.66 0

Iterations finished in: 34 steps

Updated values are:
per per.c age age.c bx1.c
1 1943 -67.11668 39 -9.54531 0.0019
2 1944 -64.24915 40 -9.34555 0.00613
3 1945 -59.06778 41 -9.27014 0.00171
4 1946 -54.10285 42 -9.03109 0.00174
5 1947 -47.71912 43 -8.79572 0.00036
6 1948 -44.96623 44 -8.64242 0.00348
7 1949 -39.87365 45 -8.4011 0.00422
8 1950 -36.46366 46 -8.35569 0.00618
9 1951 -38.65511 47 -8.08493 0.00431
```

Lee-Carter with Epi

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

Lee-Carter with ilc VIII

```
10 1952 -28.25000      48 -7.95317 0.00269
11 1953 -33.56753      49 -7.75764 0.00692
12 1954 -28.16299      50 -7.52418 0.00338
13 1955 -25.93964      51 -7.44269 0.00752
14 1956 -21.26733      52 -7.33407 0.01031
15 1957 -17.95370      53 -7.16891 0.00774
16 1958 -16.32569      54 -7.00417 0.00789
17 1959 -7.92142      55 -6.87498 0.00862
18 1960 -9.67085      56 -6.76735 0.01002
19 1961 -5.13527      57 -6.67977 0.01128
20 1962 -4.23977      58 -6.57225 0.01469
21 1963 -1.90709      59 -6.49916 0.013
22 1964 -0.65036      60 -6.4307 0.0152
23 1965 3.31265      61 -6.30139 0.0168
24 1966 4.51564      62 -6.247 0.01884
25 1967 7.16008      63 -6.20883 0.01935
26 1968 10.36382      64 -6.16206 0.02197
27 1969 10.60063      65 -6.11728 0.02439
28 1970 12.25461      66 -6.03717 0.02497
```

Lee-Carter with Epi I

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

LCa.fit convergence in 11 iterations, deviance: 8566.554 on 6084 d.f.

> LCa.Mlc

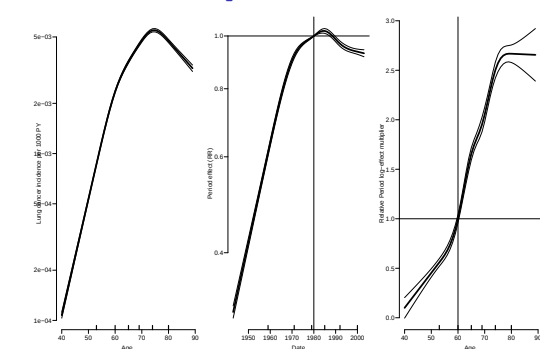
Lee-Carter model using natural splines:
log(Rate) = a(Age) + b(Age)k(Period)
with 6, 5 and 6 parameters respectively (1 aliased).
Deviance: 8566.554 on 6084 d.f.

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

Lee-Carter with ilc IX

```
29 1971 14.63642      67 -6.08387 0.028
30 1972 16.05776      68 -5.99082 0.02718
31 1973 15.53593      69 -6.00028 0.02854
32 1974 17.21334      70 -5.96719 0.02994
33 1975 17.80268      71 -5.95329 0.03323
34 1976 18.44457      72 -5.9555 0.03255
35 1977 18.71973      73 -5.9058 0.03205
36 1978 20.06082      74 -5.97665 0.03762
37 1979 20.31816      75 -6.01915 0.03916
38 1980 20.87884      76 -6.02213 0.03915
39 1981 21.61232      77 -5.99743 0.03704
40 1982 21.85089      78 -6.03741 0.03809
41 1983 22.96473      79 -6.12152 0.0405
42 1984 21.50736      80 -6.08339 0.03654
43 1985 23.22937      81 -6.12649 0.035
44 1986 20.20563      82 -6.00846 0.02978
45 1987 21.53699      83 -6.2544 0.04013
46 1988 20.54046      84 -6.08511 0.03306
47 1989 19.63340      85 -6.11129 0.02548
```

Lee-Carter with Epi



Lee-Carter and the APC-model

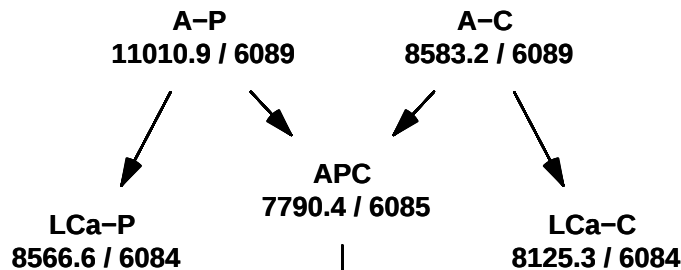
- ▶ Lee-Carter model is an interaction extension of the Age-Period model
- ▶ ... or an interaction extension of the Age-Cohort model
- ▶ Age-Period-Cohort model is:
 - ▶ interaction extension
 - ▶ the smallest **union** of Age-Period and Age-Cohort
- ▶ Extended Lee-Carter (from the `ilc` package)

$$\log(\lambda(a, p)) = f(a) + b(a) \times k(p) + c(a)m(p - a)$$
 is the union of all of these.

Lee-Carter model (LeeCarter)

282 / 327

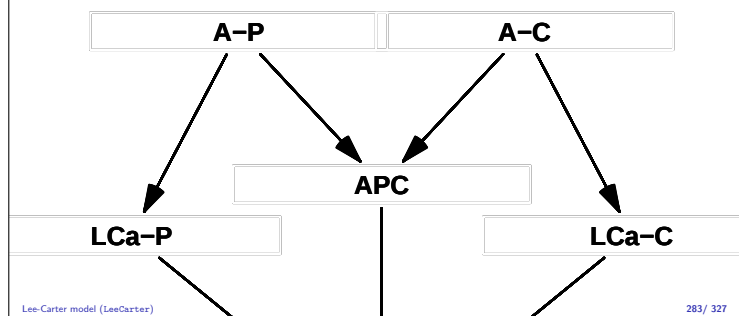
Fit L-Ca models in Epi IV



Lee-Carter model (LeeCarter)

287 / 327

Lee-Carter and the APC-model



Lee-Carter model (LeeCarter)

283 / 327

APC-models for several datasets

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

APC2

Fit L-Ca models in Epi I

```

> LCa.P <- LCa.fit( Mlc, ref.b=60, ref.t=1980 )
LCa.fit convergence in 11 iterations, deviance: 8566.554 on 6084 d.f.
> LCa.C <- LCa.fit( Mlc, ref.b=60, ref.t=1980, model="C", maxit=200, eps=10e-4 )
LCa.fit convergence in 95 iterations, deviance: 8125.318 on 6084 d.f.
> ( a.kn <- LCa.P$a.kn )
      8.333333%    25% 41.66667% 58.33333%    75% 91.66667%
      53         60         65         69         74         80
> LCa.C$a.kn
      8.333333%    25% 41.66667% 58.33333%    75% 91.66667%
      53         60         65         69         74         80
    
```

Lee-Carter model (LeeCarter)

284 / 327

Two APC-models

- ▶ APC-models for two sets of rates (men/women, types of events):

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

- ▶ Rate-ratio also an APC-model:

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

- ▶ Modeled separately and the ratio effects reported as any other APC-model.

APC-models for several datasets (APC2)

288 / 327

Fit L-Ca models in Epi II

```

> ( p.kn <- LCa.P$t.kn )
      8.333333%    25% 41.66667% 58.33333%    75% 91.66667%
      1959      1971      1979      1985      1992      2000
> ( c.kn <- LCa.C$t.kn )
      8.333333%    25% 41.66667% 58.33333%    75% 91.66667%
      1893      1904      1911      1918      1925      1935
> AP <- glm( D ~ Ns(A,knots=a.kn)+Ns(P,knots=p.kn),
+           offset=log(Y), family=poisson, data=Mlc )
> AC <- glm( D ~ Ns(A,knots=a.kn)+Ns(P-A,knots=c.kn),
+           offset=log(Y), family=poisson, data=Mlc )
> APC <- glm( D ~ Ns(A,knots=a.kn)+Ns(P,knots=p.kn)+Ns(P-A,knots=c.kn),
+           offset=log(Y), family=poisson, data=Mlc )
> c( AP$deviance, AP$df.res )
[1] 11010.88 6089.00
    
```

Lee-Carter model (LeeCarter)

285 / 327

Two sets of data I

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

```

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

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

APC-models for several datasets (APC2)

289 / 327

Fit L-Ca models in Epi III

```

> c( AC$deviance, AC$df.res )
[1] 8583.249 6089.000
> c( APC$deviance, APC$df.res )
[1] 7790.446 6085.000
> c( LCa.P$dev, LCa.P$df )
[1] 8566.554 6084.000
> c( LCa.C$dev, LCa.C$df )
[1] 8125.318 6084.000
    
```

Lee-Carter model (LeeCarter)

286 / 327

Two sets of data II

```

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

> th <- transform( th,
+                 hist = factor( hist, labels=c("Sem", "nS", "0th") ),
+                 A = age,
+                 P = diag,
+                 D = d,
+                 Y = y/10^5 ), c("A", "P", "D", "Y", "hist") )
    
```

APC-models for several datasets (APC2)

290 / 327

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

Histology	D	Y
Sem	4708.00	1275.25
nS	3632.00	1275.25
Oth	466.00	1275.25
Total	8806.00	3825.76

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

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( log(Y) ), family=po
> 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
```

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

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

Analysis of deviance for Age-Period-Cohort model

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
Age	9712	6845.8			
Age-drift	9711	6255.1	1	590.70	< 2.2e-16
Age-Cohort	9705	6210.0	6	45.09	4.500e-08
Age-Period-Cohort	9699	6184.1	6	25.90	0.0002323
Age-Period	9705	6241.9	-6	-57.75	1.289e-10
Age-drift	9711	6255.1	-6	-13.24	0.0393950

Analysis of two rates: Formal tests II

```
Model 5: D ~ Mc + Mp + Ma:type + offset(log(Y)) - 1
Resid. Df Resid. Dev Df Deviance P(>|Chi|)
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.nS <- apc.fit( subset( th, hist=="nS" ),
+                     parm = "ACP",
+                     ref.c = 1970,
+                     npar = c(A=8,P=8,C=8) )

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

Analysis of deviance for Age-Period-Cohort model

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
Age	9712	6316.4			
Age-drift	9711	5619.1	1	697.29	< 2.2e-16
Age-Cohort	9705	5675.6	6	43.51	9.243e-08
Age-Period-Cohort	9699	5502.9	6	72.75	1.117e-13
Age-Period	9705	5550.8	-6	-47.91	1.229e-08
Age-drift	9711	5619.1	-6	-68.34	8.945e-13

```
> apc.Sem$Drift
```

	exp(Est.)	2.5%	97.5%
APC (D-weights)	1.023586	1.021563	1.025614
A=d	1.023765	1.021773	1.025761

APC-model: Interactions

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

APC-int

```
> apc.nS$Drift

exp(Est.) 2.5% 97.5%
APC (D-weights) 1.029438 1.026870 1.032013
A=d 1.030162 1.027799 1.032531
```

```
> plot( apc.nS, col="blue" )
```

```
cp.offset RR.fac
1750 1
```

```
> lines( apc.Sem, col="red" )
> matlines( apc.nS$Age[,1], apc.Sem$Age[,2]/apc.nS$Age[,2],
+           lty=1, lwd=5, col="black" )
> pc.lines( apc.nS$Per[,1], apc.Sem$Per[,2]/apc.nS$Per[,2],
+           lty=1, lwd=5, col="black" )
> pc.lines( apc.nS$Coh[,1], apc.Sem$Coh[,2]/apc.nS$Coh[,2],
+           lty=1, lwd=5, col="black" )
> text( 90, 0.7, "non-Seminoma", col="blue", adj=1 )
> text( 90, 0.7~2, "Seminoma", col="red", adj=1 )
> text( 90, 0.7~3, "RR", col="black", adj=1 )
```

Analysis of DM-rates: Age×sex interaction I

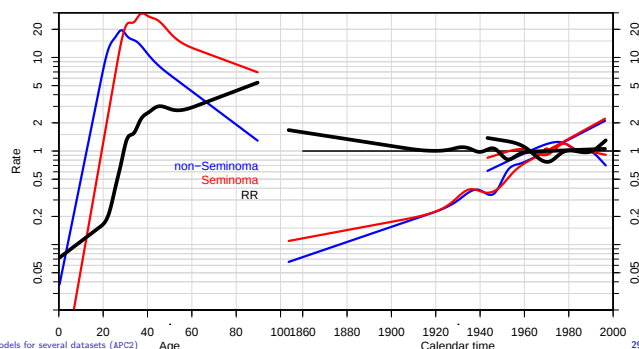
- ▶ 10 centres
- ▶ 2 sexes
- ▶ Age: 0-15
- ▶ Period 1989–1999

- ▶ Is the sex-effect the same between all centres?
- ▶ How 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.Rdata" )
dm <- dm[dm$scen=="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 )] )

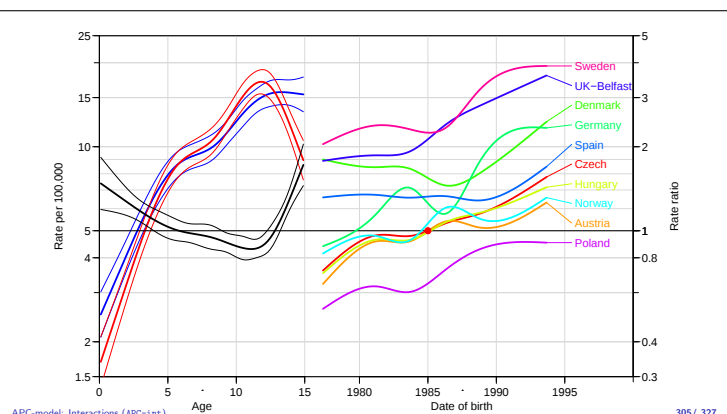
# Age-cohort model with age-sex interaction
# The model matrices for the ML fit
Ma <- ns( A, kn=A.kn, Bo=A.ok, intercept=T )
Mc <- cbind( C-c0, detrend( ns( C, kn=C.kn, Bo=C.ok ), C, weight=D ) )
Mp <- detrend( ns( P, kn=P.kn, Bo=P.ok ), P, weight=D )
# The prediction matrices
Pa <- Ma[match(A.pt,A),,drop=F]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

APC-model: Interactions (APC-Int)

300/ 327



APC-model: Interactions (APC-Int)

305/ 327

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="sexF", ctr.mat=Pa, Exp=T )[5:7]
MF.RR <- ci.lin( apcs, subset=c("sexF", "sexF"), ctr.mat=cbind(Pa,-Pa), Exp=T )[5:7]
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),
```

APC-model: Interactions (APC-Int)

301/ 327

Predicting future rates

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institut for Demographic Research, Rostock

<http://BendixCarstensen/APC/MPIDR-2016>

predict

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])
```

APC-model: Interactions (APC-Int)

302/ 327

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.
- Predictions from the model must be invariant under reparametrization.

Predicting future rates (predict)

306/ 327

Identifiability

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

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

Predicting future rates (predict)

307/ 327

Parametrization invariance

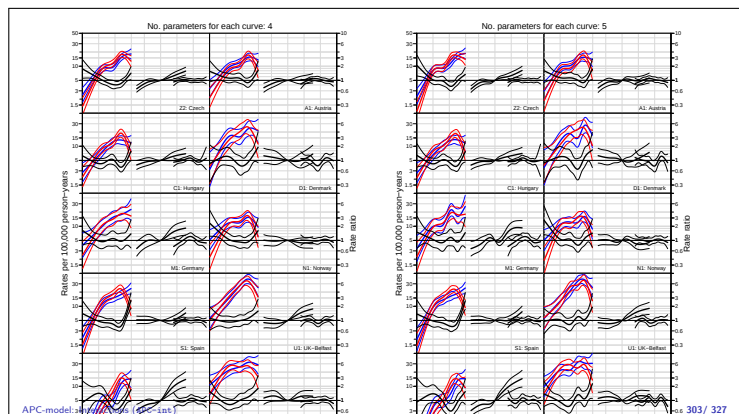
- Prediction of the future course of g and h must preserve addition of a linear term in the argument:

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

- 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.

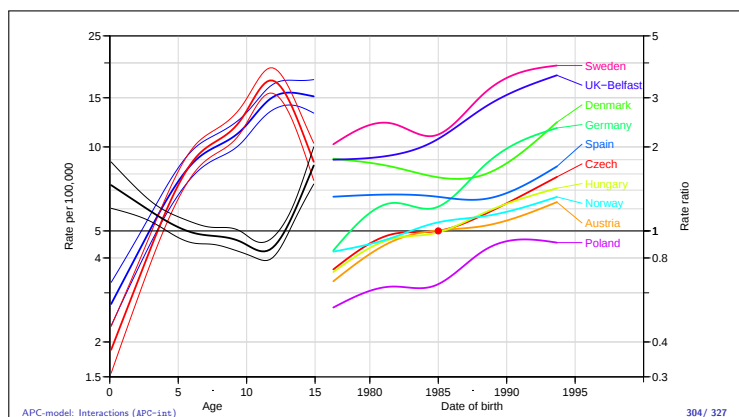
Predicting future rates (predict)

308/ 327



APC-model: Interactions (APC-Int)

303/ 327



APC-model: Interactions (APC-Int)

304/ 327

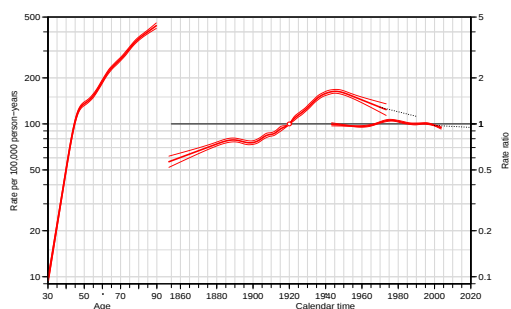
Identifiability

- 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.

Predicting future rates (predict)

309/ 327

Example: Breast cancer in Denmark



Predicting future rates (predict)

310/ 327

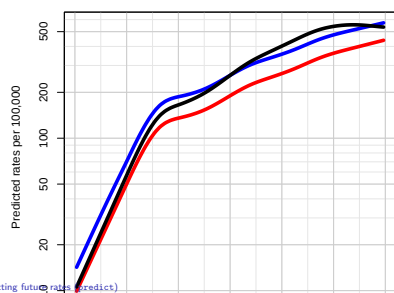
Practicalities

- 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.

Predicting future rates (predict)

311/ 327

Bresat cancer prediction



Predicting future rates (predict)

312/ 327

Predicted age-specific breast cancer rates at 2020 (black),
in the 1950 cohort (blue),
and the estimated age-effects (red).

APC-model for quantitative outcomes

- The classical model is:
$$\log(\lambda(a, p)) = f(a) + g(p) + h(p - a)$$
- In principle it would be possible to use an identity-link model:
$$\lambda(a, p) = f(a) + g(p) + h(p - a)$$
- ... or use APC-modelling for **measurement** data such as BMI, measured at different times and ages:
$$\text{BMI}_{ap} = f(a) + g(p) + h(p - a) + e_{ap}, \quad e_i \sim \mathcal{N}(0, \sigma^2)$$
- ... or more precisely:
$$\text{BMI}_i = f(a(i)) + g(p(i)) + h(p(i) - a(i)) + e_i, \quad e_i \sim \mathcal{N}(0, \sigma^2)$$

Continuous outcomes (cont)

313/ 327

APC-model for quantitative outcomes

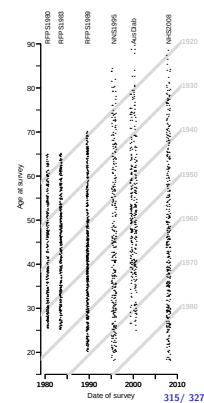
- Model:
$$\text{BMI}_i = f(a(i)) + g(p(i)) + h(p(i) - a(i)) + e_i, \quad e_i \sim \mathcal{N}(0, \sigma^2)$$
- But the identification problem is still the same:
$$c(i) = p(i) - a(i), \quad \forall i$$
- But the same machinery applies with extraction of the effects
- and plotting of predictions of
 - $E(\text{BMI})$
 - quantiles of BMI

Continuous outcomes (cont)

314/ 327

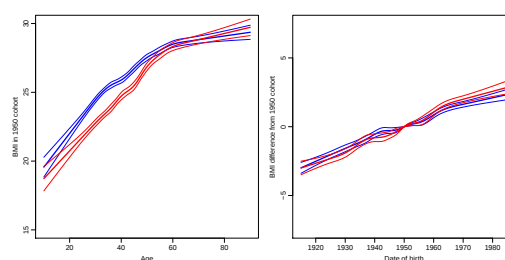
APC-model for quantitative outcomes

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



Continuous outcomes (cont)

315/ 327



Continuous outcomes (cont)

316/ 327

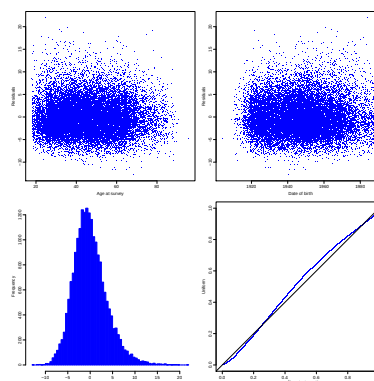
Continuous outcomes

Statistical Analysis in the
Lexis Diagram:

Age-Period-Cohort models
May 2016

Max Planck Institut for Demographic Research, Rostock
<http://BendixCarstensen/APC/MPIDR-2016>

cont



Continuous outcomes (cont)

317/ 327

