Introduction

Monday 19th, morning

Welcome

- Purpose of the course:
  - Knowledge about APC-models
  - Technical knowledge of handling them
  - Insight in the basic survival concepts

- Remedies of the course:
  - Lectures with handouts (BxC)
  - Practicals with suggested solutions (BxC + EG)
Scope of the course

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

About the lectures

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

About the practicals

- You should use you preferred R-environment.
- Epi-package for R is needed.
- Data are all on the net; but we also have them on USB-sticks.
- Try to make a text version of the answers to the exercises — it is more rewarding than just looking at output. The latter is soon forgotten.
- (An opportunity to learn R-weave?)
- A minor bug in apc.fit, apc.plot and apc.frame is fixed in Epi 1.0.11
Survival data

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

Examples of time-to-event measurements

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

Each blob a death

Study ended at 31 Dec. 2003

Ordered by date of entry

Most likely the order in your database.

Timescale changed to “Time since diagnosis”.
Patients ordered by survival time.

Survival times grouped into bands of survival.

Patients ordered by survival status within each band.
### Survival after Cervix cancer

<table>
<thead>
<tr>
<th>Year</th>
<th>Stage I</th>
<th></th>
<th>Stage II</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>D</td>
<td>L</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>110</td>
<td>5</td>
<td>5</td>
<td>234</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>7</td>
<td>7</td>
<td>207</td>
</tr>
<tr>
<td>3</td>
<td>86</td>
<td>7</td>
<td>7</td>
<td>169</td>
</tr>
<tr>
<td>4</td>
<td>72</td>
<td>3</td>
<td>8</td>
<td>129</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>0</td>
<td>7</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>54</td>
<td>2</td>
<td>10</td>
<td>85</td>
</tr>
<tr>
<td>7</td>
<td>42</td>
<td>3</td>
<td>6</td>
<td>73</td>
</tr>
<tr>
<td>8</td>
<td>33</td>
<td>0</td>
<td>5</td>
<td>62</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
<td>0</td>
<td>4</td>
<td>49</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>1</td>
<td>8</td>
<td>34</td>
</tr>
</tbody>
</table>

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

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

Life-table estimator.

### Survival function

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

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

Distribution is characterized by the survival function:

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

### Intensity or rate

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

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

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

\[\Leftrightarrow \quad S(t) = \exp \left( - \int_0^t \lambda(u) \, du \right) = \exp (-\Lambda(t)) \]

\[\Lambda(t) = \int_0^t \lambda(s) \, ds \]

is called the integrated intensity.

Not an intensity — it is dimensionless.

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

Rate and survival

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

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

Note: A cumulative measure requires an origin!

Observed survival and rate

- Survival studies: Observation of (right censored) survival time:
  \[X = \min(T, Z), \delta = 1\{X = T\} \]
  — sometimes conditional on \(T > t_0\) (left truncated).

- Epidemiological studies:
  Observation of (components of) a rate:
  \[D/Y\]

\(D\): no. events, \(Y\) no of person-years, in a prespecified time-frame.
Empirical rates for individuals

At the individual level we introduce the empirical rate: \((d, y)\),
— no. of events \((d \in \{0, 1\})\) during \(y\) risk time.
Each person contributes several obs. of \((d, y)\).
Empirical rates are responses in survival analysis.
The timescale is a covariate — varies across empirical rates from one individual:
Age, calendar time, time since diagnosis.
Don’t confuse timescale with \(y\) — difference between two points on any timescale we may choose.
Two timescales

Note that we actually have two timescales:

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

These can be shown simultaneously in a Lexis diagram.

Follow-up by calendar time and time since diagnosis:

A Lexis diagram!

Empirical rates by calendar time and time since diagnosis
Likelihood from one person

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

\[
P \{ \text{event in } (t_3, t_4) \} = P \{ \text{event in } (t_3, t_4) \mid \text{alive at } t_3 \} \times P \{ \text{survive } (t_2, t_3) \mid \text{alive at } t_2 \} \times P \{ \text{survive } (t_1, t_2) \mid \text{alive at } t_1 \} \times P \{ \text{survive } (t_0, t_1) \mid \text{alive at } t_0 \}
\]

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

Each term refers to one empirical rate \((d, y)\) — \(y = t_i - t_{i-1}\) and mostly \(d = 0\).

Likelihood for an empirical rate

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

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

\[
L(\lambda) = P \{ d \text{ events during } y \text{ time } \} = \pi^d (1 - \pi)^{1-d} = (1 - e^{-\lambda y})^d (e^{-\lambda y})^{1-d} = \left( \frac{1 - e^{-\lambda y}}{e^{-\lambda y}} \right)^d (e^{-\lambda y}) \approx (\lambda y)^d e^{-\lambda y}
\]

since the first term is equal to \(e^{-\lambda y} - 1 \approx \lambda y\).
Log-likelihood:
\[ l(\lambda) = d \log(\lambda y) - \lambda y = d \log(\lambda) + d \log(y) - \lambda y \]

The term \( d \log(y) \) does not include \( \lambda \), so the relevant part of the log-likelihood is:
\[ l(\lambda) = d \log(\lambda) - \lambda y \]

**Likelihood**

Probability of the data and the parameter:

Assuming the rate (intensity) is constant, \( \lambda \), the probability of observing 7 deaths in the course of 500 person-years:
\[
P \{ D = 7, Y = 500 | \lambda \} = \lambda^D e^{\lambda Y} \times K
= \lambda^7 e^{\lambda 500} \times K
= L(\lambda | \text{data})
\]

Best guess of \( \lambda \) is where this function is as large as possible.

Confidence interval is where it is not too far from the maximum
Likelihood-ratio function

Log-likelihood ratio

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

\[ \hat{\lambda} = \frac{7}{500} = 14 \]
\[ \hat{\lambda} \times \exp(1.96/\sqrt{7}) = (6.7, 29.4) \]

Poisson likelihood

The contributions from one individual:

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

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

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

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

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

Likelihood for follow-up of many subjects

Adding empirical rates over the follow-up of persons:

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

- Persons are assumed independent
- Contribution from the same person are conditionally independent, hence give separate contributions to the log-likelihood.
The log-likelihood is maximal for:
\[
\frac{dl(\lambda)}{d\lambda} = \frac{D}{\lambda} - Y = 0 \iff \hat{\lambda} = \frac{D}{Y}
\]

Information about \(\theta = \log(\lambda)\):
\[
l(\theta|D, Y) = D\theta - e^{\theta}Y, \quad l'_\theta = D - e^{\theta}Y, \quad l''_\theta = -e^{\theta}Y
\]
so \(I(\hat{\theta}) = e^{\hat{\theta}}Y = \hat{\lambda}Y = D\), hence \(\text{var}(\hat{\theta}) = 1/D\)

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

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

Confidence interval for a rate

A 95% confidence interval for the log of a rate is:
\[
\hat{\theta} \pm 1.96/\sqrt{D} = \log(\lambda) \pm 1.96/\sqrt{D}
\]

Take the exponential to get the confidence interval for the rate:
\[
\lambda \times \frac{\exp(1.96/\sqrt{D})}{\text{error factor, erf}}
\]

Exercise

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

Calculate the mortality rate with a 95% c.i.
Exercise – solution

The rate is computed as:

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

The confidence interval is computed as:

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

per 1000 person-years.

Ratio of two rates

If we have observations two rates $\lambda_1$ and $\lambda_0$, based on $(D_1, Y_1)$ and $(D_0, Y_0)$ the variance of the difference of the ratio of the rates, RR, is:

$$\text{var}(\log(\text{RR})) = \text{var}(\log(\lambda_1/\lambda_0))$$

$$= \text{var}(\log(\lambda_1)) + \text{var}(\log(\lambda_0))$$

$$= \frac{1}{D_1} + \frac{1}{D_0}$$

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

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

Exercises

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

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

The rate-ratio is computed as:

\[ \hat{RR} = \frac{\hat{\lambda}_1/\hat{\lambda}_0}{(D_1/Y_1)/(D_0/Y_0)} = \frac{28/632.3}{17/843.7} = 0.0443/0.0201 = 2.19 \]

The 95% confidence interval is computed as:

\[ \hat{RR} \times \text{erf} = 2.198 \times \exp(1.96\sqrt{1/17 + 1/28}) = 2.198 \times 1.837 = (1.20, 4.02) \]

Lifetables

Monday 19th, morning

Bendix Carstensen

Age-Period-Cohort models
19–21 September
University of Lisboa,
www.bendixcarstensen.com/APC/Lisboa.2011

The life table method

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

\[
\begin{array}{cccccc}
\text{interval} & \text{alive} & \text{dead} & \text{cens.} & P_i \\
i & n_i & d_i & l_i & p_i \\
1 & 77 & 5 & 2 & 5/(77 - 2/2) = 0.066 \\
2 & 70 & 7 & 4 & 7/(70 - 4/2) = 0.103 \\
3 & 59 & 8 & 1 & 8/(59 - 1/2) = 0.137 \\
\end{array}
\]

\[
p_i = P\{\text{death in interval } i\} = 1 - d_i/(n_i - l_i/2) \\
S(t) = (1 - p_1) \times \cdots \times (1 - p_t)
\]
The life table method

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

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

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

and hence the death probability:

\[
p_i = 1 - \exp(-\lambda_i \ell_i) = 1 - \exp\left(-\left(n_i - d_i/2 - l_i/2\right)\right)
\]

The modified life-table estimator.

### Population life table, DK 1997–98

<table>
<thead>
<tr>
<th>Age</th>
<th>Men</th>
<th></th>
<th></th>
<th>Women</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(S(a))</td>
<td>(\lambda(a))</td>
<td>(E[\ell_{res}(a)])</td>
<td>(S(a))</td>
<td>(\lambda(a))</td>
<td>(E[\ell_{res}(a)])</td>
</tr>
<tr>
<td>0</td>
<td>1.00000</td>
<td>567</td>
<td>73.68</td>
<td>1.00000</td>
<td>474</td>
<td>78.65</td>
</tr>
<tr>
<td>1</td>
<td>0.99433</td>
<td>67</td>
<td>73.10</td>
<td>0.99526</td>
<td>47</td>
<td>78.02</td>
</tr>
<tr>
<td>2</td>
<td>0.99366</td>
<td>38</td>
<td>72.15</td>
<td>0.99479</td>
<td>21</td>
<td>77.06</td>
</tr>
<tr>
<td>3</td>
<td>0.99329</td>
<td>25</td>
<td>71.18</td>
<td>0.99458</td>
<td>14</td>
<td>76.08</td>
</tr>
<tr>
<td>4</td>
<td>0.99304</td>
<td>25</td>
<td>70.19</td>
<td>0.99444</td>
<td>14</td>
<td>75.09</td>
</tr>
<tr>
<td>5</td>
<td>0.99279</td>
<td>21</td>
<td>69.21</td>
<td>0.99430</td>
<td>11</td>
<td>74.10</td>
</tr>
<tr>
<td>6</td>
<td>0.99258</td>
<td>17</td>
<td>68.23</td>
<td>0.99419</td>
<td>6</td>
<td>73.11</td>
</tr>
<tr>
<td>7</td>
<td>0.99242</td>
<td>14</td>
<td>67.24</td>
<td>0.99413</td>
<td>3</td>
<td>72.11</td>
</tr>
<tr>
<td>8</td>
<td>0.99227</td>
<td>15</td>
<td>66.25</td>
<td>0.99410</td>
<td>6</td>
<td>71.11</td>
</tr>
<tr>
<td>9</td>
<td>0.99213</td>
<td>14</td>
<td>65.26</td>
<td>0.99404</td>
<td>9</td>
<td>70.12</td>
</tr>
<tr>
<td>10</td>
<td>0.99199</td>
<td>17</td>
<td>64.26</td>
<td>0.99395</td>
<td>17</td>
<td>69.12</td>
</tr>
<tr>
<td>11</td>
<td>0.99181</td>
<td>19</td>
<td>63.28</td>
<td>0.99378</td>
<td>15</td>
<td>68.14</td>
</tr>
<tr>
<td>12</td>
<td>0.99162</td>
<td>16</td>
<td>62.29</td>
<td>0.99363</td>
<td>11</td>
<td>67.15</td>
</tr>
<tr>
<td>13</td>
<td>0.99147</td>
<td>18</td>
<td>61.30</td>
<td>0.99352</td>
<td>14</td>
<td>66.15</td>
</tr>
<tr>
<td>14</td>
<td>0.99129</td>
<td>25</td>
<td>60.31</td>
<td>0.99338</td>
<td>11</td>
<td>65.16</td>
</tr>
<tr>
<td>15</td>
<td>0.99104</td>
<td>45</td>
<td>59.32</td>
<td>0.99327</td>
<td>10</td>
<td>64.17</td>
</tr>
<tr>
<td>16</td>
<td>0.99059</td>
<td>50</td>
<td>58.35</td>
<td>0.99317</td>
<td>18</td>
<td>63.18</td>
</tr>
<tr>
<td>17</td>
<td>0.99009</td>
<td>52</td>
<td>57.38</td>
<td>0.99299</td>
<td>29</td>
<td>62.19</td>
</tr>
<tr>
<td>18</td>
<td>0.98957</td>
<td>85</td>
<td>56.41</td>
<td>0.99270</td>
<td>35</td>
<td>61.23</td>
</tr>
<tr>
<td>19</td>
<td>0.98873</td>
<td>79</td>
<td>55.46</td>
<td>0.99235</td>
<td>30</td>
<td>60.23</td>
</tr>
<tr>
<td>20</td>
<td>0.98795</td>
<td>70</td>
<td>54.50</td>
<td>0.99205</td>
<td>35</td>
<td>59.24</td>
</tr>
<tr>
<td>21</td>
<td>0.98726</td>
<td>71</td>
<td>53.54</td>
<td>0.99170</td>
<td>31</td>
<td>58.27</td>
</tr>
</tbody>
</table>

### Diagram

Danish life tables 1997-98

Log mortality per 10^5 (40–85 years)

Men: \(14.244 + 0.135 \text{ age}\)

Women: \(14.877 + 0.135 \text{ age}\)
Practical

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

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

**Denmark**

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log_2(\lambda(a))$</td>
<td>$-14.244 + 0.135$ age</td>
<td>$-14.877 + 0.135$ age</td>
</tr>
<tr>
<td>Doubling time</td>
<td>$1/0.135 = 7.41$ years</td>
<td></td>
</tr>
<tr>
<td>M/F rate-ratio</td>
<td>$2^{-14.244+14.877} = 2^{0.633} = 1.55$</td>
<td></td>
</tr>
<tr>
<td>Age-difference</td>
<td>$(-14.244 + 14.877)/0.135 = 4.69$ years</td>
<td></td>
</tr>
</tbody>
</table>

**Sweden**

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log_2(\lambda(a))$</td>
<td>$-15.453 + 0.146$ age</td>
<td>$-16.204 + 0.146$ age</td>
</tr>
<tr>
<td>Doubling time</td>
<td>$1/0.146 = 6.85$ years</td>
<td></td>
</tr>
<tr>
<td>M/F rate-ratio</td>
<td>$2^{-15.453+16.204} = 2^{0.751} = 1.68$</td>
<td></td>
</tr>
<tr>
<td>Age-difference</td>
<td>$(-15.453 + 16.204)/0.146 = 5.14$ years</td>
<td></td>
</tr>
</tbody>
</table>
Observations for the lifetable

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

Age-specific rates — cross-sectional!

Survival function:

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

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

Life table approach

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

It is the population experience:

\[ D: \text{ Deaths (events).} \]
\[ Y: \text{ Person-years (risk time).} \]

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

Data are collected cross-sectionally, but interpreted longitudinally.

Rates vary over time:
Finnish life tables 1986

\[
\log_2(\text{mortality per 10^5 (40−85 years)}) = -14.061 + 0.138 \text{ age}
\]

Women: \(-15.266 + 0.138 \text{ age}\)

Lifetables (lifetable) 53/238

Finnish life tables 1987

\[
\log_2(\text{mortality per 10^5 (40−85 years)}) = -14.216 + 0.140 \text{ age}
\]

Women: \(-15.384 + 0.140 \text{ age}\)

Lifetables (lifetable) 54/238

Finnish life tables 1988

\[
\log_2(\text{mortality per 10^5 (40−85 years)}) = -14.043 + 0.137 \text{ age}
\]

Women: \(-15.211 + 0.137 \text{ age}\)

Lifetables (lifetable) 55/238
Finnish life tables 1989

Men: $-14.007 + 0.136 \text{age}$

Women: $-15.156 + 0.136 \text{age}$

Finnish life tables 1990

Men: $-13.996 + 0.136 \text{age}$

Women: $-15.139 + 0.136 \text{age}$

Finnish life tables 1991

Men: $-14.093 + 0.136 \text{age}$

Women: $-15.247 + 0.136 \text{age}$
Finnish life tables 1995

log₂ (mortality per 10⁵ (40−85 years))
Men: -14.224 + 0.136 age
Women: -15.351 + 0.136 age

Finnish life tables 1996

log₂ (mortality per 10⁵ (40−85 years))
Men: -14.235 + 0.136 age
Women: -15.388 + 0.136 age

Finnish life tables 1997

log₂ (mortality per 10⁵ (40−85 years))
Men: -14.151 + 0.134 age
Women: -15.235 + 0.134 age
Finnish life tables 2001

log2( mortality per 10^5 (40−85 years) )
Men: −14.268 + 0.133 age
Women: −15.329 + 0.133 age

Finnish life tables 2002

log2( mortality per 10^5 (40−85 years) )
Men: −14.391 + 0.135 age
Women: −15.448 + 0.135 age

Finnish life tables 2003

log2( mortality per 10^5 (40−85 years) )
Men: −14.339 + 0.134 age
Women: −15.412 + 0.134 age
Empirical rates

At the individual level we introduce the empirical rate: \((d, y)\),
— the number of events \((d \in \{0, 1\})\) during \(y\) risk time.

Each individual contributes several observations.

Empirical rates are responses in survival analysis.

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

Likelihood

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

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

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

so the contribution from one individual is as the contribution from several independent Poisson observations.
The proportional hazards model

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

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

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

Who needs the Cox-model anyway? (WntCma 73/ 238)

Cox-likelihood

The partial likelihood for the regression parameters:

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

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

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

The Cox-likelihood as profile likelihood

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

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

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

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

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

Who needs the Cox-model anyway? (WntCma 74/ 238)
Log-likelihood contributions that contain information on a specific time-scale parameter $\alpha_t$ will be from:

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

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

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

$$
= \sum_{i \in R_t} \left\{ d_i (\alpha_t + \eta_i) - e^{\alpha_t + \eta_i} \right\}
$$

$$
= \alpha_t + \eta_{\text{death}} - e^{\alpha_t} \sum_{i \in R_t} e^{\eta_i}
$$

where $\eta_{\text{death}}$ is the linear predictor for the person that died.

The derivative w.r.t. $\alpha_t$ is:

$$
D_{\alpha_t} \ell(\alpha_t, \beta) = 1 - e^{\alpha_t} \sum_{i \in R_t} e^{\eta_i} = 0 \iff e^{\alpha_t} = \frac{1}{\sum_{i \in R_t} e^{\eta_i}}
$$

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.
What the Cox-model really is

Taking the life-table approach \textit{ad absurdum} by:

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

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

Sensible modelling

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

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

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

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

Just small enough to make the constant rate assumption reasonable.

Splitting the dataset

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

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

- Poisson-response $d$
- Risk time $y$
- Covariate value for the timescale (time since entry, current age, current date, ...)
- other covariates
Example: Mayo Clinic lung cancer

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

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

> tab(Lx, scale=365.25)
States:
#records: To From FALSE TRUE Sum #events: #risk time: Rate (95
FALSE 63 165 228 165 190.5352 0.8659815 0.743432

> dx <- splitLexis( Lx, "tfd", breaks=c(0,unique(Lx$time)) )
> tab( dx, scale=365.25 )
States:
#records: To From FALSE TRUE Sum #events: #risk time: Rate (
FALSE 19857 165 20022 165 190.5352 0.8659815 0.7434

Who needs the Cox-model anyway? (WntCma) 82/ 238

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 \( f(t_i) = g(t_i) + x_0^t \gamma \)
   for these points:
   \[
   \hat{f}(t_i) = B \hat{\beta}
   \]
   where \( \beta \) is the total parameter vector in the model.
3. Variance-covariance matrix of \( \hat{\beta} \): \( \hat{\Sigma} \).
4. Variance-covariance of \( \hat{f}(t_i) \): \( B \Sigma B' \).
5. Transformation to the rates is the coordinate-wise exponential function, with derivative \( \text{diag}[\exp(\hat{f}(t_i))] \)
6. Variance-covariance matrix of the rates at the points $t_i$:

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

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

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

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

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

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

---

**Mayo clinic lung cancer data**

Smoothing by natural splines with 7 parameters; knots at 0, 25, 75, 150, 250, 500, 1000 days
Computational tools for time-splitting

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

**Stata:** The function `stsplit` (part of standard Stata).
Descendant of `stlexis` written by Michael Hills & David Clayton.

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

---

## Conclusion

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

---

## Follow-up data
**Monday 19th, afternoon**

**Bendix Carstensen**

Age-Period-Cohort models
19–21 September
University of Lisboa,
www.bendixcarstensen.com/APC/Lisboa.2011
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 and $Y$ — risk time in intervals on the timescale:

**Origin:** The date where the time scale is 0:

- Age — 0 at date of birth
- Disease duration — 0 at date of diagnosis
- Occupation exposure — 0 at date of hire

**Intervals:** How should it be subdivided:

- 1-year classes? 5-year classes?
- Equal length?
Cohort with 3 persons:

<table>
<thead>
<tr>
<th>Id</th>
<th>Bdate</th>
<th>Entry</th>
<th>Exit</th>
<th>St</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14/07/1952</td>
<td>04/08/1965</td>
<td>27/06/1997</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>01/04/1954</td>
<td>08/09/1972</td>
<td>23/05/1995</td>
<td>0</td>
</tr>
</tbody>
</table>

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

Follow-up data (FU-rep-Lexis)

Splitting the follow up

<table>
<thead>
<tr>
<th></th>
<th>subj. 1</th>
<th>subj. 2</th>
<th>subj. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Entry</td>
<td>Exit</td>
<td>Status</td>
</tr>
<tr>
<td>13.06</td>
<td>18.44</td>
<td>4.54</td>
<td></td>
</tr>
<tr>
<td>44.95</td>
<td>41.14</td>
<td>11.12</td>
<td></td>
</tr>
<tr>
<td>Dead</td>
<td>Alive</td>
<td>Dead</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
Y & \quad 31.89 & \quad 22.70 & \quad 6.58 \\
D & \quad 1 & \quad 0 & \quad 1
\end{align*}
\]

<table>
<thead>
<tr>
<th>Age</th>
<th>subj. 1</th>
<th>subj. 2</th>
<th>subj. 3</th>
<th>∑</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y D</td>
<td>Y D</td>
<td>Y D</td>
<td>Y D</td>
</tr>
<tr>
<td>0–</td>
<td>0.00 0</td>
<td>0.00 0</td>
<td>5.46 0</td>
<td>5.46 0</td>
</tr>
<tr>
<td>10–</td>
<td>6.94 0</td>
<td>1.56 0</td>
<td>1.12 1</td>
<td>8.62 1</td>
</tr>
<tr>
<td>20–</td>
<td>10.00 0</td>
<td>10.00 0</td>
<td>0.00 0</td>
<td>20.00 0</td>
</tr>
<tr>
<td>30–</td>
<td>10.00 0</td>
<td>10.00 0</td>
<td>0.00 0</td>
<td>20.00 0</td>
</tr>
<tr>
<td>40–</td>
<td>4.95 1</td>
<td>1.14 0</td>
<td>0.00 0</td>
<td>6.09 1</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\sum & \quad 31.89 \quad 1 \quad 22.70 \quad 0 \quad 6.58 \quad 1 \\
& \quad 60.17 \quad 2
\end{align*}
\]
Splitting the follow-up

<table>
<thead>
<tr>
<th>id</th>
<th>Bdate</th>
<th>Entry</th>
<th>Exit</th>
<th>St</th>
<th>risk</th>
<th>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</td>
<td>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</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>14/07/1952</td>
<td>14/07/1982</td>
<td>14/07/1992</td>
<td>0</td>
<td>10.0000</td>
<td>30</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</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>01/04/1954</td>
<td>08/09/1972</td>
<td>01/04/1974</td>
<td>0</td>
<td>1.5606</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>01/04/1954</td>
<td>01/04/1974</td>
<td>31/03/1984</td>
<td>0</td>
<td>10.0000</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>01/04/1954</td>
<td>31/03/1984</td>
<td>01/04/1994</td>
<td>0</td>
<td>10.0000</td>
<td>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>1.1417</td>
<td>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</td>
<td>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</td>
<td>10</td>
</tr>
</tbody>
</table>

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

Timescales

- A timescale is a variable that varies deterministically within each person during follow-up:
  - Age
  - Calendar time
  - Time since treatment
  - Time since relapse

- All timescales advance at the same pace (1 year per year ...)

- Note: Cumulative exposure is not a timescale.

Representation of follow-up on several timescales

- The time followed is the same on all timescales.
- Only use the entry point on each time scale:
  - Age at entry.
  - Date of entry.
  - Time since treatment at entry.
  — if time of treatment is the entry, this is 0 for all.
Follow-up data in Epi: Lexis objects

A follow-up study:

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

Timescales of interest:

- Age
- Calendar time
- Time since injection

---

Definition of Lexis object

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

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

---

The looks of a Lexis object

```r
> round( thL[,c(1:8,14,15)], 2 )
   age per tfi lex.dur lex.Cst lex.Xst lex.id id
1 22.18 1938.79  0  38.00    0  1  1
2 49.55 1945.77  0  18.60    0  1  2  640
3 68.21 1955.18  0  1.40     0  1  3  3425
4 20.80 1957.61  0  34.52    0  0  4  4017
```
> plot( thL, lwd=3 )

Follow-up data (FU-rep-Lexis)

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

Follow-up data (FU-rep-Lexis)
Splitting follow-up time

```r
> spl1 <- splitLexis( thL, "age", breaks=seq(0,100,20) )
> round( spl1, 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>bi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.18</td>
<td>1938.79</td>
<td>0.00</td>
<td>17.82</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>40.00</td>
<td>1956.61</td>
<td>17.82</td>
<td>20.00</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>60.00</td>
<td>1976.61</td>
<td>37.82</td>
<td>0.18</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>49.55</td>
<td>1945.77</td>
<td>0.00</td>
<td>10.45</td>
<td>0</td>
<td>0</td>
<td>640</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>60.00</td>
<td>1956.23</td>
<td>10.45</td>
<td>8.14</td>
<td>0</td>
<td>1</td>
<td>640</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>68.21</td>
<td>1955.18</td>
<td>0.00</td>
<td>1.40</td>
<td>0</td>
<td>1</td>
<td>3425</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>20.80</td>
<td>1957.61</td>
<td>0.00</td>
<td>19.20</td>
<td>0</td>
<td>0</td>
<td>4017</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>40.00</td>
<td>1976.81</td>
<td>19.20</td>
<td>0.80</td>
<td>0</td>
<td>0</td>
<td>4017</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
```

Follow-up data (FU-rep-Lexis)

Split on a second timescale

```r
> # Split further on tfi: spl2 <- splitLexis( spl1, "tfi", breaks=c(0,1,5,20,100) )
> round( spl2, 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>bi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.18</td>
<td>1938.79</td>
<td>0.00</td>
<td>1.00</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>23.18</td>
<td>1939.79</td>
<td>1.00</td>
<td>4.00</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>27.18</td>
<td>1943.79</td>
<td>5.00</td>
<td>12.82</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>40.00</td>
<td>1956.61</td>
<td>17.82</td>
<td>2.18</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>42.18</td>
<td>1958.79</td>
<td>20.00</td>
<td>17.82</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>60.00</td>
<td>1976.61</td>
<td>37.82</td>
<td>0.18</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>49.55</td>
<td>1945.77</td>
<td>0.00</td>
<td>1.00</td>
<td>0</td>
<td>0</td>
<td>640</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>50.55</td>
<td>1946.77</td>
<td>1.00</td>
<td>4.00</td>
<td>0</td>
<td>0</td>
<td>640</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>54.55</td>
<td>1950.77</td>
<td>5.00</td>
<td>5.45</td>
<td>0</td>
<td>0</td>
<td>640</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>60.00</td>
<td>1956.23</td>
<td>10.45</td>
<td>8.14</td>
<td>0</td>
<td>1</td>
<td>640</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>68.21</td>
<td>1955.18</td>
<td>0.00</td>
<td>1.00</td>
<td>0</td>
<td>0</td>
<td>3425</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>69.21</td>
<td>1956.18</td>
<td>1.00</td>
<td>0.40</td>
<td>0</td>
<td>1</td>
<td>3425</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>20.80</td>
<td>1957.61</td>
<td>0.00</td>
<td>1.00</td>
<td>0</td>
<td>0</td>
<td>4017</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>21.80</td>
<td>1958.61</td>
<td>1.00</td>
<td>4.00</td>
<td>0</td>
<td>0</td>
<td>4017</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>25.80</td>
<td>1962.61</td>
<td>5.00</td>
<td>14.20</td>
<td>0</td>
<td>0</td>
<td>4017</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
```

Follow-up data (FU-rep-Lexis)

The Poisson likelihood for time-split data

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

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

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

The model assume that rates are constant.

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

Where are the \((d_{it}, y_{it})\) in the split data?
The Poisson likelihood for time-split data

If \( d \sim \text{Poisson}(\lambda y) \), i.e. with mean \((\lambda y)\) 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.

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

```r
> round(spl2, 2)

  lex.id age per tfi lex.dur lex.Cst lex.Xst id sex b
1    1  22.18  1938.79  0.00   1.00   0  0    1    2
2    1  23.18  1939.79  1.00   4.00   0  0    1    2
3    1  27.18  1943.79  5.00  12.82   0  0    1    2
4    1  40.00  1956.61 17.82  2.18   0  0    1    2
5    1  42.18  1958.79 20.00 17.82   0  0    1    2
6    1  60.00  1976.61 37.82  0.18   0  1    1    2
7    2  49.55  1945.77  0.00   1.00   0  0  640    2
8    2  50.55  1946.77  1.00   4.00   0  0  640    2
9    2  54.55  1950.77  5.00  5.45   0  0  640    2
10   2  60.00  1956.23 10.45  8.14   0  1  640    2
11   3  68.21  1965.18  0.00   1.00   0  0 3425    1
12   3  69.21  1966.18  1.00   0.40   0  1 3425    1
13   4  20.80  1957.61  0.00   1.00   0  0 4017    2
14   4  21.80  1958.61  1.00   4.00   0  0 4017    2
15   4  25.80  1962.61  5.00  14.20   0  0 4017    2
16   4  40.00  1976.81 19.20  0.80   0  0 4017    2
17   4  40.80  1977.61 20.00 14.52   0  0 4017    2
```
Analysis of results

- $d_i$ — events in the variable: $\text{lex.Xst}$.
- $y_i$ — 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.

Age at entry

Monday 19th, afternoon

Bendix Carstensen

Age-Period-Cohort models
19–21 September
University of Lisboa,
www.bendixcarstensen.com/APC/Lisboa.2011
Age at entry as covariate

- \( t \): time since entry
- \( e \): age at entry
- \( a = e + t \): current age

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

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

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

Non-linear effects of time-scales

Arbitrary effects of the three variables \( t, a \) and \( e \):

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

Three quantities can be arbitrarily moved between the three functions:

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

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.
Conceptual set-up

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

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

Realistic set-up

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

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

Lexis diagram

Disease registers record events.

Official statistics collect population data.

Registration of:
- cases ($D$)
- risk time, person-years ($Y$)

in subsets of the Lexis diagram.
Lexis diagram

Registration of:
cases ($D$)
risk time,
person-years ($Y$)
in subsets of the
Lexis diagram.
Rates available in
each subset.

Register data

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

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

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>773812</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>1943</td>
<td>813022</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>1943</td>
<td>790501</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>1943</td>
<td>799293</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>1943</td>
<td>769356</td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>1943</td>
<td>694073</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>1948</td>
<td>744217</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>1948</td>
<td>744706</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>1948</td>
<td>781827</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>1948</td>
<td>774542</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>1948</td>
<td>782893</td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>1948</td>
<td>754322</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>1953</td>
<td>794123</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>1953</td>
<td>721810</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>1953</td>
<td>722968</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>1953</td>
<td>769298</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>1953</td>
<td>760213</td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>1953</td>
<td>768471</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>1958</td>
<td>972853</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>1958</td>
<td>770859</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>1958</td>
<td>698612</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 levels of classifying factors.
- Output from the model will be rates and rate-ratios.
- Since we use multiplicative Poisson, usually the log rates and the log-RR are reported

Simple model for the testis cancer rates:

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

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

Deviance Residuals:
\begin{tabular}{cccc}
Min & 1Q & Median & 3Q & Max \\
-1.5991 & -0.6974 & 0.1284 & 0.6671 & 1.8904
\end{tabular}

Coefficients:
\begin{tabular}{cccccc}
Estimate & Std. Error & z value & Pr(>|z|) \\
(Intercept) & -1.4758 & 0.3267 & -4.517 & 6.26e-06 \\
factor(A)20 & 1.4539 & 0.3545 & 4.101 & 4.11e-05 \\
factor(A)25 & 2.5321 & 0.3301 & 7.671 & 1.71e-14 \\
factor(A)30 & 2.9321 & 0.3301 & 7.671 & 1.71e-14 \\
factor(A)35 & 2.9321 & 0.3301 & 7.671 & 1.71e-14 \\
factor(A)40 & 2.8613 & 0.3259 & 8.779 & < 2e-16 \\
factor(P)1948 & 0.1753 & 0.1211 & 1.447 & 0.14778 \\
factor(P)1953 & 0.3822 & 0.1163 & 3.286 & 0.00102 \\
factor(P)1958 & 0.4663 & 0.1163 & 3.286 & 0.00102
\end{tabular}

\texttt{ci.lin()} from the Epi package extracts coefficients and computes confidence intervals:

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

\begin{tabular}{ccccccc}
Estimate & StdErr & z & P & 2.5% & 97.5% \\
(Intercept) & -1.476 & 0.327 & -4.517 & 0.000 & -2.116 & -0.836 \\
factor(A)20 & 1.454 & 0.354 & 4.101 & 0.000 & 0.759 & 2.149 \\
factor(A)25 & 2.532 & 0.330 & 7.671 & 0.000 & 1.885 & 3.179 \\
factor(A)30 & 2.933 & 0.325 & 9.013 & 0.000 & 2.295 & 3.570 \\
factor(A)35 & 2.861 & 0.326 & 8.779 & 0.000 & 2.223 & 3.500 \\
factor(A)40 & 2.852 & 0.326 & 8.741 & 0.000 & 2.213 & 3.492 \\
factor(P)1948 & 0.175 & 0.121 & 1.447 & 0.148 & -0.062 & 0.413 \\
factor(P)1953 & 0.382 & 0.116 & 3.286 & 0.001 & 0.154 & 0.610 \\
factor(P)1958 & 0.466 & 0.115 & 4.052 & 0.000 & 0.241 & 0.691
\end{tabular}
Subsets of parameter estimates accessed via a character string that is grep to the names.

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

<table>
<thead>
<tr>
<th>Estimate</th>
<th>StdErr</th>
<th>z</th>
<th>P</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor(P)1948</td>
<td>0.175</td>
<td>0.121</td>
<td>1.447</td>
<td>0.148</td>
<td>-0.062</td>
</tr>
<tr>
<td>factor(P)1953</td>
<td>0.382</td>
<td>0.116</td>
<td>3.286</td>
<td>0.001</td>
<td>0.154</td>
</tr>
<tr>
<td>factor(P)1958</td>
<td>0.466</td>
<td>0.115</td>
<td>4.052</td>
<td>0.000</td>
<td>0.241</td>
</tr>
</tbody>
</table>
```

Rates / 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>StdErr</th>
<th>z</th>
<th>P</th>
<th>exp(Est.)</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor(P)1948</td>
<td>0.175</td>
<td>0.121</td>
<td>1.447</td>
<td>0.148</td>
<td>1.192</td>
<td>0.940</td>
</tr>
<tr>
<td>factor(P)1953</td>
<td>0.382</td>
<td>0.116</td>
<td>3.286</td>
<td>0.001</td>
<td>1.466</td>
<td>1.167</td>
</tr>
<tr>
<td>factor(P)1958</td>
<td>0.466</td>
<td>0.115</td>
<td>4.052</td>
<td>0.000</td>
<td>1.593</td>
<td>1.272</td>
</tr>
</tbody>
</table>
```

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

<table>
<thead>
<tr>
<th>Estimate</th>
<th>StdErr</th>
<th>z</th>
<th>P</th>
<th>exp(Est.)</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>-0.382</td>
<td>0.116</td>
<td>-3.286</td>
<td>0.001</td>
<td>0.682</td>
<td>0.543</td>
</tr>
<tr>
<td>[2,]</td>
<td>-0.207</td>
<td>0.110</td>
<td>-1.874</td>
<td>0.061</td>
<td>0.813</td>
<td>0.655</td>
</tr>
<tr>
<td>[3,]</td>
<td>0.000</td>
<td>0.000</td>
<td>NaN</td>
<td>NaN</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>[4,]</td>
<td>0.084</td>
<td>0.104</td>
<td>0.808</td>
<td>0.419</td>
<td>1.087</td>
<td>0.887</td>
</tr>
</tbody>
</table>
```
Register data - rates
Rates in “tiles” of the Lexis diagram:

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

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

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

Synthetic cohorts

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

Successively overlapping 10-year periods.
Data matrix: Testis cancer cases

Number of cases

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

Data matrix: Male risk time

1000 person-years

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>15–19</td>
<td>744.2</td>
<td>794.1</td>
<td>972.9</td>
<td>1051.5</td>
<td>961.0</td>
<td>952.5</td>
<td>1011.1</td>
<td>1005.0</td>
</tr>
<tr>
<td>20–24</td>
<td>744.7</td>
<td>721.8</td>
<td>770.9</td>
<td>960.3</td>
<td>1053.8</td>
<td>987.5</td>
<td>953.0</td>
<td>1019.7</td>
</tr>
<tr>
<td>25–29</td>
<td>781.8</td>
<td>723.0</td>
<td>698.6</td>
<td>764.8</td>
<td>962.7</td>
<td>1056.1</td>
<td>960.9</td>
<td>956.2</td>
</tr>
<tr>
<td>30–34</td>
<td>774.5</td>
<td>769.3</td>
<td>711.6</td>
<td>700.1</td>
<td>769.9</td>
<td>960.4</td>
<td>1045.3</td>
<td>955.0</td>
</tr>
<tr>
<td>35–39</td>
<td>782.9</td>
<td>760.2</td>
<td>760.5</td>
<td>711.6</td>
<td>702.3</td>
<td>767.5</td>
<td>951.9</td>
<td>1035.7</td>
</tr>
<tr>
<td>40–44</td>
<td>754.3</td>
<td>768.5</td>
<td>749.9</td>
<td>756.5</td>
<td>709.8</td>
<td>696.5</td>
<td>757.8</td>
<td>940.3</td>
</tr>
<tr>
<td>45–49</td>
<td>676.7</td>
<td>737.9</td>
<td>753.5</td>
<td>738.1</td>
<td>746.4</td>
<td>698.2</td>
<td>682.4</td>
<td>743.1</td>
</tr>
<tr>
<td>50–54</td>
<td>600.3</td>
<td>653.9</td>
<td>715.4</td>
<td>732.7</td>
<td>718.3</td>
<td>724.2</td>
<td>675.5</td>
<td>660.8</td>
</tr>
<tr>
<td>55–59</td>
<td>512.8</td>
<td>571.1</td>
<td>622.5</td>
<td>680.8</td>
<td>698.2</td>
<td>683.8</td>
<td>686.4</td>
<td>640.9</td>
</tr>
</tbody>
</table>
Data matrix: Empirical rates

Rate per 1000,000 person-years

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>15–19</td>
<td>9.4</td>
<td>16.4</td>
<td>13.4</td>
<td>14.3</td>
<td>34.3</td>
<td>36.7</td>
<td>36.6</td>
<td>48.8</td>
</tr>
<tr>
<td>20–24</td>
<td>41.6</td>
<td>63.7</td>
<td>63.6</td>
<td>57.3</td>
<td>80.7</td>
<td>113.7</td>
<td>146.9</td>
<td>148.1</td>
</tr>
<tr>
<td>25–29</td>
<td>79.3</td>
<td>87.1</td>
<td>117.4</td>
<td>113.8</td>
<td>107.0</td>
<td>144.9</td>
<td>209.2</td>
<td>223.8</td>
</tr>
<tr>
<td>30–34</td>
<td>85.2</td>
<td>106.6</td>
<td>123.7</td>
<td>147.1</td>
<td>161.1</td>
<td>170.8</td>
<td>198.0</td>
<td>218.8</td>
</tr>
<tr>
<td>35–39</td>
<td>71.5</td>
<td>73.7</td>
<td>88.1</td>
<td>139.1</td>
<td>176.6</td>
<td>185.0</td>
<td>159.7</td>
<td>181.5</td>
</tr>
<tr>
<td>40–44</td>
<td>62.3</td>
<td>84.6</td>
<td>85.3</td>
<td>88.6</td>
<td>119.8</td>
<td>147.9</td>
<td>157.0</td>
<td>128.7</td>
</tr>
<tr>
<td>45–49</td>
<td>44.3</td>
<td>50.1</td>
<td>71.7</td>
<td>61.0</td>
<td>85.7</td>
<td>90.2</td>
<td>96.7</td>
<td>123.8</td>
</tr>
<tr>
<td>50–54</td>
<td>46.6</td>
<td>33.6</td>
<td>37.7</td>
<td>62.8</td>
<td>50.1</td>
<td>69.0</td>
<td>72.5</td>
<td>92.3</td>
</tr>
<tr>
<td>55–59</td>
<td>27.3</td>
<td>28.0</td>
<td>40.2</td>
<td>38.2</td>
<td>41.5</td>
<td>40.9</td>
<td>62.6</td>
<td>65.5</td>
</tr>
</tbody>
</table>

The classical plots

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

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

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

Rates are proportional between periods:

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

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

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

Fitting the model in R

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

```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)`

Deviance Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.0925</td>
<td>-0.8784</td>
<td>0.1148</td>
<td>0.9790</td>
<td>2.7653</td>
</tr>
</tbody>
</table>

Coefficients:

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

Age-cohort model

Rates are proportional between cohorts:

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

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

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

Fit the model in R

Reference period is the 9th cohort (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

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

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

Age-Period and Age-Cohort models (AP-AC)
**Age-drift model**

**Tuesday 24th, morning**

**Bendix Carstensen**

Age-Period-Cohort models
19–21 September
University of Lisboa,
www.bendixcarstensen.com/APC/Lisboa.2011

**Linear effect of period:**

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

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

**Linear effect of cohort:**

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

that is, \( \gamma_c = \gamma(c - c_0) \)
Age and linear effect of period:

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

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

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

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

(Dispersion parameter for poisson family taken to be 1)

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

Age-drift model (Ad)

Age and linear effect of cohort:

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

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

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

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

(Dispersion parameter for poisson family taken to be 1)

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

Age-drift model (Ad)

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).
Which age-curve is period and which is cohort?

Age-Period-Cohort model
Tuesday 24th, afternoon

Bendix Carstensen

Age-Period-Cohort models
19–21 September
University of Lisboa,
www.bendixcarstensen.com/APC/Lisboa.2011

The age-period-cohort model

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

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

```r
> c1933.p <- glm( D ~ factor( A ) - 1 +
+              relevel( factor( C ), "1933" ) +
+              factor( P ) + offset( log( Y ) ), family=p)
> summary( c1933.p )
```

Coefficients: (1 not defined because of singularities)

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|---------|
| factor(A)17.5 | -4.27754 | 0.10479 | -40.819 | < 2e-16 |
| factor(A)57.5 | -2.75892 | 0.08380 | -32.922 | < 2e-16 |
| relevel(factor(C), "1933")1893 | -0.84187 | 0.28009 | -3.006 | 0.0023 |
| relevel(factor(C), "1933")1928 | -0.17922 | 0.05965 | -3.005 | 0.0026 |
| relevel(factor(C), "1933")1938 | 0.07540 | 0.05592 | 1.348 | 0.1770 |
| relevel(factor(C), "1933")1973 | 1.37438 | 0.17490 | 7.858 | 3.90e-07 |
| factor(P)1955.5 | 0.04793 | 0.07022 | 0.683 | 0.4967 |
| factor(P)1985.5 | 0.09276 | 0.04091 | 2.267 | 0.0238 |
| factor(P)1990.5 | NA | NA | NA |

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 89358.526 on 81 degrees of freedom
Residual deviance: 35.459 on 49 degrees of freedom

**No. of parameters**

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

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

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

**Relationship of models**

<table>
<thead>
<tr>
<th>Testis cancer, Denmark</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
</tr>
<tr>
<td>865.08 / 72</td>
</tr>
<tr>
<td>739.01 / 1</td>
</tr>
<tr>
<td>p=0.0000</td>
</tr>
<tr>
<td><strong>Age–drift</strong></td>
</tr>
<tr>
<td>126.07 / 71</td>
</tr>
<tr>
<td><strong>Age–Period</strong></td>
</tr>
<tr>
<td>117.7 / 64</td>
</tr>
<tr>
<td>8.37 / 7</td>
</tr>
<tr>
<td>p=0.3010</td>
</tr>
<tr>
<td>60.6 / 15</td>
</tr>
<tr>
<td>p=0.0000</td>
</tr>
<tr>
<td><strong>Age–Cohort</strong></td>
</tr>
<tr>
<td>65.47 / 56</td>
</tr>
<tr>
<td>30.01 / 7</td>
</tr>
<tr>
<td>p=0.0001</td>
</tr>
<tr>
<td><strong>Age–Period–Cohort</strong></td>
</tr>
<tr>
<td>35.46 / 49</td>
</tr>
<tr>
<td>82.24 / 15</td>
</tr>
<tr>
<td>p=0.0000</td>
</tr>
</tbody>
</table>
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></td>
</tr>
<tr>
<td>Δ</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>Δ</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>Δ</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>Δ</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 programs: Put extremes of periods or cohorts to 0, and choose a reference for the other.

Holford: Extract linear effects by regression:

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

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

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:

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

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

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

The rest is the age-effect:

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

How it adds up:

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

Only the regression on period is needed! (For this model...)
A simple practical approach

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

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

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

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

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

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

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.
Tabulation in the Lexis diagram
Tuesday 24th, morning

Bendix Carstensen

Age-Period-Cohort models
19–21 September
University of Lisboa,
www.bendixcarstensen.com/APC/Lisboa.2011

Tabulation of register data

Testis cancer cases in Denmark.

Male person-years in Denmark.
### Tabulation of register data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>109</td>
<td>110</td>
<td>111</td>
<td>112</td>
<td>113</td>
<td>114</td>
</tr>
<tr>
<td>20</td>
<td>147</td>
<td>148</td>
<td>149</td>
<td>150</td>
<td>151</td>
<td>152</td>
</tr>
<tr>
<td>25</td>
<td>185</td>
<td>186</td>
<td>187</td>
<td>188</td>
<td>189</td>
<td>190</td>
</tr>
<tr>
<td>30</td>
<td>223</td>
<td>224</td>
<td>225</td>
<td>226</td>
<td>227</td>
<td>228</td>
</tr>
<tr>
<td>35</td>
<td>261</td>
<td>262</td>
<td>263</td>
<td>264</td>
<td>265</td>
<td>266</td>
</tr>
<tr>
<td>40</td>
<td>300</td>
<td>301</td>
<td>302</td>
<td>303</td>
<td>304</td>
<td>305</td>
</tr>
<tr>
<td>45</td>
<td>338</td>
<td>339</td>
<td>340</td>
<td>341</td>
<td>342</td>
<td>343</td>
</tr>
</tbody>
</table>

Tabulation in the Lexis diagram (Lexis-tab) 168/238

### Tabulation of register data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>720.8</td>
<td>742.1</td>
<td>764.4</td>
<td>786.7</td>
<td>809.0</td>
<td>831.3</td>
</tr>
<tr>
<td>31</td>
<td>676.7</td>
<td>708.0</td>
<td>739.3</td>
<td>770.6</td>
<td>801.9</td>
<td>833.2</td>
</tr>
<tr>
<td>32</td>
<td>632.1</td>
<td>663.4</td>
<td>694.7</td>
<td>726.0</td>
<td>757.3</td>
<td>788.6</td>
</tr>
<tr>
<td>33</td>
<td>588.5</td>
<td>620.8</td>
<td>653.1</td>
<td>685.4</td>
<td>717.7</td>
<td>749.0</td>
</tr>
<tr>
<td>34</td>
<td>544.9</td>
<td>577.2</td>
<td>609.5</td>
<td>641.8</td>
<td>674.1</td>
<td>706.4</td>
</tr>
<tr>
<td>35</td>
<td>501.3</td>
<td>533.6</td>
<td>565.9</td>
<td>598.2</td>
<td>630.5</td>
<td>662.8</td>
</tr>
<tr>
<td>36</td>
<td>457.7</td>
<td>490.0</td>
<td>522.3</td>
<td>554.6</td>
<td>586.9</td>
<td>619.2</td>
</tr>
<tr>
<td>37</td>
<td>414.1</td>
<td>446.4</td>
<td>478.7</td>
<td>511.0</td>
<td>543.3</td>
<td>575.6</td>
</tr>
<tr>
<td>38</td>
<td>370.5</td>
<td>402.8</td>
<td>435.1</td>
<td>467.4</td>
<td>499.7</td>
<td>532.0</td>
</tr>
<tr>
<td>39</td>
<td>327.0</td>
<td>359.3</td>
<td>391.6</td>
<td>423.9</td>
<td>456.2</td>
<td>488.5</td>
</tr>
<tr>
<td>40</td>
<td>283.5</td>
<td>315.8</td>
<td>348.1</td>
<td>380.4</td>
<td>412.7</td>
<td>445.0</td>
</tr>
</tbody>
</table>

Testis cancer cases in Denmark.
Male person-years in Denmark.

Tabulation in the Lexis diagram (Lexis-tab) 169/238

### Tabulation of register data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>300</td>
<td>301</td>
<td>302</td>
<td>303</td>
<td>304</td>
<td>305</td>
</tr>
<tr>
<td>31</td>
<td>265</td>
<td>266</td>
<td>267</td>
<td>268</td>
<td>269</td>
<td>270</td>
</tr>
<tr>
<td>32</td>
<td>223</td>
<td>224</td>
<td>225</td>
<td>226</td>
<td>227</td>
<td>228</td>
</tr>
<tr>
<td>33</td>
<td>181</td>
<td>182</td>
<td>183</td>
<td>184</td>
<td>185</td>
<td>186</td>
</tr>
<tr>
<td>34</td>
<td>139</td>
<td>140</td>
<td>141</td>
<td>142</td>
<td>143</td>
<td>144</td>
</tr>
<tr>
<td>35</td>
<td>98.2</td>
<td>99.5</td>
<td>100.8</td>
<td>102.1</td>
<td>103.4</td>
<td>104.7</td>
</tr>
<tr>
<td>36</td>
<td>57.0</td>
<td>58.3</td>
<td>59.6</td>
<td>60.9</td>
<td>62.2</td>
<td>63.5</td>
</tr>
<tr>
<td>37</td>
<td>16.0</td>
<td>17.3</td>
<td>18.6</td>
<td>19.9</td>
<td>21.2</td>
<td>22.5</td>
</tr>
<tr>
<td>38</td>
<td>14.0</td>
<td>15.3</td>
<td>16.6</td>
<td>17.9</td>
<td>19.2</td>
<td>20.5</td>
</tr>
<tr>
<td>39</td>
<td>12.0</td>
<td>13.3</td>
<td>14.6</td>
<td>15.9</td>
<td>17.2</td>
<td>18.5</td>
</tr>
<tr>
<td>40</td>
<td>10.0</td>
<td>11.3</td>
<td>12.6</td>
<td>13.9</td>
<td>15.2</td>
<td>16.5</td>
</tr>
</tbody>
</table>

Testis cancer cases in Denmark.
Male person-years in Denmark.

Tabulation in the Lexis diagram (Lexis-tab) 170/238
Tabulation of register data

Testis cancer cases in Denmark.

Male person-years in Denmark.

Subdivision by year of birth (cohort).

Major sets in the Lexis diagram

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

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

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

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

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

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

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

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

Lower triangles: Classification by age and cohort, last born cohort. (⊇)
Mean time in triangles

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

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

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

Upper triangles (∨), A:

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

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

\[
E_A(c) = \frac{1}{3} - \frac{2}{3} = -\frac{1}{3}
\]
Lower triangles (\(\triangle\)), B:

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

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

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

---

Tabulation by age, period and cohort

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

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

---

Population figures

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

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

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

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

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

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

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

\[
\int_{p=0}^{p=1} \int_{a=0}^{a=p} 2(p - a) \, da \, dp = \int_{p=0}^{p=1} \left[ 2pa - a^2 \right]_{a=0}^{a=p} \, dp = \int_{p=0}^{p=1} p^2 \, dp = \frac{1}{3}
\]
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:</th>
<th>B:</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>$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}y$</td>
<td>$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}y$</td>
</tr>
<tr>
<td>Dead in B:</td>
<td>$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}y$</td>
<td>$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}y$</td>
</tr>
<tr>
<td>$\sum$</td>
<td>$(\frac{1}{3}\ell_{a,p} + \frac{1}{6}\ell_{a+1,p+1}) \times 1y$</td>
<td>$(\frac{1}{6}\ell_{a,p} + \frac{1}{3}\ell_{a+1,p+1}) \times 1y$</td>
</tr>
</tbody>
</table>

**Population as of 1. January from Statistics Denmark:**

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>2000 2001 2002</td>
<td>2000 2001 2002</td>
</tr>
<tr>
<td>22</td>
<td>33435 33540 32272</td>
<td>32637 32802 31709</td>
</tr>
<tr>
<td>23</td>
<td>35357 35379 33742</td>
<td>34163 32853 33156</td>
</tr>
<tr>
<td>24</td>
<td>38199 35400 33674</td>
<td>37803 34353 33070</td>
</tr>
<tr>
<td>25</td>
<td>37958 38257 35499</td>
<td>37318 37955 34526</td>
</tr>
<tr>
<td>26</td>
<td>38194 38048 38341</td>
<td>37292 37371 38119</td>
</tr>
<tr>
<td>27</td>
<td>39891 38221 38082</td>
<td>39273 37403 37525</td>
</tr>
</tbody>
</table>
Exercise:

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

Summary:

Population risk time:

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

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

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

APC-model for triangular data

Wednesday 25th, morning

Bendix Carstensen

Age-Period-Cohort models
19–21 September
University of Lisboa,
www.bendixcarstensen.com/APC/Lisboa.2011
Model for triangular data

- One parameter per distinct value on each timescale.
- Example: 3 age-classes and 3 periods:
  - 6 age parameters
  - 6 period parameters
  - 10 cohort parameters
- Model:
  \[ \lambda_{ap} = \alpha_a + \beta_p + \gamma_c \]

Problem: Disconnected design!

Log-likelihood contribution from one triangle:

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

The log-likelihood can be separated:

\[ \sum_{a,p \in \triangledown} D_{ap}\log(\lambda_{ap}) - \lambda_{ap}Y_{ap} + \sum_{a,p \in \Delta} D_{ap}\log(\lambda_{ap}) - \lambda_{ap}Y_{ap} \]

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

Illustration by lung cancer data

```R
> library(Epi)
> data(lungDK)
> lungDK[1:10,]
A5  P5  C5  up  Ax  Px  Cx  D  Y
1  40 1943 1898  1 43.3333 1944.667 1901.333 52 336233.8
2  40 1943 1903  0 41.6667 1946.333 1904.667 28 357812.7
3  40 1948 1903  1 43.3333 1949.667 1906.333 51 363783.7
4  40 1948 1908  0 41.6667 1951.333 1909.667 30 390985.8
5  40 1953 1908  1 43.3333 1954.667 1911.333 50 391925.3
6  40 1953 1913  0 41.6667 1956.333 1914.667 23 377515.3
7  40 1958 1913  1 43.3333 1959.667 1916.333 56 365575.5
8  40 1958 1918  0 41.6667 1961.333 1919.667 43 383689.0
9  40 1963 1918  1 43.3333 1964.667 1921.333 44 385878.5
10 40 1963 1923  0 41.6667 1966.333 1924.667 38 371361.5
```
Fill in the number of cases (D) and person-years (Y) from previous slide.

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

Mark mean date of birth for these.

---

**APC-model with “synthetic” cohorts**

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

... 

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

No. parameters: 220 - 182 = 38.

A = 10, P = 11, C = 20 \Rightarrow A + P + C - 3 = 38
```
APC-model with “correct” cohorts

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

... null deviance: 1.0037e+08 on 220 degrees of freedom
Residual deviance: 2.8473e+02 on 144 degrees of freedom

No. parameters: 220 - 144 = 76 (= 38 x 2).

A = 20,   P = 22,   C = 40  ⇒   A + P + C - 3 = 79

We have fitted two age-period-cohort models separately to upper and lower triangles.
```
Now, explicitly fit models for upper and lower triangles:

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

APC-model for triangular data (APC-tri)

---

**APC-model: Parametrization**

**Wednesday 25th, afternoon**

Bendix Carstensen

Age-Period-Cohort models
19–21 September
University of Lisboa,
www.bendixcarstensen.com/APC/Lisboa.2011
What’s the problem?

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

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

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

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

The identifiability problem still exists:

\[ c = p - a \iff p - a - c = 0 \]

\[ \lambda_{ap} = f(a) + g(p) + h(c) = f(a) + g(p) + h(c) + \gamma(p - a - c) = f(a) - \mu_a - \gamma a + g(p) + \mu_a + \mu_c + \gamma p + h(c) - \mu_c - \gamma c \]

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

Smooth functions

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

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

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

All of these contain the linear effect as special case.
**Parametrization of effects**

There are still three “free” parameters:

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

Choose \(\mu_a, \mu_c\) and \(\gamma\) 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?

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?
Implementation:

1. Obtain any set of parameters $f(a)$, $g(p)$, $h(c)$.
2. Extract the trend from the period effect:
   \[ \tilde{g}(p) = \hat{g}(p) - (\mu + \beta p) \]
3. Use the functions:
   \[
   \begin{align*}
   \tilde{f}(a) &= \hat{f}(a) + \mu + \beta a + \hat{h}(c_0) + \beta c_0 \\
   \tilde{g}(p) &= \hat{g}(p) - \mu - \beta p \\
   \tilde{h}(c) &= \hat{h}(c) + \beta c - \hat{h}(c_0) - \beta c_0
   \end{align*}
   \]
   These functions fulfill the criteria.

“Extract the trend”

Not a well-defined concept:

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

How do we get the standard errors?

Matrix-algebra! Projections!

Parametric function

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

Example: 2nd order polynomial:

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

nrow($M$) is the number of observations in the dataset.
Extract the trend from $g$:

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

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

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

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

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

(Orthogonality requires an inner product!)

---

**Practical parametization**

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

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

---

APC-model: Parametrisation (APC-par)
**Information in the data and inner product**

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

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

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

Two relevant inner products:

\[
\langle m_j | m_k \rangle = \sum_i m_{ij} m_{ik} \quad \langle m_j | w m_k \rangle = \sum_i m_{ij} w_i m_{ik}
\]

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

---

**How to?**

Implemented in \texttt{apc.fit}:

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

Consult the help page for:
\texttt{apc.fit} to see options for weights in inner product, type of function, variants of parametrization etc. \texttt{apc.plot}, \texttt{apc.lines} and \texttt{apc.frame} to see how to plot the results.
APC-model: Parametrization (APC-par)

Weighted drift: 2.58 (2.42 - 2.74) %/year

APC-model: Parametrization (APC-par)
APC-models for several datasets
Wednesday 25th, afternoon

Bendix Carstensen

Age-Period-Cohort models
19–21 September
University of Lisboa,
www.bendixcarstensen.com/APC/Lisboa.2011

Two sets of data
Example: Testis cancer in Denmark, Seminoma and non-Seminoma cases.

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

<table>
<thead>
<tr>
<th>Histology</th>
<th>D</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4708.00</td>
<td>127.53</td>
</tr>
<tr>
<td>2</td>
<td>3632.00</td>
<td>127.53</td>
</tr>
<tr>
<td>3</td>
<td>466.00</td>
<td>127.53</td>
</tr>
<tr>
<td>Total</td>
<td>8806.00</td>
<td>382.58</td>
</tr>
</tbody>
</table>

First step is separate analyses for each subtype.
APC-models for several datasets (APC-2)
Analysis of two rates: Formal tests I

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

Analysis of Deviance Table

| Resid. Df | Resid. Dev | Df | Deviance | P(>|Chi|) |
|-----------|------------|----|----------|---------|
| 1          | 10737      | 10553.7 | 19   | 185.7   | 2.278e-29 |
| 2          | 10718      | 10553.7 | 14   | 168.3   | 1.513e-28 |
| 3          | 10704      | 10553.7 | -19  | -309.0  | 2.832e-54 |
| 4          | 10723      | 10553.7 | -14  | -45.0   | 4.042e-05 |

Analysis of two rates: Formal tests II

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10737</td>
<td>10553.7</td>
</tr>
<tr>
<td>2</td>
<td>10718</td>
<td>10367.9</td>
</tr>
<tr>
<td>3</td>
<td>10704</td>
<td>10199.6</td>
</tr>
<tr>
<td>4</td>
<td>10723</td>
<td>10508.6</td>
</tr>
<tr>
<td>5</td>
<td>10737</td>
<td>10553.7</td>
</tr>
</tbody>
</table>

APC-model: Interactions

Friday 27th, morning

Bendix Carstensen

Age-Period-Cohort models
19–21 September
University of Lisboa,
www.bendixcarstensen.com/APC/Lisboa.2011
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.

II

```r
library(Epi)
library(splines)
load(file="c:/Bendix/Artikler/A_P_C/IDDM/Eurodiab/data/tri.Rdat"

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

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

APC-model: Interactions (APC-int)

III

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

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

# The prediction matrices
Pa <- Ma[match(A.pt,A),drop=F]
Pc <- Mc[match(C.pt,C),drop=F]
Pp <- Mp[match(P.pt,P),drop=F]

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

summary(apcs)
```

APC-model: Interactions (APC-int)
Analysis of DM-rates: Age × sex interaction

\[ \text{ci.lin} \left( \text{apcs} \right) \]
\[ \text{ci.lin} \left( \text{apcs, subset="sexF"}, \text{Exp=T} \right) \]
\[ \text{ci.lin} \left( \text{apcs, subset="sexF"}, \text{ctr.mat=Pa, Exp=T} \right) \]

# Extract the effects
\[ \text{F.inc} \leftarrow \text{ci.lin} \left( \text{apcs, subset="sexF"}, \text{ctr.mat=Pa, Exp=T} \right)[,5:7] \]
\[ \text{M.inc} \leftarrow \text{ci.lin} \left( \text{apcs, subset="sexM"}, \text{ctr.mat=Pa, Exp=T} \right)[,5:7] \]
\[ \text{MF.RR} \leftarrow \text{ci.lin} \left( \text{apcs, subset=c("sexM","sexF"), ctr.mat=chind(Pa}, \text{Exp=T} \right)[,5:7] \]
\[ \text{c.RR} \leftarrow \text{ci.lin} \left( \text{apcs, subset="Mc"}, \text{ctr.mat=Pc, Exp=T} \right)[,5:7] \]
\[ \text{p.RR} \leftarrow \text{ci.lin} \left( \text{apcs, subset="Mp"}, \text{ctr.mat=Pp, Exp=T} \right)[,5:7] \]

# plt( paste( "DM-DK" ), width=11 )
\[ \text{par( mar=c(4,4,1,4), mgp=c(3,1,0)/1.6, las=1 )} \]

# The the frame for the effects
\[ \text{fr} \leftarrow \text{apc.frame}( \text{a.lab=c(0,5,10,15),} \]
\[ \quad \text{a.tic=c(0,5,10,15),} \]
\[ \quad \text{r.lab=c(c(1,1.5,3,5),c(1,1.5,3,5)*10),} \]
\[ \quad \text{r.tic=c(c(1,1.5,2,5),c(1,1.5,2,5)*10),} \]
\[ \quad \text{cp.lab=seq(1980,2000,10),} \]
\[ \quad \text{cp.tic=seq(1975,2000,5),} \]

APC-model: Interactions (APC-int)

Analysis of DM-rates: Age × sex interaction

\[ \text{rr.ref=5,} \]
\[ \text{gap=1,} \]
\[ \text{col.grid=gray(0.9),} \]
\[ \text{a.txt="",} \]
\[ \text{cp.txt="",} \]
\[ \text{r.txt="",} \]
\[ \text{rr.txt=""} \]

# Draw the estimates
\[ \text{matlines( A.pt, F.inc, lwd=c(3,1,1), lty=1, col="blue" )} \]
\[ \text{matlines( A.pt, M.inc, lwd=c(3,1,1), lty=1, col="red" )} \]
\[ \text{matlines( C.pt - fr[1], c.RR \times fr[2],} \]
\[ \quad \text{lwd=c(3,1,1), lty=1, col="black" )} \]
\[ \text{matlines( P.pt - fr[1], p.RR \times fr[2],} \]
\[ \quad \text{lwd=c(3,1,1), lty=1, col="black" )} \]
\[ \text{matlines( A.pt, MF.RR \times fr[2],} \]
\[ \quad \text{lwd=c(3,1,1), lty=1, col=gray(0.6) )} \]
\[ \text{abline(h=fr[2])} \]
Predicting future rates

Friday 27th, morning

Bendix Carstensen

Age-Period-Cohort models
19–21 September
University of Lisboa,

www.bendixcarstensen.com/APC/Lisboa.2011
Prediction of future rates

Model:

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

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

Identifiability

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

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

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

\[
\begin{align*}
\text{pred}(g(p) + \gamma p) &= \text{pred}(g(p)) + \gamma p \\
\text{pred}(h(c) - \gamma c) &= \text{pred}(h(c)) - \gamma c
\end{align*}
\]

If this is met, the predictions made will not depend on the parametrization chosen.
If one of the conditions does not hold, the prediction will depend on the parametrization chosen.

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

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

Example: Breast cancer in Denmark

Practicalities

- Long term predictions notoriously unstable.
- Decreasing slopes are possible, the requirement is that at any future point changes in the parametrization should cancel out in the predictions.
Predicted age-specific breast cancer rates at 2020 (black),
in the 1950 cohort (blue),
and the estimated age-effects (red).