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

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

MEB, Karolinska Institutet, Stockholm
May 2010

www.biostat.ku.dk/~bxc/APC/MEB-2010

Introduction
Monday 3rd, afternoon

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm
`www.biostat.ku.dk/~bxc/APC/MEB-2010`

Welcome

- Purpose of the course:
  - Knowledge about APC-models
  - Technincal knowledge of handling them — reporting
- Remedies of the course:
  - Lectures with handouts (BxC)
  - Practicals with suggested solutions

About the practicals

- You should use you preferred **R**-enviroment.
- `Epi`-package for **R** is needed.
- 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.

Essence of APC-models

The literature has mixed up the different components:

- Tabulation of data.
- Specification of model.
- Parametrization of model.

These concepts can *and should* be treated separately.

The purpose of this workshop is to make this clear.

Matrix algebra
Monday 3rd, afternoon

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm
`www.biostat.ku.dk/~bxc/APC/MEB-2010`

Matrices

An $r \times c$ matrix A (reads "an r times c matrix") is a table with r $r$ows og c $c$olumns

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1c} \\ a_{21} & a_{22} & \dots & a_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \dots & a_{rc} \end{bmatrix}$$

Matrices in R

```
> A <- matrix(c(1,3,2,2,8,9),ncol=3)
> A

     [,1] [,2] [,3]
[1,]    1    2    8
[2,]    3    2    9
```

But this is unreadable code; use instead `rbind`:

```
> A <- rbind( c(1,2,8),
+             c(1,2,9) )
> A

     [,1] [,2] [,3]
[1,]    1    2    8
[2,]    1    2    9
```

Transpose of a matrix

Transposing simply interchanges rows and columns.

```
> A

     [,1] [,2] [,3]
[1,]    1    2    8
[2,]    1    2    9

> t(A)

     [,1] [,2]
[1,]    1    1
[2,]    2    2
[3,]    8    9

> t(t(A))

     [,1] [,2] [,3]
[1,]    1    2    8
[2,]    1    2    9
```

Matrix storage

A matrix in **R** is a vector with a `dim` attribute:

```
> z <- 1:8
> z

[1] 1 2 3 4 5 6 7 8

> dim(z) <- c(2,4)
> z

     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8
```

Note that values are stored columnwise:

```
> z

     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8

> as.vector(z)

[1] 1 2 3 4 5 6 7 8

> as.vector(t(z))

[1] 1 3 5 7 2 4 6 8
```

Matrices in R

Multiplication of matrix by a number:

```
> 7*A
```

```
     [,1] [,2] [,3]
[1,]    7   14   56
[2,]    7   14   63
```

Multiplication of matrix by a vector:

```
> A * c(2,7)
```

```
     [,1] [,2] [,3]
[1,]    2    4   16
[2,]    7   14   63
```

Only works if the vector has same length as no. rows in `A`:

```
> as.vector(A)

[1] 1 1 2 2 8 9

> as.vector(A) * c(2,7)

[1]  2  7  4 14 16 63
```

What happens is that `c(2,7)` is recycled to the same length as `A`, i.e. to `c(2,7,2,7,2,7)` and then is multiplied to `A` as vector.

Illustration of storage order of matrices:

```
> A * c(2,7,10)

     [,1] [,2] [,3]
[1,]    2   20   56
[2,]    7    4   90

> t( t(A)*c(2,7,10) )

     [,1] [,2] [,3]
[1,]    2   14   80
[2,]    2   14   90
```

Multiplying a matrix by a vector

$$\begin{bmatrix} 1 & 2 \\ 3 & 8 \\ 2 & 9 \end{bmatrix} \begin{bmatrix} 5 \\ 8 \end{bmatrix} = \begin{bmatrix} 1 \cdot 5 + 2 \cdot 8 \\ 3 \cdot 5 + 8 \cdot 8 \\ 2 \cdot 5 + 9 \cdot 8 \end{bmatrix} = \begin{bmatrix} 21 \\ 79 \\ 82 \end{bmatrix}$$

```
> ( A <- cbind(c(1,3,2),c(2,8,9)) )

     [,1] [,2]
[1,]    1    2
[2,]    3    8
[3,]    2    9

> A %*% c(5,8)

     [,1]
[1,]   21
[2,]   79
[3,]   82
```

Matrix multiplication and multiplication

```
> A %*% c(5,8)

     [,1]
[1,]   21
[2,]   79
[3,]   82

> A * c(5,8)

     [,1] [,2]
[1,]    5   16
[2,]   24   40
[3,]   10   72
```

Multiplying a matrix by a matrix

$$\begin{bmatrix} 1 & 2 \\ 3 & 8 \\ 2 & 9 \end{bmatrix} \begin{bmatrix} 5 & 4 \\ 8 & 2 \end{bmatrix} = \left[\begin{pmatrix} 1 & 2 \\ 3 & 8 \\ 2 & 9 \end{pmatrix} \begin{pmatrix} 5 \\ 8 \end{pmatrix} \vdots \begin{pmatrix} 1 & 2 \\ 3 & 8 \\ 2 & 9 \end{pmatrix} \begin{pmatrix} 4 \\ 2 \end{pmatrix} \right]$$

$$= \begin{bmatrix} 1 \cdot 5 + 2 \cdot 8 & 1 \cdot 4 + 2 \cdot 2 \\ 3 \cdot 5 + 8 \cdot 8 & 3 \cdot 4 + 8 \cdot 2 \\ 2 \cdot 5 + 9 \cdot 8 & 2 \cdot 4 + 9 \cdot 2 \end{bmatrix}$$

$$= \begin{bmatrix} 21 & 8 \\ 79 & 28 \\ 82 & 26 \end{bmatrix}$$

```
> A <- cbind(c(1,3,2),c(2,8,9))
> B <- cbind(c(5,8),c(4,2))
> A
     [,1] [,2]
[1,]    1    2
[2,]    3    8
[3,]    2    9

> B
     [,1] [,2]
[1,]    5    4
[2,]    8    2

> A%*%B
     [,1] [,2]
[1,]   21    8
[2,]   79   28
[3,]   82   26
```

Inverse of a matrix

```
> B

     [,1] [,2]
[1,]    5    4
[2,]    8    2

> ( iB <- solve(B) )

            [,1]       [,2]
[1,] -0.0909091  0.1818182
[2,]  0.3636364 -0.2272727

> iB %*% B

              [,1] [,2]
[1,] 1.000000e+00    0
[2,] 2.220446e-16    1

> round( iB %*% B, 6 )

     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

Generalized inverse of a matrix

There is also a (non-unique) inverse for singular matrices (the Moore-Penrose inverse):

```
> library( MASS )
> A

     [,1] [,2]
[1,]    1    2
[2,]    3    8
[3,]    2    9

> ( iA <- ginv(A) )

            [,1]       [,2]       [,3]
[1,]  0.4066667  0.6333333 -0.6533333
[2,] -0.1066667 -0.1333333  0.2533333

> iA %*% A

     [,1]          [,2]
[1,]    1 -7.771561e-16
[2,]    0  1.000000e+00
```

A simple confidence interval

Suppose you have fitted a model and want to report a confidnece interval for your estimate of β, and that you have:

$$\hat{\beta} = 2.57, \quad \text{s.e.}(\hat{\beta}) = 0.87$$

Then the confidence interval is

$$(\hat{\beta} - 1.96 \times \text{s.e.}(\hat{\beta}); \hat{\beta} + 1.96 \times \text{s.e.}(\hat{\beta}))$$

A simple confidence interval

The estimate and the confidence interval can be recognized as the matrix multiplication: the matrix:

$$(\hat{\beta}, \text{s.e.}(\hat{\beta})) \begin{pmatrix} 1 & 1 \\ 0 & -1.96 & 1.96 \end{pmatrix}$$
$$= (\hat{\beta}, \hat{\beta} - 1.96 \times \text{s.e.}(\hat{\beta}), \hat{\beta} + 1.96 \times \text{s.e.}(\hat{\beta}))$$

Several confidence intervals

This generalizes to the confidence interval for several parameters in one go:

$$\begin{pmatrix} \hat{\beta}_1 & \text{s.e.}(\hat{\beta}_1) \\ \hat{\beta}_2 & \text{s.e.}(\hat{\beta}_2) \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & -1.96 & 1.96 \end{pmatrix}$$

$$= \begin{pmatrix} \hat{\beta}_1 & \hat{\beta}_1 - 1.96 \times \text{s.e.}(\hat{\beta}_1) & \hat{\beta}_1 + 1.96 \times \text{s.e.}(\hat{\beta}_1) \\ \hat{\beta}_2 & \hat{\beta}_2 - 1.96 \times \text{s.e.}(\hat{\beta}_2) & \hat{\beta}_2 + 1.96 \times \text{s.e.}(\hat{\beta}_2) \end{pmatrix}$$

The `ci.mat()`

The `Epi` package has the function `ci.mat` which returns the relevant matrix. It has the significance level as argument

```
> library(Epi)
> ci.mat()

     Estimate       2.5%    97.5%
[1,]        1  1.000000 1.000000
[2,]        0 -1.959964 1.959964

> ci.mat(0.1)

     Estimate       5.0%    95.0%
[1,]        1  1.000000 1.000000
[2,]        0 -1.644854 1.644854
```

This is used to produce confidence intervals:

```
> data(births)
> m1 <- lm(bweight~gestwks+matage,data=births)
> ( beta <- summary(m1)$coef[-1,2:3] )

         Std. Error     t value
gestwks   8.799868 22.38340601
matage    5.222086  0.00735004

> beta %*% ci.mat()

         Estimate      2.5%     97.5%
gestwks 8.799868 -35.07080 52.670538
matage  5.222086   5.20768  5.236491
```

Confidence intervals for curves
Monday 3rd, afternoon

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm
`www.biostat.ku.dk/~bxc/APC/MEB-2010`

Confidence limits for a line

$$\hat{\beta}_0 + \hat{\beta}_1 x, \quad x = x_1, \ldots, x_n$$

With s.e.:

$$\sqrt{\text{var}(\beta_0) + \text{var}(\beta_1)x^2 + x\text{cov}(\beta_0, \beta_1)}$$

or

$$\sqrt{(1 \quad x)\Sigma \begin{pmatrix} 1 \\ x \end{pmatrix}}$$

where Σ is the variance-covariance matrix of (β_0, β_1). Available using `predict` or `fitted`

Confidence limits for contrasts on a line

$$\hat{\beta}_0 + \hat{\beta}_1 x - (\hat{\beta}_0 + \hat{\beta}_1 x_{\text{ref}}) = \hat{\beta}_1 (x - x_{\text{ref}})$$

which has standard error:

$$\text{s.e.}(\hat{\beta}_1)(x - x_{\text{ref}})$$

Not available from `predict` — it is not a prediction; it is a contrast (or rather, many contrasts, one for each x_i where we want an estimate of the contrast)

Confidence limits for contrasts on a curve

$$\begin{aligned}
&\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 - (\hat{\beta}_0 + \hat{\beta}_1 x_{\text{ref}} + \hat{\beta}_2 x_{\text{ref}}^2) \\
&= \hat{\beta}_1 (x - x_{\text{ref}}) + \hat{\beta}_1 (x^2 - x_{\text{ref}}^2) \\
&= (x - x_{\text{ref}} \quad x^2 - x_{\text{ref}}^2) \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}
\end{aligned}$$

which has standard error:

$$\sqrt{(x - x_{\text{ref}} \quad x^2 - x_{\text{ref}}^2) \Sigma \begin{pmatrix} x - x_{\text{ref}} \\ x^2 - x_{\text{ref}}^2 \end{pmatrix}}$$

Not available from `predict` — it is not a prediction; it is a contrast (or rather, many contrasts, one for each x_i)

Confidence limits for contrasts on a curve

The matrix that we pre- multiply to the parameter estimates and pre- and post-multiply to the variance-covariance matrix is:

$$(x - x_{\text{ref}} \quad x^2 - x_{\text{ref}}^2) = (x \quad x^2) - (x_{\text{ref}} \quad x_{\text{ref}}^2)$$

The first one is the design matrix for the curve —evaluated at the prediction points,
the second one is the design matrix for the curve —evaluated at the reference point.

Confidence limits for contrasts on a curve

Needed:

- The parameters, (β_1, β_2)
- Their variance-covariance matrix, Σ
- The contrast matrix — *i.e.* the desired linear function of the parameters.

`ci.lin` from the `Epi` package provides the tools.

Confidence limits for a curve

The `births` example: Birthweight as a function of gestational age:

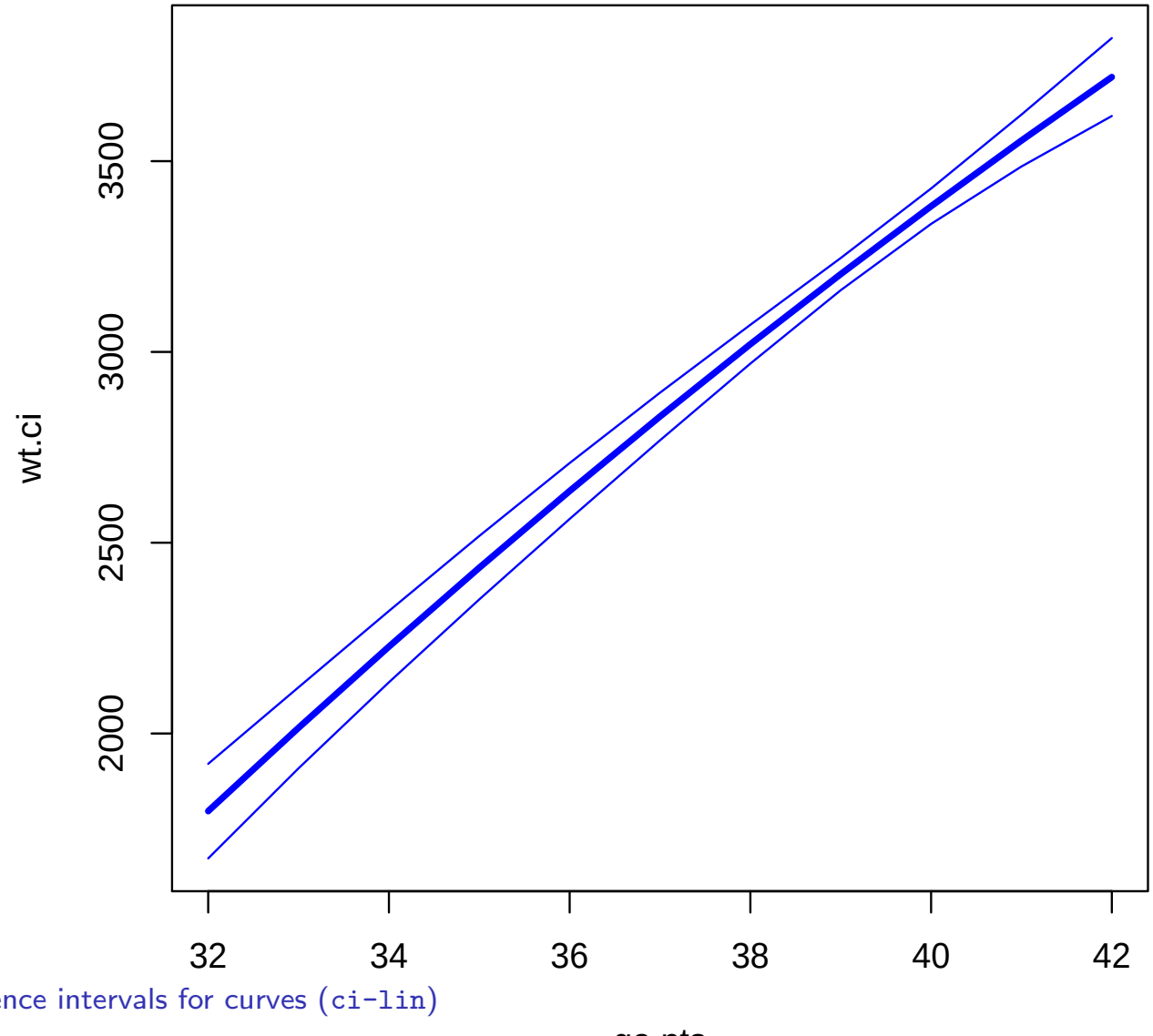
```
> library(Epi)
> data(births)
> ma <- lm( bweight ~ gestwks + I(gestwks^2), data=births )
> summary( ma )$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8247.693621	2303.735213	-3.580140	0.0003778475
gestwks	406.461711	127.296455	3.193032	0.0014989838
I(gestwks^2)	-2.893052	1.753789	-1.649601	0.0996694332

Confidence limits for a curve

```
> ga.pts <- 32:42
> G <- cbind( 1, ga.pts, ga.pts^2 )
> wt.ci <- ci.lin( ma, ctr.mat=G )[,c(1,5,6)]
> matplot( ga.pts, wt.ci, type="l", lty=1, lwd=c(3,1,1), col="bl
```

Confidence limits for a curve



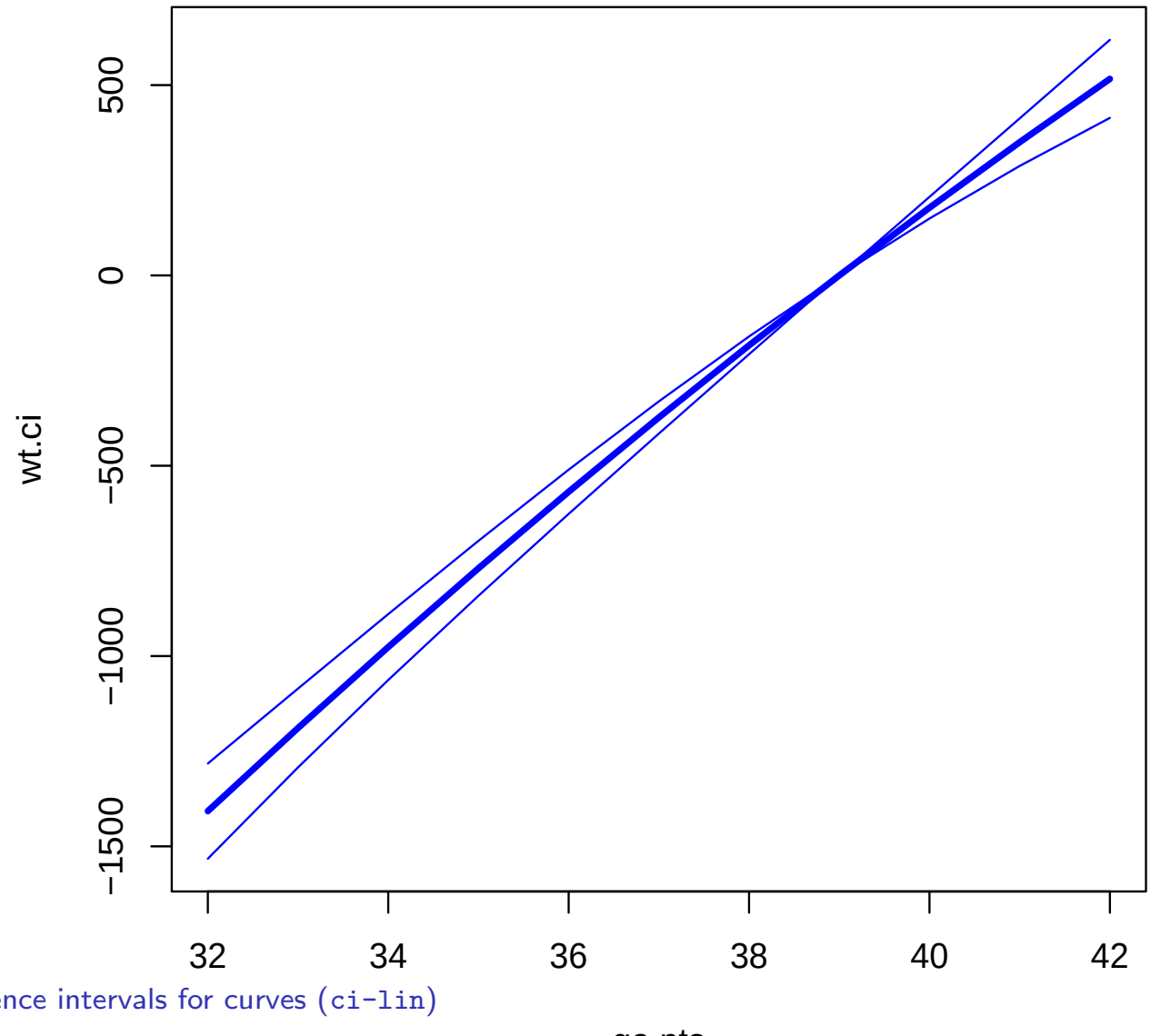
Confidence limits for contrasts on a curve

If we want the differences relative to a reference of say 39 weeks of gestation:

- set up a row with this value.
- subtract a matrix with identical replicates of this from the first contrast matrix.

```
> G.ref <- cbind( 1, 39, 39^2 )
> wt.ci <- ci.lin( ma, ctr.mat=G-G.ref[rep(1,nrow(G)),] )[,c(1,5
> matplot( ga.pts, wt.ci, type="l", lty=1, lwd=c(3,1,1), col="bl
```

Confidence limits for contrasts on a curve



Likelihood for rates
Monday 3rd, afternoon

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm
`www.biostat.ku.dk/~bxc/APC/MEB-2010`

Likelihood from one person

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

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

The likelihood is a product of *conditional* probabilities, not of *independent* entities.

Each term refers to one empirical rate (d, y):
$d = 1\{\text{event in } (t_{i-1}, t_i)\}$ and $y = t_i - t_{i-1}$.

Likelihood for an empirical rate

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

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

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

since the term in $()$ equal to $\mathrm{e}^{\lambda y} - 1 \approx \lambda y$.

Log-likelihood:

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

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

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

Note that the log-likelihood for one Poisson observation d with mean λy is:

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

Poisson likelihood

The contributions from **one** individual:

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

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

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

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

Likelihood for follow-up of many subjects

Adding empirical rates over the follow-up of persons:

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

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

The log-likelihood is maximal for:

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

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

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

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

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

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

Confidence interval for a rate

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

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

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

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

Calculation of rate

Suppose you have 15 events during 5532 person-years. The rate and c.i. is then computed by:

```
> D <- 15
> Y <- 5532
> rate <- D / Y
> erf <- exp( 1.96 / sqrt(D) )
> c( rate, rate/erf, rate*erf )

[1] 0.002711497 0.001634654 0.004497720

> exp( c( log(D/Y), 1/sqrt(D) ) %*% ci.mat() )

        Estimate         2.5%       97.5%
[1,] 0.002711497 0.001634669 0.004497678
```

Calculation of rate

This can also be achieved in a Poisson model:

```
> mm <- glm( D ~ 1, offset=log(Y), family=poisson )
> ci.lin( mm, E=T)[,5:7]

  exp(Est.)         2.5%       97.5%
0.002711497 0.001634669 0.004497678

> #
> # Use an additive model
> ma <- glm( D/Y ~ 1, weight=Y, family=poisson(link=identity) )
> ci.lin( ma )[,c(1,5,6)]

   Estimate         2.5%       97.5%
0.002711497 0.001339315 0.004083678
```

Why are the confidence limits not the same?

Ratio of two rates

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

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

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

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

Calculation of rate-ratio

Suppose you have 15 events during 5532 person-years in the un-exposed group and 28 events during 4783 person-years in the the exposed group:

The rate-ratio and c.i. is then computed by:

```
> D0 <- 15   ; D1 <- 28
> Y0 <- 5532 ; Y1 <- 4783
> RR <- (D1/Y1)/(D0/Y0)
> erf <- exp( 1.96 * sqrt(1/D0+1/D1) )
> c( RR, RR/erf, RR*erf )

[1] 2.158980 1.153146 4.042153

> exp( c( log(RR), sqrt(1/D0+1/D1) ) %*% ci.mat() )

     Estimate     2.5%    97.5%
[1,] 2.158980 1.153160 4.042106
```

Calculation of RR and RD

This can also be achieved in a Poisson model:

```
> D <- c(D0,D1) ; Y <- c(Y0,Y1); xpos <- 0:1
> mm <- glm( D ~ factor(xpos), offset=log(Y), family=poisson )
> ci.lin( mm, E=T)[,5:7]

                 exp(Est.)        2.5%       97.5%
(Intercept)   0.002711497 0.001634669 0.004497678
factor(xpos)1 2.158979720 1.153159560 4.042106222

> # Use an additive model to get rate-difference
> ma <- glm( D/Y ~ factor(xpos), weight=Y,
+            family=poisson(link=identity) )
> ci.lin( ma )[,c(1,5,6)]

                  Estimate         2.5%       97.5%
(Intercept)   0.002711497 0.0013393153 0.004083678
factor(xpos)1 0.003142570 0.0005765288 0.005708611
```

Calculation of rates and RR

We can use `ctr.mat` to get it all in one go:

```
> CM <- rbind( c(1,0), c(1,1), c(0,1) )
> rownames( CM ) <- c("rate 0","rate 1","RR 1 vs. 0")
> CM

           [,1] [,2]
rate 0        1    0
rate 1        1    1
RR 1 vs. 0    0    1

> mm <- glm( D ~ factor(xpos), offset=log(Y), family=poisson )
> ci.lin( mm, ctr.mat=CM, E=T)[,5:7]

             exp(Est.)        2.5%       97.5%
rate 0     0.002711497 0.001634669 0.004497678
rate 1     0.005854066 0.004041994 0.008478512
RR 1 vs. 0 2.158979720 1.153159560 4.042106222

> round( ci.lin( mm, ctr.mat=CM, E=T), 3 )

           Estimate StdErr       z     P exp(Est.)  2.5% 97.5%
rate 0       -5.910  0.258 -22.890 0.000      0.003 0.002 0.004
rate 1       -5.141  0.189 -27.202 0.000      0.006 0.004 0.008
RR 1 vs. 0    0.770  0.320   2.405 0.016      2.159 1.153 4.042
```

Calculation of rates and RD

We can use `ctr.mat` to get it all in one go:

```
> rownames( CM ) <- c("rate 0","rate 1","RD 1 vs. 0")
> ma <- glm( D/Y ~ factor(xpos), weight=Y,
+            family=poisson(link=identity) )
> ci.lin( ma, ctr.mat=CM )[,c(1,5,6)]

               Estimate          2.5%       97.5%
rate 0     0.002711497 0.0013393153 0.004083678
rate 1     0.005854066 0.0036857298 0.008022403
RD 1 vs. 0 0.003142570 0.0005765288 0.005708611

> round( ci.lin( ma, ctr.mat=CM ), 3 )

           Estimate StdErr     z     P  2.5% 97.5%
rate 0        0.003  0.001 3.873 0.000 0.001 0.004
rate 1        0.006  0.001 5.292 0.000 0.004 0.008
RD 1 vs. 0    0.003  0.001 2.400 0.016 0.001 0.006
```

Confidence intervals for rates
Monday 3rd, afternoon

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm
`www.biostat.ku.dk/~bxc/APC/MEB-2010`

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

'data.frame': 55 obs. of  4 variables:
 $ age   : num  25 25 25 25 25 30 30 30 30 30 ...
 $ period: num  1955 1960 1965 1970 1975 ...
 $ D     : num  3 3 1 4 12 16 17 11 8 8 ...
 $ Y     : num  1e+07 1e+07 1e+07 1e+07 1e+07 ...

> bl <- transform( blcaIT, A=age+2.5, P=period+2.5 )
> head( bl )

  age period  D        Y    A      P
1  25   1955  3 10000000 27.5 1957.5
2  25   1960  3 10000000 27.5 1962.5
3  25   1965  1 10000000 27.5 1967.5
4  25   1970  4 10000000 27.5 1972.5
5  25   1975 12 10000000 27.5 1977.5
6  30   1955 16  9411765 32.5 1957.5
```

```
> lAP <- glm( D ~ -1 + factor(A) + P,
+             offset=log(Y/10^5),
+             family=poisson,
+             data=bl )
> round( ci.lin( lAP ), 3 )

               Estimate StdErr       z P    2.5%   97.5%
factor(A)27.5   -59.868  1.408 -42.525 0 -62.627 -57.108
factor(A)32.5   -58.843  1.398 -42.084 0 -61.584 -56.103
factor(A)37.5   -57.864  1.395 -41.483 0 -60.598 -55.130
factor(A)42.5   -56.743  1.393 -40.720 0 -59.474 -54.012
factor(A)47.5   -55.750  1.393 -40.023 0 -58.480 -53.019
factor(A)52.5   -54.841  1.393 -39.378 0 -57.571 -52.111
factor(A)57.5   -54.127  1.392 -38.872 0 -56.856 -51.398
factor(A)62.5   -53.565  1.393 -38.454 0 -56.295 -50.835
factor(A)67.5   -53.107  1.393 -38.118 0 -55.838 -50.377
factor(A)72.5   -52.794  1.393 -37.897 0 -55.525 -50.064
factor(A)77.5   -52.544  1.393 -37.722 0 -55.274 -49.813
P                 0.029  0.001  40.802 0   0.027   0.030
```

```
> # Model with quadratic effect of period
> qAP <- glm( D ~ -1 + factor(A) + P + I(P^2),
+             offset=log(Y/10^5),
+             family=poisson,
+             data=bl )
> #
> # Prediction points for age and period
> A.pr <- unique(bl$A)
> P.pr <- 1955:1980
> #
> # Period reference - ALWAYS in a variable, you may want to cha
> P.ref <- 1970
> #
> # Contrast matrix (design matrix for age effects
> CA <- model.matrix( ~ -1 + factor(A.pr) )
> #
> # Contrast matrices for period effects
>   CP     <- cbind( P.pr , P.pr^2 )
> ( CP.ref <- cbind( P.ref, P.ref^2 ) )

     P.ref
[1,]  1970 3880900
```

The rates in 1970 — note how the `CP.ref` is expandend to match the no. of rows in `CA`.

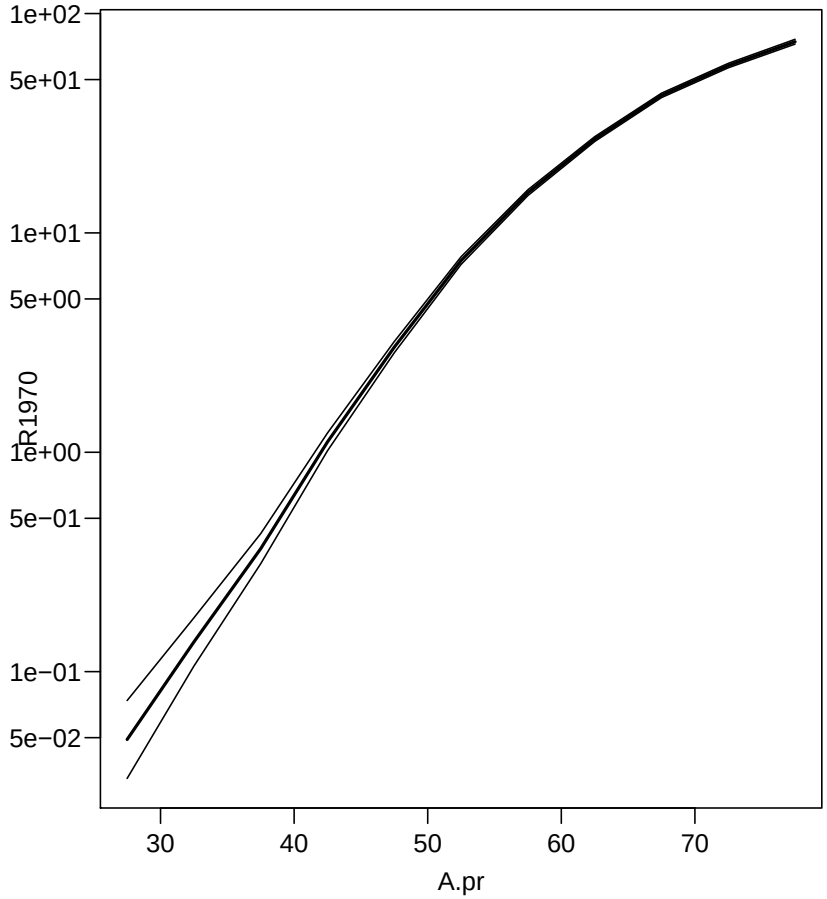
```
> R1970 <-
+ ci.lin( qAP, subset=c("A","P"),
+          ctr.mat=cbind(CA,CP.ref[rep(1,nrow(CA)),]),
+          Exp=TRUE )[,5:7]
```

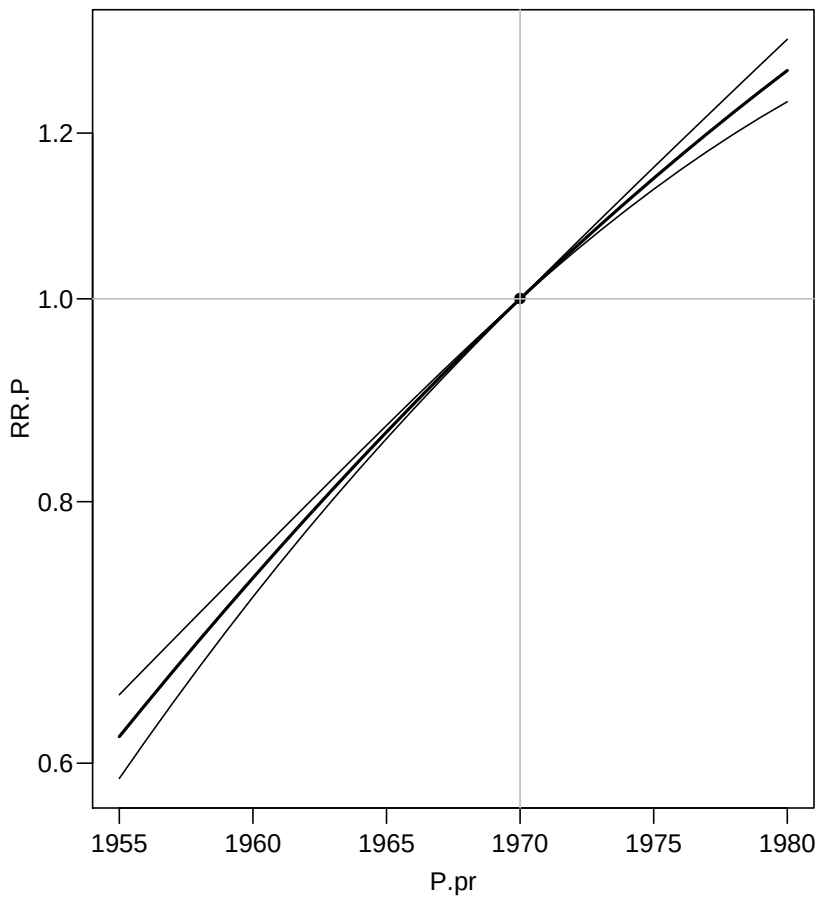
The RR relative to 1970 — note how the `CP.ref` is expandend to match the no. of rows in `CP`.

```
> RR.P  <-
+ ci.lin( qAP, subset=       "P" ,
+          ctr.mat=      CP-CP.ref[rep(1,nrow(CP)),] ,
+          Exp=TRUE )[,5:7]
```

```
> # Now plot the rates and the RRs
> #
> # Set up a 1 by two layout for the graphs
> par(mfrow=c(1,2),mar=c(3,3,1,1),mgp=c(3,1,0)/1.6,las=1)
> #
> # First plot is the rates in 1970
> matplot( A.pr, R1970,
+          type="l", col="black", lty=1, log="y", lwd=c(2,1,1) )
> #
> # Second plot is the RRs relative to 1970
> matplot( P.pr,  RR.P,
+          type="l", col="black", lty=1, log="y", lwd=c(2,1,1) )
> points(1970,1,pch=16)
> abline(h=1,v=1970,col="gray")
```

Predicted rates and RRs





Age-drift model
Tuesday 4th, morning

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm
`www.biostat.ku.dk/~bxc/APC/MEB-2010`

Linear effect of period:

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

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

Linear effect of cohort:

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

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

Age and linear effect of period:

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

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

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

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

(Dispersion parameter for poisson family taken to be 1)

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

Age and linear effect of cohort:

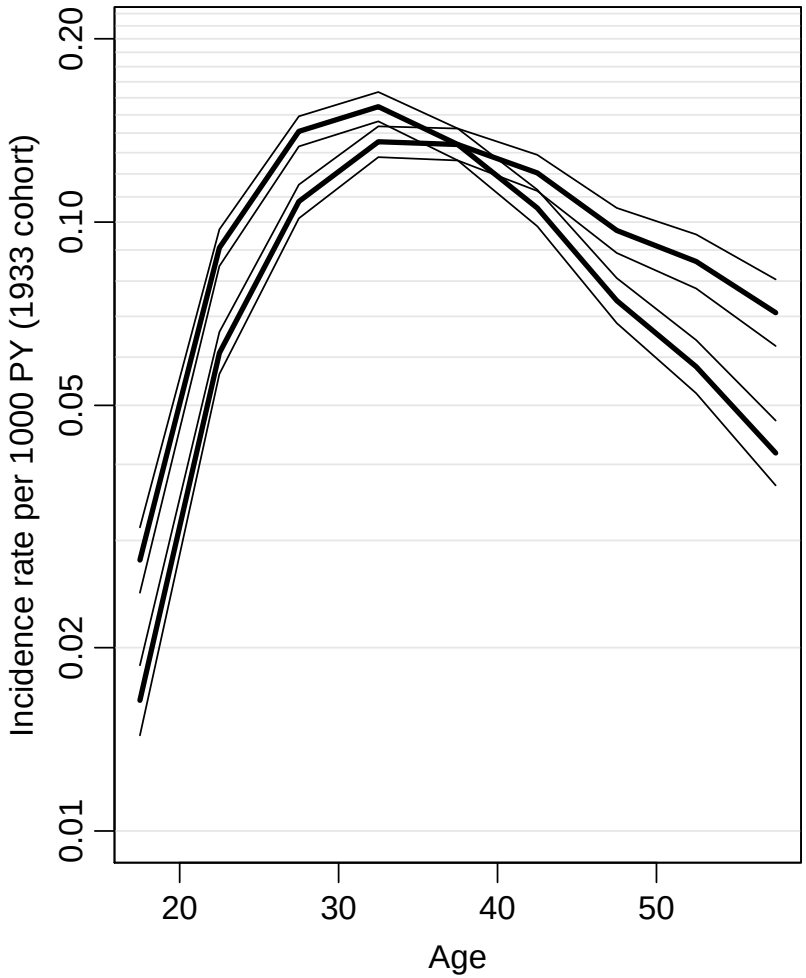
```
> acd <- glm( D ~ factor( A ) - 1 + I(C-1933) +
+             offset( log( Y ) ),
+             family=poisson )
> summary( acd )

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

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

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

(Dispersion parameter for poisson family taken to be 1)

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

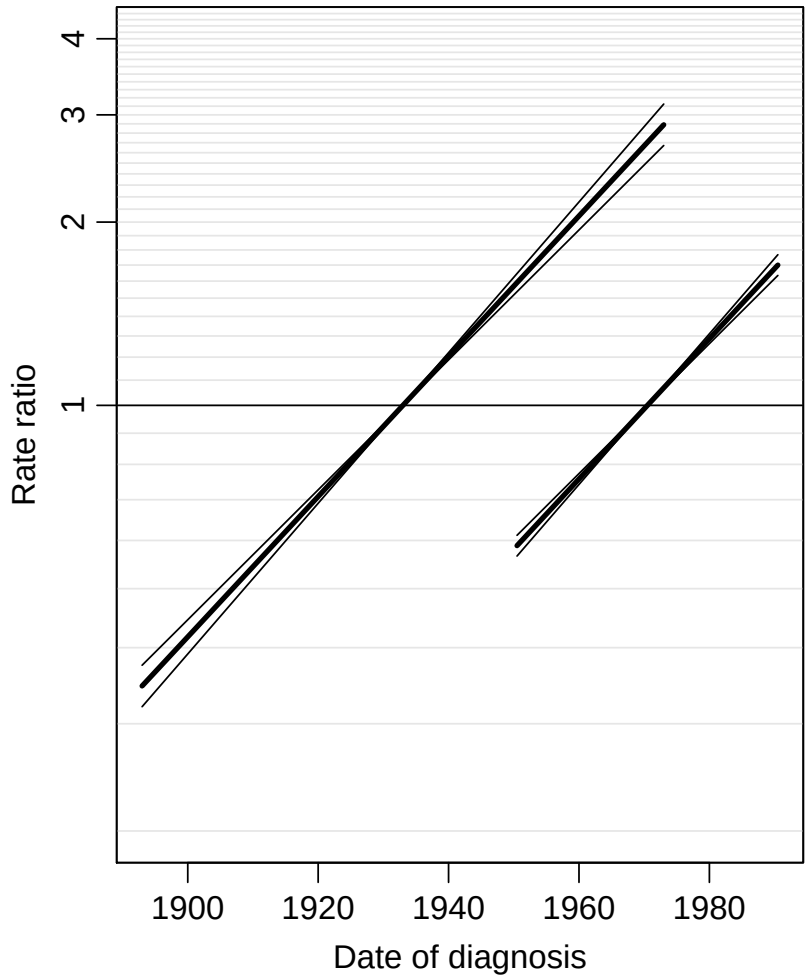
What goes on?

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

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

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





Which age-curve is period and which is cohort?

Age-Period-Cohort model
Tuesday 4th, morning

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm
`www.biostat.ku.dk/~bxc/APC/MEB-2010`

The age-period-cohort model

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

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

Fitting the model in R

```
> c1933.p <- glm( D ~ factor( A ) - 1 +
+                     relevel( factor( C ), "1933" ) +
+                     factor( P ) + offset( log( Y ) ), family=p
> summary( c1933.p )
Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|
factor(A)17.5                 -4.27754    0.10479 -40.819  < 2e
...
factor(A)57.5                 -2.75892    0.08380 -32.922  < 2e
relevel(factor(C), "1933")1893 -0.84187    0.28009  -3.006  0.00
...
relevel(factor(C), "1933")1928 -0.17922    0.05965  -3.005  0.00
relevel(factor(C), "1933")1938  0.07540    0.05592   1.348  0.17
...
relevel(factor(C), "1933")1973  1.37438    0.17490   7.858 3.90e
factor(P)1955.5                 0.04793    0.07022   0.683  0.49
...
factor(P)1985.5                 0.09276    0.04091   2.267  0.02
factor(P)1990.5                      NA         NA      NA

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 89358.526  on 81  degrees of freedom
Residual deviance:    35.459  on 49  degrees of freedom
```

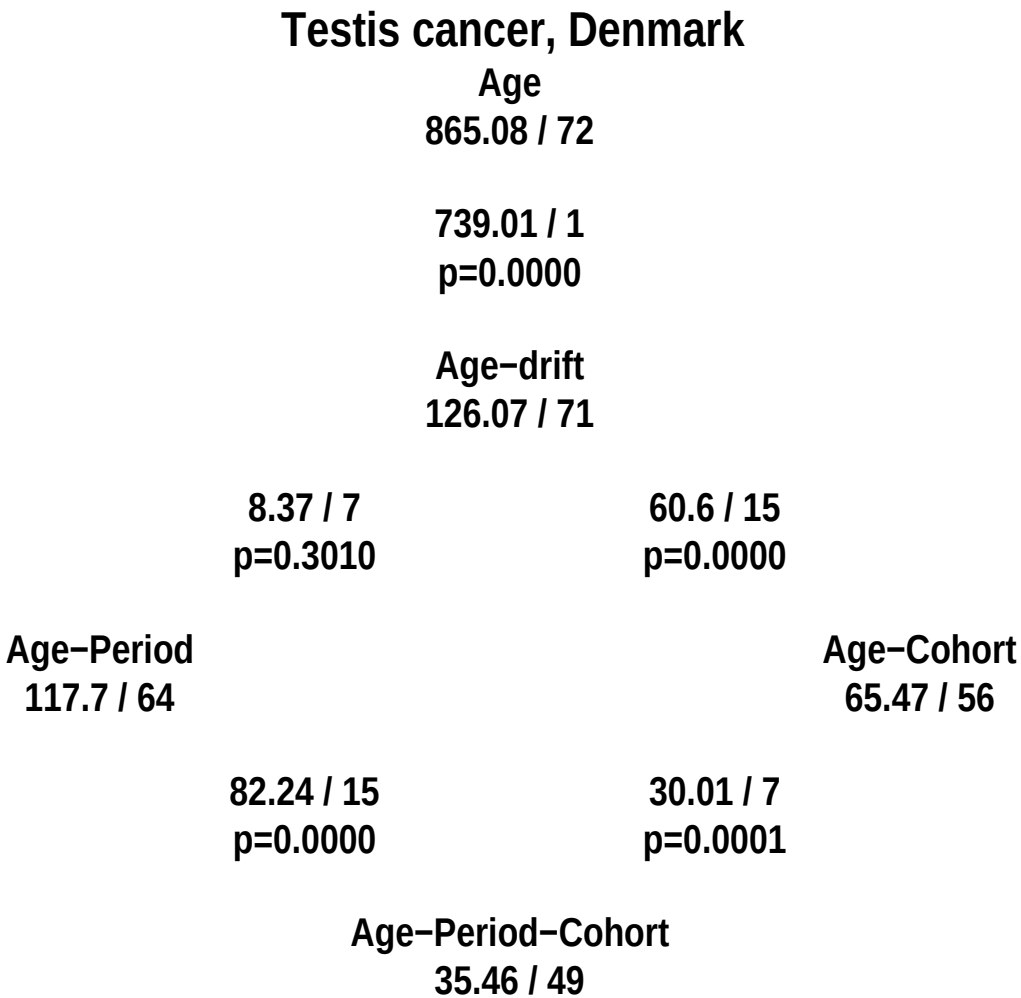
No. of parameters

A has 9 levels
P has 9 levels
C has 17 levels

Age-drift model has $A + 1 = 10$ parameters
Age-period model has $A + P - 1 = 17$ parameters
Age-cohort model has $A + C - 1 = 25$ parameters
Age-period-cohort model has $A + P + C - 3 = 32$ parameters

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

Relationship of models



Test for effects

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

How to choose a parametrization

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

Holford: Extract linear effects by regression:

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

Identifiable drift

Since $p - a - c = 0$, we can freely add $\zeta(p - a - c)$ to the r.h.s. of the model specification, changing the slopes of each of the terms, and so get:

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

Hence, the "identifiable" drifts are $\delta_a + \delta_p$ and $\delta_p + \delta_c$.

The latter is often taken as the "true" secular drift.

Putting it together again

Assumptions are needed, e.g.:

- Age is the major time scale.
- Cohort is the secondary time scale (the major secular trend).
- c_0 is the reference cohort.
- Period is the residual time scale: 0 on average, 0 slope.

Period effect, on average 0, slope is 0:

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

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

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

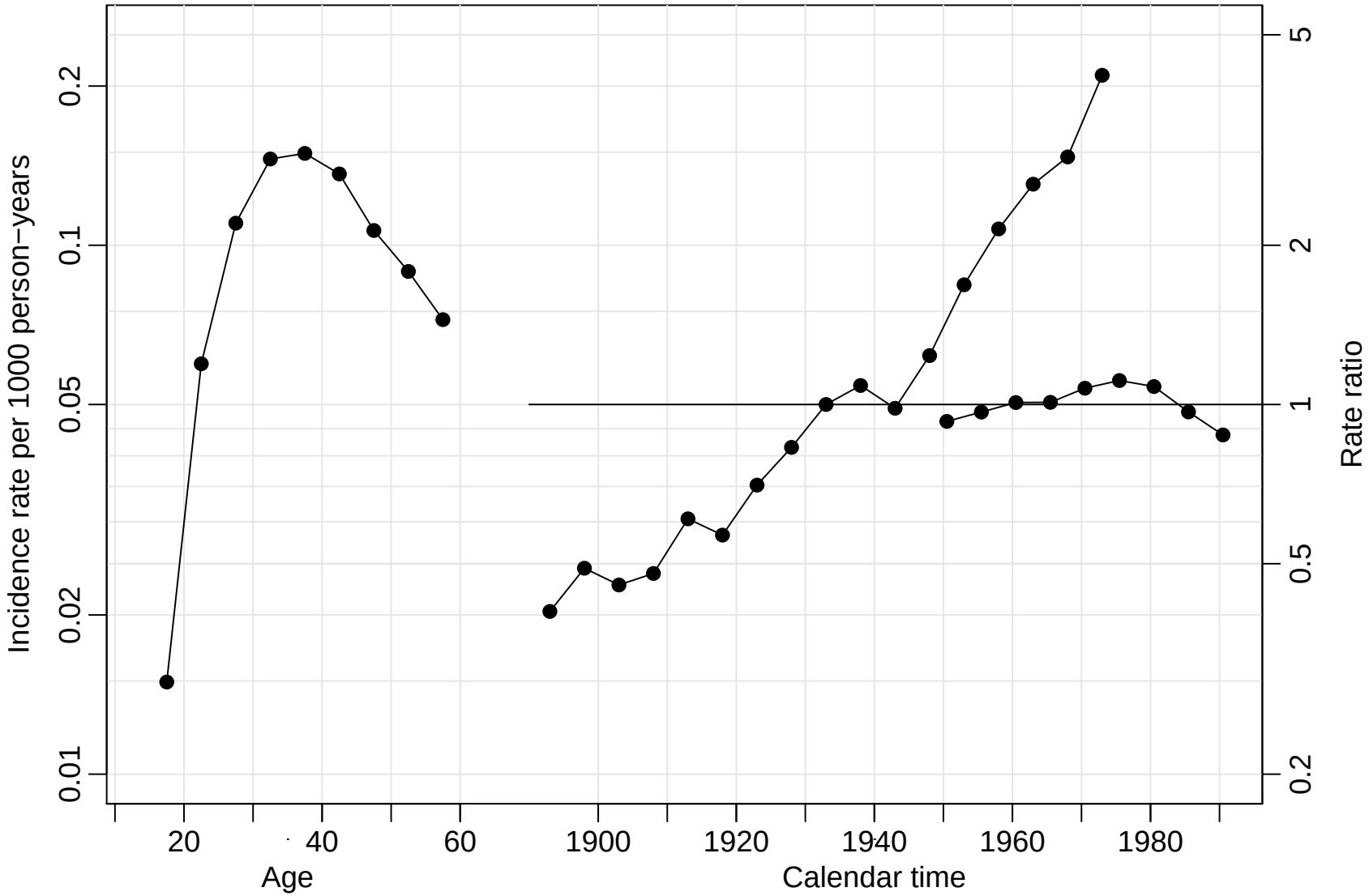
The rest is the age-effect:

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

How it adds up:

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

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


Incidence rate per 1000 person−years
0.01
0.02
0.05
0.1
0.2
20
40
60
Age
1900
1920
1940
1960
1980
Calendar time
0.2
0.5
1
2
5
Rate ratio

A simple practical approach

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

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

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

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

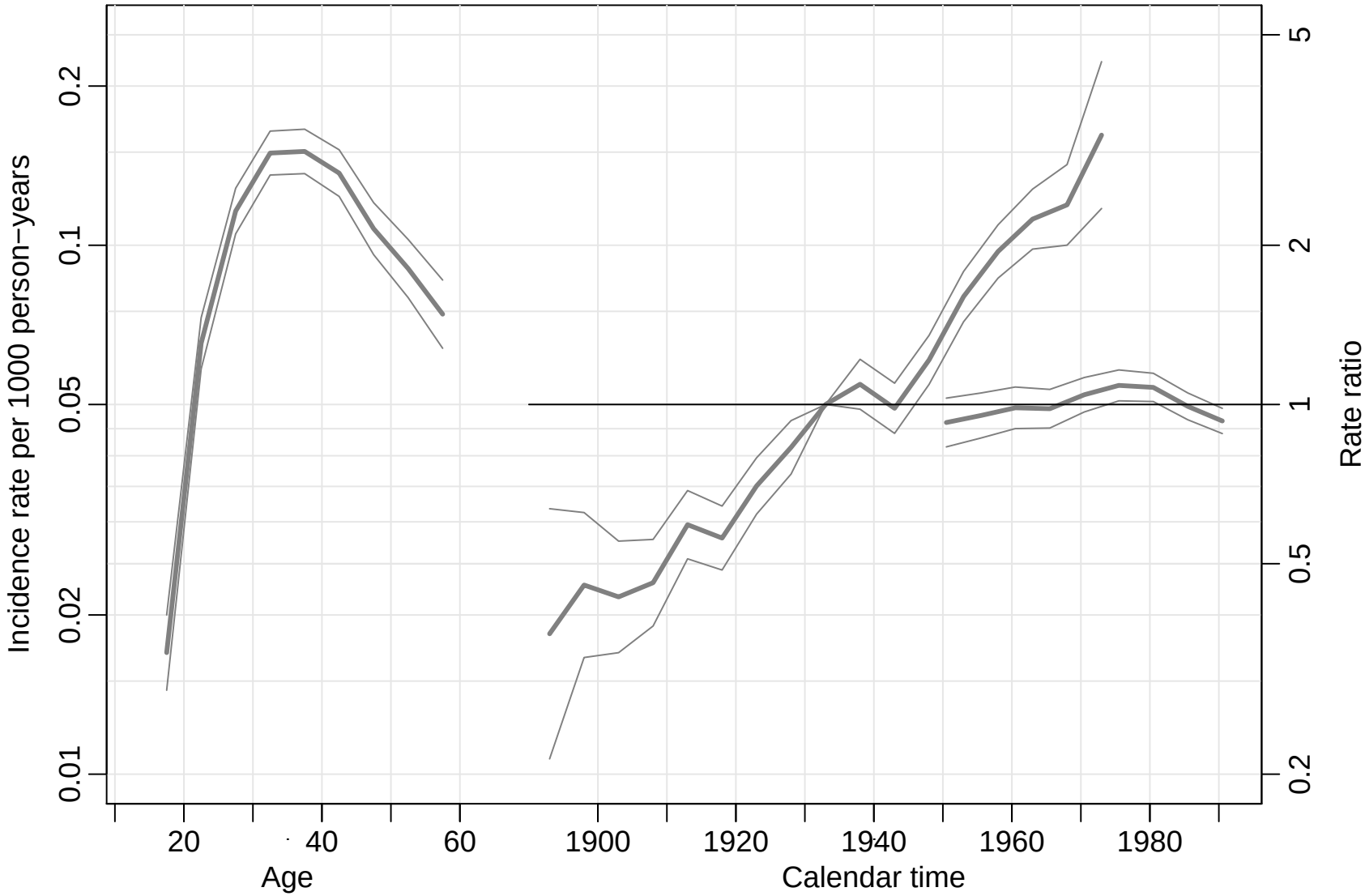
The residual period effect can be estimated if we note that for the number of cases we have:

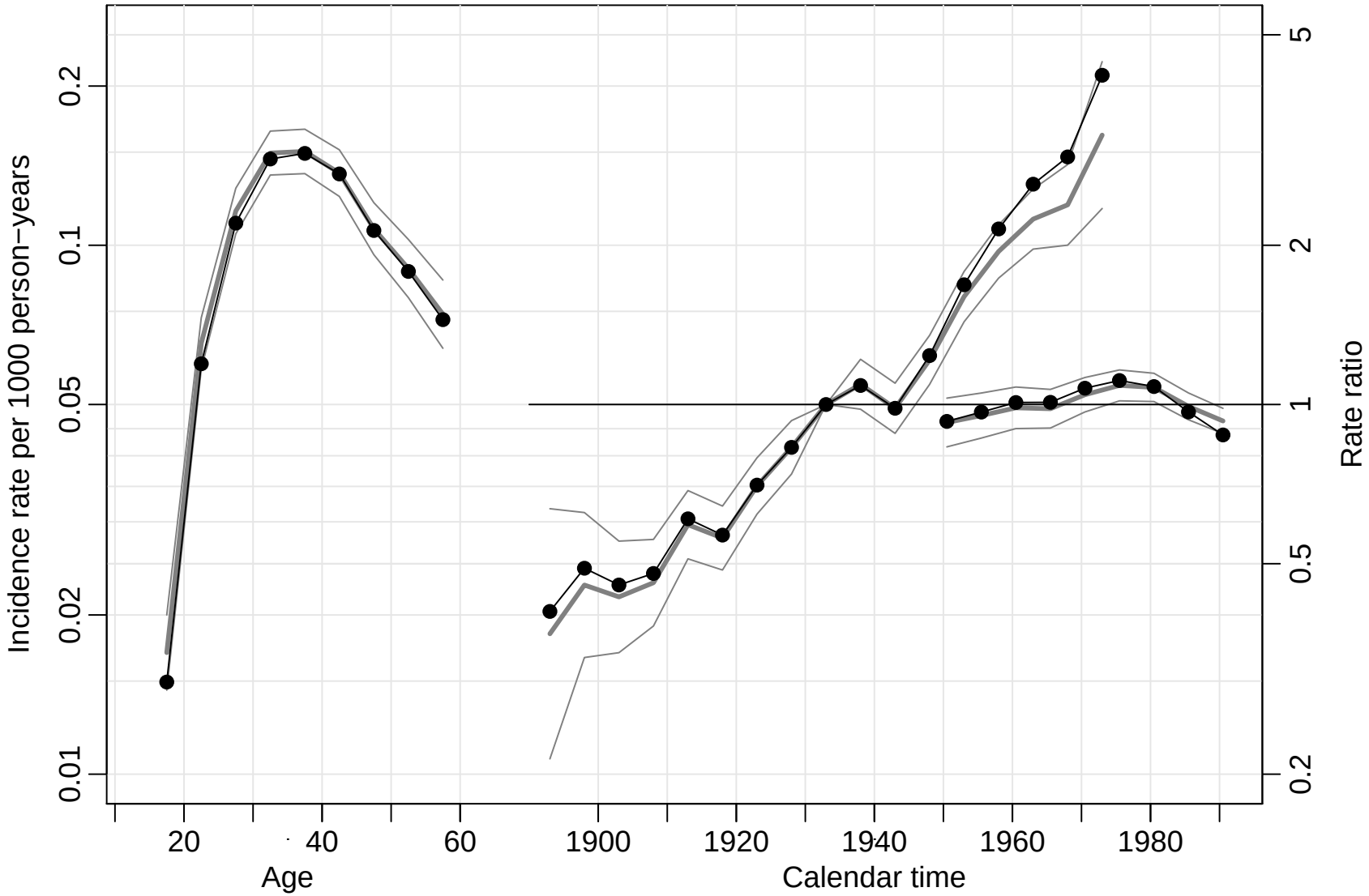
$$\log(\text{expected cases}) = \log[\lambda(a,p)\,Y] = \underbrace{\hat{\alpha}_a + \hat{\gamma}_c + \log(Y)}_{\text{“known”}}$$

This is analogous to the expression for a Poisson model in general, but now is the offset not just $\log(Y)$ but $\hat{\alpha}_a + \hat{\gamma}_c + \log(Y)$, the log of the fitted values from the age-cohort model.

β_ps are estimated in a Poisson model with this as offset.

Advantage: We get the standard errors for free..


Incidence rate per 1000 person−years
0.01
0.02
0.05
0.1
0.2
20
40
60
Age
1900
1920
1940
1960
1980
Calendar time
0.2
0.5
1
2
5
Rate ratio


Incidence rate per 1000 person−years
0.01
0.02
0.05
0.1
0.2
20
40
60
Age
1900
1920
1940
1960
1980
Calendar time
0.2
0.5
1
2
5
Rate ratio

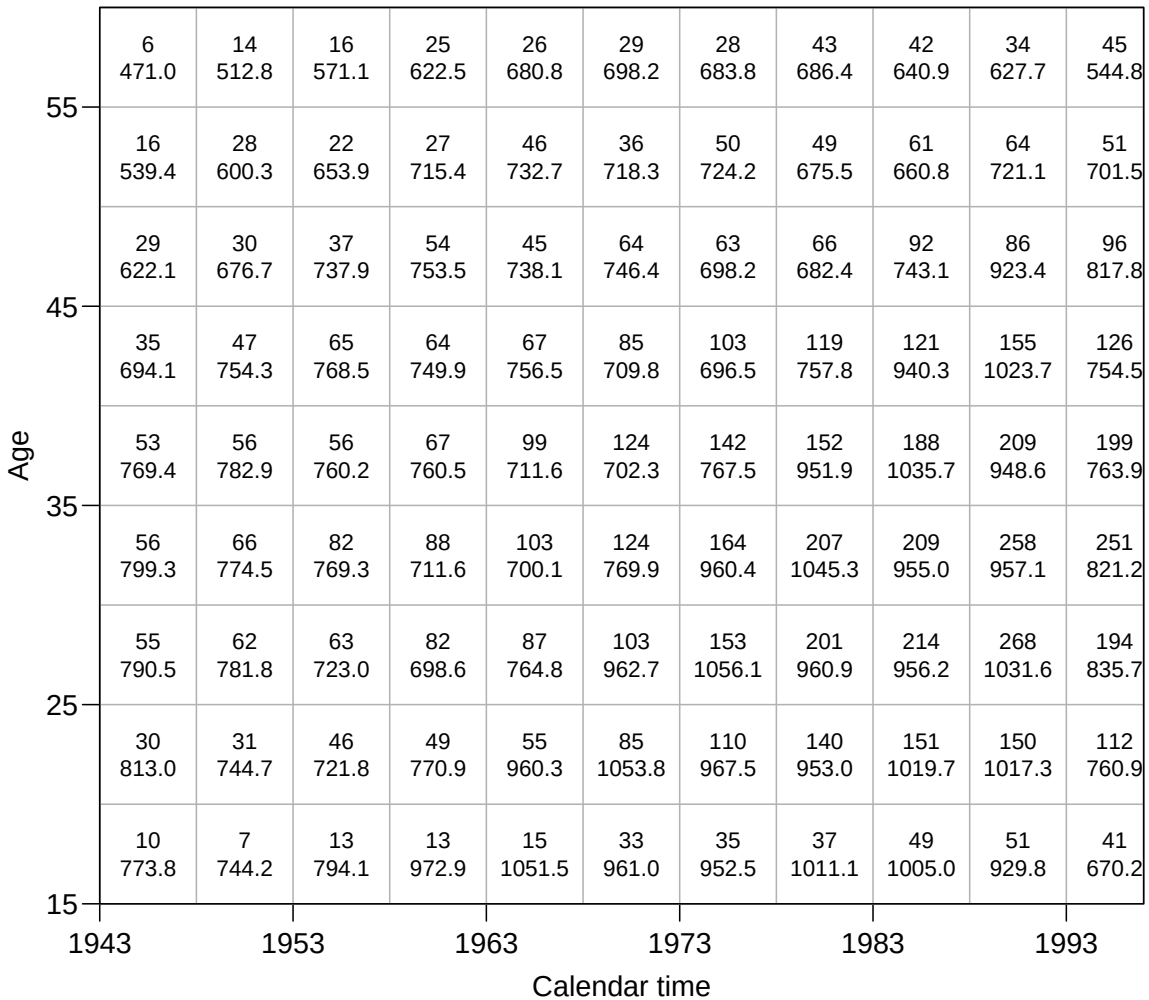
Tabulation in the Lexis diagram

Tuesday 4th, afternoon

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm
www.biostat.ku.dk/~bxc/APC/MEB-2010

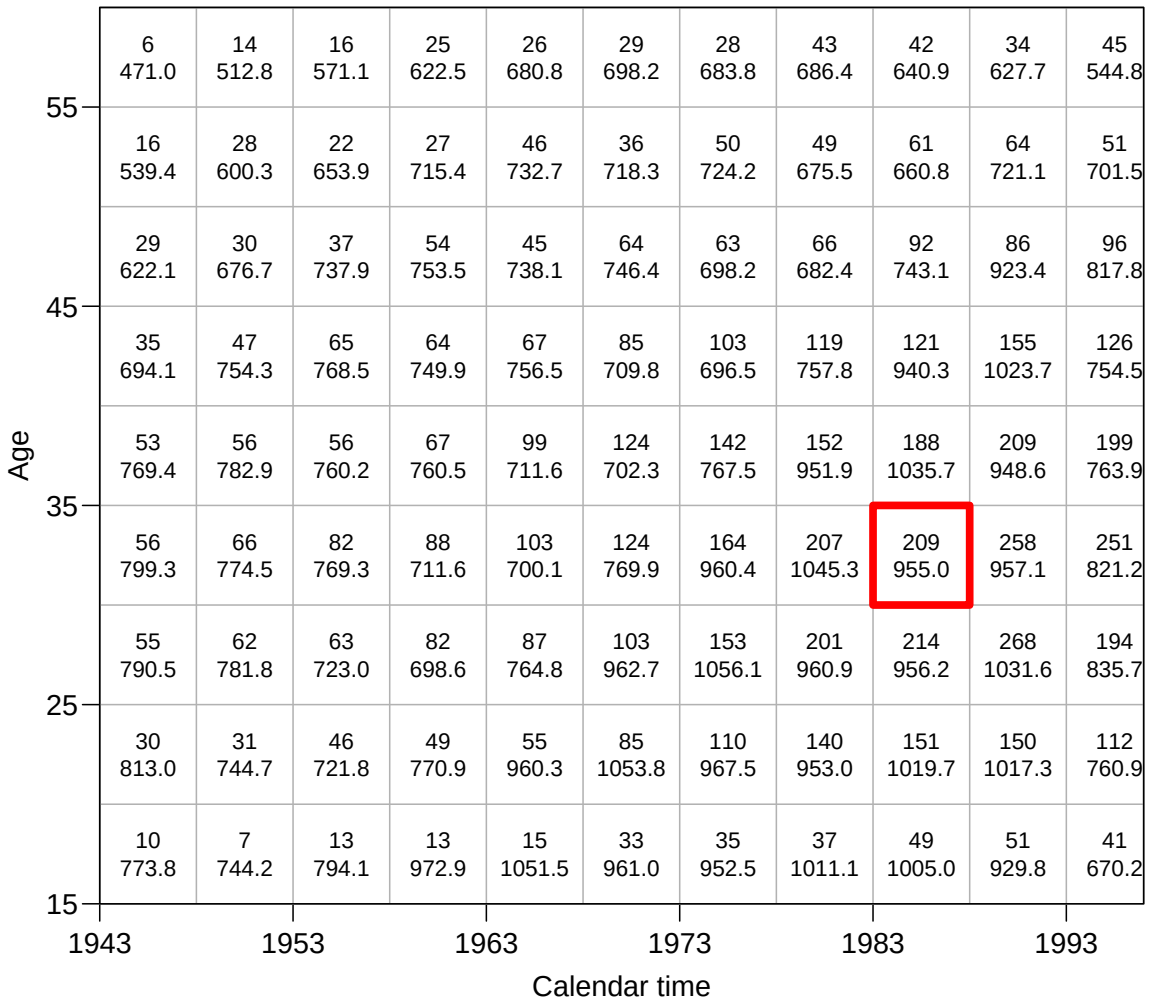
Tabulation of register data



Testis cancer cases in Denmark.

Male person-years in Denmark.

Tabulation of register data

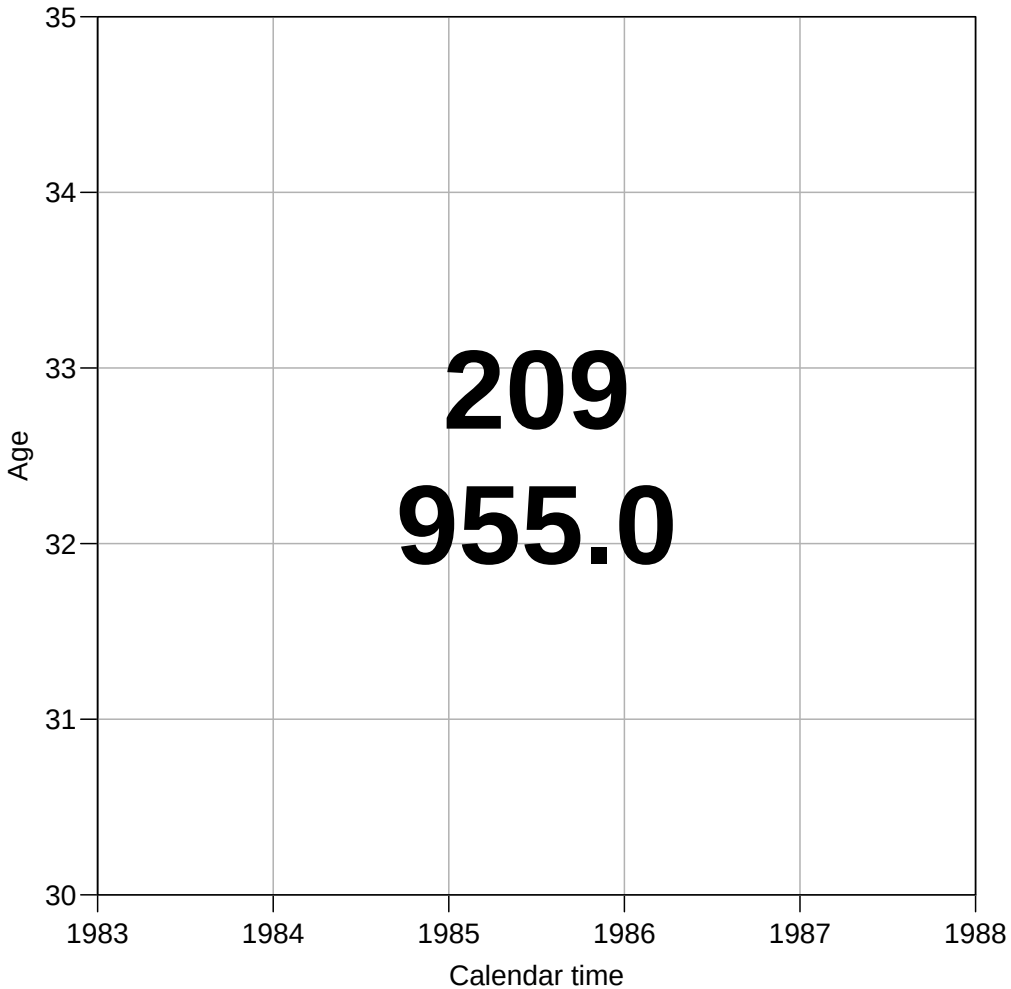


Age \ Calendar time	1943–1953	1953–1963	1963–1973	1973–1983	1983–1993
55–60	6 471.0	14 512.8	16 571.1	25 622.5	26 680.8	29 698.2	28 683.8	43 686.4	42 640.9	34 627.7	45 544.8
50–55	16 539.4	28 600.3	22 653.9	27 715.4	46 732.7	36 718.3	50 724.2	49 675.5	61 660.8	64 721.1	51 701.5
45–50	29 622.1	30 676.7	37 737.9	54 753.5	45 738.1	64 746.4	63 698.2	66 682.4	92 743.1	86 923.4	96 817.8
40–45	35 694.1	47 754.3	65 768.5	64 749.9	67 756.5	85 709.8	103 696.5	119 757.8	121 940.3	155 1023.7	126 754.5
35–40	53 769.4	56 782.9	56 760.2	67 760.5	99 711.6	124 702.3	142 767.5	152 951.9	188 1035.7	209 948.6	199 763.9
30–35	56 799.3	66 774.5	82 769.3	88 711.6	103 700.1	124 769.9	164 960.4	207 1045.3	209 955.0	258 957.1	251 821.2
25–30	55 790.5	62 781.8	63 723.0	82 698.6	87 764.8	103 962.7	153 1056.1	201 960.9	214 956.2	268 1031.6	194 835.7
20–25	30 813.0	31 744.7	46 721.8	49 770.9	55 960.3	85 1053.8	110 967.5	140 953.0	151 1019.7	150 1017.3	112 760.9
15–20	10 773.8	7 744.2	13 794.1	13 972.9	15 1051.5	33 961.0	35 952.5	37 1011.1	49 1005.0	51 929.8	41 670.2

Testis cancer cases in Denmark.

Male person-years in Denmark.

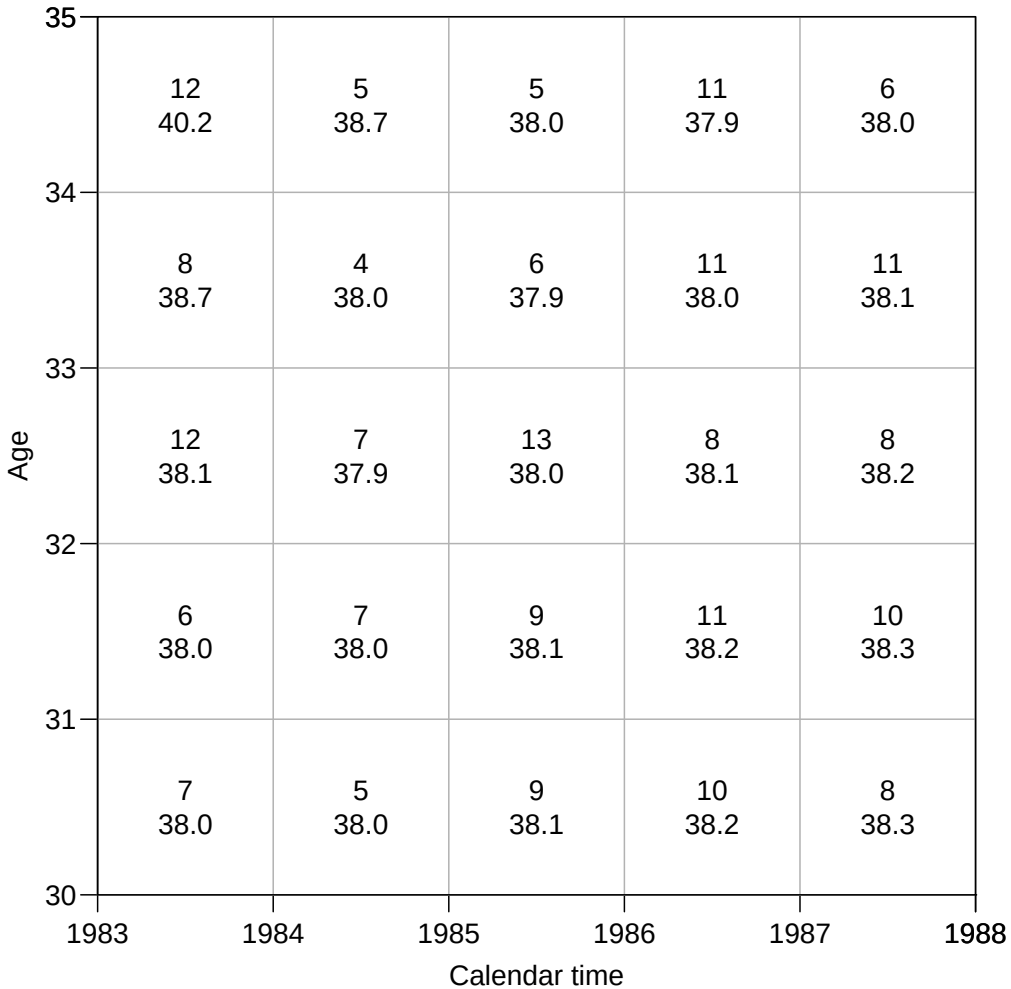
Tabulation of register data



Testis cancer cases in Denmark.

Male person-years in Denmark.

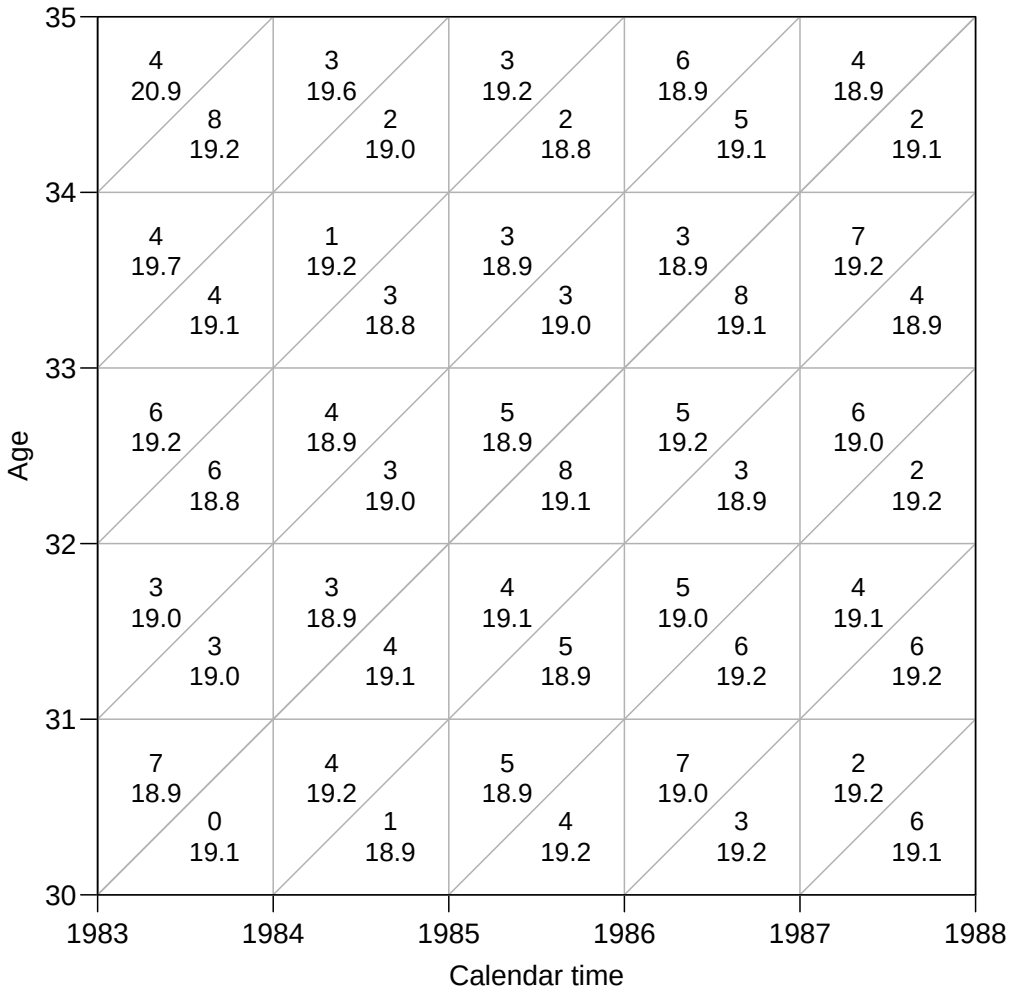
Tabulation of register data



Testis cancer cases in Denmark.

Male person-years in Denmark.

Tabulation of register data



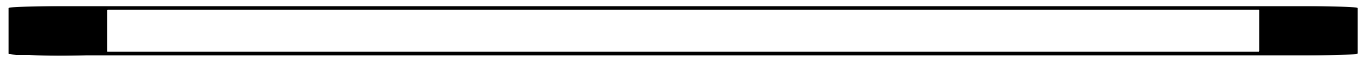
Testis cancer cases in Denmark.

Male person-years in Denmark.

Subdivision by year of birth (cohort).

Major sets in the Lexis diagram

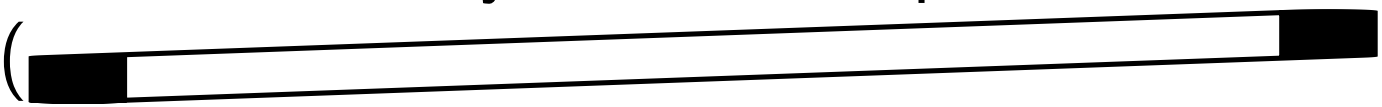
A-sets: Classification by age and period.

()

B-sets: Classification by age and cohort.

()

C-sets: Classification by cohort and period.

()

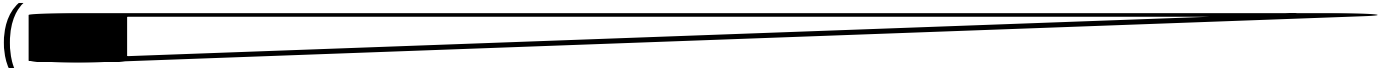
The mean age, period and cohort for these sets is just the mean of the tabulation interval.

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

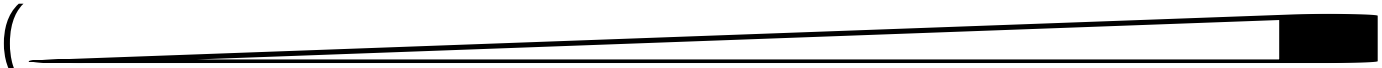
Analysis of rates from a complete observation in a Lexis digram need not be restricted to these classical sets classified by two factors.

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.

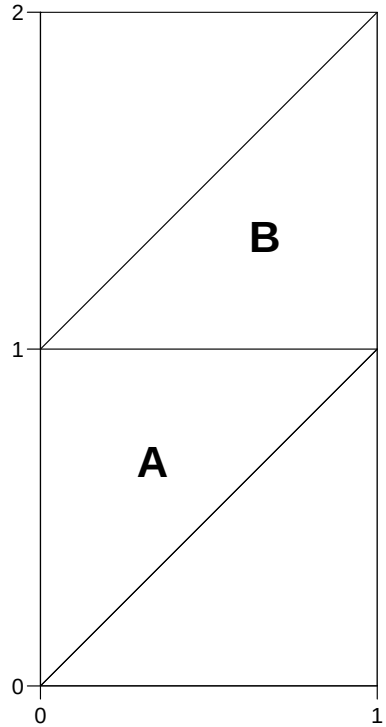
()

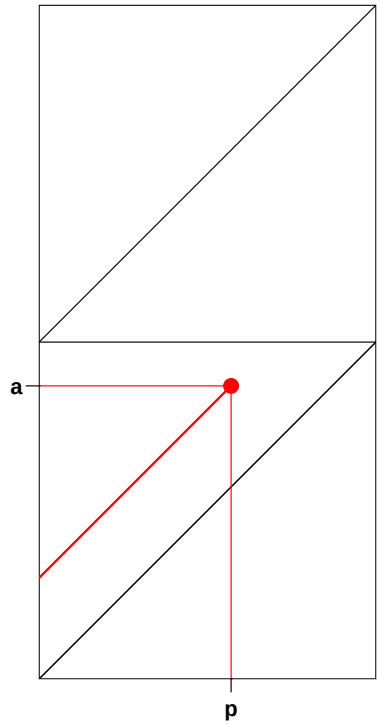
Mean time in triangles

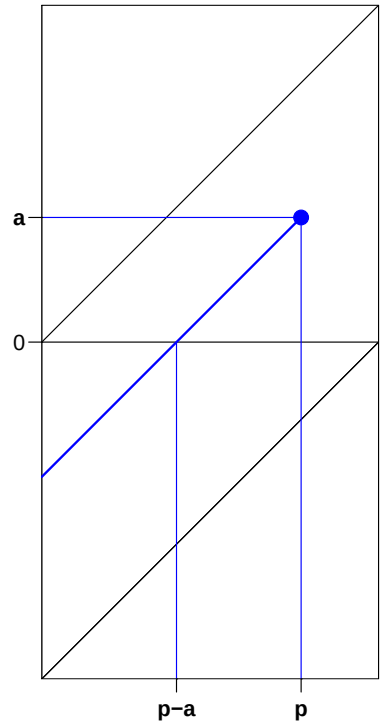
Modelling requires that each set (=observation in the dataset) be assigned a value of age, period and cohort. So for each triangle we need:

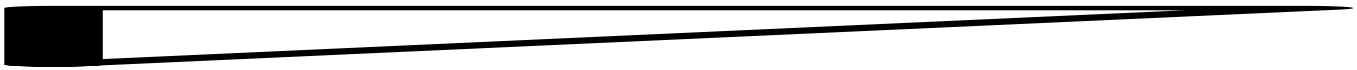
- mean age at risk.
- mean date at risk.
- mean cohort at risk.

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







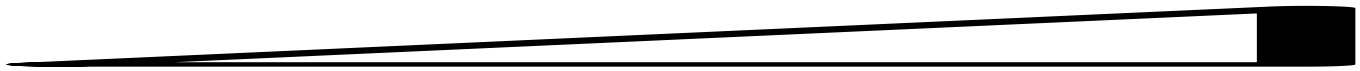
Upper triangles (), A:

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

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

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

Lower triangles



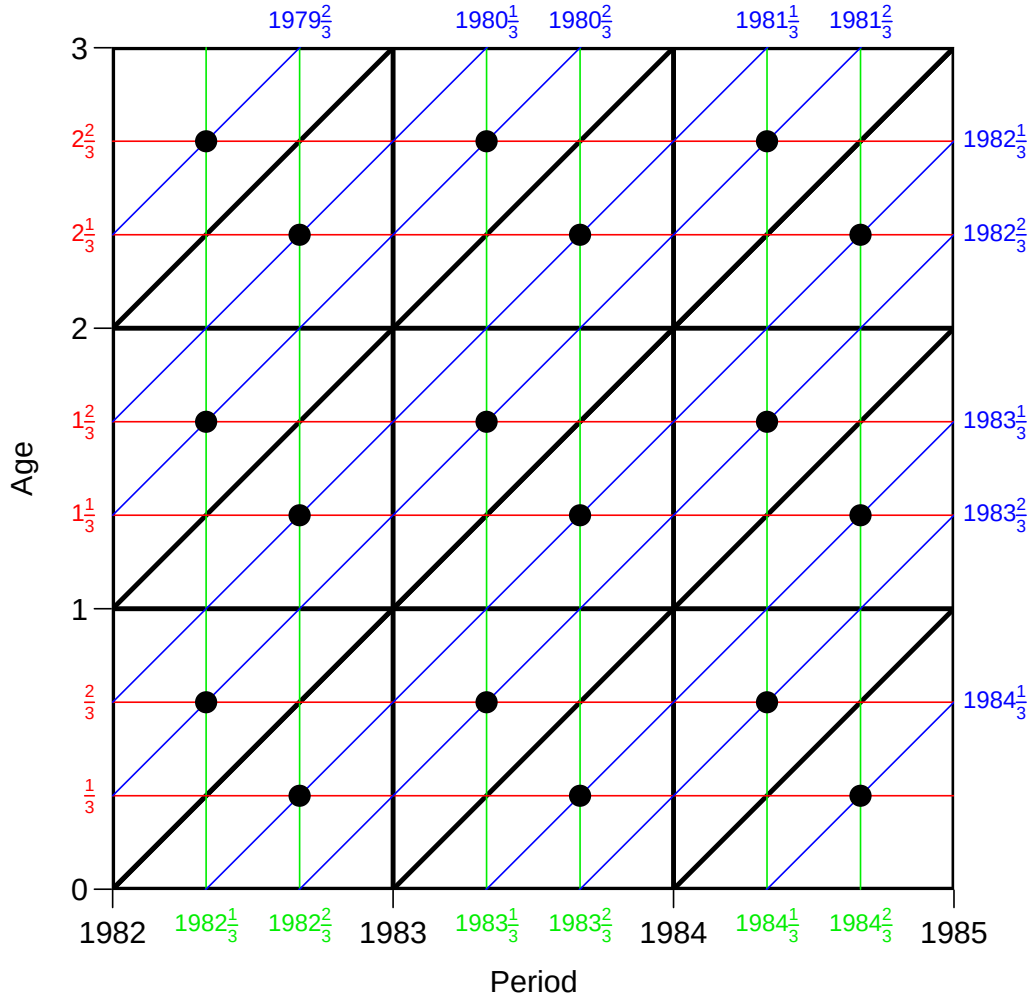
), B:

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

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

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

Tabulation by age, period and cohort



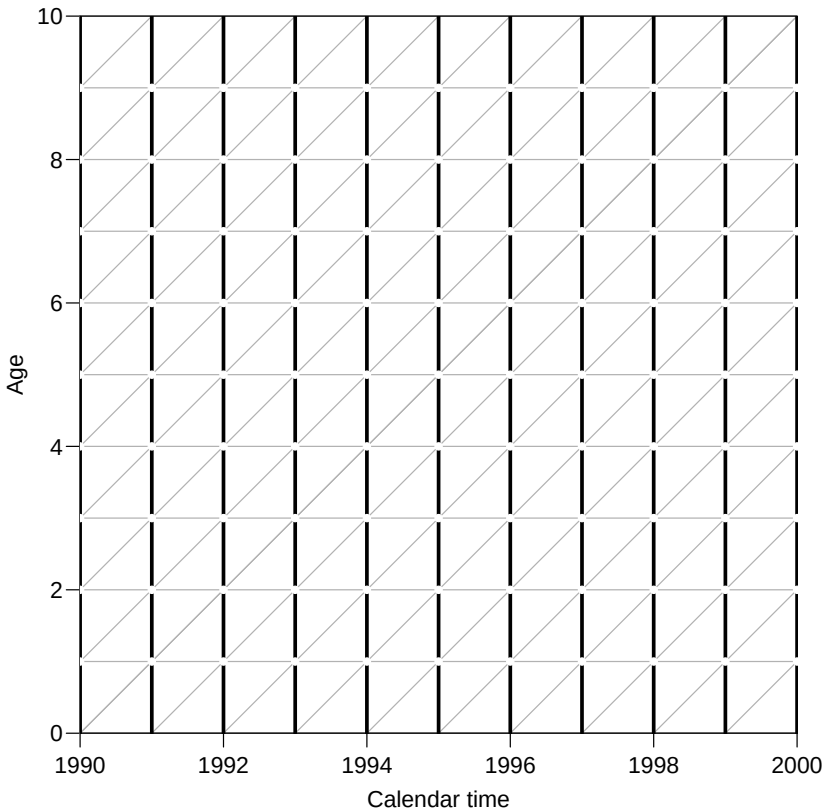
Gives triangular sets with differing mean age, period and cohort:

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

Population figures

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

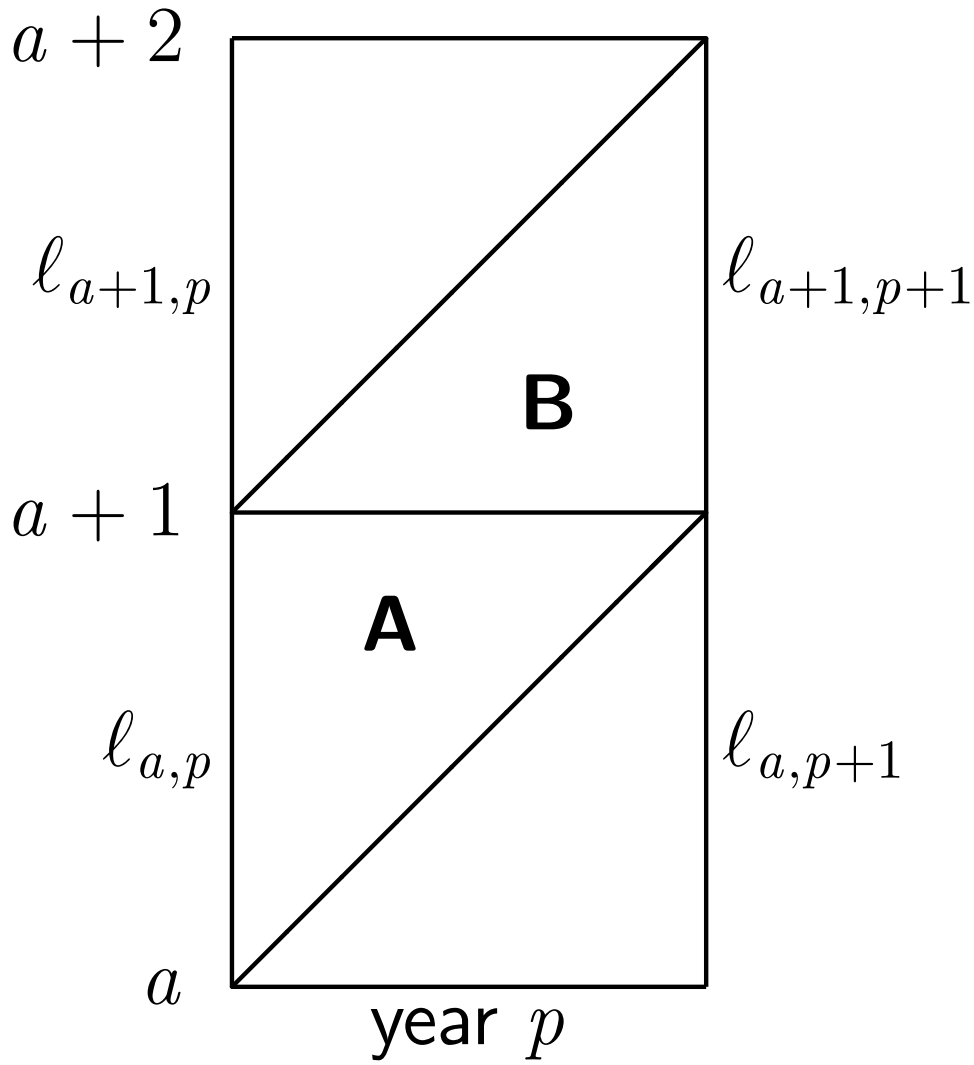
This corresponds to population sizes along the vertical lines indicated in the diagram. We want risk time figures for the population in the squares and triangles in the diagram.



Prevalent population figures

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

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

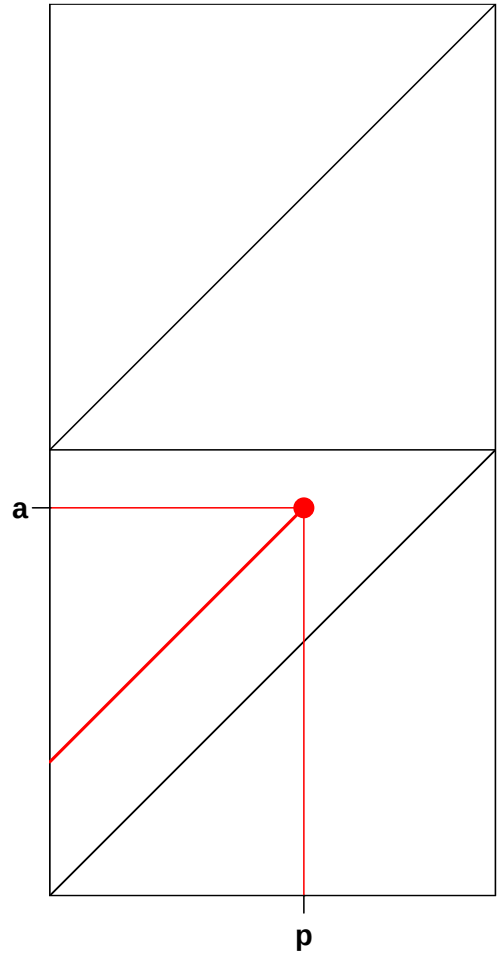


A person dying in age a at date p in **A** contributes p risk time, so the average will be:

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

$$= \int_{p=0}^{p=1} 2p(1-p) \; \mathrm{d}p$$

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

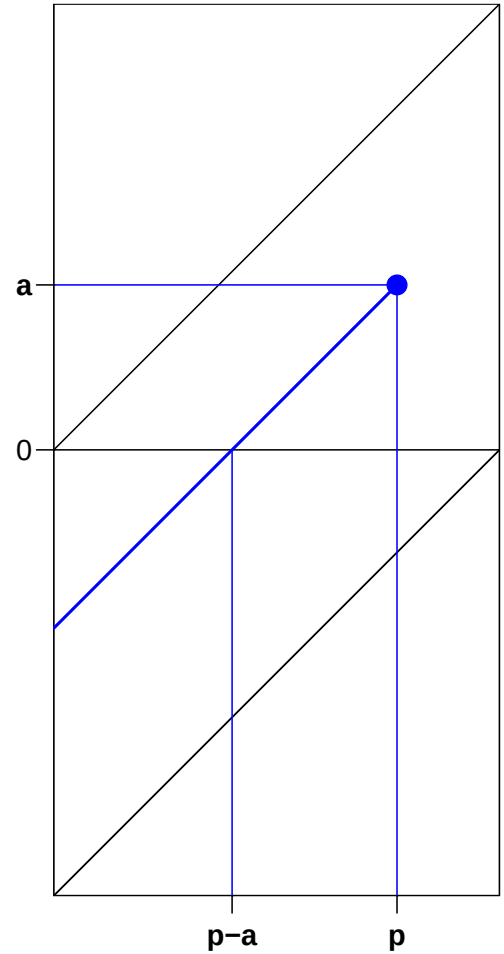


A person dying in age a at date p in **B** contributes $p - a$ risk time in **A**, so the average will be:

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

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

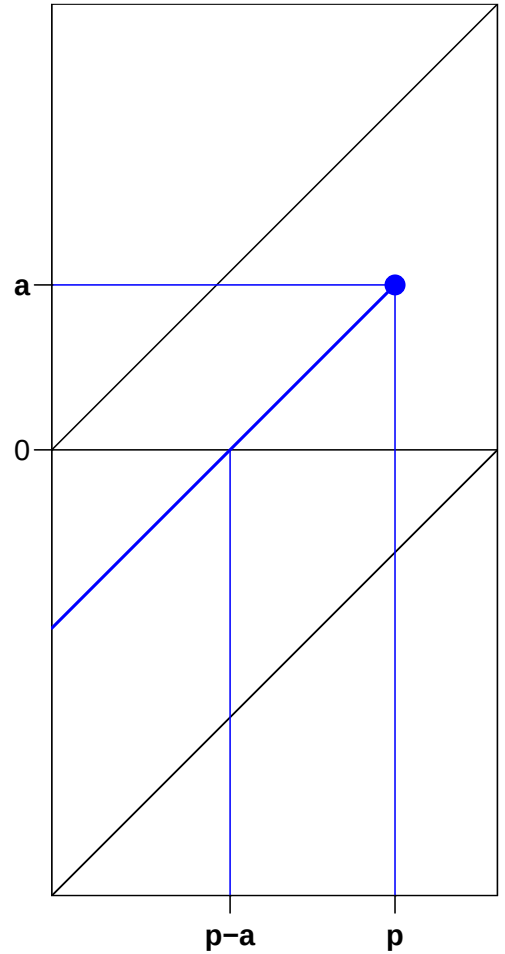
$$= \int_{p=0}^{p=1} p^2\,\mathrm{d}p = \frac{1}{3}$$



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

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

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



Contributions to risk time in A and B:

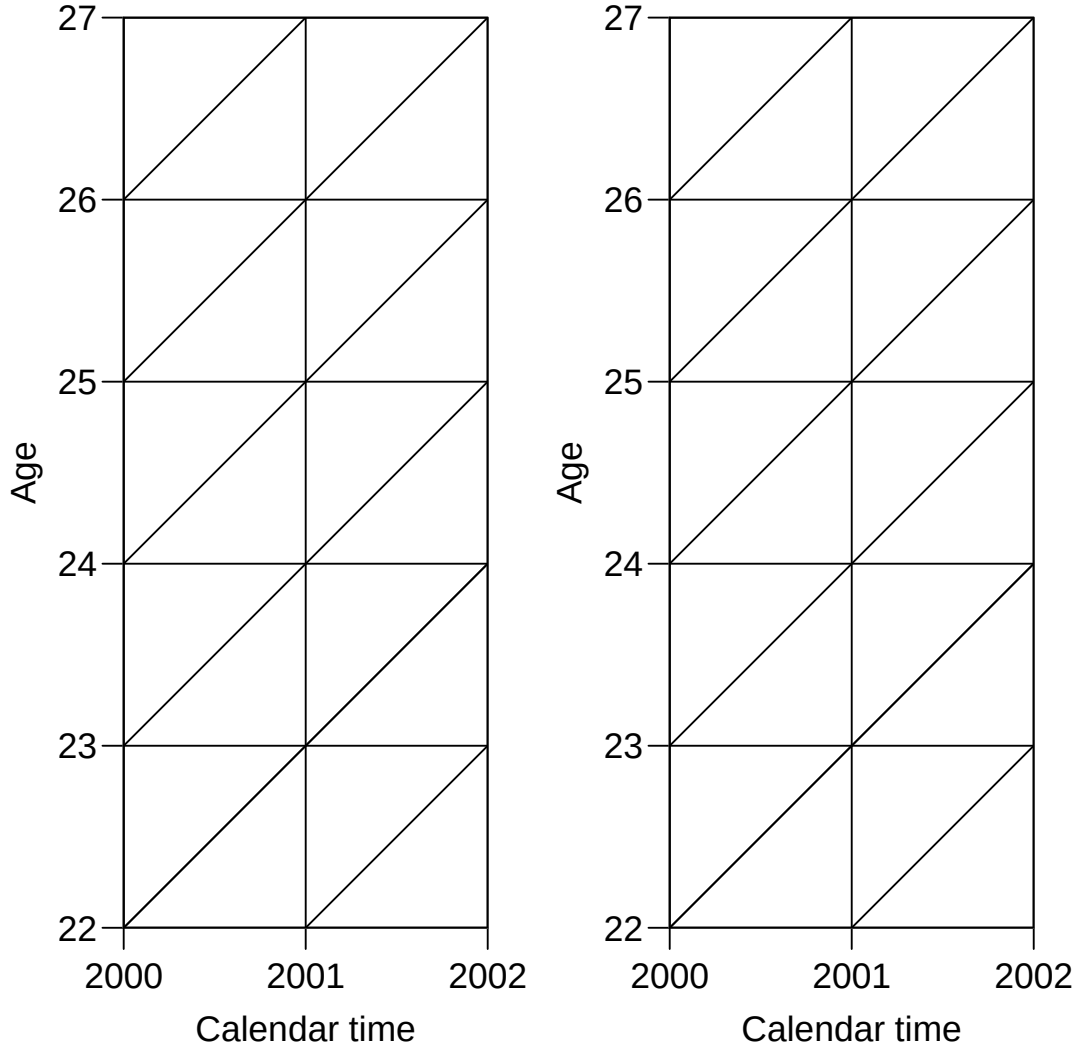
	A:	**B**:
Survivors:	$\ell_{a+1,p+1} \times \frac{1}{2}$y	$\ell_{a+1,p+1} \times \frac{1}{2}$y
Dead in **A**:	$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}$y	
Dead in **B**:	$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}$y	$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times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

Population as of 1. January from Statistics Denmark:

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

Exercise:

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



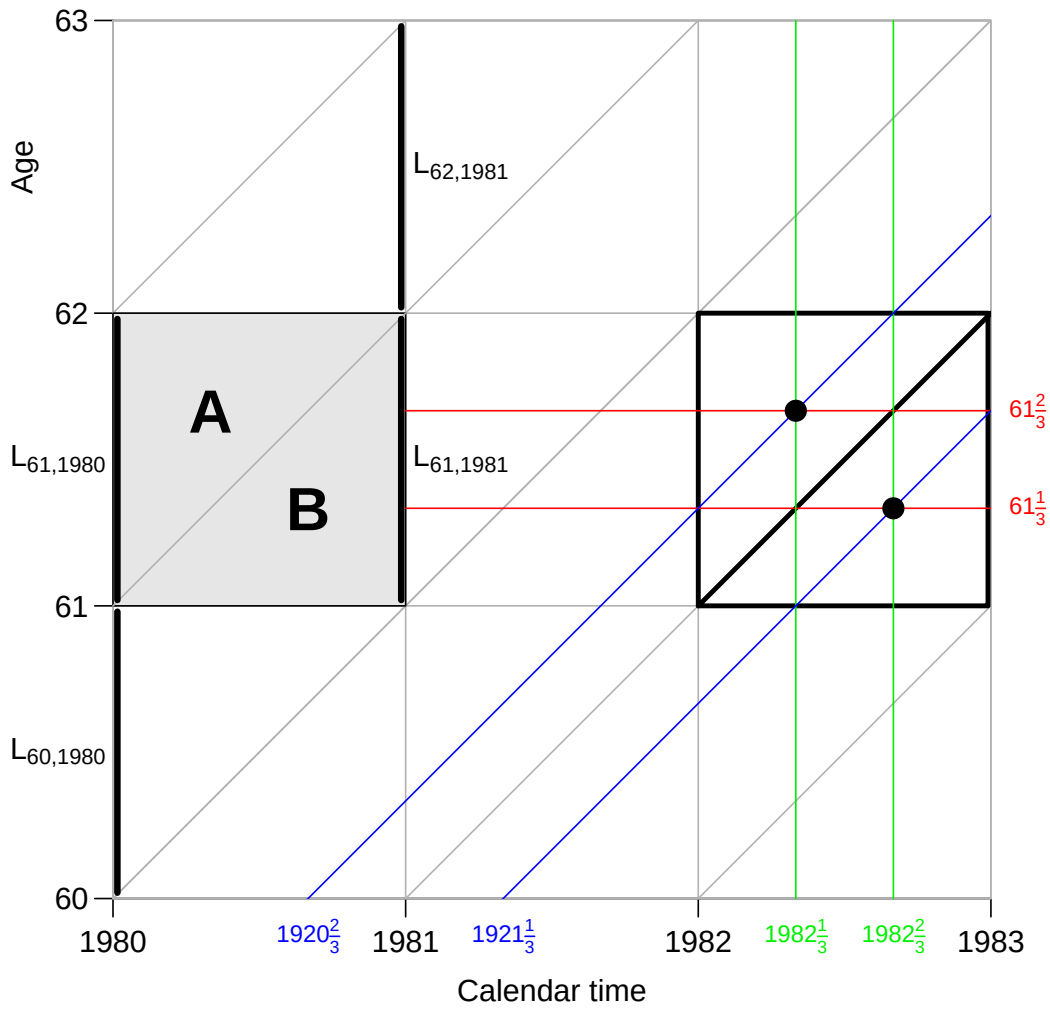
Summary:

Population risk time:

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

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

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



APC-model for triangular data

Tuesday 4th, afternoon

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm

`www.biostat.ku.dk/~bxc/APC/MEB-2010`

Model for triangular data

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

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

Problem: Disconnected design!

Log-likelihood contribution from one triangle:

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

The log-likelihood can be separated:

$$\sum_{a,p\in} D_{ap}\log(\lambda_{ap}$$

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

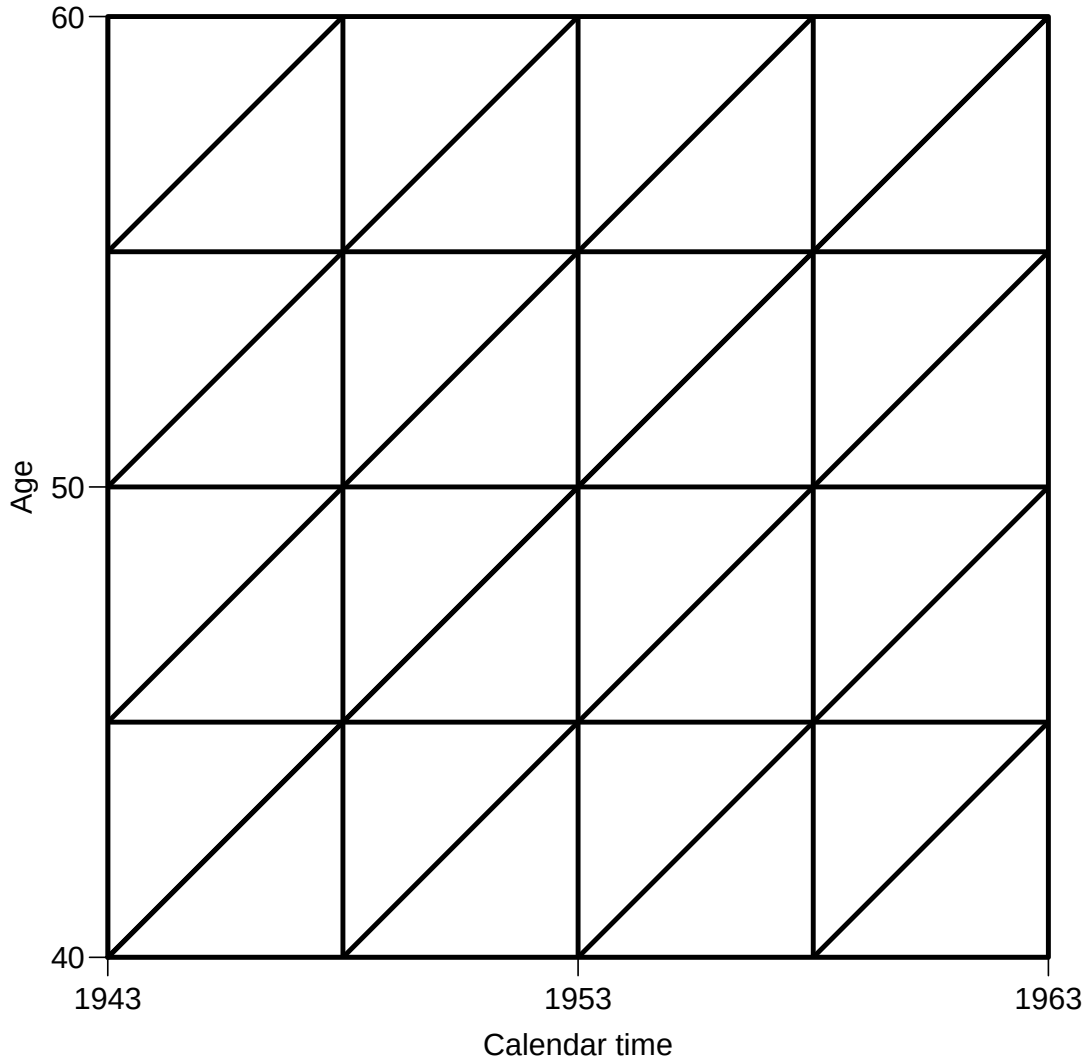
Illustration by lung cancer data

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

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

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

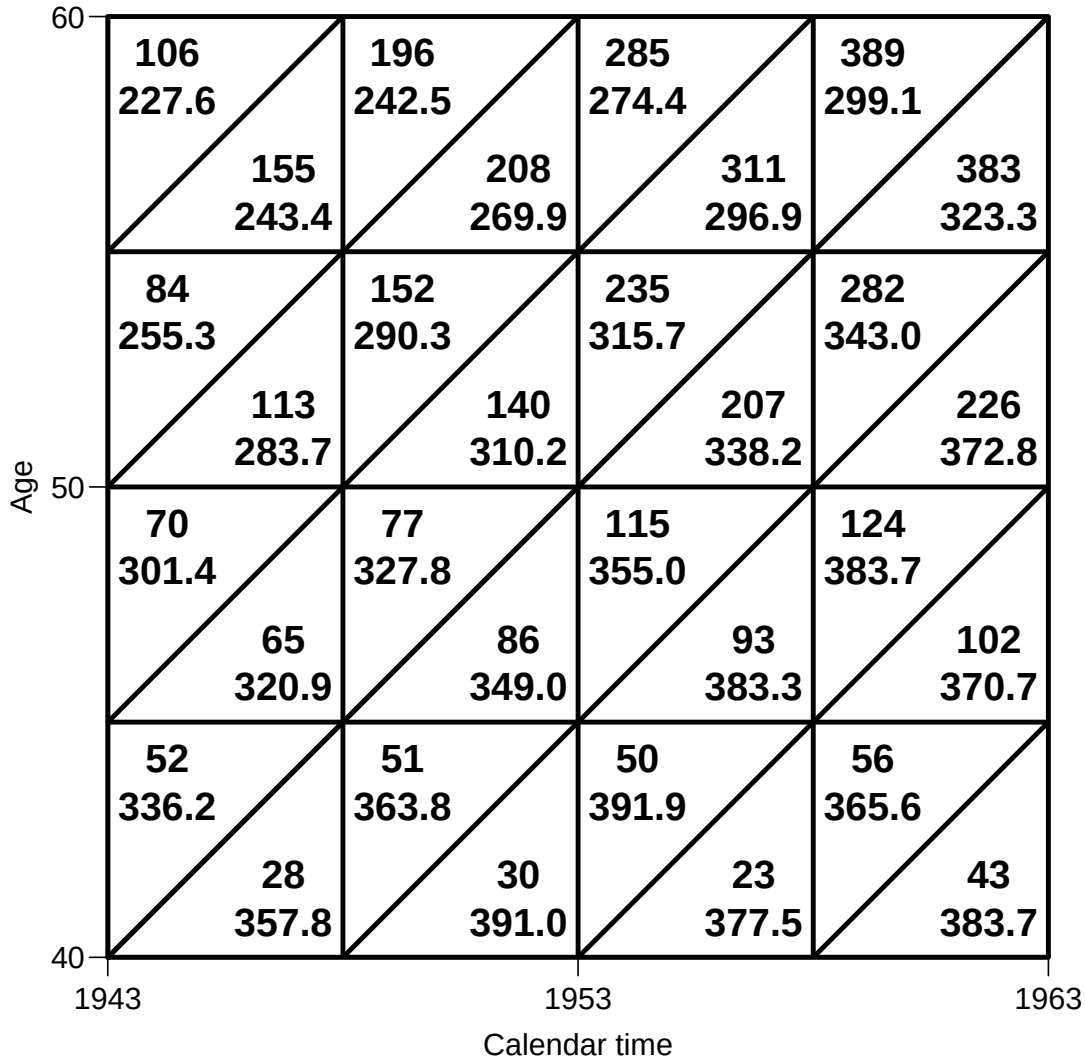
Mark mean date of birth for these.



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

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

Mark mean date of birth for these.



APC-model with "synthetic" cohorts

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

...

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

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

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

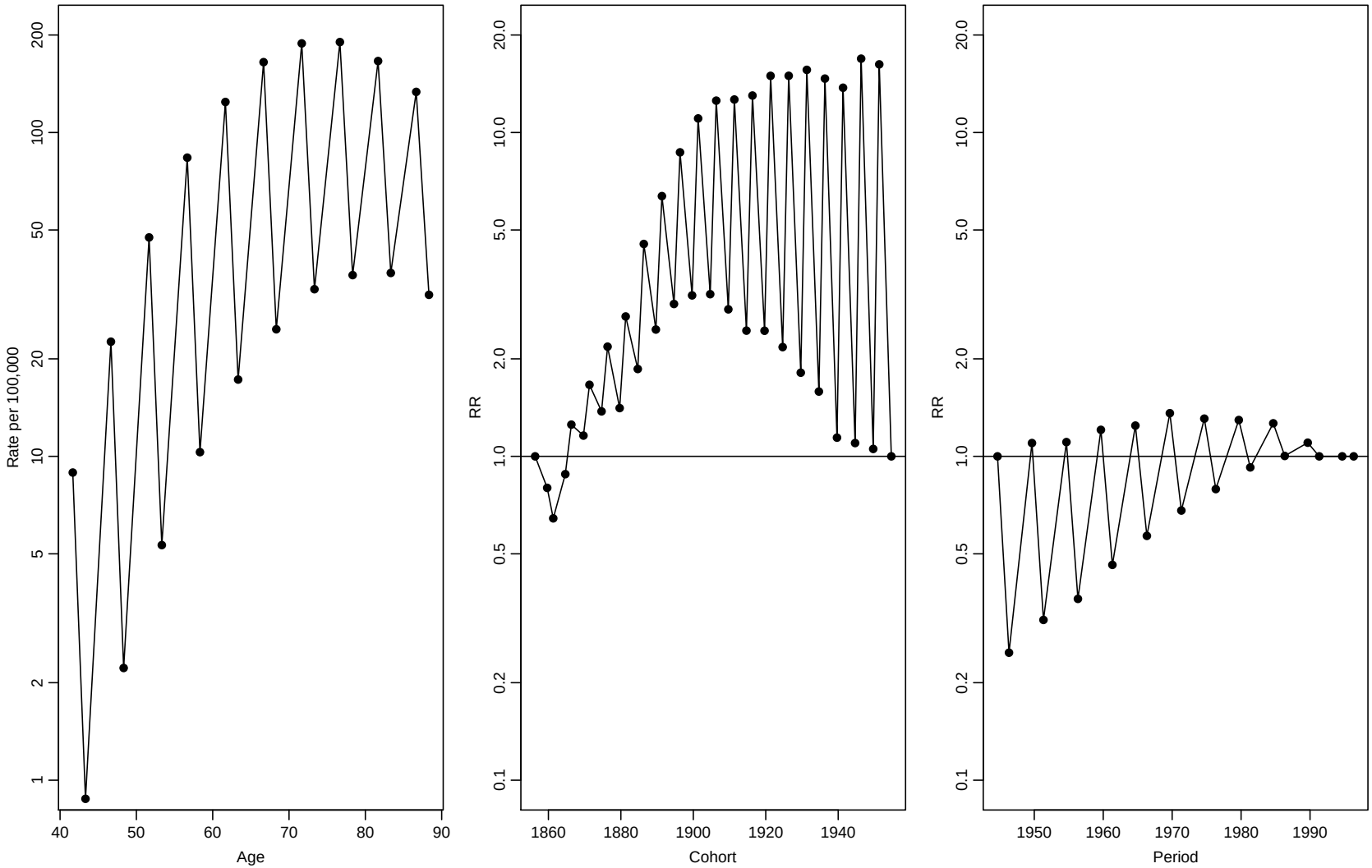
APC-model with "correct" cohorts

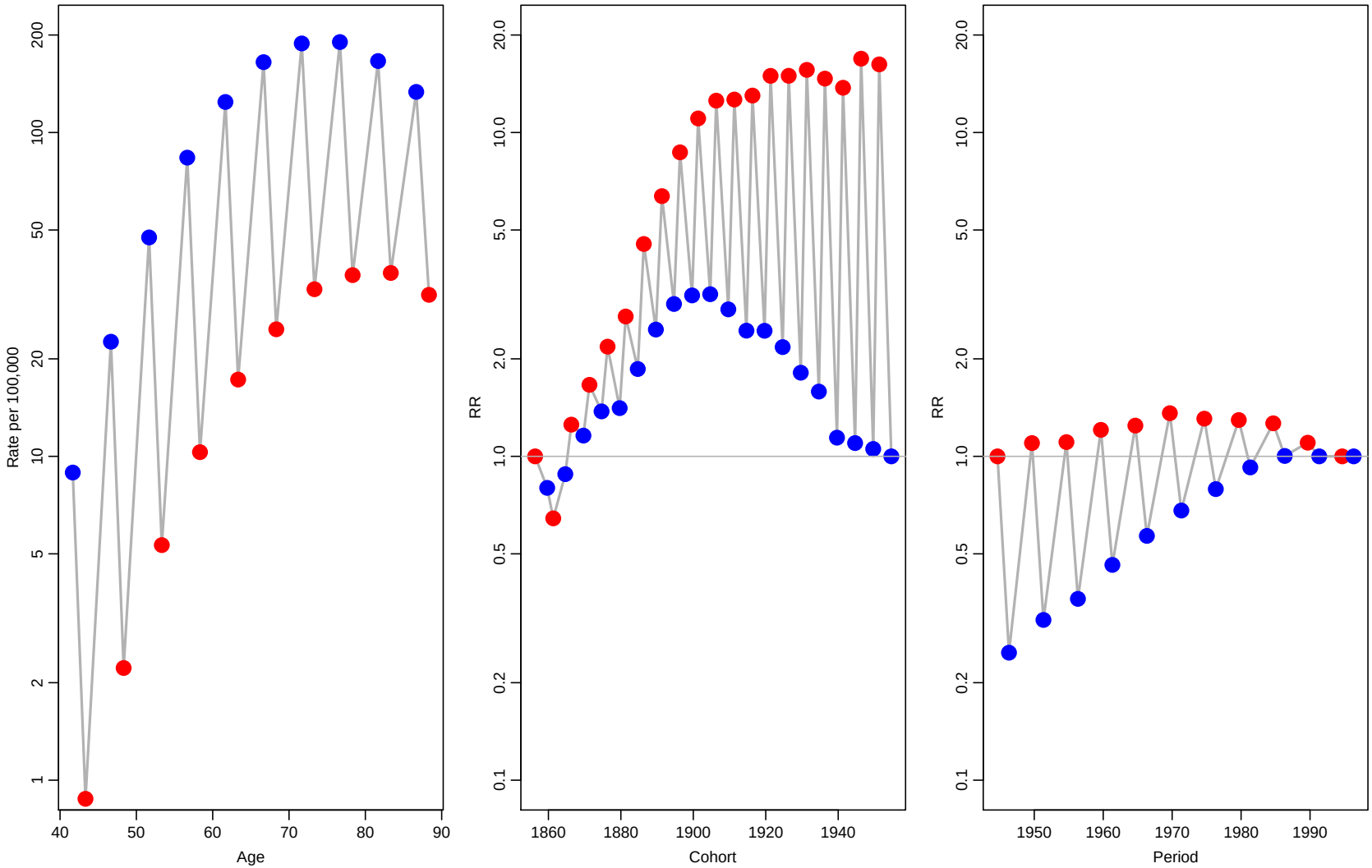
```
> mx <- glm( D ~ factor(Ax) - 1 +
+                factor(Cx) +
+                factor(Px) + offset( log( Y ) ),
+            family=poisson )
> summary( mx )
...
    Null deviance: 1.0037e+08  on 220  degrees of freedom
Residual deviance: 2.8473e+02  on 144  degrees of freedom
```

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

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

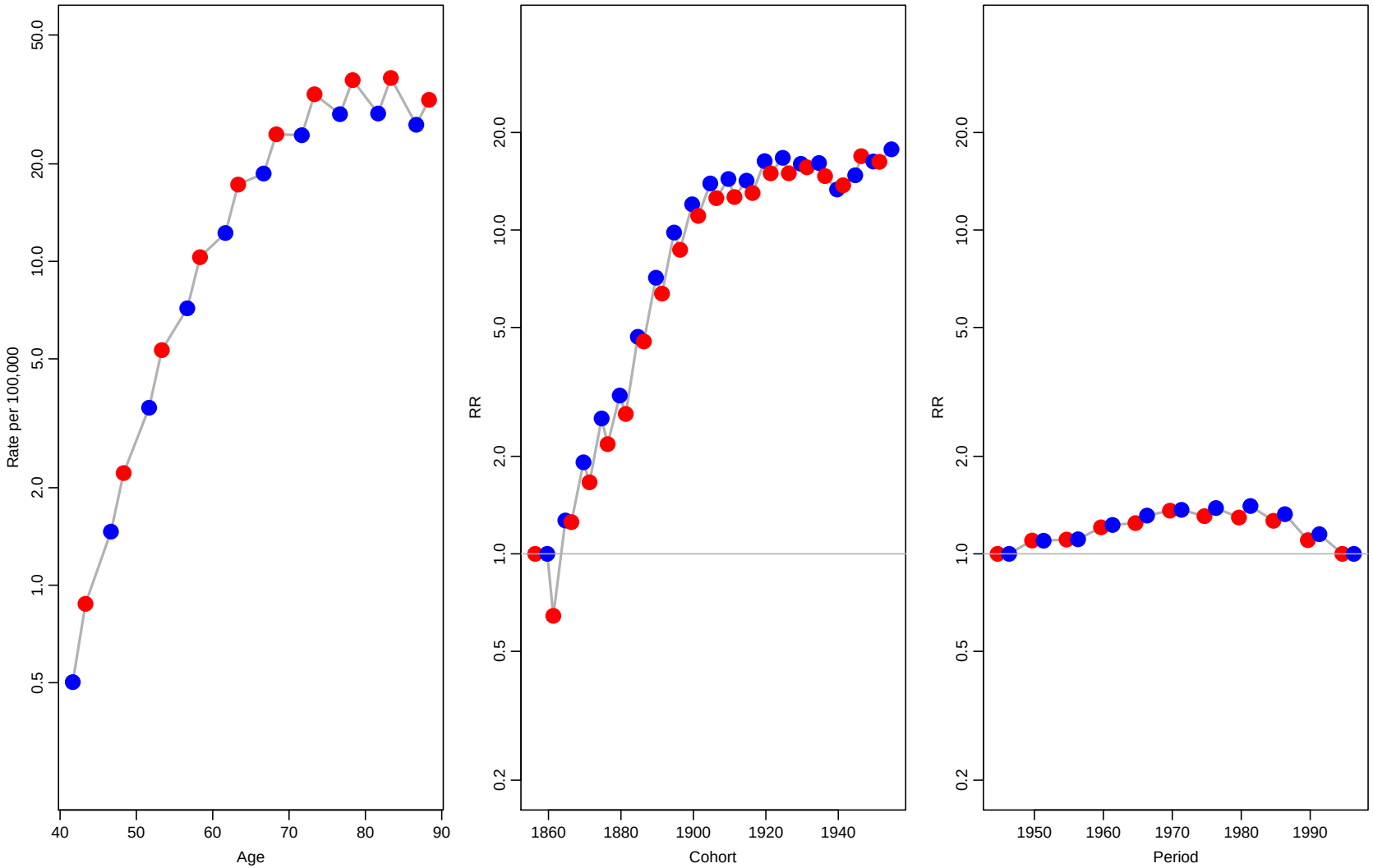
We have fitted two age-period-cohort models separately to upper and lower triangles.


Rate per 100,000
200
100
50
20
10
5
2
1
40
50
60
70
80
90
Age
RR
20.0
10.0
5.0
2.0
1.0
0.5
0.2
0.1
1860
1880
1900
1920
1940
Cohort
RR
20.0
10.0
5.0
2.0
1.0
0.5
0.2
0.1
1950
1960
1970
1980
1990
Period


Rate per 100,000
200
100
50
20
10
5
2
1
40
50
60
70
80
90
Age
RR
20.0
10.0
5.0
2.0
1.0
0.5
0.2
0.1
1860
1880
1900
1920
1940
Cohort
RR
20.0
10.0
5.0
2.0
1.0
0.5
0.2
0.1
1950
1960
1970
1980
1990
Period

Now, explicitly fit models for upper and lower triangles:

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


Rate per 100,000
50.0
20.0
10.0
5.0
2.0
1.0
0.5
40
50
60
70
80
90
Age
RR
20.0
10.0
5.0
2.0
1.0
0.5
0.2
1860
1880
1900
1920
1940
Cohort
RR
20.0
10.0
5.0
2.0
1.0
0.5
0.2
1950
1960
1970
1980
1990
Period

APC-model: Parametrization

Tuesday 4th, afternoon

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm

`www.biostat.ku.dk/~bxc/APC/MEB-2010`

What's the problem?

One parameter is assigned to each distinct value of the timescales, the ordering of the variables is not used.

The solution is to "tie together" the points on the scales together with smooth functions of the *mean* age, period and cohort with three functions:

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

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

The identifiability problem still exists:

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

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

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

Smooth functions

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

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

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

All of these contain the linear effect as special case.

Parametrization of effects

There are still three "free" parameters:

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

Choose μ_a, μ_c and γ according to some criterion for the functions.

Parametrization principle

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

Longitudinal or cohort age-effects.

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

Alternatively, the period function could be constrained to be 0 at a reference date, p_0.

Then, age-effects at $a_0 = p_0 - c_0$ would equal the fitted rate for period p_0 (and cohort c_0), and the period effects would be residual log-RRs relative to p_0.

Cross-sectional or period age-effects?

Bureaucratically interpretable — whats is seen at a particular date?

Implementation:

1. Obtain any set of parameters $f(a)$, $g(p)$, $h(c)$.
2. Extract the trend from the period effect:

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

3. Use the functions:

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

These functions fulfill the criteria.

"Extract the trend"

Not a well-defined concept:

- Regress $\hat{g}(p)$ on p for all units in the dataset.
- Regress $\hat{g}(p)$ on p for all different values of p.
- Weighted regression?

How do we get the standard errors?

Matrix-algebra! Projections!

Parametic function

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

Example: 2nd order polymomial:

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

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

Extract the trend from g:

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

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

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

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

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

(Orthogonality requires an inner product!)

Practical parametization

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

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

Information in the data and inner product

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

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

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

Two relevant inner products:

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

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

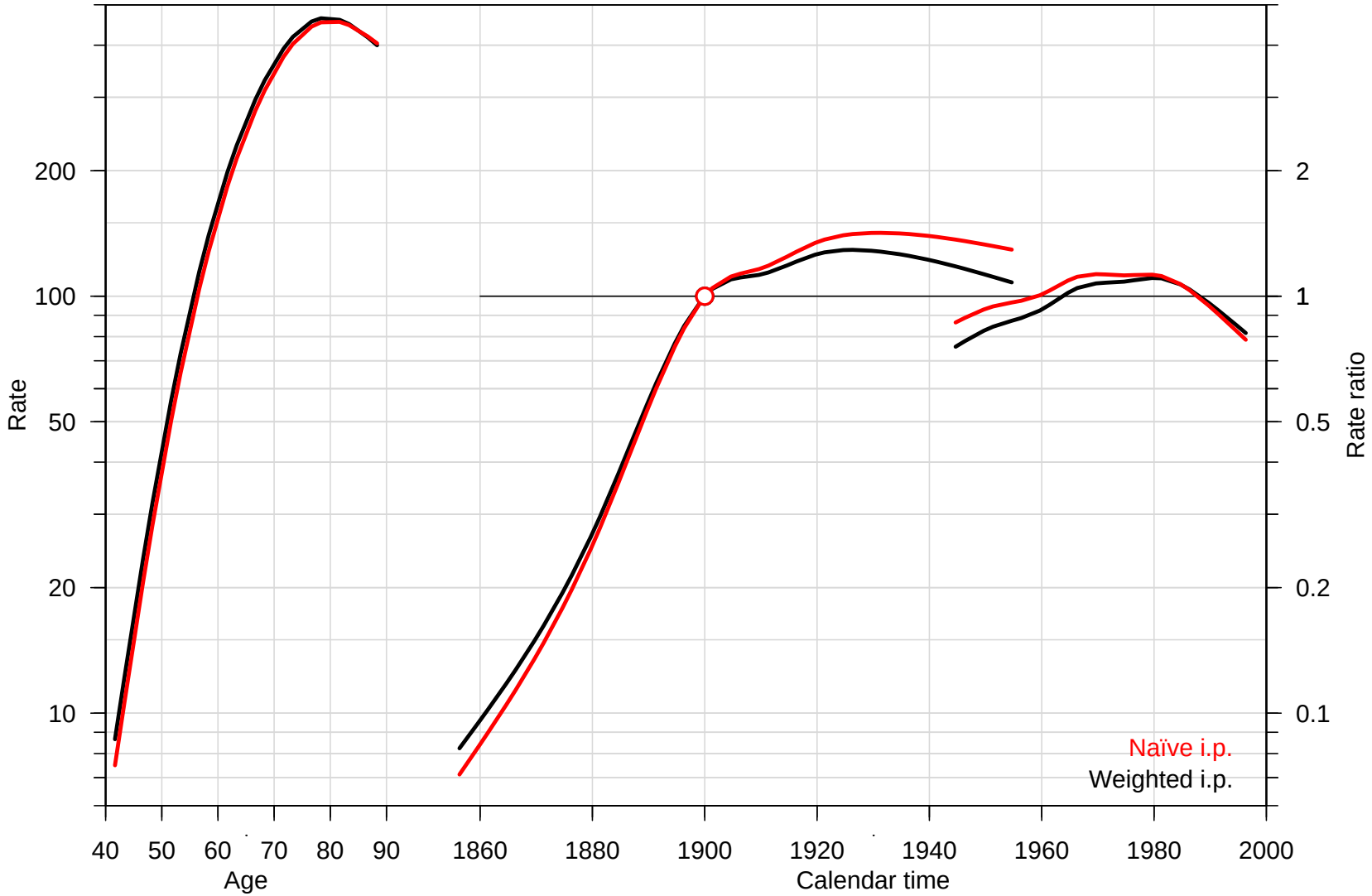
How to?

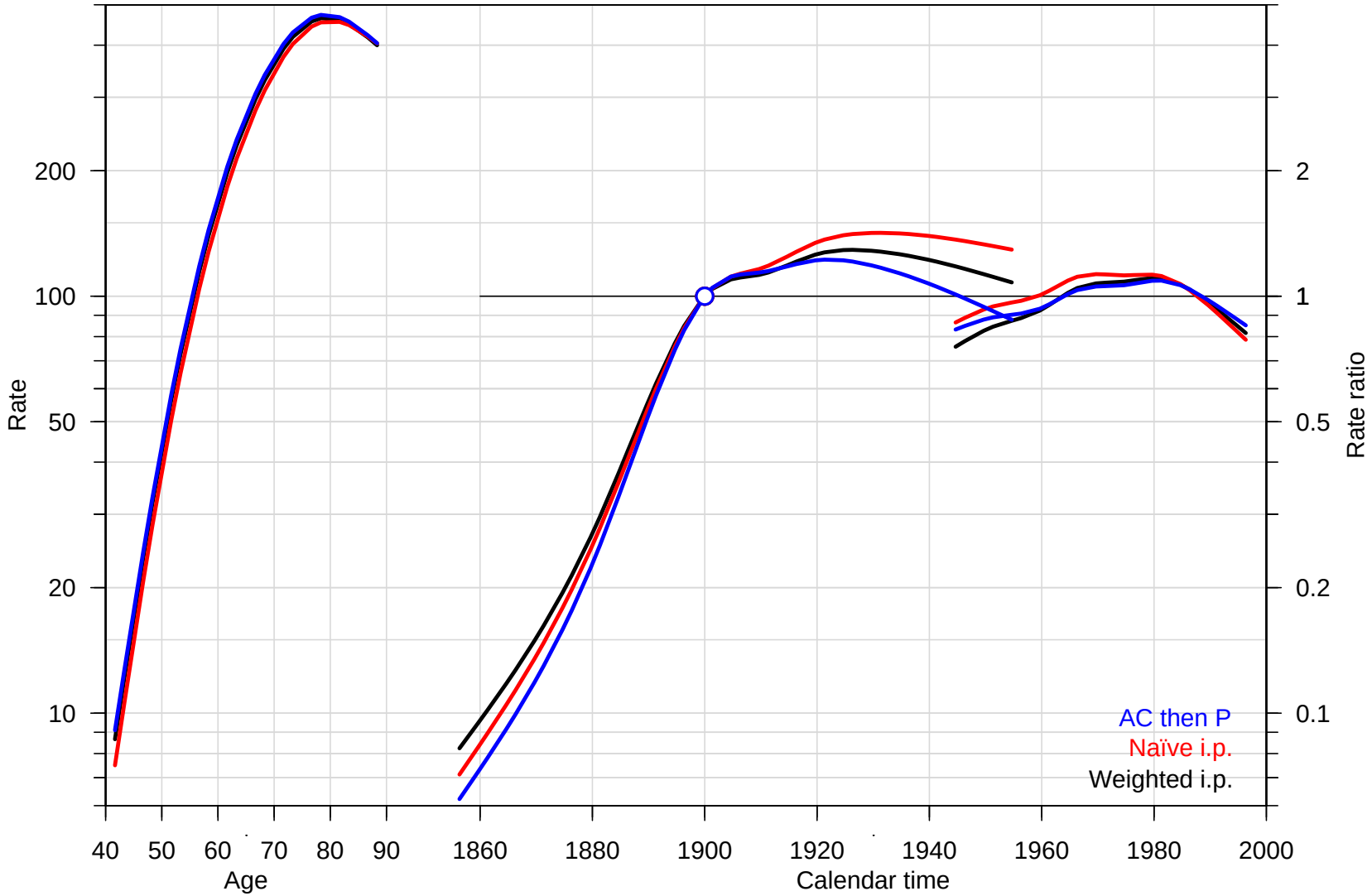
Implemented in `apc.fit`:

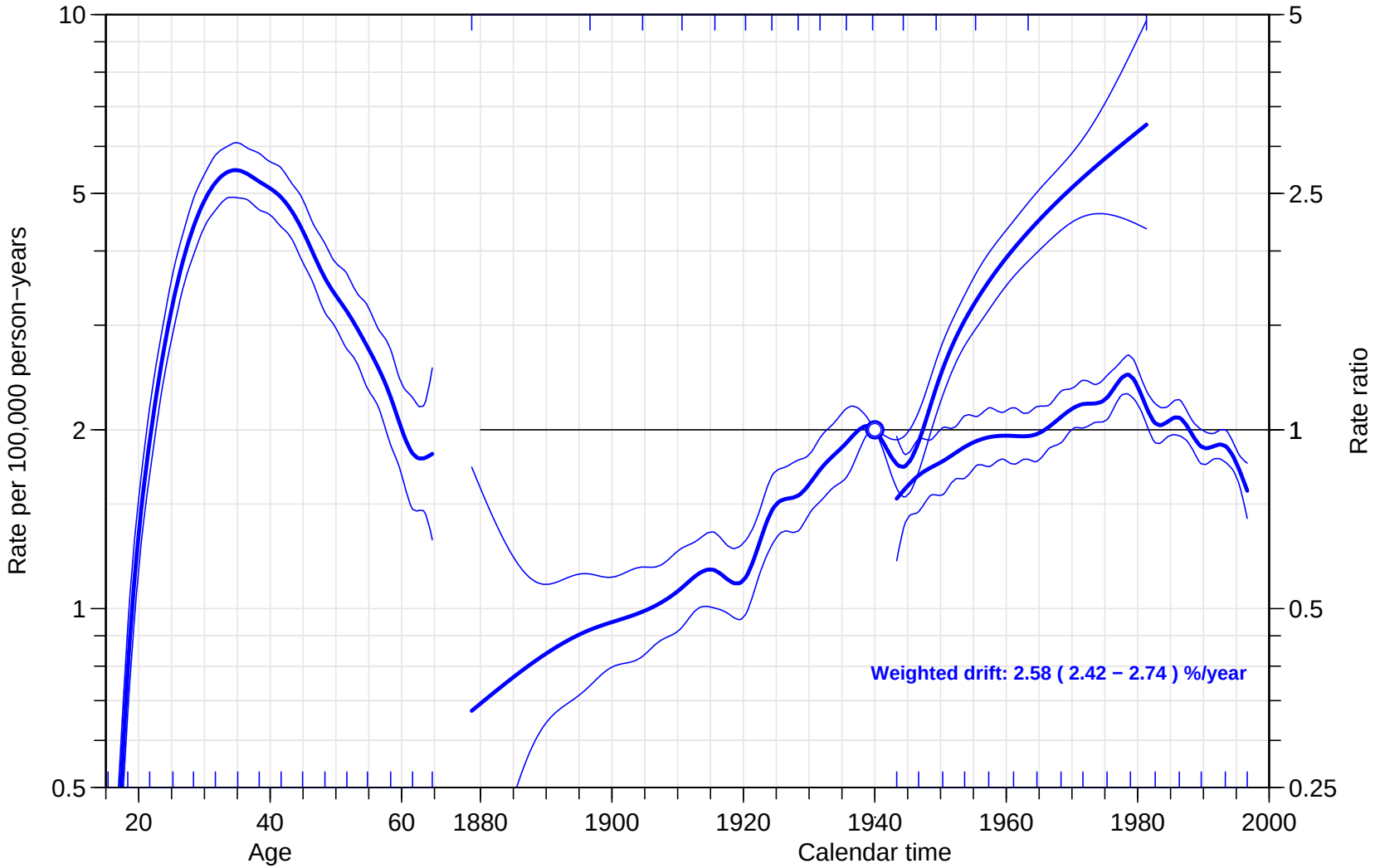
```
m1 <- apc.fit( A=lungDK$Ax,
               P=lungDK$Px,
               D=lungDK$D,
               Y=lungDK$Y/10^5,
           ref.c=1900 )
apc.plot( m1 )
```

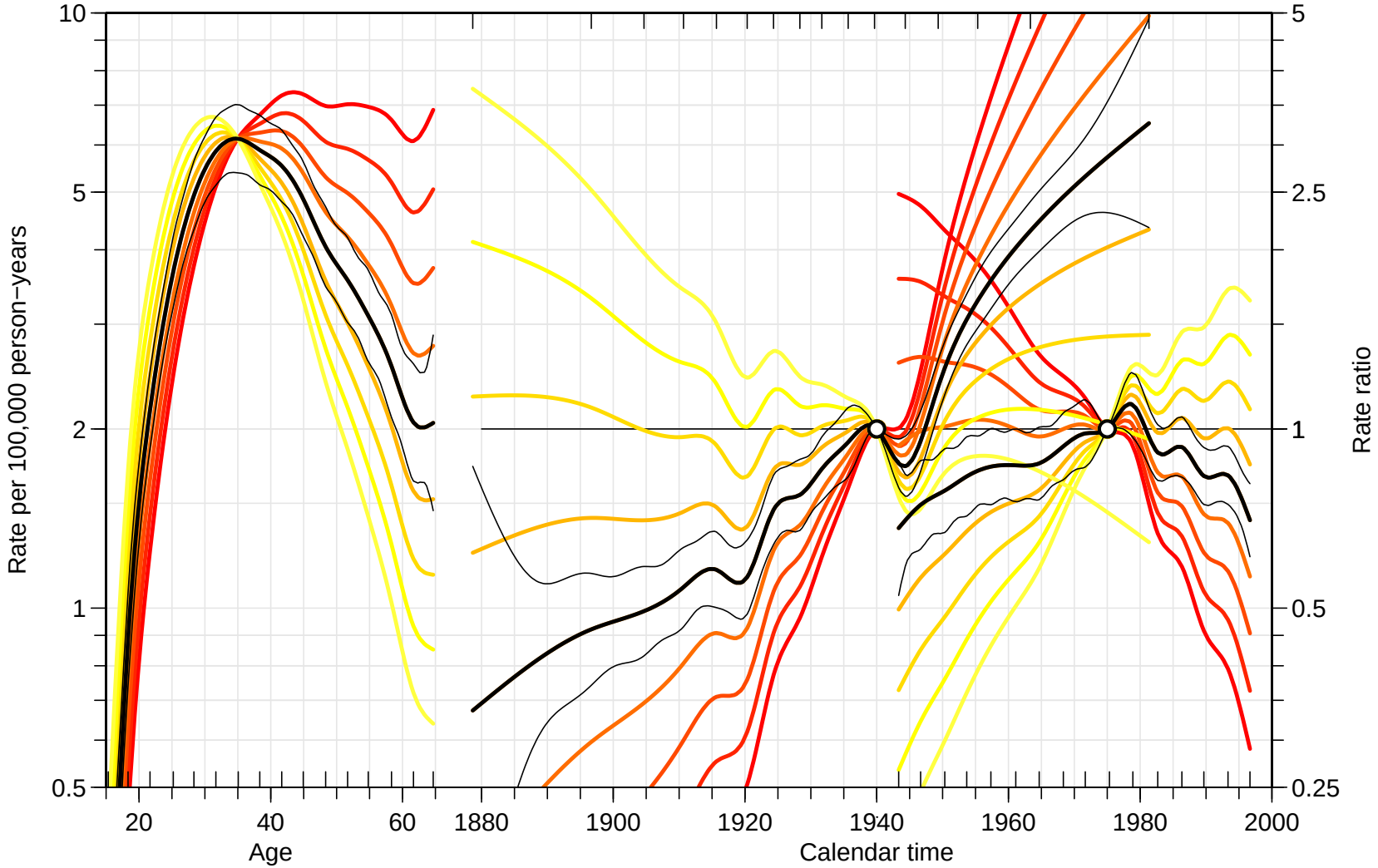
Consult the help page for:
`apc.fit` to see options for weights in inner product, type of function, variants of parametrization etc.
`apc.plot`, `apc.lines` and `apc.frame` to see how to plot the results.


Rate
200
100
50
20
10
Rate ratio
2
1
0.5
0.2
0.1
Weighted i.p.
40
50
60
70
80
90
Age
1860
1880
1900
1920
1940
1960
1980
2000
Calendar time


Rate
200
100
50
20
10
Rate ratio
2
1
0.5
0.2
0.1
40
50
60
70
80
90
Age
1860
1880
1900
1920
1940
1960
1980
2000
Calendar time
Naïve i.p.
Weighted i.p.


Rate
200
100
50
20
10
Rate ratio
2
1
0.5
0.2
0.1
40
50
60
70
80
90
Age
1860
1880
1900
1920
1940
1960
1980
2000
Calendar time
AC then P
Naïve i.p.
Weighted i.p.


Rate per 100,000 person−years
10
5
2
1
0.5
Rate ratio
5
2.5
1
0.5
0.25
20
40
60
Age
1880
1900
1920
1940
1960
1980
2000
Calendar time
Weighted drift: 2.58 (2.42 − 2.74) %/year


Rate per 100,000 person−years
10
5
2
1
0.5
20
40
60
Age
1880
1900
1920
1940
1960
1980
2000
Calendar time
5
2.5
1
0.5
0.25
Rate ratio

APC-models for several datasets

Wednesday 5th, morning

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm

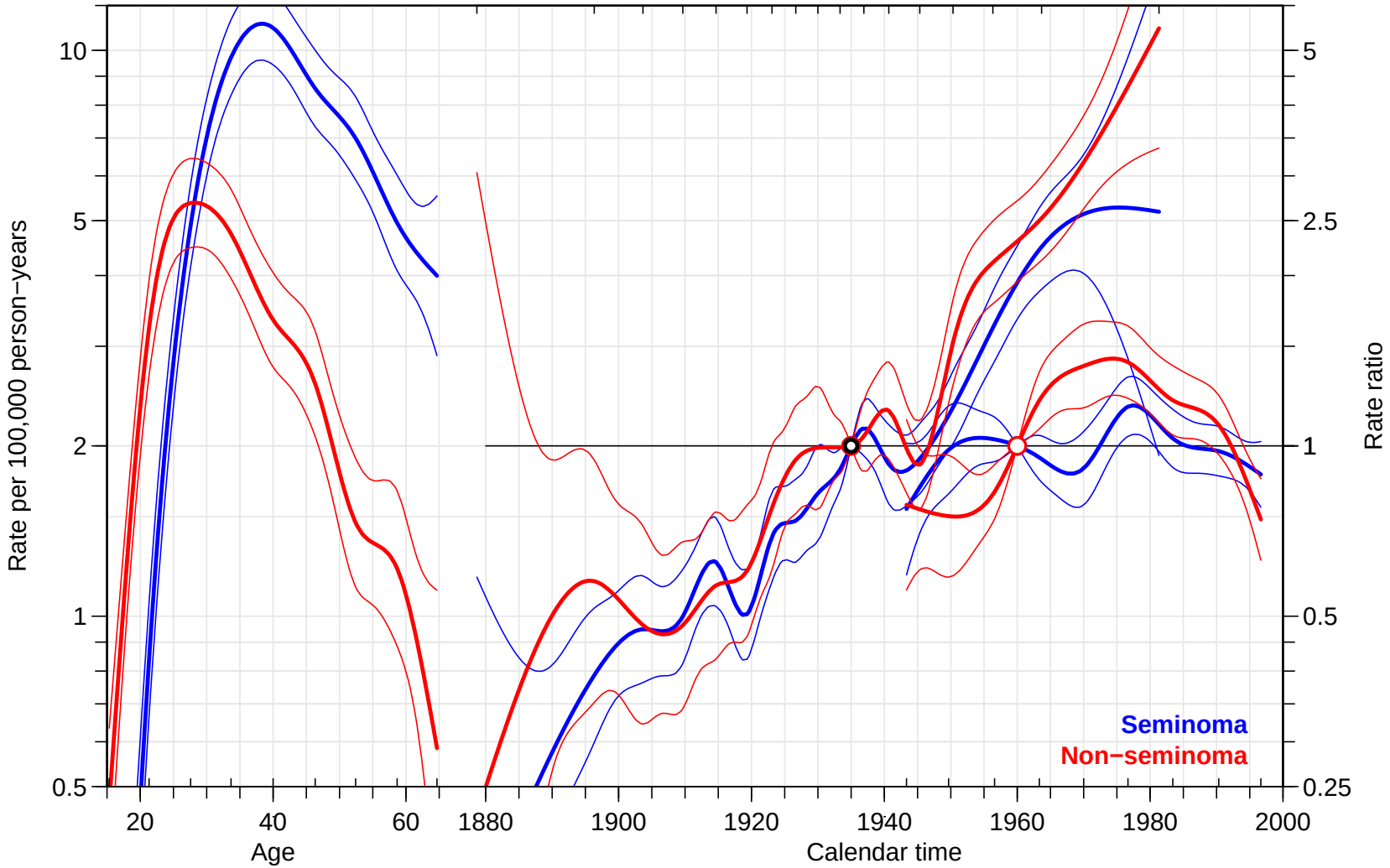
`www.biostat.ku.dk/~bxc/APC/MEB-2010`

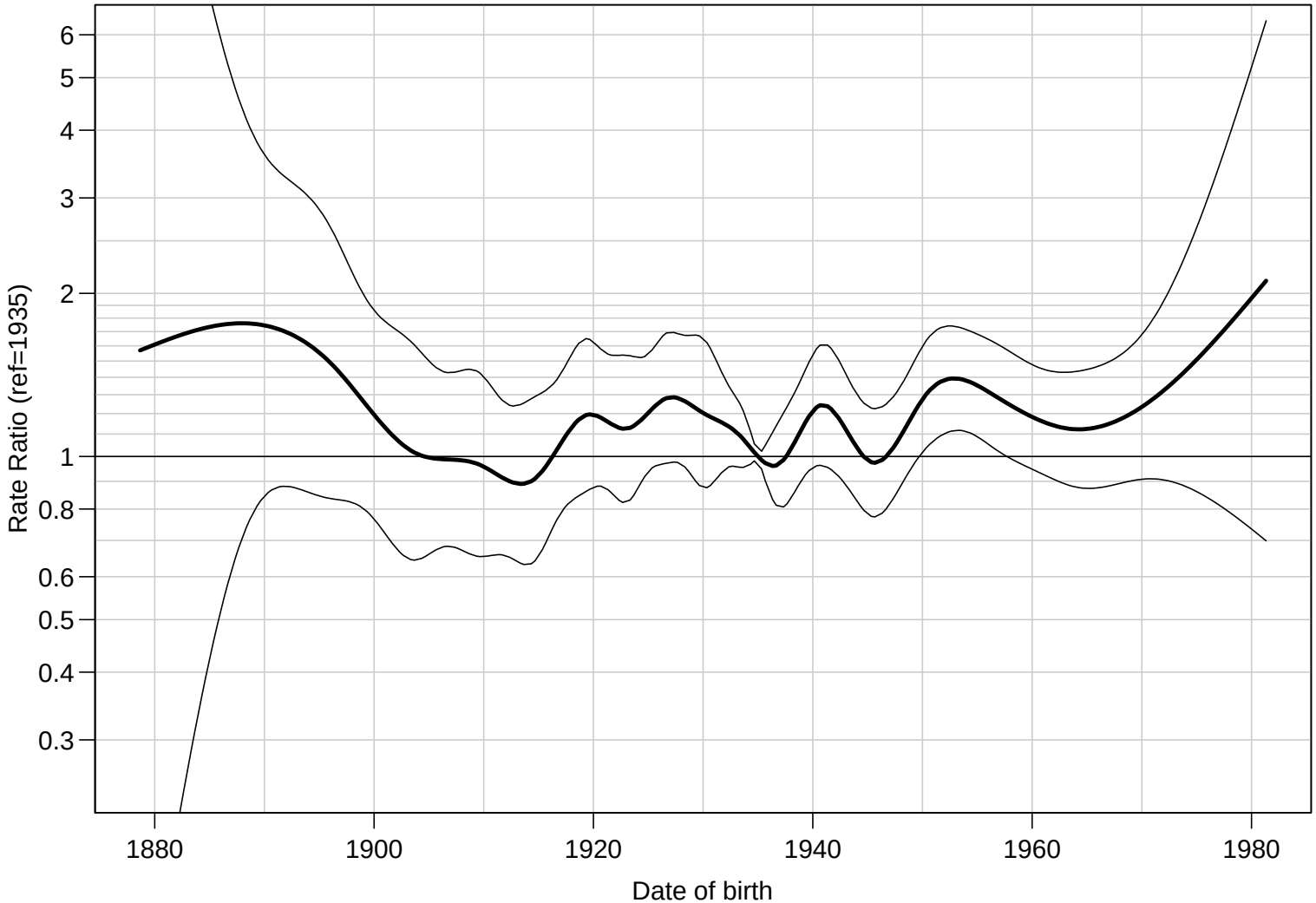
Two sets of data

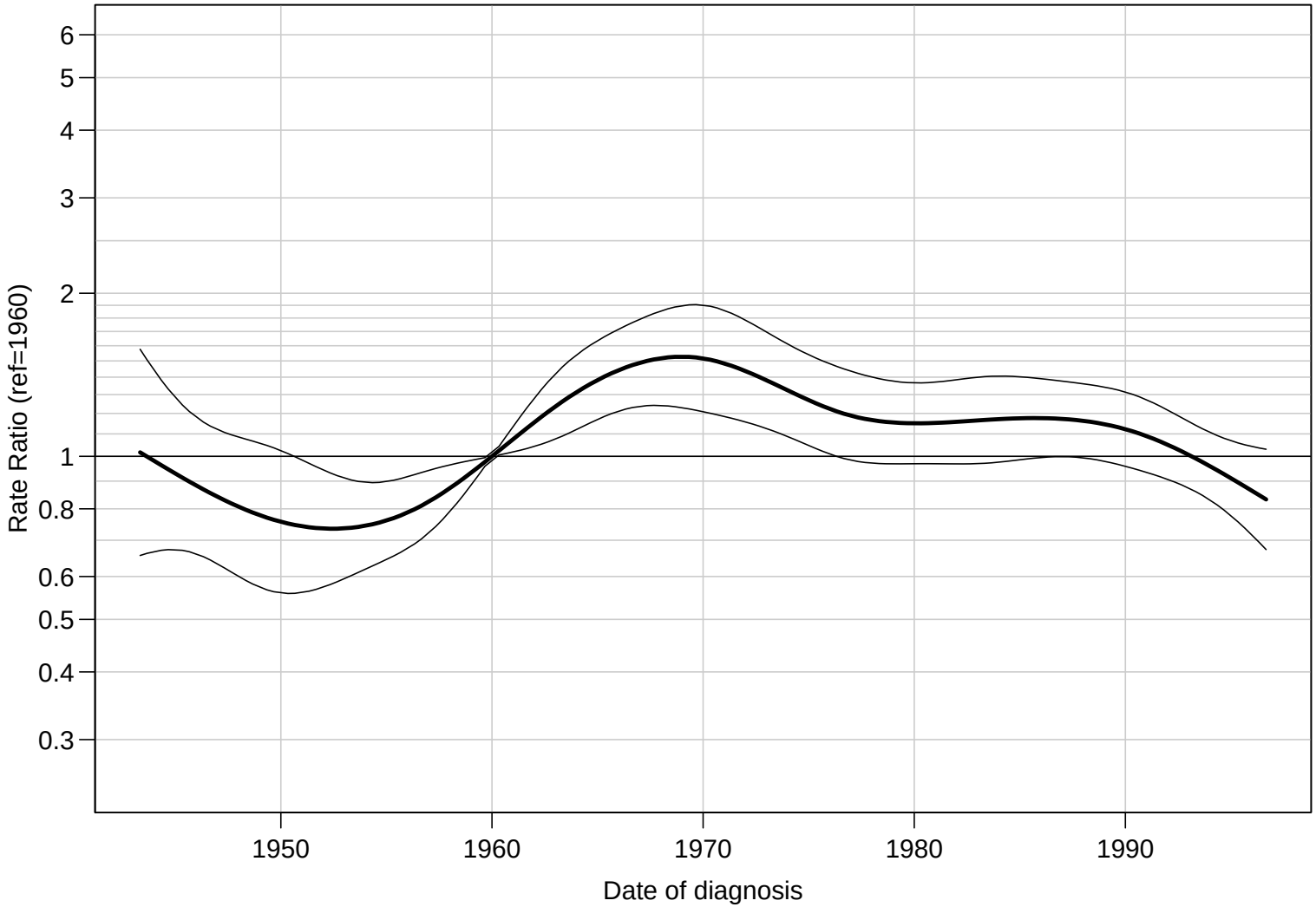
Example: Testis cancer in Denmark, Seminoma and non-Seminoma cases.

```
> stat.table( list( Histology=hist ),
+             list( D=sum(d), Y=sum(y/10^6) ),
+             margins = TRUE )
 ----------------------------
 Histology          D       Y
 ----------------------------
 1             4708.00  127.53
 2             3632.00  127.53
 3              466.00  127.53

 Total         8806.00  382.58
 ----------------------------
```

First step is separate analyses for each subtype.


Rate per 100,000 person−years
10
5
2
1
0.5
20
40
60
Age
1880
1900
1920
1940
1960
1980
2000
Calendar time
5
2.5
1
0.5
0.25
Rate ratio
Seminoma
Non−seminoma


Rate Ratio (ref=1935)
6
5
4
3
2
1
0.8
0.6
0.5
0.4
0.3
1880
1900
1920
1940
1960
1980
Date of birth


Rate Ratio (ref=1960)
6
5
4
3
2
1
0.8
0.6
0.5
0.4
0.3
1950
1960
1970
1980
1990
Date of diagnosis

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( l
> m.ap  <- update( m.apc, . ~ . - Mc:type + Mc )
> m.ac  <- update( m.apc, . ~ . - Mp:type + Mp )
> m.a   <- update( m.ap , . ~ . - Mp:type + Mp )
>
> anova( m.a, m.ac, m.apc, m.ap, m.a, test="Chisq")
Analysis of Deviance Table

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

Analysis of two rates: Formal tests II

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

APC-model: Interactions
Wednesday 5th, morning

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm
`www.biostat.ku.dk/~bxc/APC/MEB-2010`

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.Rdat
dm <- dm[dm$cen=="D1: Denmark",]

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

Analysis of DM-rates: Age×sex interaction III

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

# Age-cohort model with age-sex interaction
# The model matrices for the ML fit
Ma <-                          ns( A, kn=A.kn, Bo=A.ok,
                                   intercept=T )
Mc <- cbind( C-c0, detrend( ns( C, kn=C.kn, Bo=C.ok ),
                                C, weight=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) ),
```

Analysis of DM-rates: Age×sex interaction IV

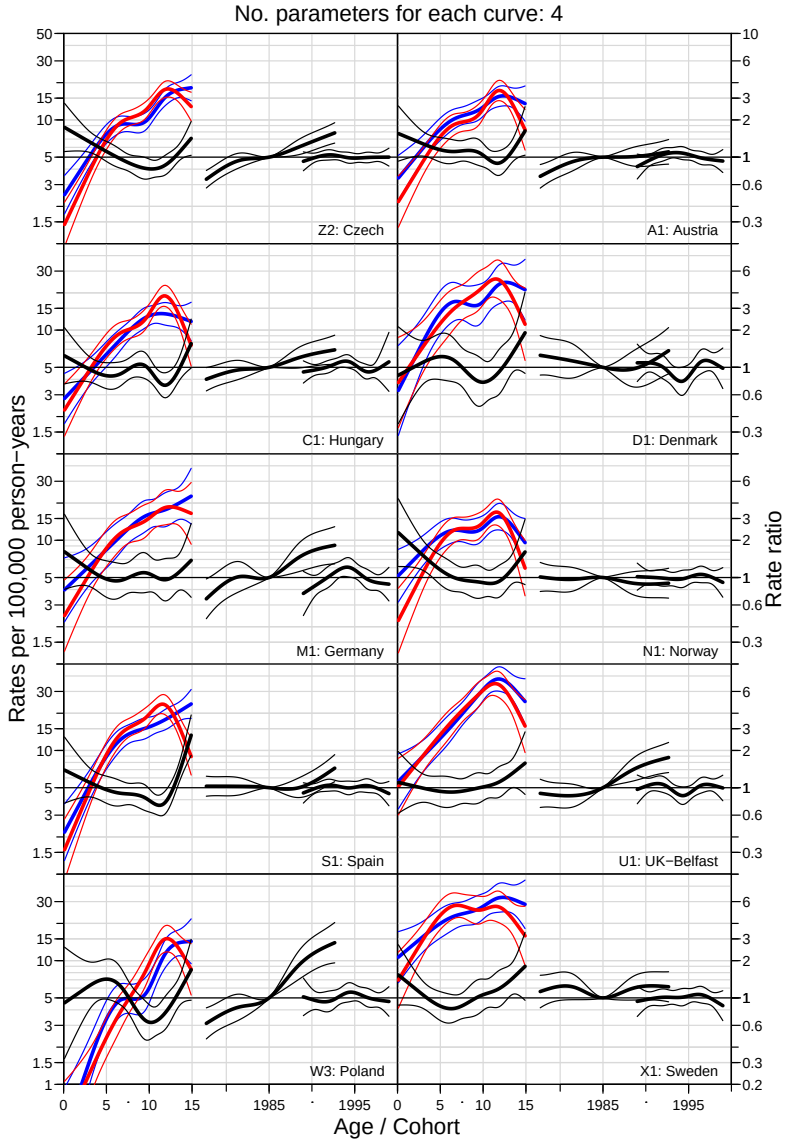
```
                  family=poisson,
                  data=dm )
summary( apcs )
ci.lin( apcs )
ci.lin( apcs, subset="sexF", Exp=T)
ci.lin( apcs, subset="sexF", ctr.mat=Pa, Exp=T)

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

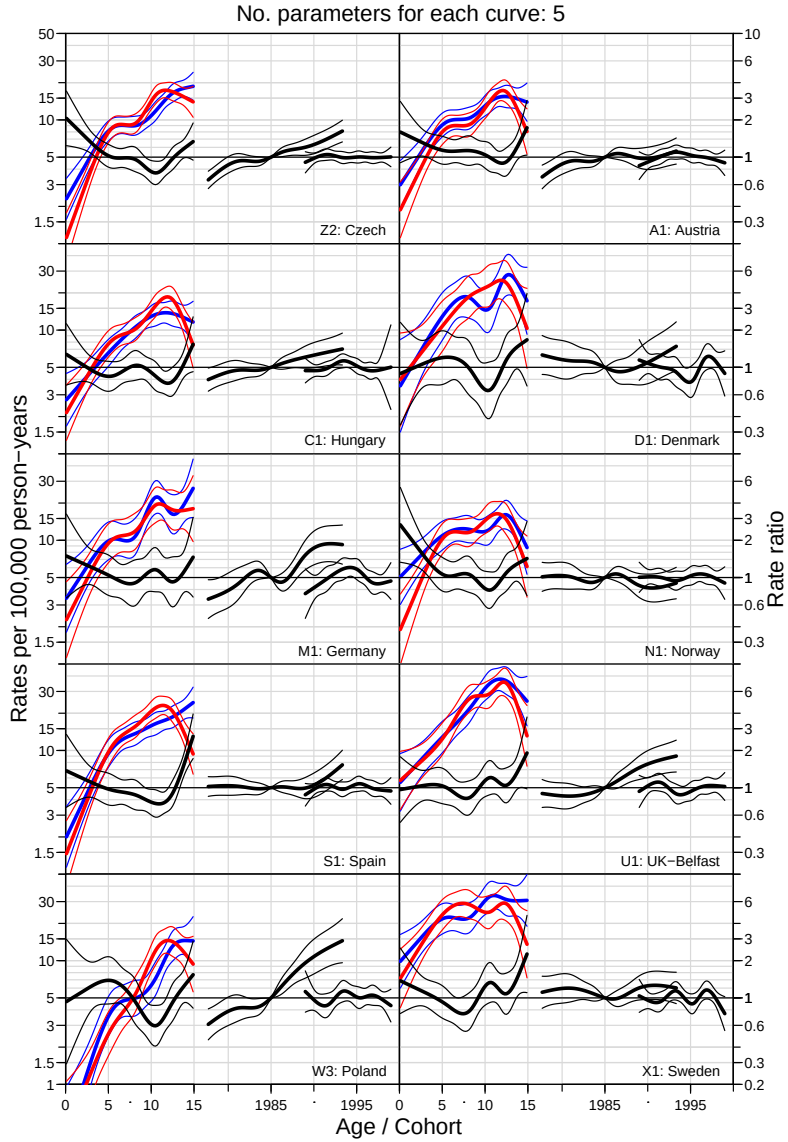
Analysis of DM-rates: Age×sex interaction V

```
# 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),
                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" )
```

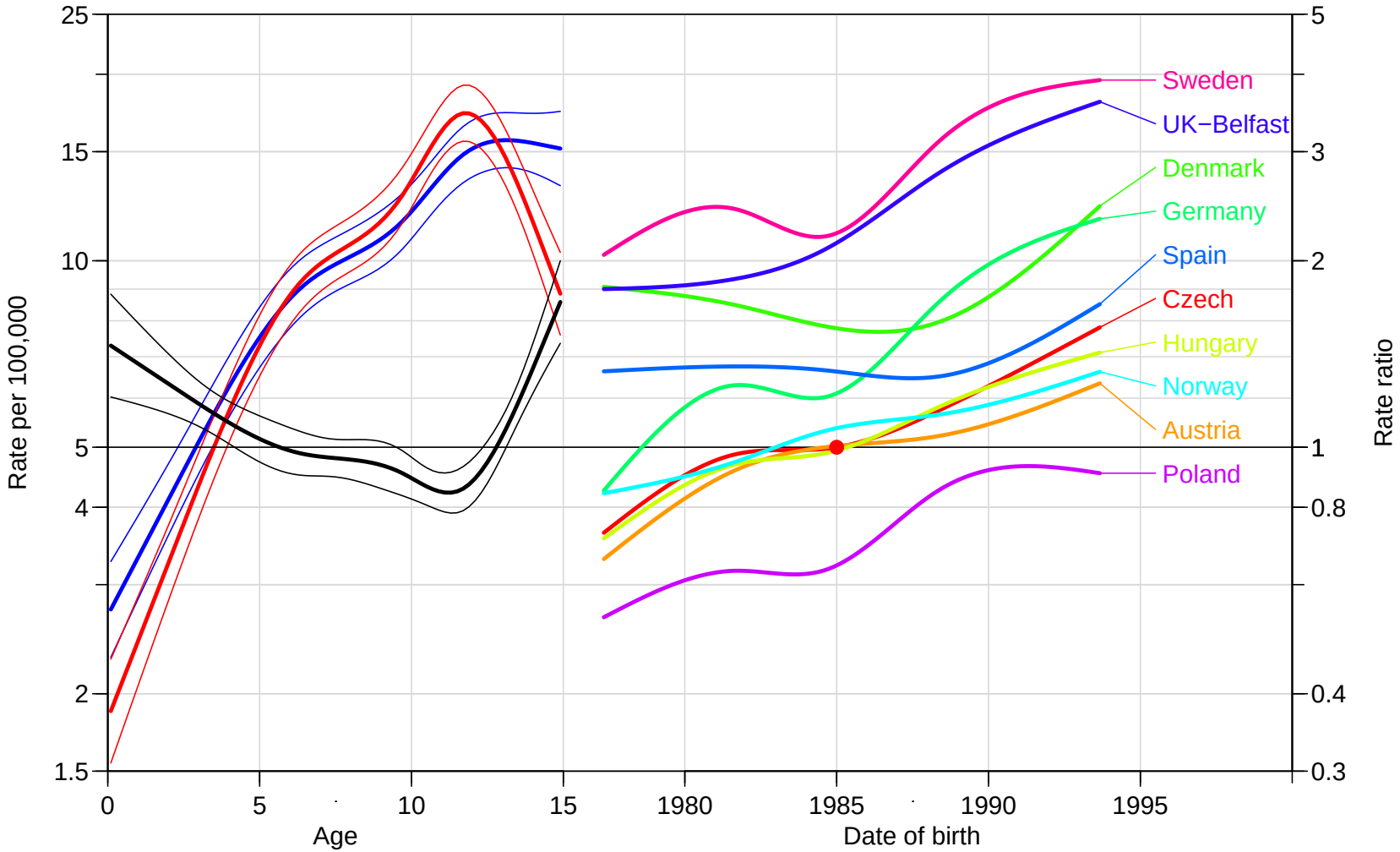
Analysis of DM-rates: Age×sex interaction VI

```
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])
```

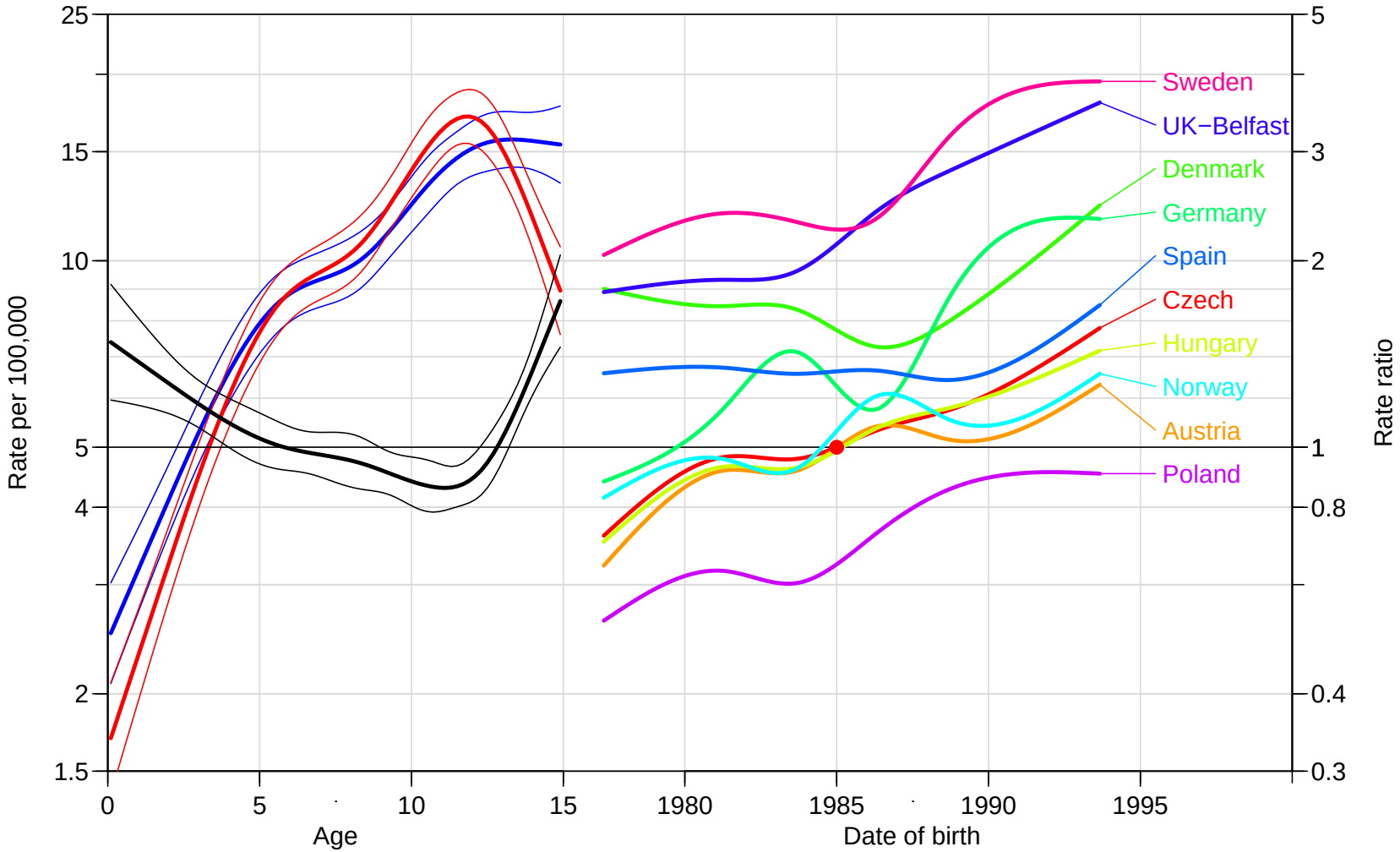

No. parameters for each curve: 4
Rates per 100,000 person−years
Rate ratio
Age / Cohort
Z2: Czech
A1: Austria
C1: Hungary
D1: Denmark
M1: Germany
N1: Norway
S1: Spain
U1: UK−Belfast
W3: Poland
X1: Sweden


No. parameters for each curve: 5
Rates per 100,000 person−years
Rate ratio
Age / Cohort
Z2: Czech
A1: Austria
C1: Hungary
D1: Denmark
M1: Germany
N1: Norway
S1: Spain
U1: UK−Belfast
W3: Poland
X1: Sweden

5 knots for age and coh (4 d.f)



6 knots for age and coh (5 d.f)



Predicting future rates
Wednesday 5th, morning

Bendix Carstensen

Age-Period-Cohort models
May 2010
MEB, Karolinska Institutet, Stockholm
`www.biostat.ku.dk/~bxc/APC/MEB-2010`

Prediction of future rates

Model:

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

- Why not just extend the estimated functions into the future?
- The parametrization curse — the model as stated is not uniquely parametrized.
- Prediction must be invariant under reparametrization.

Identifiability

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

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

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

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

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

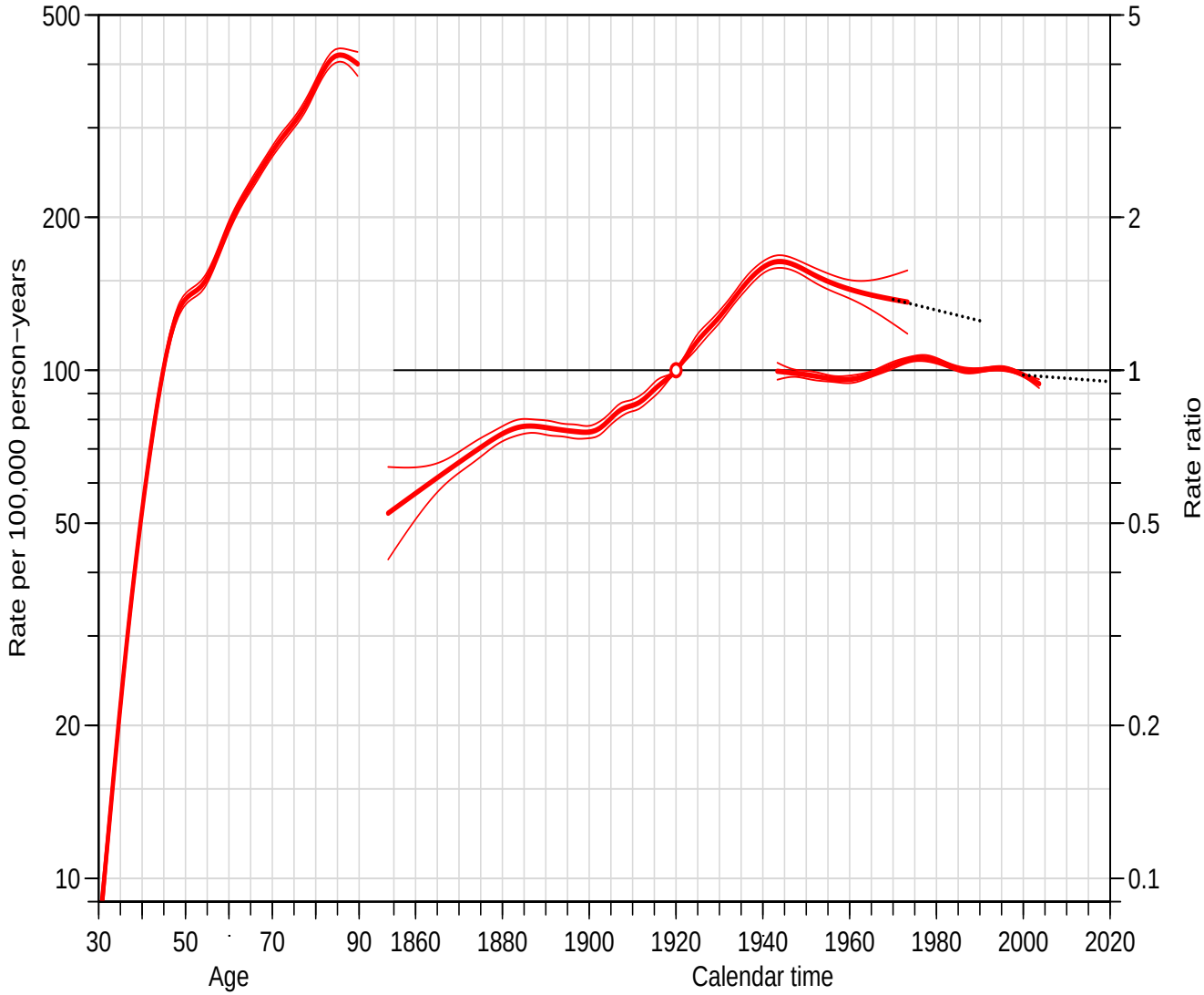
If one of the conditions does *not* hold, the prediction wil depend on the parametrization chosen.

Any linear combination of (known) function values of $g(p)$ and $h(c)$ will work.

Identifiability

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

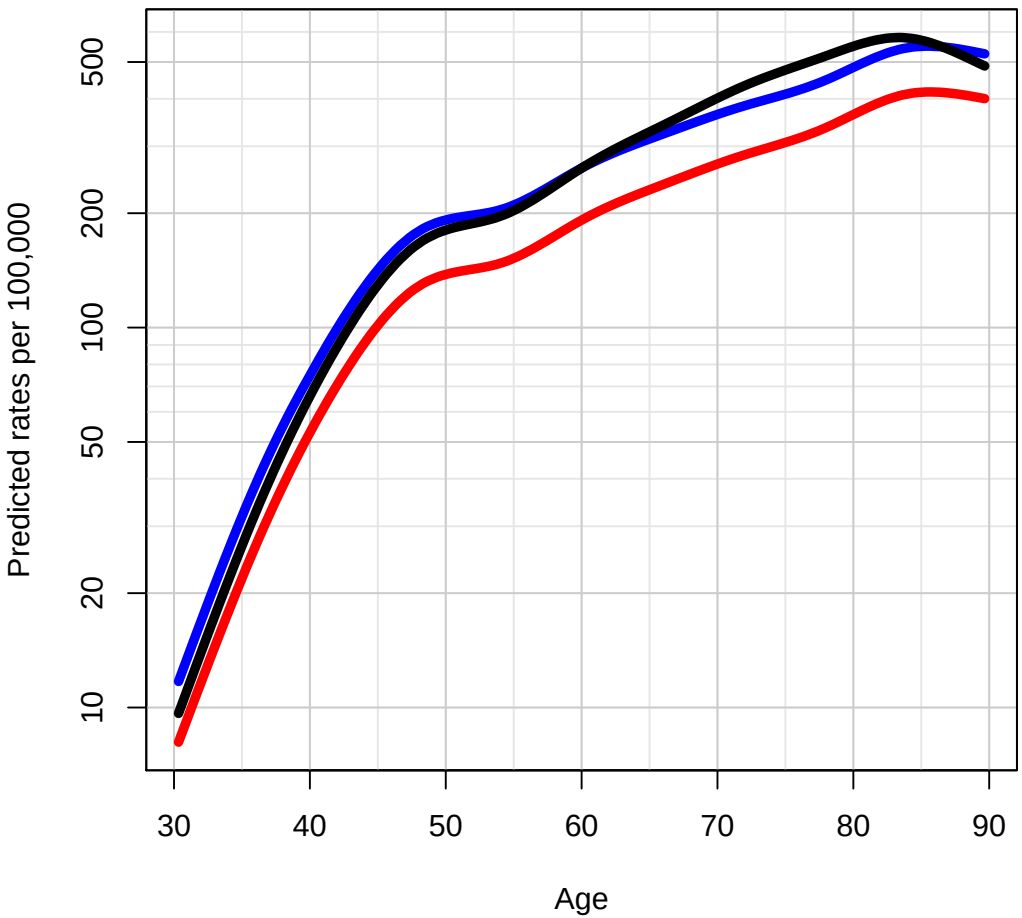
Example: Breast cancer in Denmark



Practicalities

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

Bresat cancer prediction



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

in the 1950 cohort (blue),

and the estimated age-effects (red).