Introduction

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

Welcome

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.

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.

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?

Rates and Survival

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

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

Survival after Cervix cancer

<table>
<thead>
<tr>
<th>Year</th>
<th>Stage I</th>
<th>Stage II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>D</td>
</tr>
<tr>
<td>1</td>
<td>110</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>86</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>72</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>54</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>42</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>33</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>1</td>
</tr>
</tbody>
</table>

Estimated risk in year 1 for Stage I women is \( \frac{5}{107.5} = 0.0465 \)
Estimated 1 year survival is \( 1 - 0.0465 = 0.9535 \) — Life-table estimator.
### Survival function

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

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

Distribution is characterized by the survival function:

$$S(t) = \Pr \{ \text{survival at least till } t \} = 1 - \Pr \{ T \leq t \} = 1 - F(t)$$

### Intensity or rate

$$\lambda(t) = \Pr \{ \text{event in } (t, t+h] \mid \text{alive at } t \} / h$$

$$= \frac{F(t+h) - F(t)}{S(t) \times h}$$

$$= - \frac{S(t+h) - S(t)}{S(t) h} \int_{t}^{t+h} \frac{d\log S(t)}{dt}$$

This is the intensity or hazard function for the distribution.

Characterizes the survival distribution as does $f$ or $F$.

Theoretical counterpart of a rate.

### Relationships

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

$$S(t) = \exp \left( - \int_{0}^{t} \lambda(u) \, du \right) = \exp (-\Lambda(t))$$

$\Lambda(t) = \int_{0}^{t} \lambda(s) \, ds$ is called the integrated intensity or cumulative hazard.

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

### 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!

### 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.

### 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.

### Empirical rates by time since diagnosis.
Two timescales

Note that we actually have two timescales:
- Time since diagnosis (i.e. since entry into the study)
- Calendar time.
These can be shown simultaneously in a Lexis diagram.

Likelihood contribution from one person

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

\[ P(\text{event at } t_i \text{ alive at } t_0) = P(\text{event at } t_i \text{ alive at } t_1) \times P(\text{survive } (t_2, t_2) \text{ alive at } t_2) \times P(\text{survive } (t_1, t_2) \text{ alive at } t_1) \times P(\text{survive } (t_0, t_1) \text{ alive at } t_0) \]

Likelihood contribution from one individual is a product of terms.
Each term refers to one empirical rate \((d, y)\)
\[-\log y = t_i - t_{i-1} \text{ (mostly } d = 0)\]

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.

\[
L(\lambda|y,d) = P(\text{survive } y) \times P(\text{event}^d) = e^{-\lambda y} \times (\lambda dt)^d = \lambda^d e^{-\lambda y} \\
\ell(\lambda|y,d) = d \log(\lambda) - \lambda y
\]

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

Empirical rates by calendar time and time since diagnosis

Follow-up by calendar time and time since diagnosis:
A Lexis diagram!
Aim of dividing time into bands:
- Compute rates in different bands of:
  - age
  - calendar time
  - disease duration
  - ...
- Allow rates to vary along the timescale:

\[
\begin{align*}
0 \log(\lambda) - \lambda y_1 &= 0 \log(\lambda) - \lambda y_1 \\
+ 0 \log(\lambda) - \lambda y_2 &\quad \rightarrow + 0 \log(\lambda_2) - \lambda_2 y_2 \\
+ d \log(\lambda) - \lambda y_3 &\quad + d \log(\lambda_3) - \lambda_3 y_3
\end{align*}
\]

Likelihood-ratio function

Log-likelihood from more persons
- One person, \( p \): \( \sum_i (d_i \log(\lambda_i) - \lambda_i y_{i|p}) \)
- More persons: \( \sum_i \sum_p (d_i \log(\lambda_i) - \lambda_i y_{i|p}) \)
- Collect terms with identical values of \( \lambda_i \):

\[
\sum_i \sum_p (d_i \log(\lambda_i) - \lambda_i y_{i|p}) = \sum_i \left( \sum_p d_i \log(\lambda_i) - \lambda_i \left( \sum_p y_{i|p} \right) \right) = \sum_i \left( \sum_p D_i \log(\lambda_i) - \lambda_i Y_i \right)
\]

- All events in interval \( t \) ("at" time \( t \), \( D_i \))
- All exposure time in interval \( t \) ("at" time \( t \), \( Y_i \))

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

\[
P \{ D = 7, Y = 500 | \lambda \} = \frac{\lambda^7 e^{\lambda Y} \times K}{\lambda^7 e^{\lambda Y} \times K} = L(\lambda | \text{data})
\]
- Best guess of \( \lambda \) is where this function is as large as possible.
- Confidence interval is where it is not too far from the maximum
Likelihood for follow-up of many subjects

Adding empirical rates over the follow-up of persons:

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

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

The log-likelihood is maximal for:

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

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

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

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

Standard error of log-rate: \( \frac{1}{\sqrt{D}} \).

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:

\[ \hat{\lambda} \pm \frac{1.96}{\sqrt{D}} \exp(1.96/\sqrt{D}) \]

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

\[ \lambda \pm 1.96\sqrt{D}/Y \]
**Exercise**

Suppose we have 17 deaths during 843.6 years of follow-up. Calculate the mortality rate with a 95% c.i.

**Rates with glm**

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

**The life table method**

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

\[ \lambda_i = \frac{n_i}{D_i - d_i/2 - l_i/2} \times t_i \]

and hence the death probability:

\[ p_i = 1 - \exp(-\lambda_i t_i) = 1 - \exp \left( - \frac{d_i}{n_i - d_i/2 - l_i/2} \right) \]

The modified life-table estimator.

**Exercise**

Suppose we 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.

\[ RR = \exp \left( 1.96 \frac{1}{D_1} + \frac{1}{D_0} \right) \]

**Population life table, DK 1997–98**

<table>
<thead>
<tr>
<th>Year</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1000000</td>
<td>567</td>
</tr>
<tr>
<td>1</td>
<td>100023</td>
<td>67</td>
</tr>
<tr>
<td>2</td>
<td>100144</td>
<td>74</td>
</tr>
<tr>
<td>3</td>
<td>100271</td>
<td>81</td>
</tr>
<tr>
<td>4</td>
<td>100434</td>
<td>88</td>
</tr>
<tr>
<td>5</td>
<td>101582</td>
<td>95</td>
</tr>
<tr>
<td>6</td>
<td>101734</td>
<td>102</td>
</tr>
<tr>
<td>7</td>
<td>101837</td>
<td>109</td>
</tr>
<tr>
<td>8</td>
<td>101956</td>
<td>116</td>
</tr>
<tr>
<td>9</td>
<td>102085</td>
<td>123</td>
</tr>
<tr>
<td>10</td>
<td>102221</td>
<td>130</td>
</tr>
<tr>
<td>11</td>
<td>102350</td>
<td>137</td>
</tr>
<tr>
<td>12</td>
<td>102485</td>
<td>144</td>
</tr>
<tr>
<td>13</td>
<td>102629</td>
<td>151</td>
</tr>
<tr>
<td>14</td>
<td>102774</td>
<td>158</td>
</tr>
<tr>
<td>15</td>
<td>102928</td>
<td>165</td>
</tr>
<tr>
<td>16</td>
<td>103081</td>
<td>172</td>
</tr>
<tr>
<td>17</td>
<td>103233</td>
<td>179</td>
</tr>
<tr>
<td>18</td>
<td>103384</td>
<td>186</td>
</tr>
</tbody>
</table>
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.

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:

\[
S(t) = e^{-\int_0^t \lambda(s) \, ds} = e^{-\sum \lambda(s)}
\]

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

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?

<table>
<thead>
<tr>
<th>Country</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denmark</td>
<td>( \log_2(\lambda(a)) = -14.244 + 0.135 \text{ age} )</td>
<td>( \log_2(\lambda(a)) = -14.877 + 0.135 \text{ age} )</td>
</tr>
<tr>
<td></td>
<td>Doubling time: 1/0.135 = 7.41 years</td>
<td>Doubling time: 1/0.135 = 7.41 years</td>
</tr>
<tr>
<td></td>
<td>M/F rate-ratio: 2(^{14.244+0.135} = 2^{14.380} = 1.55 )</td>
<td>M/F rate-ratio: 2(^{14.877+0.135} = 2^{15.012} = 1.55 )</td>
</tr>
<tr>
<td></td>
<td>Age-difference: ((-14.244 + 14.877)/0.135 = 4.69 ) years</td>
<td>Age-difference: ((-14.244 + 14.877)/0.135 = 4.69 ) years</td>
</tr>
<tr>
<td>Sweden</td>
<td>( \log_2(\lambda(a)) = -15.453 + 0.146 \text{ age} )</td>
<td>( \log_2(\lambda(a)) = -16.204 + 0.146 \text{ age} )</td>
</tr>
<tr>
<td></td>
<td>Doubling time: 1/0.146 = 6.85 years</td>
<td>Doubling time: 1/0.146 = 6.85 years</td>
</tr>
<tr>
<td></td>
<td>M/F rate-ratio: 2(^{-15.453+0.146} = 2^{-15.599} = 1.68 )</td>
<td>M/F rate-ratio: 2(^{-16.204+0.146} = 2^{-16.350} = 1.68 )</td>
</tr>
<tr>
<td></td>
<td>Age-difference: ((-15.453 + 16.204)/0.146 = 5.14 ) years</td>
<td>Age-difference: ((-15.453 + 16.204)/0.146 = 5.14 ) years</td>
</tr>
</tbody>
</table>
### 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](http://BendixCarstensen/APC/MPIDR-2016)

### 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.

### Cox-likelihood

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

\[
\ell(\beta) = \sum_{\text{death times}} \log \left( \frac{e^{\eta_{\text{death}}}}{\sum_{i \in R_i} e^{\eta_i}} \right)
\]

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

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

### The Cox-likelihood as profile likelihood

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

\[
\log(\lambda(t, x)) = \log(\lambda_0(t)) + \beta_1 x_{1i} + \cdots + \beta_j x_{ji} = \alpha_t + \eta_i
\]

- Profile likelihood:
  - Derive estimates of \( \alpha_t \) as function of data and \( \beta \)
  - assuming constant rate between death times
  - Insert in likelihood, now only a function of data and \( \beta \)
  - Turns out to be Cox’s partial likelihood

### Note:

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

\[
\ell_t(\alpha_t, \beta) = \sum_{i \in R_t} d_i \log(\lambda_i(t)) - \lambda_i(t) y_i
\]

- Suppose the time scale has been divided into small intervals with at most one death in each.
- Assume w.l.o.g. the \( y_i \)s in the empirical rates all are 1.
- Log-likelihood contributions that contain information on a specific time-scale parameter \( \alpha_t \) 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 \( t \).

The derivative w.r.t. \( \alpha_t \) is:

\[
D_{\alpha_t} \ell(\alpha_t, \beta) = 1 - d \sum_{i \in R_t} e^{\eta_i} = 0 \quad \Leftrightarrow \quad e^{\eta_i} = \frac{1}{\sum_{i \in R_t} e^{\eta_i}}
\]

If this estimate is fed back into the log-likelihood for \( \alpha_t \), we get the profile likelihood (with \( \alpha_t \) ‘profiled out’):

\[
\log \left( \frac{1}{\sum_{i \in R_t} e^{\eta_i}} \right) + \eta_{\text{death}} - 1 = \log \left( \frac{e^{\eta_{\text{death}}}}{\sum_{i \in R_t} e^{\eta_i}} \right) - 1
\]

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**

```
> 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 <- chbind( 1, exp( seq(10,1000,10)-5, knots=t.kn ), 60, 1 )
> 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 \( \alpha_t \) 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
  - ...  

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:

\[
S(t) = \exp\left(-\sum_{\tau<t} \exp(g(\tau) + x_0^T \gamma)\right)
\]

This is just a non-linear function of the parameters in the model, \( g \) and \( \gamma \). So the variance can be computed using the \( \delta \)-method.

\( \delta \)-method for survival function

1. Select timepoints \( t_i \) (fairly close).
2. Get estimates of log-rates \( \hat{f}(t_i) = g(t_i) + x_0^T \hat{\gamma} \) for these points:

\[
\hat{f}(t_i) = B \hat{\beta}
\]

where \( \hat{\beta} \) is the total parameter vector in the model.
3. Variance-covariance matrix of \( \hat{\beta} \): \( \Sigma \).
4. Variance-covariance of \( \hat{f}(t_i) \): \( BB' \).
5. Transformation to the rates is the coordinate-wise exponential function, with derivative \( \text{diag}(\exp(\hat{f}(t_i))) \)

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

\[
\text{diag}(e^{\hat{f}(t_i)}) BB' \text{diag}(e^{\hat{f}(t_i)})'
\]

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

\[
\ell \times\begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 & 0
\end{bmatrix} \begin{bmatrix}
e^{\hat{f}(t_i)} \\
e^{\hat{f}(t_i)} \\
e^{\hat{f}(t_i)} \\
e^{\hat{f}(t_i)}
\end{bmatrix} = L \begin{bmatrix}
e^{\hat{f}(t_i)} \\
e^{\hat{f}(t_i)} \\
e^{\hat{f}(t_i)} \\
e^{\hat{f}(t_i)}
\end{bmatrix}
\]

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

\[
L \text{diag}(e^{\hat{f}(t_i)}) BB' \text{diag}(e^{\hat{f}(t_i)})' L'
\]

This is all implemented in the \texttt{ci.cum()} function in \texttt{Epi}.

Practical: Cox and Poisson modelling

\( \text{(non)-Linear models: Estimates and predictions} \)

Statistical Analysis in the
Lexis Diagram:
Age-Period-Cohort models
May 2016
Max Planck Institut for Demographic Research, Rostock
http://BendixCarstensen/APC/MPIDR-2016

\( \text{lin-mod} \)

\( \text{library( Epi )} \)
\( \text{data( diet )} \)
\( \text{names( diet )} \)
\( \text{[1] "id" "sex" "age" "year" "dose" "dose2" "dose3" "sex2" "sex3" } \)
\( \text{[8] "month" "energy" "height" "weight" "fat" "fibre" } \)
\( \text{[15] "chd" } \)
\( \text{with( diet, plot( weight ~ height, pch=16 ) )} \)
\( \text{abline( lm( weight ~ height, data=diet ), col="red", lwd=2 )} \)
\( \text{lm( weight ~ height, data=diet, col="red", lwd=2 }) \)
\( \text{(non)-Linear models: Estimates and predictions (lin-mod)} \)

\( \text{nlm( model ~ x + y, data=yourdata, start=c(x=1, y=2) )} \)
\( \text{with( yourdata, plot( x ~ y ) )} \)
\( \text{abline( lm( x ~ y, data=yourdata ), col="red", lwd=2 )} \)
\( \text{lm( x ~ y, data=yourdata, col="red", lwd=2 }) \)
Linear models, extracting estimates

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

Residuals:
   Min      1Q  Median      3Q     Max
-31.858  -0.8252  9.2616  0.6025  3.9260

Coefficients:
            Estimate Std. Error t value
(Intercept) -31.858     0.0825   -38.16
height        0.7642     0.0825    9.26

Residual standard error: 9.626 on 330 degrees of freedom
Multiple R-squared:  0.9776, Adjusted R-squared:  0.9775
```

Cases, PY and rates

```r
> stat.table( list(A=floor(A/10)*10,
+                 P=floor(P/10)*10),
+                 list( D=sum(D),
+                 Y=sum(Y/1000),
+                 cases( D,P,Y ),
+                 py( D,P,Y ),
+                 rates(D,P,Y) ) )
```

Linear effects in glm

```r
> ml <- glm( D ~ A, offset=log(Y), family=poisson, data=testisDK )
> summary(ml)
Call:
  glm(formula = D ~ A, family = poisson, offset = log(Y), data = testisDK)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-10.375   -0.072    0.203   0.227   3.046

Coefficients: (1 not estimable)
            Estimate Std. Error z value
(Intercept)    1.0055     0.0055   187.49
A              0.0055     0.0064    0.89

Dispersion Parameter for Poisson family taken to be 1
Null deviance: 1874.92 on 18738 degrees of freedom
Residual deviance: 1873.73 on 18736 degrees of freedom
AIC: 5159.48
```

Testis cancer in Denmark:

```r
> library( Epi )
> data( testisDK )
> str(testisDK)
'data.frame': 4860 obs. of 4 variables:
  $ P: int 5376 5348 4628 5969 5471 3914
  $ Y: int 2940 2930 2920 2910 2900 2890
  $ C: int 1 2 3 4 5 6

> head(testisDK)
   A     P     Y     C
1 1940 5376 2940 1
2 1950 5348 2930 2
3 1960 4628 2920 3
4 1970 5969 2910 4
5 1980 5471 2900 5
6 1990 3914 2890 6
```

Linear effects in glm

```r
> ml <- glm( D ~ A, offset=log(Y), family=poisson, data=testisDK )
> summary(ml)
Call:
  glm(formula = D ~ A, family = poisson, offset = log(Y), data = testisDK)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-10.375   -0.072    0.203   0.227   3.046

Coefficients: (1 not estimable)
            Estimate Std. Error z value
(Intercept)    1.0055     0.0055   187.49
A              0.0055     0.0064    0.89

Dispersion Parameter for Poisson family taken to be 1
Null deviance: 1874.92 on 18738 degrees of freedom
Residual deviance: 1873.73 on 18736 degrees of freedom
AIC: 5159.48
```

Non-linear models, prediction

```r
> ml <- glm( D ~ A, offset=log(Y), family=poisson, data=testisDK )
> pr <- predict(ml, newdata=nd, type="link", se.fit=TRUE)
> with(nd, plot(D ~ pr, type="l", lty=1, lwd=c(3,1,1), col="black", log="y") )
```

Linear effects in glm

```r
> ml <- glm( D ~ A, offset=log(Y), family=poisson, data=testisDK )
> summary(ml)
Call:
  glm(formula = D ~ A, family = poisson, offset = log(Y), data = testisDK)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-10.375   -0.072    0.203   0.227   3.046

Coefficients: (1 not estimable)
            Estimate Std. Error z value
(Intercept)    1.0055     0.0055   187.49
A              0.0055     0.0064    0.89

Dispersion Parameter for Poisson family taken to be 1
Null deviance: 1874.92 on 18738 degrees of freedom
Residual deviance: 1873.73 on 18736 degrees of freedom
AIC: 5159.48
```
Linear effects in glm

Quadratic effect in glm

How do rates depend on age?

Quadratic effects in glm

Spline effects in glm
Adding a linear period effect

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

The period effect

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

A quadratic period effect

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

A spline period effect

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

A spline period effect

```r
> msp <- glm( D ~ Ns(A,knots=seq(15,65,10)) + P + I(P^2),
+ offset=log(Y), family=poisson, data=testisDK )
> round( ci.lin( msp ), 3 )
```
Recap 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.
- $\ell_{\text{Poisson}} = d \log(\lambda(t)) - \lambda y$; obs: $(d, y)$, rate par: $\lambda$
- $\ell_{\text{Poisson}} - \ell_{\text{Poisson}} = d \log(\lambda(t))$ does not involve $\lambda$
- Use either to find m.i.e. of $\lambda$
- Poisson model is for $\log(\mu) = \log(\lambda(t)) = \log(\lambda) + \log(y)$
- Hence offset=10log(Y)
- Once rates are known, we can construct survival curves and derivatives of that.

Recap Monday — models
- Empirical rate $(d, y)$ 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)

Follow-up 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
FU-rep-Lexis

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

In a cohort study we have records of:

Events and Risk time.

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

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

Specific for each type of outcome.

Aim of dividing time into bands:

Put

\[ D \] — events
\[ Y \] — risk time

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?

Note: Cumulative exposure is not a timescale.

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 . . .)

- Treat each segment as a separate unit of observation.

- Keep track of exit status in each interval.

Cohort with 3 persons:

<table>
<thead>
<tr>
<th>id</th>
<th>Bdate</th>
<th>Entry</th>
<th>Exit</th>
<th>St</th>
<th>risk int</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14/07/1952</td>
<td>03/08/1965</td>
<td>14/07/1972</td>
<td>0</td>
<td>6.9432 10</td>
</tr>
<tr>
<td>1</td>
<td>14/07/1952</td>
<td>14/07/1972</td>
<td>14/07/1982</td>
<td>0</td>
<td>10.0000 20</td>
</tr>
<tr>
<td>1</td>
<td>14/07/1952</td>
<td>14/07/1992</td>
<td>27/06/1997</td>
<td>1</td>
<td>4.9528 40</td>
</tr>
<tr>
<td>2</td>
<td>01/04/1954</td>
<td>08/09/1972</td>
<td>01/04/1995</td>
<td>0</td>
<td>1.5606 10</td>
</tr>
<tr>
<td>2</td>
<td>01/04/1954</td>
<td>01/04/1974</td>
<td>31/03/1994</td>
<td>0</td>
<td>10.0000 20</td>
</tr>
<tr>
<td>2</td>
<td>01/04/1954</td>
<td>31/03/1994</td>
<td>01/04/1994</td>
<td>0</td>
<td>10.0000 30</td>
</tr>
<tr>
<td>2</td>
<td>01/04/1954</td>
<td>01/04/1994</td>
<td>23/05/1995</td>
<td>0</td>
<td>1.4417 40</td>
</tr>
<tr>
<td>3</td>
<td>10/06/1987</td>
<td>23/12/1991</td>
<td>09/06/1997</td>
<td>0</td>
<td>5.4634 0</td>
</tr>
<tr>
<td>3</td>
<td>10/06/1987</td>
<td>09/06/1997</td>
<td>24/07/1998</td>
<td>1</td>
<td>1.1211 10</td>
</tr>
</tbody>
</table>

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

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.

### Splitting the follow-up

<table>
<thead>
<tr>
<th>subj. 1</th>
<th>subj. 2</th>
<th>subj. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at Entry:</td>
<td>13.06 18.44 4.54</td>
<td></td>
</tr>
<tr>
<td>Age at Exit:</td>
<td>44.95 41.14 11.12</td>
<td></td>
</tr>
<tr>
<td>Status at exit:</td>
<td>Dead Alive Dead</td>
<td></td>
</tr>
<tr>
<td>[ Y ]</td>
<td>31.89 22.70 6.58</td>
<td></td>
</tr>
<tr>
<td>[ D ]</td>
<td>1 0 1</td>
<td></td>
</tr>
</tbody>
</table>
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 0 1938.79 0 22 1976.79 1
2  2 40 1956.61 0 1956.61 20 1964.37 1
3  3 34.52 1955.18 0 1.40 0 1 3 3425 1956.59 1
4  4 4017 2 1936.81 20 1957.61 0 1992.14 2

Timescales of interest:
- Age
- Calendar time
- Time since injection

Definition of Lexis object

> thL <- Lexis( entry = list( age=injecdat-birthdat,
+ per=injecdat,
+ tfi=0 ),
+ exit = list( per=exitdat ),
+ data = th )

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

The looks of a Lexis object

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

<table>
<thead>
<tr>
<th>lex.id</th>
<th>age</th>
<th>per</th>
<th>tfi</th>
<th>lex.dur</th>
<th>lex.Cst</th>
<th>lex.Xst</th>
<th>id</th>
<th>sex</th>
<th>birthdat</th>
<th>contrast</th>
<th>injecdat</th>
<th>volume</th>
<th>exitdat</th>
<th>exitstat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.18</td>
<td>1938.79</td>
<td>0</td>
<td>38.00</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1976.79</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>49.55</td>
<td>1945.77</td>
<td>0</td>
<td>18.60</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>640</td>
<td>1964.37</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>368.21</td>
<td>1955.18</td>
<td>0</td>
<td>1.40</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3425</td>
<td>1956.59</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>40.80</td>
<td>1957.61</td>
<td>0</td>
<td>34.52</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4017</td>
<td>1992.14</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Splitting follow-up time

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

<table>
<thead>
<tr>
<th>lex.id</th>
<th>age</th>
<th>per</th>
<th>tfi</th>
<th>lex.dur</th>
<th>lex.Cst</th>
<th>lex.Xst</th>
<th>id</th>
<th>sex</th>
<th>birthdat</th>
<th>contrast</th>
<th>injecdat</th>
<th>volume</th>
<th>exitdat</th>
<th>exitstat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>1916.61</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>40</td>
<td>1916.61</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>1916.61</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td>1916.61</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td>1916.61</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td></td>
<td>1916.61</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td>1916.61</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td></td>
<td>1916.61</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Split on a second timescale

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

> plot( thL, lwd=3 )
> points( thL, 2:1, pch=c(NA,3)[thL$lex.Xst+1],lwd=3, cex=1.5 )
> plot( thL, lwd=3 )
> plot( thL, 2:1, lwd=5, col=c("red","blue")[thL$contrast], grid=T )
> points( thL, 2:1, pch=c(NA,3)[thL$lex.Xst+1],lwd=3, cex=1.5 )
> points( spl2, 2:1, pch=c(NA,3)[thL$lex.Xst+1],lwd=3, cex=1.5 )
> plot( spl2, lwd=3 )
> plot( spl2, 2:1, lwd=5, col=c("red","blue")[thL$contrast], grid=T )
> points( spl2, 2:1, pch=c(NA,3)[thL$lex.Xst+1],lwd=3, cex=1.5 )
> plot( spl2, lwd=3 )
> plot( spl2, 2:1, lwd=5, col=c("red","blue")[thL$contrast], grid=T )
> points( spl2, 2:1, pch=c(NA,3)[thL$lex.Xst+1],lwd=3, cex=1.5 )
The Poisson likelihood for time-split data

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

\[
D \log(\lambda) - \lambda Y = \sum_{i,t} (d_i \log(\lambda) - \lambda y_i)
\]

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

The model assumes that rates are constant.

But the split data allows relaxing this to models that assume different rates for different \((d_i, y_i)\).

Where are the \((d_i, y_i)\) in the split data?

The Poisson likelihood for time-split data

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

\[
d \log(\lambda y) - \lambda y
\]

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

\[
\log(\mu) = \log(\lambda y) = \log(\lambda) + \log(y)
\]

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

Analysis of results

- \(d_i\) — events in the variable: \(\text{lex.Xst}\).
- \(y_i\) — risk time: \(\text{lex.dur}\) (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 \(\text{glm}\) — no difference between time-scales and other covariates.

Poisson model for split data

- Each interval contribute \(\lambda Y\) to the log-likelihood.
- All intervals with the same set of covariate values (age, exposure, . . . ) have the same \(\lambda\).
- 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.

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/MPIDR-2016

tab-mod

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

> head(spl2, 2)

<table>
<thead>
<tr>
<th>lex.id</th>
<th>age</th>
<th>per</th>
<th>tfi</th>
<th>lex.dur</th>
<th>lex.Cat</th>
<th>lex.Xst</th>
<th>id</th>
<th>sex</th>
<th>birthdat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1916.61</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1916.61</td>
</tr>
</tbody>
</table>

Conceptual set-up

Follow-up of the entire (male) population from 1943–2006 w.r.t. occurrence of testiscancer:
- Split follow-up time for all about 4 mio. men in 1-year classes by age and calendar time \((y)\).
- Allocate testis cancer event \((d = 0, 1)\) to each.
- Analyse all 200,000,000 records by a Poisson model.
Realistic set-up
- Tabulate the follow-up time and events by age and period.
- 100 age-classes.
- 65 periods (single calendar years).
- 6500 aggregate records of \((D, Y)\).
- Analyze by a Poisson model.

Practical set-up
- Tabulate only events (as obtained from the cancer registry) by age and period.
- 100 age-classes.
- 65 periods (single calendar years).
- 6500 aggregate records of \(D\).
- Estimate the population follow-up based on census data from Statistics Denmark.
- Or get it from the human mortality database.
- Analyse by Poisson model.

Register data
Classification of cases \((Dap)\) by age at diagnosis and date of diagnosis, and population \((Yap)\) by age at risk and date at risk, in compartments of the Lexis diagram, e.g.:

<table>
<thead>
<tr>
<th>Age</th>
<th>1943</th>
<th>1948</th>
<th>1953</th>
<th>1958</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>7</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>25</td>
<td>28</td>
<td>23</td>
<td>26</td>
<td>35</td>
</tr>
<tr>
<td>30</td>
<td>28</td>
<td>43</td>
<td>49</td>
<td>51</td>
</tr>
<tr>
<td>35</td>
<td>36</td>
<td>42</td>
<td>39</td>
<td>44</td>
</tr>
<tr>
<td>40</td>
<td>24</td>
<td>32</td>
<td>46</td>
<td>53</td>
</tr>
<tr>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
</tr>
</tbody>
</table>

Lexis diagram

Registration of:
- cases \((D)\)
- risk time, person-years \((Y)\)
in subsets of the Lexis diagram.

Rates available in each subset.

Reshape data to analysis form:

<table>
<thead>
<tr>
<th>A</th>
<th>P</th>
<th>D</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>1943</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>1943</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>1943</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>1943</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>1943</td>
<td>36</td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>1943</td>
<td>32</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>1948</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>1948</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>1948</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>1948</td>
<td>43</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>1948</td>
<td>42</td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>1948</td>
<td>32</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>1953</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>1953</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>1953</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>1953</td>
<td>49</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>1953</td>
<td>39</td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>1953</td>
<td>46</td>
</tr>
</tbody>
</table>

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 level of classifying factors.
- Output from the model will be rates and rate-ratios.
- Since we use multiplicative Poisson, usually the log rates and the log-RR are reported.
Simple model for the testis cancer rates:

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

Coefficients:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>z</th>
<th>P</th>
<th>exp(Est.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.4758</td>
<td>0.327</td>
<td>-4.517</td>
<td>0.000</td>
</tr>
<tr>
<td>factor(A)20</td>
<td>1.4539</td>
<td>0.345</td>
<td>4.135</td>
<td>0.000</td>
</tr>
<tr>
<td>factor(A)25</td>
<td>2.5321</td>
<td>0.330</td>
<td>7.671</td>
<td>0.000</td>
</tr>
<tr>
<td>factor(A)30</td>
<td>2.9330</td>
<td>0.325</td>
<td>9.013</td>
<td>0.000</td>
</tr>
<tr>
<td>factor(A)35</td>
<td>2.8613</td>
<td>0.325</td>
<td>8.779</td>
<td>0.000</td>
</tr>
<tr>
<td>factor(A)40</td>
<td>2.8531</td>
<td>0.328</td>
<td>8.743</td>
<td>0.000</td>
</tr>
<tr>
<td>factor(P)1948</td>
<td>0.1753</td>
<td>0.121</td>
<td>1.447</td>
<td>0.148</td>
</tr>
<tr>
<td>factor(P)1953</td>
<td>0.3822</td>
<td>0.116</td>
<td>3.286</td>
<td>0.001</td>
</tr>
<tr>
<td>factor(P)1958</td>
<td>0.466</td>
<td>0.115</td>
<td>4.052</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Residual deviance: 17.532 on 15 degrees of freedom
AIC: 149.53

Models for tabulated data (tab-mod) 146/327

Register data - rates
Rates in "tiles" of the Lexis diagram:

\[ \lambda(a, p) = \frac{D_{ap}}{Y_{ap}} \]

Descriptive epidemiology based on disease registers:

- 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

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

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

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>z</th>
<th>P</th>
<th>exp(Est.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor(P)1948</td>
<td>0.175</td>
<td>0.121</td>
<td>1.447</td>
<td>0.192</td>
</tr>
<tr>
<td>factor(P)1953</td>
<td>0.382</td>
<td>0.116</td>
<td>3.286</td>
<td>0.001</td>
</tr>
<tr>
<td>factor(P)1958</td>
<td>0.466</td>
<td>0.115</td>
<td>4.052</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Rate ratios are computed on the fly by `Exp=TRUE`:

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

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>z</th>
<th>P</th>
<th>exp(Est.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor(P)1948</td>
<td>0.175</td>
<td>0.121</td>
<td>1.447</td>
<td>0.192</td>
</tr>
<tr>
<td>factor(P)1953</td>
<td>0.382</td>
<td>0.116</td>
<td>3.286</td>
<td>0.001</td>
</tr>
<tr>
<td>factor(P)1958</td>
<td>0.466</td>
<td>0.115</td>
<td>4.052</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Models for tabulated data (tab-mod) 150/327

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.
Lexis diagram: data
Testis cancer cases in Denmark.
Male person-years in Denmark.

Data matrix: Testis cancer cases

<table>
<thead>
<tr>
<th>Date of diagnosis (year)</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1940</td>
<td>0-4</td>
</tr>
<tr>
<td>1941</td>
<td>5-9</td>
</tr>
<tr>
<td>1942</td>
<td>10-14</td>
</tr>
<tr>
<td>1943</td>
<td>15-19</td>
</tr>
<tr>
<td>1944</td>
<td>20-24</td>
</tr>
<tr>
<td>1945</td>
<td>25-29</td>
</tr>
<tr>
<td>1946</td>
<td>30-34</td>
</tr>
<tr>
<td>1947</td>
<td>35-39</td>
</tr>
<tr>
<td>1948</td>
<td>40-44</td>
</tr>
<tr>
<td>1949</td>
<td>45-49</td>
</tr>
<tr>
<td>1950</td>
<td>50-54</td>
</tr>
<tr>
<td>1951</td>
<td>55-59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of cases</th>
<th>1000 person-years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td></td>
<td>48-52</td>
</tr>
<tr>
<td></td>
<td>53-57</td>
</tr>
<tr>
<td></td>
<td>58-62</td>
</tr>
<tr>
<td></td>
<td>63-67</td>
</tr>
<tr>
<td></td>
<td>68-72</td>
</tr>
<tr>
<td></td>
<td>73-77</td>
</tr>
<tr>
<td></td>
<td>78-82</td>
</tr>
<tr>
<td></td>
<td>83-87</td>
</tr>
<tr>
<td></td>
<td>88-92</td>
</tr>
</tbody>
</table>

Data matrix: Male risk time

<table>
<thead>
<tr>
<th>Date of diagnosis (1900)</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1940</td>
<td>0-4</td>
</tr>
<tr>
<td>1941</td>
<td>5-9</td>
</tr>
<tr>
<td>1942</td>
<td>10-14</td>
</tr>
<tr>
<td>1943</td>
<td>15-19</td>
</tr>
<tr>
<td>1944</td>
<td>20-24</td>
</tr>
<tr>
<td>1945</td>
<td>25-29</td>
</tr>
<tr>
<td>1946</td>
<td>30-34</td>
</tr>
<tr>
<td>1947</td>
<td>35-39</td>
</tr>
<tr>
<td>1948</td>
<td>40-44</td>
</tr>
<tr>
<td>1949</td>
<td>45-49</td>
</tr>
<tr>
<td>1950</td>
<td>50-54</td>
</tr>
<tr>
<td>1951</td>
<td>55-59</td>
</tr>
</tbody>
</table>

Data matrix: Empirical rates

<table>
<thead>
<tr>
<th>Date of diagnosis (1900)</th>
<th>Rate per 1000000 person-years</th>
</tr>
</thead>
<tbody>
<tr>
<td>1940</td>
<td>0-4</td>
</tr>
<tr>
<td>1941</td>
<td>5-9</td>
</tr>
<tr>
<td>1942</td>
<td>10-14</td>
</tr>
<tr>
<td>1943</td>
<td>15-19</td>
</tr>
<tr>
<td>1944</td>
<td>20-24</td>
</tr>
<tr>
<td>1945</td>
<td>25-29</td>
</tr>
<tr>
<td>1946</td>
<td>30-34</td>
</tr>
<tr>
<td>1947</td>
<td>35-39</td>
</tr>
<tr>
<td>1948</td>
<td>40-44</td>
</tr>
<tr>
<td>1949</td>
<td>45-49</td>
</tr>
<tr>
<td>1950</td>
<td>50-54</td>
</tr>
<tr>
<td>1951</td>
<td>55-59</td>
</tr>
</tbody>
</table>
**Fitting the model in R**

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

```R
> 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 = poisson)
```

**Coefficients:**

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| `factor(A)17.5` | -3.56605 | 0.07249 | -49.194 | < 2e-16 |
| `factor(A)22.5` | -2.35347 | 0.04992 | -47.766 | < 2e-16 |
| `factor(A)27.5` | -1.94005 | 0.04827 | -41.204 | < 2e-16 |
| `factor(A)32.5` | -1.99514 | 0.04917 | -41.176 | < 2e-16 |
| `factor(A)37.5` | -0.18734 | 0.05044 | -37.176 | < 2e-16 |
| `factor(A)42.5` | -0.09134 | 0.04944 | -18.517 | < 2e-16 |
| `factor(A)47.5` | -0.28743 | 0.05279 | -54.917 | < 2e-16 |

**Deviance Residuals:**

- Min 1Q Median 3Q Max
- -3.0926 -0.8784 0.1144 0.9790 2.7653

**Age-period model**

Rates are proportional between periods:

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

Choose \( p_0 \) as reference period, where \( \beta_{p_0} = 0 \):

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

**Fit the model in R**

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

```R
> 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 = poisson)
```

**Coefficients:**

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| `factor(C)1960` | -0.72942 | 0.05683 | -12.983 | < 2e-16 |
| `factor(C)1970` | -0.50614 | 0.05046 | -10.104 | < 2e-16 |
| `factor(C)1980` | -0.38362 | 0.04966 | -7.728 | < 2e-16 |
| `factor(C)1990` | -0.26112 | 0.04857 | -5.387 | < 2e-16 |
| `factor(C)2000` | -0.14862 | 0.04748 | -3.125 | < 2e-16 |
| `factor(C)2010` | -0.03612 | 0.04638 | -0.783 | 0.4376 |

**Deviance Residuals:**

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

**Age-cohort model**

Rates are proportional between cohorts:

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

Choose \( c_0 \) as reference cohort, where \( \gamma_{c_0} = 0 \):

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

Age and linear effect of period:
\[ \log[\lambda(a,p)] = \alpha_a + \beta(p - p_0) \]
that is, \( \beta_p = \beta(p - p_0) \).

Linear effect of cohort:
\[ \log[\lambda(a,p)] = \alpha_a + \gamma_c = \alpha_a + \gamma(c - c_0) \]
that is, \( \gamma_c = \gamma(c - c_0) \)

What goes on?
\[ \alpha_a + \beta(p - p_0) = \alpha_a + \beta(a + c - (a_0 + c_0)) = \alpha_a + \beta(a - a_0) + \beta(c - c_0) \]

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 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

Rate ratio
Age-at-entry
Non-linear effects of time-scales

Arbitrary effects of the three variables \( t \), \( a \) and \( e \): \( \rightarrow \) genuine extension of the model.

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

Three quantities can be arbitrarily moved between the three functions:

\[
\begin{align*}
\tilde{f}(t) &= f(a) - \mu_a - \mu_t + \gamma t \\
\tilde{g}(a) &= g(p) + \mu_a - \gamma a \\
\tilde{h}(e) &= h(c) + \mu_e + \gamma e
\end{align*}
\]

because \( t - a + e = 0 \).

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

“Controlling for age”

— is not a well defined statement.

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

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

This would require modelling of multiple timescales.

Which is best accomplished by splitting time.

Age-Period-Cohort model

Statistical Analysis in the Lexis Diagram:

Age-Period-Cohort models

May 2016

Max Planck Institute for Demographic Research, Rostock

http://BendixCarstensen/APC/MPIDR-2016

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).

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 \( C \)

Fitting the model in R I

\[
> \text{library(}\text{Epi})
> \text{load( file="../data/testisDK.Rda")}
> \text{head( testisDK )}
\]

Fitting the model in R II

\[
> \text{glm(formula = D ~ factor(A) + factor(P) + factor(P - A), family = poisson, data = testisDK, offset = log(Y))}
\]

Fitting the model in R III

\[
> \text{m.apc <- glm( D ~ factor( A ) + factor( P ) + factor( P - A ), family = poisson, data = testisDK )}
\]

Age-Period-Cohort model (APC-cat)
Fitting the model in R IV

No. of parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age - drift</td>
<td>126.07</td>
<td>71</td>
<td>0.000</td>
</tr>
<tr>
<td>∆</td>
<td>60.60</td>
<td>15</td>
<td>0.000</td>
</tr>
<tr>
<td>Age - cohort</td>
<td>65.47</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>∆</td>
<td>30.01</td>
<td>7</td>
<td>0.000</td>
</tr>
<tr>
<td>Age - period - cohort</td>
<td>35.46</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>∆</td>
<td>82.24</td>
<td>15</td>
<td>0.000</td>
</tr>
<tr>
<td>Age - period</td>
<td>117.70</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>∆</td>
<td>8.37</td>
<td>7</td>
<td>0.301</td>
</tr>
<tr>
<td>Age - drift</td>
<td>126.07</td>
<td>71</td>
<td></td>
</tr>
</tbody>
</table>

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:
  \[
  \lambda(a, p) = \tilde{\alpha}_a + \tilde{\gamma}_c
  \]

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:

\[
g(p) = \tilde{\beta}_p = \beta_p - \tilde{\delta}_p
\]

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

\[
h(c) = \gamma_c - \gamma_{c_0} + \tilde{\delta}_p(c - c_0)
\]

The rest is the age-effect:

\[
f(a) = \hat{\alpha}_a + \hat{\mu}_p + \hat{\delta}_p a + \hat{\delta}_p c_0 + \gamma_{c_0}
\]

How it all adds up:

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

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

Test for effects

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age - drift</td>
<td>126.07</td>
<td>71</td>
<td>0.000</td>
</tr>
<tr>
<td>∆</td>
<td>60.60</td>
<td>15</td>
<td>0.000</td>
</tr>
<tr>
<td>Age - cohort</td>
<td>65.47</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>∆</td>
<td>30.01</td>
<td>7</td>
<td>0.000</td>
</tr>
<tr>
<td>Age - period - cohort</td>
<td>35.46</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>∆</td>
<td>82.24</td>
<td>15</td>
<td>0.000</td>
</tr>
<tr>
<td>Age - period</td>
<td>117.70</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>∆</td>
<td>8.37</td>
<td>7</td>
<td>0.301</td>
</tr>
<tr>
<td>Age - drift</td>
<td>126.07</td>
<td>71</td>
<td></td>
</tr>
</tbody>
</table>

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:

- $\alpha$ has 9 levels
- $\beta$ has 9 levels
- $\gamma$ has 17 levels
- $\delta$ has 9 levels
- $\mu$ has 17 levels
- $\nu$ has 17 levels
- $\rho$ has 9 levels

The missing parameter is because of the identifiability problem.
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:
  \[
  \log[\lambda(a, p)] = \hat{\alpha}_a + \hat{\gamma}_c + \beta_p
  \]

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

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

- 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.
- \( \beta_p \)'s are estimated in a Poisson model with this as offset.
- Advantage: We get the standard errors for free.
Using `contr.2nd IV`:

```r
taxplot( sort(unique(testisDK$A)),
  cbind(ci.exp(mp,subset="(A")$),
  log=TRUE, xlab="Age", ylab="Incidence rate per 100,000 PY",
  type="l",lty=1.3,lwd=c(3,1.7),col=rep(c("red","blue"),each=2) )
```

**Tabulation of register data**

Testis cancer cases in Denmark.

Male person-years in Denmark.

**Tabulation in the Lexis diagram**

Statistical Analysis in the Lexis Diagram:

- Age-Period-Cohort models
- May 2016
- Max Planck Institut for Demographic Research, Rostock

**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 in the Lexis diagram**

**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 $a = p - e$. 
Analysis of rates from a complete observation in a Lexis diagram need not be restricted to these classical sets classified by two factors.

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

- **Upper triangles**: Classification by age and period, earliest born cohort. (\( \Gamma \))
- **Lower triangles**: Classification by age and cohort, last born cohort. (\( \triangle \))

**Mean time in triangles**

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

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

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

**Upper triangles (\( \Gamma \)), A:**

\[
E_a(\alpha) = \int_{p=0}^{p=1} \int_{a=0}^{a=1} a \times 2 \, da \, dp = \int_{p=0}^{p=1} 1 - p^2 \, dp = \frac{1}{2}
\]

\[
E_a(\rho) = \int_{a=0}^{a=1} \int_{p=0}^{p=1} p \times 2 \, dp \, da = \int_{a=0}^{a=1} a^2 \, da = \frac{1}{6}
\]

\[
E_a(\zeta) = \frac{1}{2} - \frac{1}{3} = -\frac{1}{3}
\]

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

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

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

\[
E_B(\zeta) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}
\]

**Tabulation by age, period and cohort**

Given 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 bureaus.

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

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

**Prevalent population figures**

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

The aim is to compute person-years for the triangles \( A \) and \( B \), respectively.
The area of the triangle is $\frac{1}{2}$, 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:

$$\int_{p=0}^{p=1} \int_{a=0}^{a=p} 2p \, da \, dp = \int_{p=0}^{p=1} 2p - 2p^2 \, dp = \left[ \frac{p^2}{\frac{1}{2}} - \frac{p^3}{3} \right]_{p=0}^{p=1} = \frac{1}{3}$$

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

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

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

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

**Contributions to risk time in $A$ and $B$:**

<table>
<thead>
<tr>
<th></th>
<th>$A$: $\ell_{a+1,p+1} \times \frac{1}{2}y$</th>
<th>$B$: $\ell_{a+1,p+1} \times \frac{1}{2}y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survivors:</td>
<td>$\ell_{a+1,p+1} \times \frac{1}{2}y$</td>
<td>$\ell_{a+1,p+1} \times \frac{1}{2}y$</td>
</tr>
<tr>
<td>Dead in $A$:</td>
<td>$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{2}y$</td>
<td>$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{2}y$</td>
</tr>
<tr>
<td>Dead in $B$:</td>
<td>$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{2}y$</td>
<td>$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{2}y$</td>
</tr>
<tr>
<td>$\sum$</td>
<td>$(\ell_{a,p} + \ell_{a+1,p+1}) \times \frac{1}{2}y$</td>
<td>$(\ell_{a,p} + \ell_{a+1,p+1}) \times \frac{1}{2}y$</td>
</tr>
</tbody>
</table>

**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
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:
\[ D_{ap}(\lambda_{ap}) = \lambda_{ap} Y_{ap} = \lambda_{ap} \log((\alpha_a + \beta_p + \gamma_c) - (\alpha_a + \beta_p + \gamma_c) Y_{ap}) \]
The log-likelihood can be separated:
\[ \sum_{a,p \in \Delta} D_{ap}(\lambda_{ap}) - \lambda_{ap} Y_{ap} + \sum_{a,p \in \Delta} D_{ap}(\lambda_{ap}) - \lambda_{ap} Y_{ap} \]
No common parameters between terms — we have two separate models:
One for upper triangles, one for lower.

Illustration by lung cancer data
\[
\begin{array}{cccccc}
& A5 & P5 & C5 & up & Ax \ 
1 & 40 & 1943 & 1958 & 1 & 67 \\
2 & 40 & 1943 & 1903 & 1 & 67 \\
3 & 40 & 1944 & 1906 & 1 & 67 \\
4 & 40 & 1948 & 1906 & 1 & 67 \\
5 & 40 & 1953 & 1911 & 1 & 67 \\
6 & 40 & 1953 & 1911 & 1 & 67 \\
7 & 40 & 1953 & 1911 & 1 & 67 \\
8 & 40 & 1953 & 1911 & 1 & 67 \\
9 & 40 & 1953 & 1911 & 1 & 67 \\
10 & 40 & 1953 & 1911 & 1 & 67 \\
\end{array}
\]

APC-model for triangular data (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 with “synthetic” cohorts
\[
\begin{align*}
\text{> mc } & \leftarrow \text{ glm( D } \sim \text{ factor(A5) } + 1 + \\
& \text{ factor(P5) } + \text{ offset( log( Y ) ),} \\
& \text{ family=poisson) }
\end{align*}
\]
\[
\begin{align*}
& \text{summary( mc )} \\
& \text{Null deviance: 1.0037e+02 on 220 degrees of freedom} \\
& \text{Residual deviance: 8.6866e+02 on 182 degrees of freedom}
\end{align*}
\]
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
\[
\begin{align*}
\text{> mx } & \leftarrow \text{ glm( D } \sim \text{ factor(Ax) } + 1 + \\
& \text{ factor(Cx) } + \text{ factor(Px) } + \text{ offset( log( Y ) ),} \\
& \text{ family=poisson) }
\end{align*}
\]
\[
\begin{align*}
& \text{summary( mx )} \\
& \text{Null deviance: 1.0037e+02 on 220 degrees of freedom} \\
& \text{Residual deviance: 2.8473e+02 on 144 degrees of freedom}
\end{align*}
\]
No. parameters: 220 – 144 = 76 ( = 38 × 2).
\[
A = 20, \quad P = 22, \quad C = 40 \quad \Rightarrow \quad A + P + C - 3 = 79 \neq 76!
\]
We have fitted two age-period-cohort models separately to upper and lower triangles.
The identifiability problem still exists:
Rate per 100,000
Age-Period-Cohort models
May 2016
Max Planck Institut for Demographic Research, Rostock
http://BendixCarstensen/APC/MPIDR-2016

What’s the problem?
- One parameter is assigned to each distinct value of the timescales, the scale of the variables is not used.
- The solution is to “tie together” the points on the scales together with smooth functions of the mean age, period and cohort with three functions:
  \[ \lambda_{ap} = f(a) + g(p) + h(c) \]
- The practical problem is how to choose a reasonable parametrization of these functions, and how to get estimates.

The identifiability problem still exists:
\[ c = p - a \quad \Rightarrow \quad p - a - c = 0 \]
\[ \lambda_{ap} = f(a) + g(p) + h(c) \]
\[ = f(a) + g(p) + h(c) + \gamma(p - a - c) \]
\[ = f(a) - \mu_a \quad - \quad \gamma a + g(p) + \mu_c + \gamma p + h(c) \quad - \quad \mu_c - \gamma c \]

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

Smooth functions

**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

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 of effects
There are still three “free” parameters:
\[ f(a) = f(a) - \mu_a - \gamma a \]
\[ g(p) = g(p) + \mu_a + \mu_c + \gamma p \]
\[ h(c) = h(c) - \mu_c - \gamma c \]

Any set of 3 numbers, \( \mu_a, \mu_c, \gamma \) will produce effects with teh same sum. Choose \( \mu_a, \mu_c, \gamma \) 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?

**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...

**Implementation:**

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

These functions fulfill the criteria.

**Parametric function**

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

Example: 2nd order polynomial:

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

\( \text{nrow(M) is the no. of observations in the dataset, ncol(M) is the no. of parameters} \)

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

**Extract the trend from \( f \):**

1. Set up model matrices for age, period and cohort, \( M_a, M_p, \) and \( M \). 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 \( \tilde{M}_{c_0} \).

**Practical parametization**

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

4. Use:
   - \( M_a \) for the age-effects,
   - \( M_p \) for the period effects and
   - \( |c - c_0| M_{c_0} \) for the cohort effects.
5. Value of \( \tilde{f}(a) \) is \( M_a\tilde{f}_a \), similarly for the other two effects.

Variance is found by \( M_a^\prime\Sigma_a M_a \), where \( \Sigma_a \) is the variance-covariance matrix of \( \tilde{f}_a \).
Information in the data and inner product

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

\[
l(\theta | D, Y) = D\theta - e^\theta Y, \quad i^o = D - e^\theta Y, \quad i'' = -e^\theta Y
\]

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

Log-likelihood for an observation \((D, Y)\), with rate \(\lambda\):

\[
l(\lambda | D, Y) = D\lambda - \lambda Y, \quad i^o = D/\lambda - Y, \quad i'' = -D/\lambda^2,
\]

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

How to? III

**NOTE:** \(\text{spar}\) is specified as: A P C

8 8 8

1 "ML of APC-model Poisson with log(Y) offset : ( ACP ) : ^n\n
Analysis of deviance for Age-Period-Cohort model

<table>
<thead>
<tr>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Df</th>
<th>Deviance</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>212</td>
<td>15468.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age-drift</td>
<td>211</td>
<td>6858.9</td>
<td>1</td>
<td>8609.7 &lt; 2.2e-16</td>
</tr>
<tr>
<td>Age-Cohort</td>
<td>205</td>
<td>1034.7</td>
<td>6</td>
<td>5824.1 &lt; 2.2e-16</td>
</tr>
<tr>
<td>Age-Period</td>
<td>205</td>
<td>3082.6</td>
<td>-6</td>
<td>-2659.4 &lt; 2.2e-16</td>
</tr>
<tr>
<td>Age-drift</td>
<td>211</td>
<td>6858.9</td>
<td>-6</td>
<td>-3776.3 &lt; 2.2e-16</td>
</tr>
</tbody>
</table>

> plot( mw )

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
Age-Cohort 205 1034.7
Age-Period 205 3082.6
Age-drift 211 6858.9

How to? IV

Consult the help page \texttt{apc.fit} to see options for weights in inner product, type of function, variants of parametrization etc.

\texttt{apc.plot, apc.lines} and \texttt{apc.frame} to see how to plot the results.

How to? I

Implemented in \texttt{apc.fit} in the Epi package:

\[
> \text{library( Epi )}
\]

\[
> \text{sessionInfo()}
\]

Results.

Other models I

consult the help page \texttt{apc.fit} to see options for weights in inner product, type of function, variants of parametrization etc.

\texttt{apc.plot, apc.lines} and \texttt{apc.frame} to see how to plot the results.

Other models II

Consult the help page \texttt{apc.fit} to see options for weights in inner product, type of function, variants of parametrization etc.

\texttt{apc.plot, apc.lines} and \texttt{apc.frame} to see how to plot the results.
> ml <- 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: APC

$\text{exp(Est.)} \pm 2.5\%$ BCI:

- Age 212 $15468.6$
- Age-drift 211 $6868.9$ $1$ $8609.7$ $< 2.2e-16$
- 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 $3082.6$ $-6$ $-2655.4$ $< 2.2e-16$
- Age-drift 211 $6868.9$ $-6$ $-3776.3$ $< 2.2e-16$

% drift:

- Age 212 $15468.6$ $1.023487$ $1.022971$ $1.024003$
- Age-drift 211 $6868.9$ $1.019662$ $1.019062$ $1.020263$
- Age-Cohort 205 $1034.7$ $6$ $5824.1$
- Age-Period-Cohort 199 $423.2$ $6$ $611.6$
- Age-Period 205 $3082.6$ $-6$ $-2655.4$
- Age-drift 211 $6868.9$ $-6$ $-3776.3$

> mw$Drift

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

Lee-Carter model (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 $b_0$ and $k_0$ cannot be determined
- $k_0$ only determined up to an affine transformation:
  $$a_x + b_x (k_1 + m) = (a_x + b_x m) = \bar{a}_x + \bar{b}_x k_t$$
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:
  \[ (f(a) + b(a)k(t) + m) = (f(a) + b(a)m) = f(a) + b(a)k(t) \]

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 \( k(p_{ref}) = 1 \)
    - for persons aged \( a_{ref} \) where \( b(a_{ref}) = 0 \)
  - \( b(a) \) is an age-specific multiplier for the RR
  - Choose \( p_{ref} \) and \( a_{ref} \) a priori.

Danish lung cancer data I

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

```r
sex A P C D Y
1 1 0 0 940 0 1945
2 1 0 0 940 0 1945
3 1 0 1 944 0 1944
4 1 0 0 943 0 1943
```

```r
> ltab <- xtabs( cbind(D,Y) ~ A + P, data=subset(lung,sex==1) )
> str( ltab )
```

```r
'xtabs' num [1:80, 1:61, 1:2] 0 0 0 0 0 0 0 0 0 0 ...
   ..$ : chr [1:2] "D" "Y"
   ..$ A: chr [1:61] "0" "1" "2" "3" ...
   ..$ P: chr [1:61] "1943" "1944" "1945" "1946" ...
```

```r
> lcM <- demogdata( data = as.matrix(ltab[,1:10,"D"] / ltab[,1:10,"Y"])
```

```r
[1] 40 50 60 70 80
```

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:

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

Lee-Carter with demography I

```r
> library(demography)
> lcD <- demogdata( data = as.matrix(ltab[40:90,"D"] / ltab[40:90,"Y"])
```

```r
[3,
  [1] "Ns(P, k = p.kn)2"
  [2] "Ns(A, k = a.kn, i = T)5"
  [1] "Ns(P, k = p.kn)2"
```

```r
> # Only A by P classification - and only ages over 40
> lcM <- demogdata( data = as.matrix(ltab[40:90,"D"] / ltab[40:90,"Y"])
```

```r
[1,] "(Intercept)"
[2,] "Ns(A, k = a.kn, i = T)2"
[3,] "Ns(A, k = a.kn, i = T)3"
[4,] "Ns(A, k = a.kn, i = T)4"
```

```r
> # Set A by P classification
```

```r
[3,] "Ns(A, k = a.kn, i = T)5:Ns(P, k = p.kn)2"
[2,] "Ns(A, k = a.kn, i = T)3:Ns(P, k = p.kn)2"
[1,] "Ns(A, k = a.kn, i = T)1:Ns(P, k = p.kn)1"
```

```r
> mD <- with( dfr, model.matrix( ~ Ns(A,k=a.kn,i=T):Ns(P,k=p.kn) + Ns(A,k=a.kn,i=T) -1 ) )
```

```r
[4,] "Ns(A, k = a.kn, i = T)4:Ns(P, k = p.kn)1"
[3,] "Ns(A, k = a.kn, i = T)5:Ns(P, k = p.kn)1"
[2,] "Ns(A, k = a.kn, i = T)5:Ns(P, k = p.kn)2"
```

```r
> mD <- with( dfr, model.matrix( ~ Ns(A,k=a.kn,i=T):Ns(P,k=p.kn) -1 ) )
```

```r
[8,] "Ns(A, k = a.kn, i = T)1:Ns(P, k = p.kn)1"
[7,] "Ns(P, k = p.kn)2"
[6,] "Ns(A, k = a.kn, i = T)2:Ns(P, k = p.kn)1"
```

Danish lung cancer data II

```r
> xtabs( formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1) )
```

```r
attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```

```r
> attr(*, "call")= language xtabs(formula = cbind(D,Y) ~ A + P, data=subset(lung,sex==1))
```
Lee-Carter with demography

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

Original sample: Mortality data for Denmark

```
Series: Male
Years: 1943 - 2003
Age: 39 - 89
```

Age interaction with between age and both period and/or cohort (=period-age)

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

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

```
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 -9.287 0.02
4 1946 0 42 -9.087 0.02
5 1947 0 43 -8.887 0.02
6 1948 0 44 -8.687 0.02
7 1949 0 45 -8.487 0.02
8 1950 0 46 -8.287 0.02
9 1951 0 47 -8.087 0.02
10 1952 0 48 -7.887 0.02
11 1953 0 49 -7.687 0.02
12 1954 0 50 -7.487 0.02
13 1955 0 51 -7.287 0.02
14 1956 0 52 -7.087 0.02
15 1957 0 53 -6.887 0.02
16 1958 0 54 -6.687 0.02
17 1959 0 55 -6.487 0.02
```

```
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 between age and both period and/or cohort (=period-age)

```
Extension of APC-model:
\[
b(a) = 1 \text{ and } a(a) = 1 \iff \text{APC model.}
\]

Lee-Carter with ilc II

```
$ year : num [1:61] 1943 1944 1945 1946 1947 ...
$ age : num [1:51] 39 40 41 42 43 ...
$ lambda: num 1
- attr(*, "class")= chr "demogdata"
```

```
$ type : chr "Lung cancer incidence"
$ pop :List of 1
  .. ..$ : chr [1:61] "1943" "1944" "1945" "1946" ...
  .. ..$ : chr [1:51] "39" "40" "41" "42" ...
  .. ..- attr(*, "dimnames")=List of 2
```

```
* Age interaction with between age and both period and/or cohort (=period-age)
* Extension of APC-model:
```
```
b(a) = 1 \text{ and } a(a) = 1 \iff \text{APC model.}
```

Lee-Carter with ilc III

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

Original sample: Mortality data for Denmark

```
Series: Male
Years: 1943 - 2003
Age: 39 - 89
```

Applied sample: Mortality data for Denmark (Corrected: interpolate)
```
Series: Male
Years: 1943 - 2003
Age: 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.
```

```
Note: 0 cells have 0/NA deaths and 0 have 0/NA exposure
```

```
> abline(h=1)
```
```
+ xlab="Age", type="l", lty=1, lwd=4 )
```
```
> matplot( dmg.lcM$age, dmg.lcM$bx*50,
```
```
> plot( NA, xlim=0:1, ylim=0:1, axes=FALSE, xlab="", ylab="" )
```
```
> matplot( dmg.lcM$age, exp(dmg.lcM$ax+dmg.lcM$bx*20)*1000,
```
```
> abline(h=0)
```
```
+ xlab="Date", type="l", lty=1, lwd=4 )
```
```
+ ylab="Age effect",
```
```
+ xlab="Age", type="l", yty=1, lwd=4 )
```
```
> matplot( dmg.lcM$age, exp(dmg.lcM$ax)*1000,
```
```
> par( mfcol=c(2,2) )
```
```
> ilc.lcM <- lca.rh( mrt(lcM), model="lc", interpolate=TRUE )
```

```
$ year : num [1:61] 1943 1944 1945 1946 1947 ...
$ age : num [1:51] 39 40 41 42 43 ...
$ label : chr "Denmark"
```

```
$ type : chr "Lung cancer incidence"
$ pop :List of 1
  .. ..$ : chr [1:61] "1943" "1944" "1945" "1946" ...
  .. ..$ : chr [1:51] "39" "40" "41" "42" ...
  .. ..- attr(*, "dimnames")=List of 2
```

```
$ Male: num [1:51, 1:61] 28488 28152 27363 26791 26092 ...
```

```
$ Female: num [1:51, 1:61] 26415 26301 25495 24878 24298 ...
```

```
$ Male: num [1:51, 1:61] 28488 28152 27363 26791 26092 ...
```

```
$ Female: num [1:51, 1:61] 26415 26301 25495 24878 24298 ...
```

```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
Lee-Carter with ilc III

<table>
<thead>
<tr>
<th>Year</th>
<th>Age</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1890</td>
<td>0</td>
<td>56.798</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1960</td>
<td>0</td>
<td>56.798</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lee-Carter with ilc IV

<table>
<thead>
<tr>
<th>Year</th>
<th>Age</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970</td>
<td>0</td>
<td>64.194</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2070</td>
<td>0</td>
<td>64.194</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lee-Carter with ilc V

<table>
<thead>
<tr>
<th>Year</th>
<th>Age</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970</td>
<td>0</td>
<td>64.194</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2070</td>
<td>0</td>
<td>64.194</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lee-Carter with ilc VI

<table>
<thead>
<tr>
<th>Year</th>
<th>Age</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960</td>
<td>0</td>
<td>56.798</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1970</td>
<td>0</td>
<td>56.798</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lee-Carter with ilc VII

<table>
<thead>
<tr>
<th>Year</th>
<th>Age</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1950</td>
<td>0</td>
<td>3805.368</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1960</td>
<td>0</td>
<td>3804.512</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lee-Carter with ilc VIII

<table>
<thead>
<tr>
<th>Year</th>
<th>Age</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1950</td>
<td>0</td>
<td>3805.368</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1960</td>
<td>0</td>
<td>3804.512</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lee-Carter with ilc IX

<table>
<thead>
<tr>
<th>Year</th>
<th>Age</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>0</td>
<td>3803.66</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1991</td>
<td>0</td>
<td>3803.66</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lee-Carter with ilc X

<table>
<thead>
<tr>
<th>Year</th>
<th>Age</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>0</td>
<td>3803.66</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1991</td>
<td>0</td>
<td>3803.66</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Updated values are:

- b0 = 9.0365
- b1 = 0.02
- itx = 0
- kt = 0

Iterations finished in: 34 steps

Updated values are:

- b0 = 9.0365
- b1 = 0.02
- itx = 0
- kt = 0

Iterations finished in: 34 steps

Updated values are:

- b0 = 9.0365
- b1 = 0.02
- itx = 0
- kt = 0

Iterations finished in: 34 steps

Updated values are:

- b0 = 9.0365
- b1 = 0.02
- itx = 0
- kt = 0

Iterations finished in: 34 steps

Updated values are:

- b0 = 9.0365
- b1 = 0.02
- itx = 0
- kt = 0

Iterations finished in: 34 steps

Updated values are:

- b0 = 9.0365
- b1 = 0.02
- itx = 0
- kt = 0

Iterations finished in: 34 steps
Lee-Carter with ilc

- 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 iic 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 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 Epi I

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

Lee-Carter with Epi II

```r
> (p.kn <- LCa.P$t.kn)
```

Fit L-Ca models in Epi I

```r
> 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=1e-4)
LCa.fit convergence in 95 iterations, deviance: 8125.318 on 6084 d.f.
```

Fit L-Ca models in Epi II

```r
> (c.kn <- LCa.C$t.kn)
```

Lee-Carter model (LeeCarter) 278/327

Extended Lee-Carter (from the iic package)

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

Lee-Carter model (LeeCarter) 279/327

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 iic package)
  \[ \log(\lambda(a,p)) = f(a) + b(a) \times k(p) + c(a)m(p-a) \]
  is the union of all of these.
Two sets of data I

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

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

> `data.frame`
  `6 0 1943 1943 20796.5 0.3333333 1943.667 1943.333 3 0
6 0 1943 1943 20796.5 0.3333333 1943.667 1943.333 2 0
5 0 1943 1942 18853.0 0.6666667 1943.333 1942.667 3 0
5 0 1943 1942 18853.0 0.6666667 1943.333 1942.667 2 1
```

> head(th)
>a p c y age diag birth hist d
1 0 1943 1942 18853.0 0.6666667 1943.333 1942.667 1 ... P = diag,
+ D = d,
+ Y = y/10^5 )[,c("A","P","D","Y","hist")]
```

> `hist()`
```r
+ Histology = hist ),
+ Y = sum(Y ),
+ data = th )
```

```r
> hist(th ,
+ hist= factor(hist, labels=c("Sem","nS","OTH" ) ,
+ A = age,
+ P = diag,
+ D = d,
+ Y = y/10^5 )[,c("A","P","D","Y","hist")]
```

> `library( Epi )`
> `stat.table( list( Histology = hist ),
+ Y = sum(Y ) ,
+ data = th )`
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 two rates: Formal tests I

Model 1: D ~ Mc + Mp + Ma:type + offset(log(Y)) - 1
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1 10737 10853.7 -19 -309.0 2.832e-54
2 10737 10853.7 -14 -45.0 6.042e-05

APC-int

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

Analysis of two rates: Formal tests II

Model 5: D ~ Mc + Mp + Ma:type + offset(log(Y)) - 1
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1 10737 10853.7 -19 -309.0 2.832e-54
2 10737 10853.7 -14 -45.0 6.042e-05

Analysis of DM-rates: Age × sex interaction II

- 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 two rates: Formal tests II

Model 6: D ~ Mc + Mp + Ma:type + offset(log(Y)) - 1
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1 10737 10853.7 -19 -309.0 2.832e-54
2 10737 10853.7 -14 -45.0 6.042e-05

APC-int

Analysis of DM-rates: Age × sex interaction II

library( Epi )
load( file="c:/Bendix/Artikler/A_P_C/IDDM/Eurodiab/data/tri.Rdata" )
# Define knots and points of prediction
a.a <- 5
a.c <- 8
m.p <- seq(1/(3*m.a),1-1/(3*m.a),m.a )
g.c <- seq(1/(3*m.c),1-1/(3*m.c),m.P )
g.d <- seq(1/(3*m.p),1-1/(3*m.p),m.C )
co <- NA
attach( dm, warn.conflicts=FALSE )
A.km <- quantile( rep( A, D ), probs=c(1,m.k) )
A.km <- quantile( rep( A, D ), probs=c(1,m.k) )
A.px <- sort( A[match( unique(A), A ) ])
C.uk <- quantile( rep( C, D ), probs=c(1,m.c) )
C.uk <- quantile( rep( C, D ), probs=c(1,m.c) )
C.px <- sort( C[match( unique(C), C ) ])

Analysis of two rates: Formal tests I

Model 1: D ~ Mc + Mp + Ma:type + offset(log(Y)) - 1
Model 2: D ~ Mc + Mp + Ma:type + offset(log(Y)) - 1
Model 3: D ~ Mc + Ma:type + offset(log(Y)) - 1
Model 4: D ~ Mc + Ma:type + offset(log(Y)) - 1

Analysis of two rates: Formal tests II

Model 5: D ~ Mc + Mp + Ma:type + offset(log(Y)) - 1
Model 6: D ~ Mc + Mp + Ma:type + offset(log(Y)) - 1

Analysis of DM-rates: Age × sex interaction II

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 III

- **Date of birth**
- **1985**
- **1980**
- **1985**
- **1985**

- **APC-model: Interactions (APC-int)**

- **Prediction of future rates**

- **Analysis of DM-rates: Age × sex interaction**

- **IV**

- **Model matrices for the ML fit**

- **Rate per 100,000 person−years**

- **Summary**

- **prediction**

- **http://BendixCarstensen/APC/MPIDR-2016**

- **Max Planck Institut for Demographic Research, Rostock**

- **May 2016**

- **Age-Period-Cohort models**

- **Lexis Diagram:**

- **Statistical Analysis in the Lexis Diagram:**

- **Age-Period-Cohort models**

- **Why not just extend the estimated functions into the future?**

- **Predictions from the model must be invariant under reparametrization.**
### Identifiability

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

\[
\log(\lambda(a, p)) = \tilde{f}(a) + \tilde{g}(p) + \tilde{h}(c) = (f(a) - \gamma a) + (g(p) + \gamma p) + (h(c) - \gamma c)
\]

### 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.

### Parametrization invariance

- Prediction of the future course of \( g \) and \( h \) must preserve addition of a linear term in the argument:
  
  \[
  \text{pred}(g(p) + \gamma p) = \text{pred}(g(p)) + \gamma p
  \]
  
  \[
  \text{pred}(h(c) - \gamma c) = \text{pred}(h(c)) - \gamma c
  \]

- 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.

### Bresat cancer prediction

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

### 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

### 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:
  
  \[
  \text{BMI}_{ap} = f(a) + g(p) + h(p - a) + e_{ap}, \quad e_{ap} \sim \mathcal{N}(0, \sigma^2)
  \]

- ... or more precisely:
  
  \[
  \text{BMI} = f(a(i)) + g(p(i)) + h(p(i) - a(i)) + e_i, \quad e_i \sim \mathcal{N}(0, \sigma^2)
  \]
APC-model for quantitative outcomes

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

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 (\( \text{E(BMI)} \))
- Quantile regression (median, quantile)
  - the latter is not a model