Survival analysis and medical demography

`www.biostat.ku.dk/~bxc/Melbourne/staff`

Bendix Carstensen

Clinical Epidemiology and Biostatistics Unit, Royal Children's Hospital, Melbourne
Department of Public Health, University of Melbourne
Steno Diabetes Center, Copenhagen
Department of Biostatistics, University of Copenhagen

e-mail: `bxc@steno.dk` homepage: `www.biostat.ku.dk/~bxc`

Melbourne, 2003

Introduction

- Data analysis.
- Concepts behind follow-up studies. (Probability theory).
- Empirical demography.
- Assignments / exercises.
- Your ID.

Survival analysis and medical demography

Basics of survival data

20 March 2003

Bendix Carstensen

Simple survival

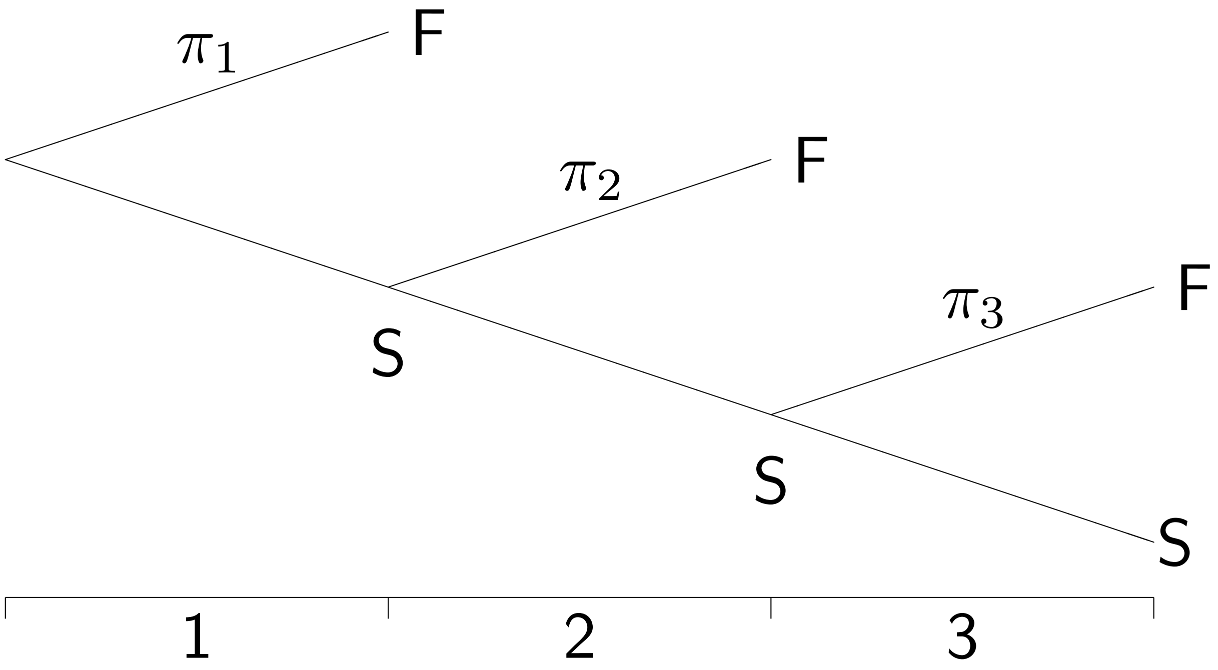
In the simplest case, all persons start at time 0, the designated origin of the timescale, which could be:

Age: Date of birth.

Time on study: Date of entry (randomization, . . .).

Suppose we divide time into bands, and estimate the probability of death in each band, $\pi_1, \pi_2, \pi_3, \ldots$

Survival tree

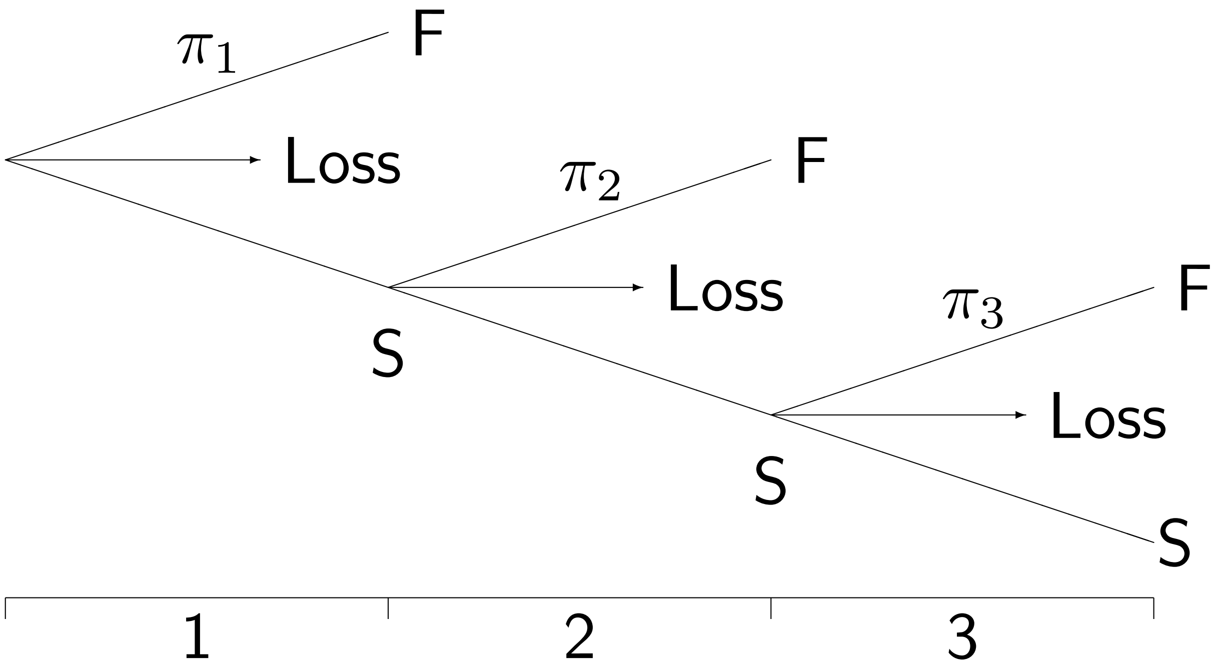


$$\hat{\pi}_i = D_i/N_i \quad i = 1, \ldots, 3$$

D_i deaths in, N_i no. alive at start of interval i.

π_i **conditional** probability **given** survival till start of interval.

Loss to follow–up (censoring)



Subjects lost to follow–up in band 1 make no contribution to estimation of π_2 or π_3, but what about the estimate of π_1?

Compensating loss to follow-up

With N subjects observed, D failures, and L lost to follow–up, estimate depends on *when* losses occurred:
— if at the end of the band: D/N
— if at the start of the band: $D/(N-L)$
— unknown then compromise: $D/(N-L/2)$ or:

$$\pi = \frac{D + \pi L/2}{N} \quad \Longrightarrow \quad \pi = D/(N-L/2)$$

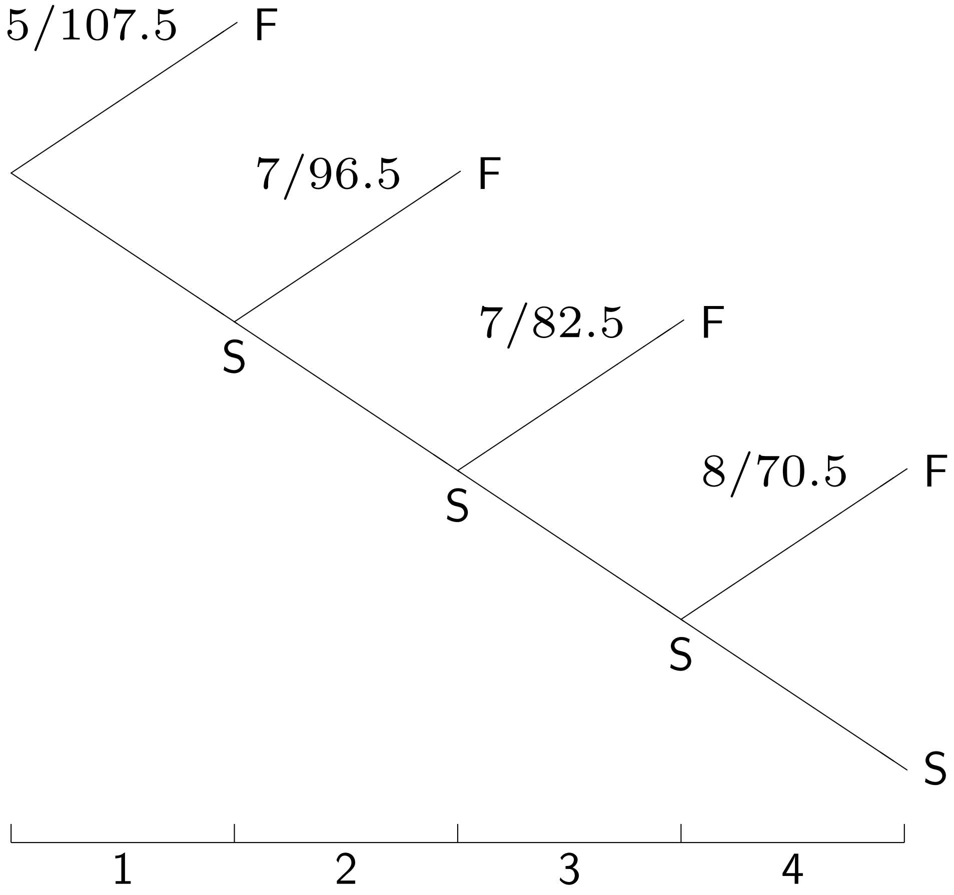
Known as the **actuarial** or **life table** method.

Survival after Cervix cancer (C& H)

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

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

Life table calculations year 1–4, stage I



N	D	L	$N-\frac{1}{2}L$	π	$1-\pi$
110	5	5	107.5	0.047	0.953
100	7	7	96.5	0.073	0.927
86	7	7	82.5	0.085	0.915
72	8	3	70.5	0.113	0.887

Cumulative survival probabilities:

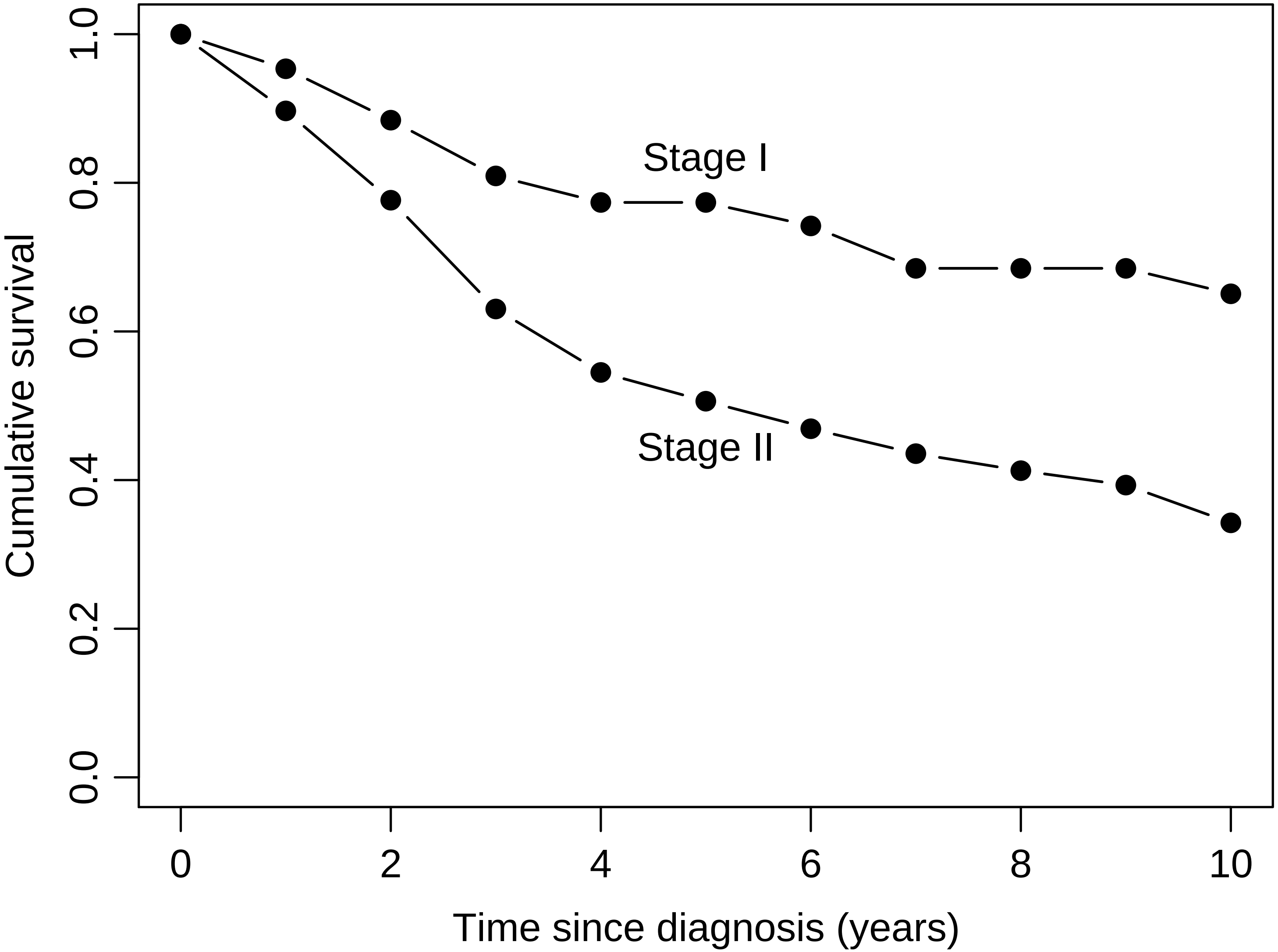
1 year : 0.953

2 years: $0.953 \times 0.927 = 0.884$

3 years: $0.884 \times 0.915 = 0.809$

4 years: $0.809 \times 0.887 = 0.635$

Life table estimates for both stages

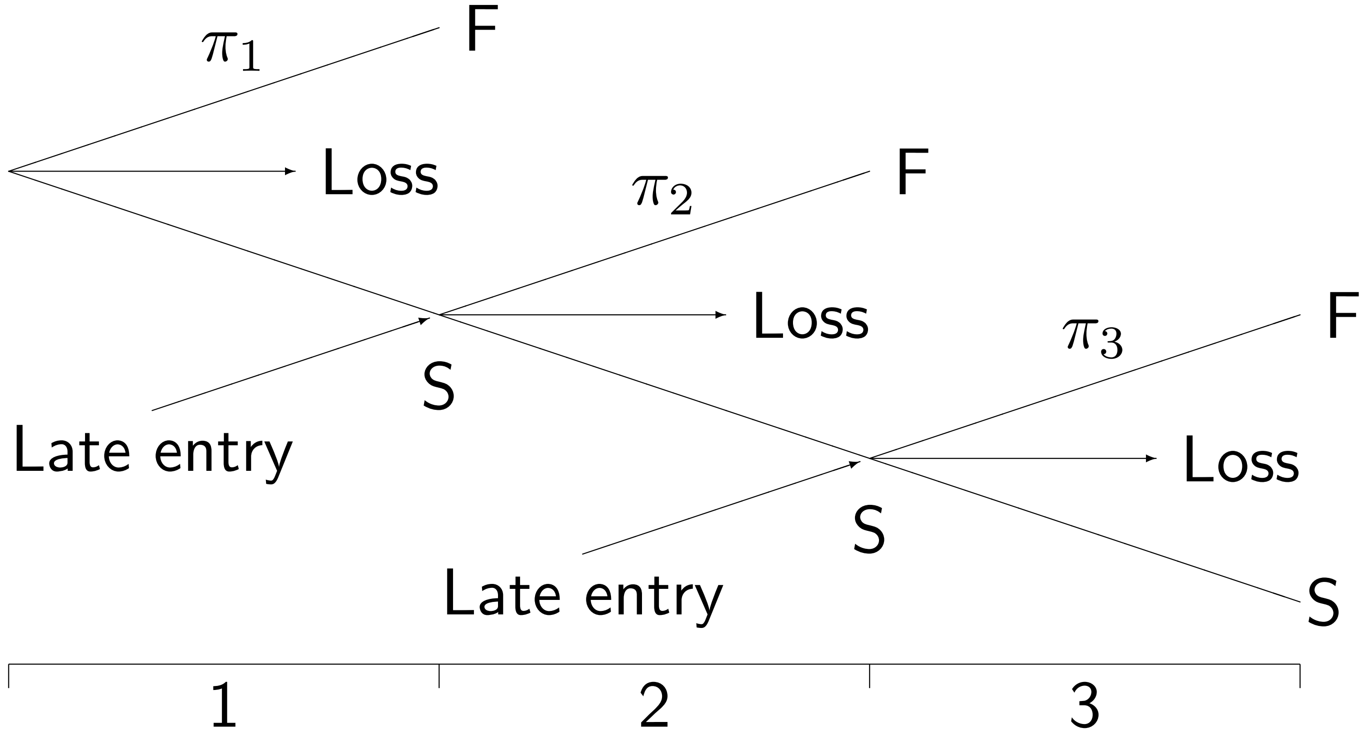


Life tables in Stata: `ltable`

```
. use gcervix, clear
. list
          time          N     status      stage
  1.         0          5          0          1
  2.         0          5          1          1
  3.         1          7          0          1
  4.         1          7          1          1
  5.         2          7          0          1
  6.         2          7          1          1
  7.         3          3          0          1
  8.         3          8          1          1
...
```

```
. ltable time status [freq=N] if stage==1

           Beg.                               Std.
Interval  Total Deaths Lost  Survival  Error    [95% C.I.]
--------------------------------------------------------------
 0     1    110      5    5   0.9535  0.0203  0.8919 0.9804
 1     2    100      7    7   0.8843  0.0314  0.8052 0.9326
 2     3     86      7    7   0.8093  0.0395  0.7170 0.8741
 3     4     72      8    3   0.7175  0.0465  0.6146 0.7974
 4     5     61      7    0   0.6351  0.0505  0.5273 0.7247
 5     6     54     10    2   0.5153  0.0533  0.4064 0.6137
 6     7     42      6    3   0.4390  0.0538  0.3321 0.5406
 7     8     33      5    0   0.3724  0.0532  0.2694 0.4753
 8     9     28      4    0   0.3192  0.0518  0.2211 0.4215
 9    10     24      8    1   0.2106  0.0463  0.1282 0.3068
10    11     15     15    0   0.0000       .       .      .
--------------------------------------------------------------
```

Late entry (left truncation)



- Observation does not start at $t = 0$ for all subjects

- Example:
  In an Observational study of survival after operation, some may enter the study some time after operation, because their record had not been available earlier.

- Example:
  Immigrants enter Australia in their mid-20s.

- Late entries in band 3 do not contribute to the estimates of π_1 and π_2

- Why not (we know they didn't die)?

Independent censoring and truncation

- The comparison would be biased.
  We would miss observation time and deaths among those who died before they came to our knowledge.

- Only include observation time where an event occurrence would have been recorded.

- Studies with censoring and truncation could give biased answers, if not handled properly.

- Bias could be caused by:
  - Selective removal (censoring) of high risk subjects
  - Selective import (late entry) of low risk subjects

- In epidemiology these would be termed *selection biases*

- If censoring or truncation has no effect on later failure rates it is said to be *independent* censoring / truncation

The Kaplan–Meier method

- Divide time into clicks, i.e very small time intervals, so no censorings occur in the same click as events.

- Clicks with censorings contribute 1 to the cumulative survival function.

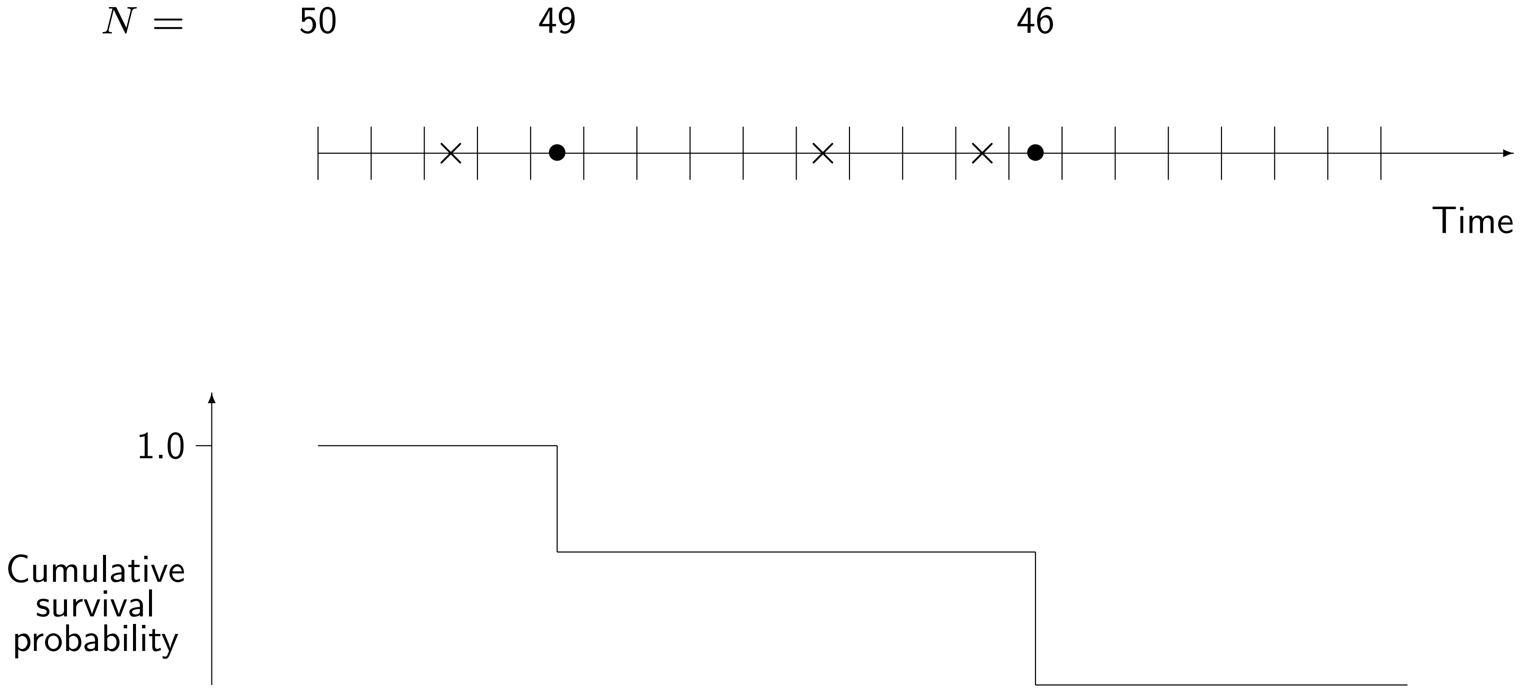
- Clicks with 1 event each contribute $1 - 1/N$ to the cumulative survival function.

- Tied events poses no problem:

$$1-\frac{2}{N}=\frac{N-2}{N}=\frac{N-2}{N-1}\times\frac{N-1}{N}=(1-\frac{1}{N-1})\times(1-\frac{1}{N})$$

- Censorings and events tied:
  Convention is that events come before censorings.

The Kaplan–Meier method



Steps caused by multiplying by $(1 - 1/49)$ and $(1 - 1/46)$ respectively

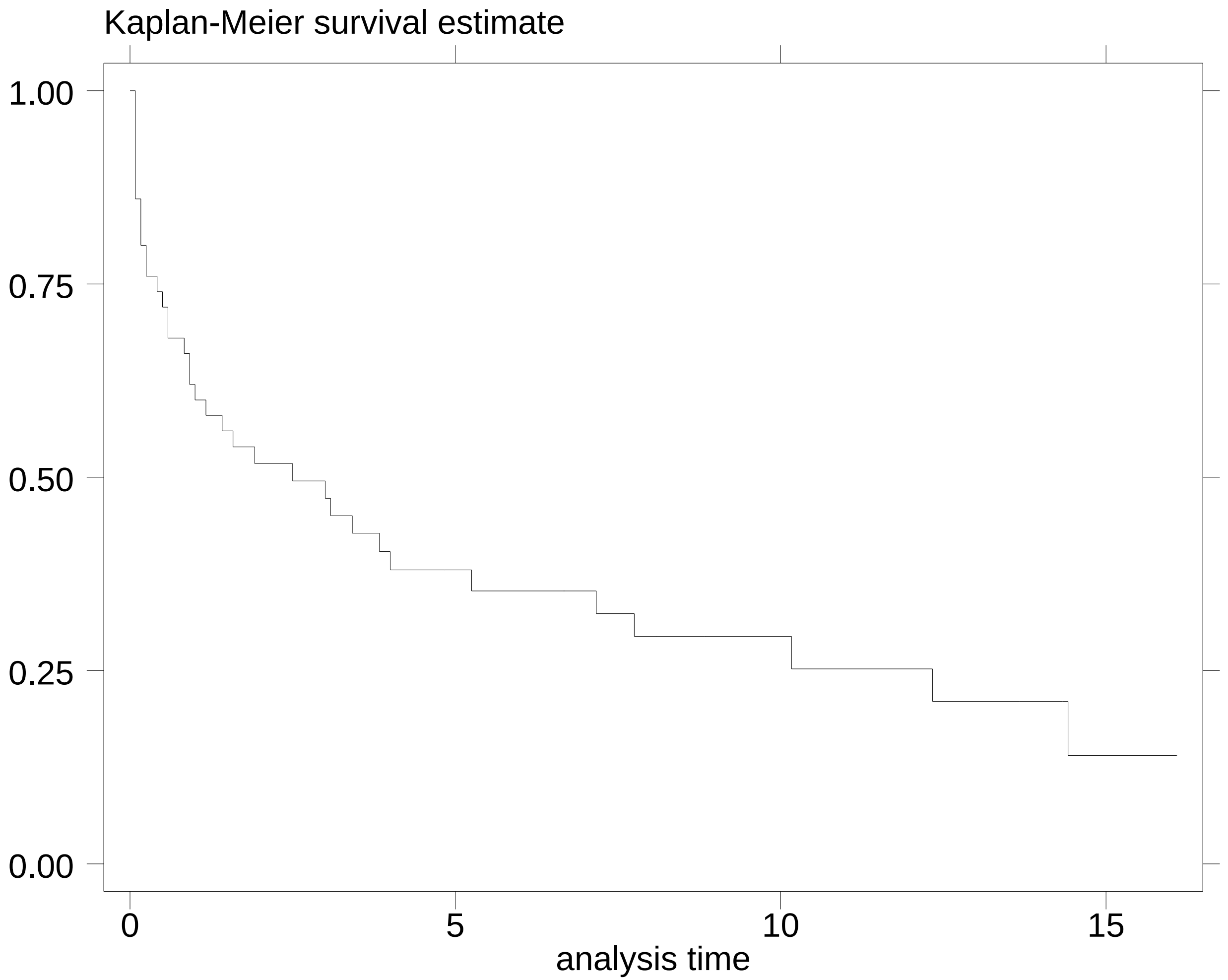
Example calculation in Stata

```
. stset survmm, fail(status==1,2) scale(12)

     failure event:  status == 1 2
obs. time interval:  (0, survmm]
 exit on or before:  failure
    t for analysis:  time/12
----------------------------------------------------------
       53  total obs.
        3  obs. end on or before enter()
----------------------------------------------------------
       50  obs. remaining, representing
       36  failures in single record/single failure data
 197.4167  total analysis time at risk, at risk from t = 0
                      earliest observed entry t = 0
                           last observed exit t = 16.08333
```

Late entries (left truncation) is handled by `stset` through the option `enter`.

```
. sts list

         failure _d:  status == 1 2
   analysis time _t:  survmm/12
         Beg.          Net   Survivor      Std.
  Time  Total  Fail  Lost   Function     Error     [95% Cf. Int.]
--------------------------------------------------------------------
 .0833     50     7     0     0.8600    0.0491     0.7286   0.9307
 .1667     43     3     0     0.8000    0.0566     0.6602   0.8870
   .25     40     2     0     0.7600    0.0604     0.6163   0.8559
...
 3.083     21     1     0     0.4502    0.0716     0.3073   0.5828
 3.417     20     1     0     0.4277    0.0715     0.2867   0.5613
   3.5     19     0     1     0.4277    0.0715     0.2867   0.5613
 3.833     18     1     0     0.4039    0.0714     0.2649   0.5386
     4     17     1     1     0.3801    0.0710     0.2436   0.5156
```

Kaplan–Meier plot using `sts graph`



Survival data: Censoring and truncation

- **Right**-censoring: It is only known that a person has lived until time t_x. The survival time is censored at time t_x. The time of death is in (t_x, ∞).

- **Left**-truncaton: A person is only known if he has survived until time t_i. Had he died before t_i he would not have been known.

  Bias: Inclusion of survival time prior to the point where a death would have been registered. Those who died did not contribute anything, so only survival not death is included.

Survival time distribution

Survival time is a stochastic variable T, with a distribution characterized by the cumulative distribution function, F, and density, f:

$$F(t) = \mathrm{P}\left\{\text{death} \leq t\right\}, \qquad f(t) = F'(t) = \frac{\mathrm{d}F}{\mathrm{d}t}$$

In survival terms it is more of interest to have the probability to survive at least to time t, the survival function[1]

$$S(t) = 1 - F(t)$$

[1] Pocock SJ, Clayton TC & Altman DG argue to use F in: Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. Lancet. 2002 May 11;359(9318):1686-9.

For persons known to be alive at age t_0, we have the **conditional** survival function, given they are alive at time t_i:

$$\mathrm{P}\left\{\text{Alive at } t \mid \text{alive at } t_i\right\} = \frac{S(t)}{S(t_i)}$$

If we want to know something specific about mortality around age t, the limit:

$$\mathrm{P}\left\{\text{Alive at } t+h \mid \text{alive at } t\right\} = \frac{S(t+h)}{S(t)} \xrightarrow[h \to 0]{} 1$$

is not very interesting for obvious reasons.

Intensity or rate

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

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

Theoretical counterpart of a **rate**.

Relationships

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

$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s$ is called the **integrated intensity**.

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

The integrated intensity and cumulative risk

The integrated intensity

$$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s$$

is **not** an intensity, it is dimensionless.

The empirical counterpart of the integrated intensity is known as the cumulative rate.

The cumulative risk of an event (to time t) is:

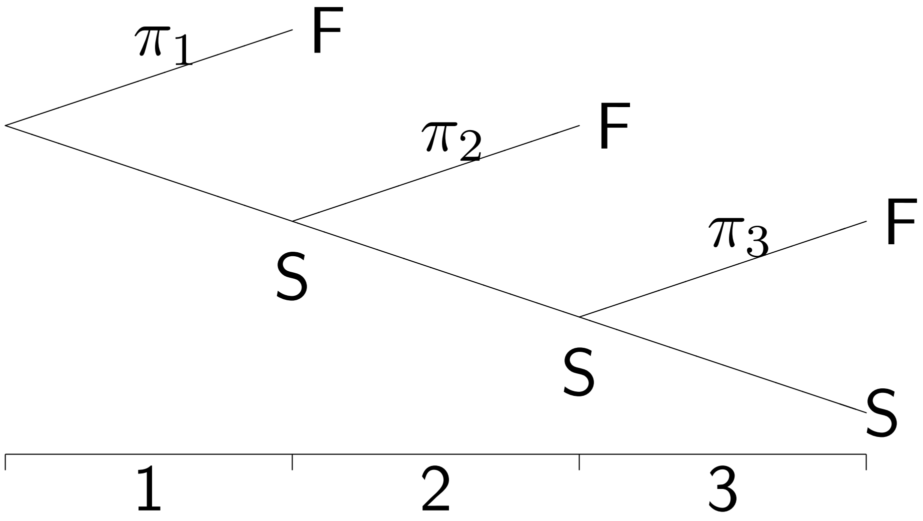
$$\mathrm{P}\left\{\text{Event before time } t\right\} = 1 - S(t) = 1 - \mathrm{e}^{-\Lambda(t)}$$

For small $|x|$ (< 0.05), we have that $1 - \mathrm{e}^{-x} \approx x$, so for small values of the integrated intensity:

$$\text{Cumulative risk to time } t \approx \Lambda(t) = \text{Cumulative rate}$$

Estimation of the integrated intensity

Consider again the timescale in intervals:



A naïve counterpart of the actuarial estimator of the survival function would be (no name):

$$\Lambda(t_n) = \sum_{i=1}^{n} \frac{D_i}{N_i - L_i/2}$$

The Nelson-Aalen estimator

(“Aa” in Norwegian is the old (pre-1950) version of “Å”).

- Divide time into clicks, i.e very small time intervals, so no censorings occur in the same click as events.

- Clicks with censorings contribute 0 to the integrated hazard function.

- Clicks with 1 event each contribute $1/N$ to the integrated hazard function.

- Tied events: $2/N < 1/N + 1/(N-1)$.

  The former is used.

- Censorings and events tied:
  Convention is that events come before censorings.

Comparison to the Kaplan-Meier estimator

For each click with an event:

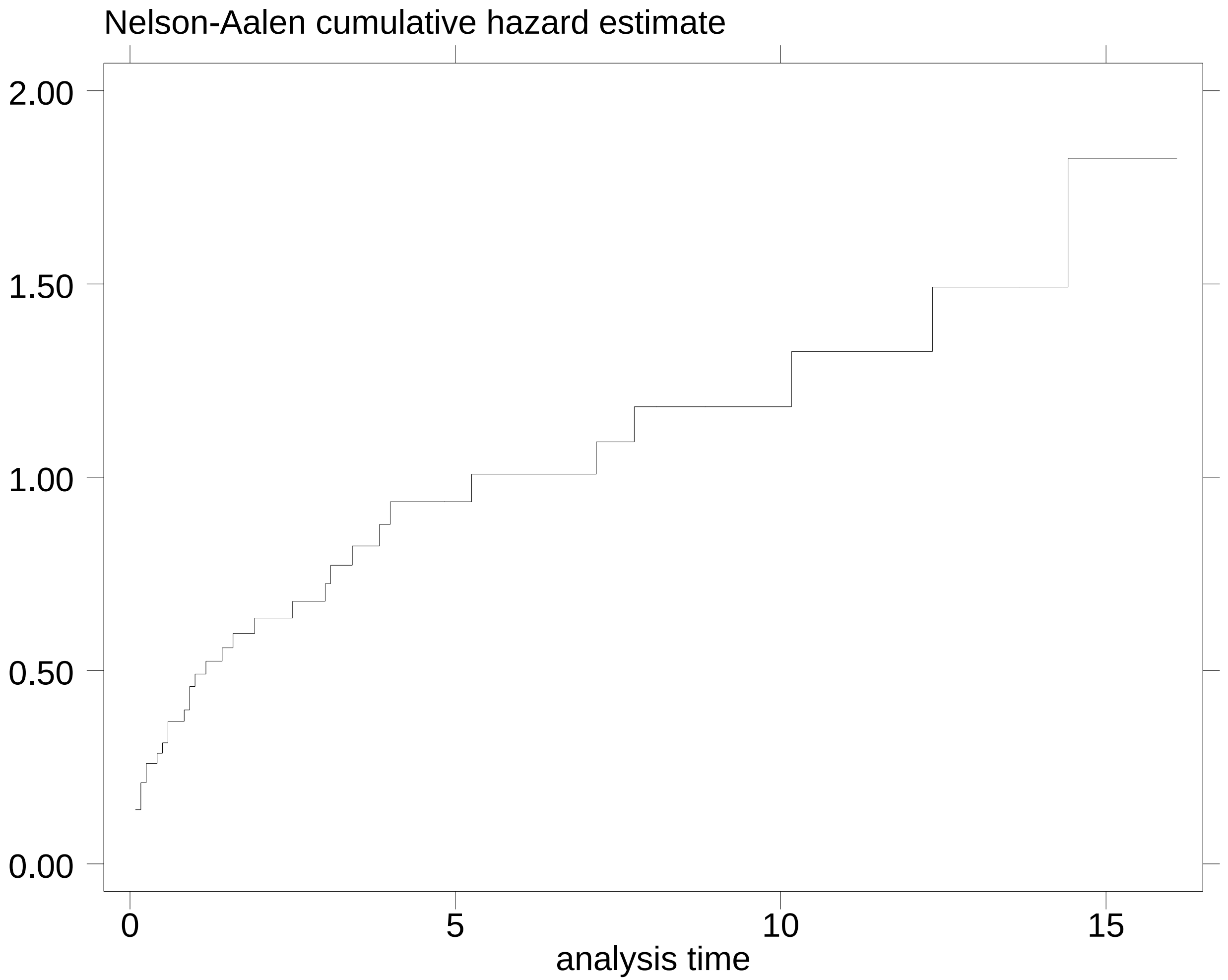
- KM multiplies $1 - 1/N$ to the cumulative survival.
- NA adds $1/N$ to the integrated intensity, i.e. multiplies the cumulative survival by $\mathrm{e}^{-1/N}$.

  Note that $\mathrm{e}^{-1/N} > 1 - 1/N$, by the convexity of the exponential.
- In practical terms they are alike.

Nelson-Aalen estimator in Stata

```
. sts list, na

         failure _d:  status == 1 2
   analysis time _t:  survmm/12

         Beg.         Net  Nelson-Aalen    Std.
  Time   Total  Fail  Lost   Cum. Haz.     Error     [95% Cf. Int.]
------------------------------------------------------------------
 .0833     50     7     0     0.1400     0.0529    0.0667   0.2937
 .1667     43     3     0     0.2098     0.0665    0.1127   0.3905
   .25     40     2     0     0.2598     0.0753    0.1472   0.4585
...
 3.083     21     1     0     0.7723     0.1545    0.5218   1.1432
 3.417     20     1     0     0.8223     0.1624    0.5584   1.2111
   3.5     19     0     1     0.8223     0.1624    0.5584   1.2111
 3.833     18     1     0     0.8779     0.1716    0.5984   1.2879
     4     17     1     1     0.9367     0.1814    0.6408   1.3693
```

Nelson-Aalen plot using `sts graph, na`



Survival analysis and medical demography

Population life tables

20 March 2003

Bendix Carstensen

Expected lifetime

Distribution of death times (age at death) has density $f(a)$, so the expectation of age at death, i.e. expected life time is:

$$\begin{aligned} \int_0^\infty a f(a) \mathrm{d}a &= -\int_0^\infty a(-f(a)) \mathrm{d}a \\ &= -\left[aS(a)\right]_0^\infty + \int_0^\infty S(a) \mathrm{d}a \\ &= \int_0^\infty S(a) \mathrm{d}a \end{aligned}$$

by integration by parts.

This is used in population life tables to compute expected lifetime at birth, but also expected residual life time **given** survival to age a:

$$\mathrm{E}[\ell_{\textsf{res}}(a)] = \int_a^\infty S(t|\textsf{alive at } a)\mathrm{d}t = \int_a^\infty S(t)/S(a)\mathrm{d}t$$

In practice the integral is computed as a sum.

Population life tables (DS, ABS)

These are **cross-sectional life tables**, based on the age-specific mortality in one or two calendar years.

- Mortality rates, typically per $100,000$ p.y.
- Survival function.
- Expected residual life time at the beginning of the age-class.

  The expected residual life time for a new born is a much used figure for comparing mortality between populations.

  And for assessing population mortality trends.

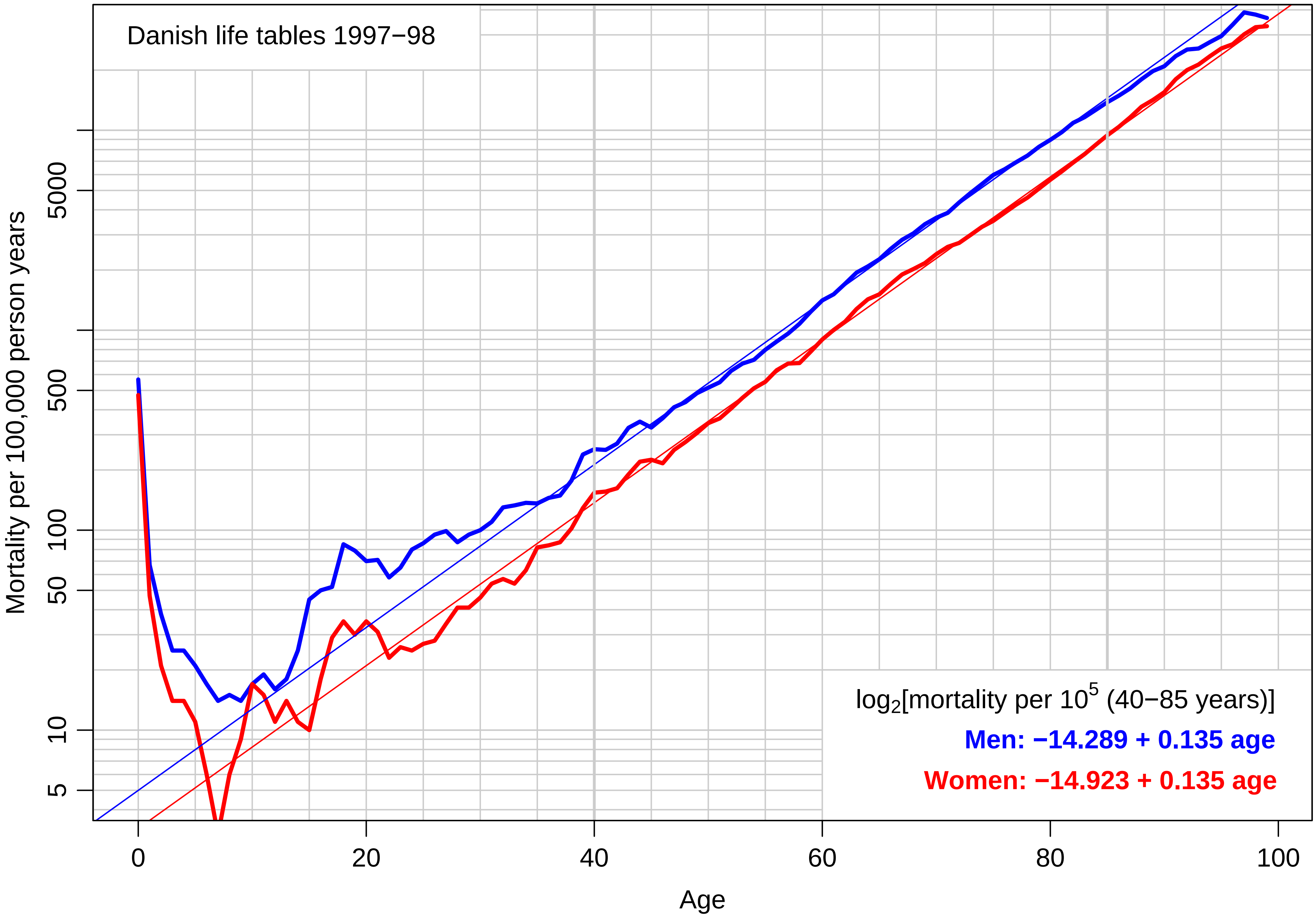
Population life table, DK 1997–98

	Men			Women		
a	$S(a)$	$\lambda(a)$	$\mathrm{E}[\ell_{\mathsf{res}}(a)]$	$S(a)$	$\lambda(a)$	$\mathrm{E}[\ell_{\mathsf{res}}(a)]$
0	1.00000	567	73.68	1.00000	474	78.65
1	0.99433	67	73.10	0.99526	47	78.02
2	0.99366	38	72.15	0.99479	21	77.06
3	0.99329	25	71.18	0.99458	14	76.08
4	0.99304	25	70.19	0.99444	14	75.09
5	0.99279	21	69.21	0.99430	11	74.10
6	0.99258	17	68.23	0.99419	6	73.11
7	0.99242	14	67.24	0.99413	3	72.11
8	0.99227	15	66.25	0.99410	6	71.11
9	0.99213	14	65.26	0.99404	9	70.12
10	0.99199	17	64.26	0.99395	17	69.12
11	0.99181	19	63.28	0.99378	15	68.14
12	0.99162	16	62.29	0.99363	11	67.15
13	0.99147	18	61.30	0.99352	14	66.15
14	0.99129	25	60.31	0.99338	11	65.16
15	0.99104	45	59.32	0.99327	10	64.17
16	0.99059	50	58.35	0.99317	18	63.18
17	0.99009	52	57.38	0.99299	29	62.19
18	0.98957	85	56.41	0.99270	35	61.21
19	0.98873	79	55.46	0.99235	30	60.23

20	0.98795	70	54.50	0.99205	35	59.24
21	0.98726	71	53.54	0.99170	31	58.27
22	0.98656	58	52.58	0.99139	23	57.28
23	0.98599	65	51.61	0.99117	26	56.30
24	0.98535	80	50.64	0.99091	25	55.31
25	0.98456	86	49.68	0.99066	27	54.32
26	0.98371	95	48.72	0.99039	28	53.34
27	0.98277	99	47.77	0.99012	34	52.35
28	0.98180	87	46.81	0.98978	41	51.37
29	0.98094	95	45.86	0.98938	41	50.39
30	0.98002	100	44.90	0.98898	46	49.41
31	0.97903	110	43.94	0.98852	54	48.43
32	0.97795	130	42.99	0.98799	57	47.46
33	0.97668	133	42.05	0.98743	54	46.49
34	0.97537	137	41.10	0.98689	63	45.51
35	0.97403	136	40.16	0.98627	82	44.54
36	0.97271	145	39.21	0.98546	84	43.58
37	0.97130	149	38.27	0.98463	87	42.61
38	0.96985	177	37.32	0.98377	102	41.65
39	0.96813	239	36.39	0.98277	129	40.69
40	0.96582	254	35.48	0.98150	154	39.74
41	0.96336	252	34.56	0.97999	156	38.80
42	0.96093	271	33.65	0.97846	162	37.86
43	0.95833	325	32.74	0.97687	190	36.93
44	0.95522	349	31.85	0.97502	220	35.99
45	0.95189	326	30.96	0.97287	225	35.07
46	0.94879	363	30.06	0.97068	216	34.15
47	0.94534	412	29.16	0.96858	252	33.22

48	0.94145	437	28.28	0.96614	276	32.31
49	0.93733	484	27.40	0.96347	306	31.39
50	0.93279	516	26.53	0.96053	343	30.49
51	0.92798	549	25.67	0.95723	362	29.59
52	0.92289	626	24.81	0.95376	406	28.70
53	0.91711	681	23.96	0.94989	459	27.81
54	0.91087	711	23.12	0.94553	512	26.94
55	0.90439	798	22.28	0.94068	552	26.08
56	0.89717	878	21.46	0.93549	631	25.22
57	0.88929	962	20.65	0.92959	681	24.37
58	0.88074	1077	19.84	0.92326	685	23.54
59	0.87126	1240	19.05	0.91693	783	22.70
60	0.86045	1410	18.28	0.90975	899	21.87
61	0.84831	1513	17.54	0.90157	1003	21.07
62	0.83548	1709	16.80	0.89253	1104	20.27
63	0.82120	1940	16.08	0.88268	1276	19.49
64	0.80527	2086	15.39	0.87142	1428	18.74
65	0.78848	2264	14.71	0.85898	1512	18.00
66	0.77063	2551	14.04	0.84599	1702	17.27
67	0.75097	2833	13.39	0.83159	1900	16.56
68	0.72969	3052	12.77	0.81580	2024	15.87
69	0.70743	3390	12.15	0.79929	2166	15.19
70	0.68344	3650	11.56	0.78197	2401	14.52
71	0.65850	3863	10.98	0.76320	2611	13.86
72	0.63306	4352	10.40	0.74327	2732	13.22
73	0.60551	4855	9.86	0.72297	2993	12.58
74	0.57611	5379	9.33	0.70133	3286	11.95
75	0.54512	5974	8.83	0.67828	3523	11.34

	Men			Women		
a	$S(a)$	$\lambda(a)$	$\mathrm{E}[\ell_{\mathsf{res}}(a)]$	$S(a)$	$\lambda(a)$	$\mathrm{E}[\ell_{\mathsf{res}}(a)]$
76	0.51255	6385	8.36	0.65439	3870	10.74
77	0.47983	6921	7.90	0.62906	4245	10.15
78	0.44662	7470	7.45	0.60236	4606	9.58
79	0.41326	8262	7.01	0.57461	5104	9.01
80	0.37911	8953	6.60	0.54528	5654	8.47
81	0.34517	9778	6.20	0.51445	6219	7.95
82	0.31142	10890	5.82	0.48246	6891	7.44
83	0.27751	11624	5.47	0.44921	7590	6.96
84	0.24525	12660	5.12	0.41512	8473	6.49
85	0.21420	13808	4.79	0.37994	9437	6.04
86	0.18462	14881	4.47	0.34409	10388	5.62
87	0.15715	16188	4.17	0.30834	11604	5.21
88	0.13171	17999	3.88	0.27256	13118	4.83
89	0.10800	19766	3.62	0.23681	14174	4.49
90	0.08666	20911	3.39	0.20324	15504	4.14
91	0.06853	23496	3.15	0.17173	18000	3.81
92	0.05243	25337	2.97	0.14082	20009	3.54
93	0.03915	25634	2.80	0.11264	21312	3.30
94	0.02911	27599	2.60	0.08864	23462	3.06
95	0.02108	29587	2.39	0.06784	25649	2.84
96	0.01484	33753	2.19	0.05044	26944	2.65
97	0.00983	38851	2.05	0.03685	30178	2.44
98	0.00601	37882	2.04	0.02573	32773	2.28
99	0.00373	36433	1.98	0.01730	33149	2.15


Danish life tables 1997−98
Mortality per 100,000 person years
5
10
50
100
500
5000
0
20
40
60
80
100
Age
log$_2$[mortality per 10^5 (40−85 years)]
Men: −14.289 + 0.135 age
Women: −14.923 + 0.135 age

- Why was the logarithm to base 2 chosen for modelling the mortality rates?

- 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?

- The log to base 2 was chosen, so one could easily compute the doubling time for mortality:

  If mortality should double, $\log_2(\text{mortality})$ must increase by 1, so age must increase by $1/0.135 = 7.41$ years.

- $\log_2(\text{rate-ratio})$ between men and women is $-14.289 - (-14.923) = 0.634$, so the rate ratio is $2^{0.634} = 1.55$.

- In order to advance 0.634 in $\log_2$-mortality a woman must advance $0.634/0.135 = 4.70$ years in age. Thus men have the same mortality as women that are 4.70 years older.

Australian life tables

- If mortality should double, $\log_2(\text{mortality})$ must increase by 1, so age must increase by $1/0.144 = 6.94$ years.

- $\log_2(\text{rate-ratio})$ between men and women is $-15.293 - (-16.052) = 0.759$, so the rate ratio is $2^{0.634} = 1.69$.

- In order to advance 0.759 in $\log_2$-mortality a woman must advance $0.759/0.144 = 5.27$ years in age. Thus men have the same mortality as women that are 5.27 years older.

At 40 years, the survival for men is 0.96582. So we want to find the age at which the survival is half of this, i.e. 0.48291.

Survival at 76 is 0.51255 and at 77 0.47983.

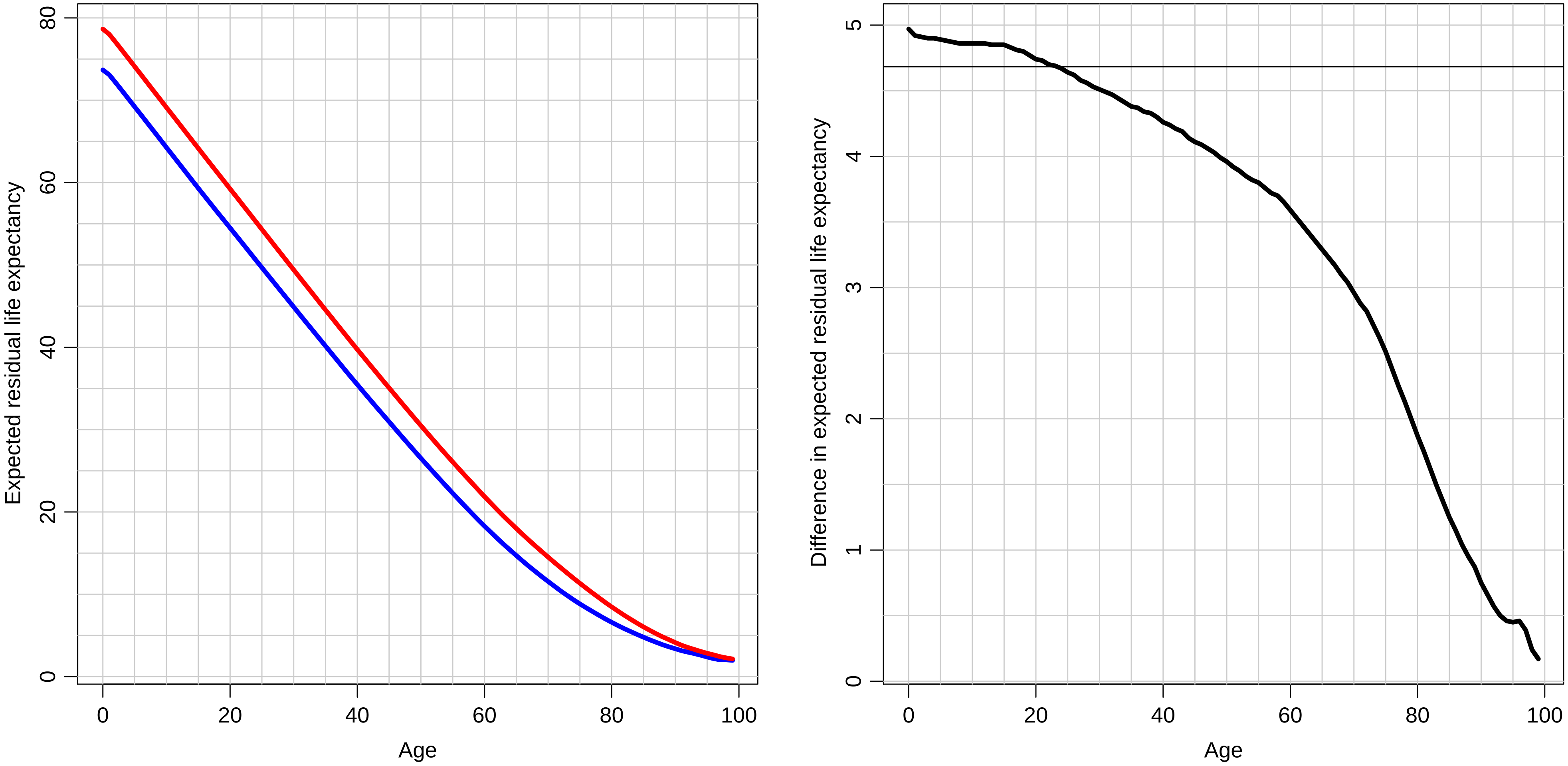
The survival probability for the age-claas 76 is $0.47983/0.51255 = 0.93616 = \mathrm{e}^{-\lambda 1}$ so the mortality rate is $-\log(0.93616) = 0.06597$.

If it takes t to reduce 51255 to 48291 we have:

$$48291/51255 = 0.94217 = \mathrm{e}^{-\lambda t} = \mathrm{e}^{-0.06597t}$$

then $t = -\log(0.94217)/0.06597 = 0.90298$, so the age sought is 76.903.

Expected residual life time



Survival analysis and medical demography

Likelihoods for rates

27 March 2003

Bendix Carstensen

Definition of likelihood

Likelihood is the probability of the observed data given the probability model which gave rise to these data.

It is used to compare candidate values for the parameters of the model; the greater the probability of the observed data, the more *likely* the parameter value.

Consider 10 persons follwed for the **same** period of time:

The **data**: F F S F S S S F S S

The **model**: $\mathrm{P}\left\{\mathsf{F}\right\} = \pi, \quad \mathrm{P}\left\{\mathsf{S}\right\} = 1 - \pi$

Contribution to the **likelihood** is:

- $\times\pi$ for a subject who fails
- $\times(1-\pi)$ for a subject who survives

Total likelihood for π is:

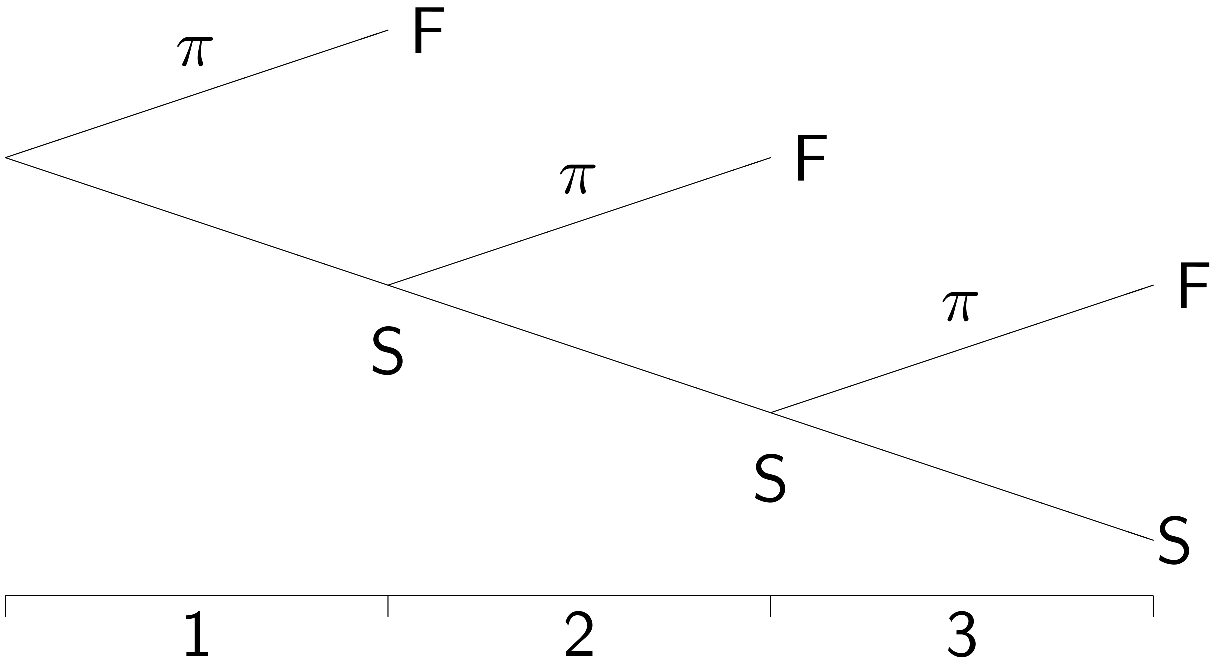
$$\pi^4(1-\pi)^6$$

Function of model [parameter(s)] and data.

Consecutive time bands

- Assume risk does not vary with time:

- Subdivide time into *bands*, and model as a sequence of consecutive Bernoulli trials:



- They are not independent, but the likelihood contribution is a product, for example:

$$\begin{aligned}(1-\pi)^2\pi &= \mathrm{P}\left\{\text{S 1st band}\right\}\\ &\times \mathrm{P}\left\{\text{S 2nd band}\middle|\text{ alive at start of 2nd}\right\}\\ &\times \mathrm{P}\left\{\text{S 3rd band}\middle|\text{ alive at start of 3rd}\right\}\end{aligned}$$

- Observations of one subject through one band behave as if they were independent “atoms” of data.

- Break up your data into little pieces of follow-up for each person and base inference on that.

Estimating a rate heuristically

Total follow-up time Y years, with a total of D failures.

Divide time into clicks of length h years. The total number of clicks is then $N = Y/h$, and D of them end in failure.

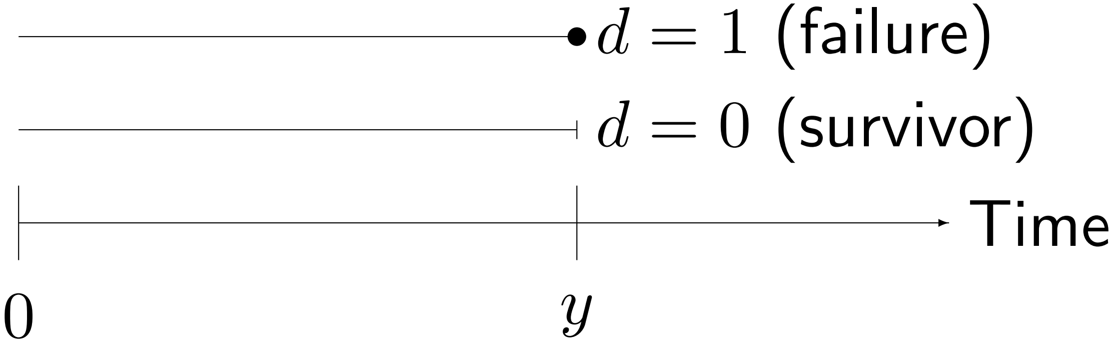
The estimated value of π, the probability of failure during any one click, is D/N.

The estimated value of λ, **the rate**, the probability of failure **per unit time** is:

$$\hat{\lambda} = \frac{\pi}{h} = \frac{D}{Nh} = \frac{D}{Y}$$

Likelihood for a constant rate

Follow-up of 2 subjects:



Probability of **surviving** y years is $\exp(-\lambda y)$ so the survivor contributes $-\lambda y$ to the log likelihood.

Probability of failure in the last click of length h is $\pi = \lambda h$, so the contribution for the failure is

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

The last term does not involve λ, so is irrelevant.

All subjects contribute $-\lambda y$ to the log-likelihhod.
Failures contribute an additional $\log(\lambda)$

Adding these over a group of persons gives

$$D\log(\lambda) - \lambda Y,$$

Y is the total follow-up time, and
D the total number of failures.

The log-likelihood is maximal for:

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

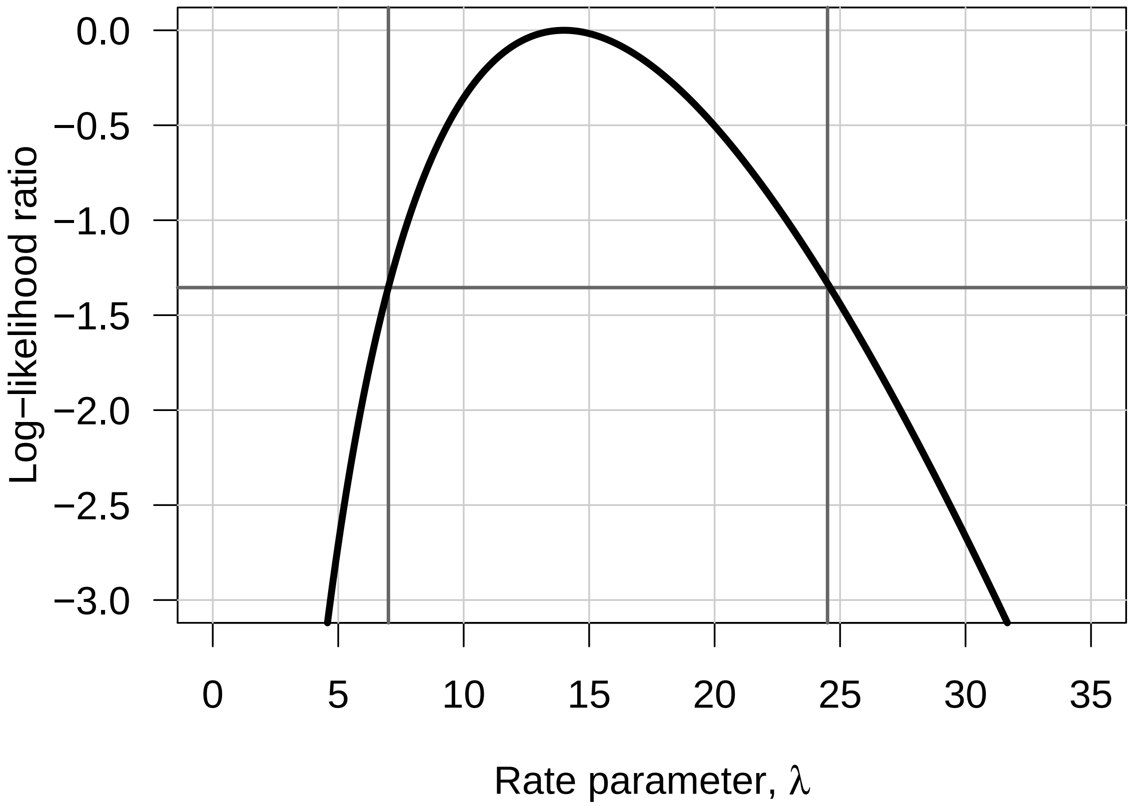
Example: 7 failures in 500 person–years

The log likelihood is

$$7\log(\lambda) - 500\lambda$$

The maximum value of the log likelihood occurs at

$\lambda = 7/500 = 0.014$ per person-year.



A 90% c.i. for λ may be found by reading off the values of λ at which the log likelihood ratio has reduced to -1.353, i.e. $(7.0; 24.6) \times 10^{-3}$ per person-year. $(1.353 = \chi^2_{0.90}(1)/2)$

Empirical rates

The epidemiological definition of an empirical rate is:

$$\frac{\text{No. events}}{\text{risk-time}}$$

Small time-intervals for single individuals $\Rightarrow$ almost never any events, so empirical rates will either be 0, or very large.

For statistical modelling we define an empirical rate as a **pair:** (d, y), the number of events ($d \in \{0, 1\}$) and the length of the interval (y).

The log-likelihood contribution from an empirical rate is $d \log(\lambda) - \lambda y$

The modified life-table estimator

If deaths and censorings occur uniformly spaced over an interval, the total risk time of N individuals with D deaths and L censorings over an interval of length y is $(N - D/2 - L/2)y$ thus the rate is estimated by:

$$\hat{\lambda} = \frac{D}{(N - D/2 - L/2)y}$$

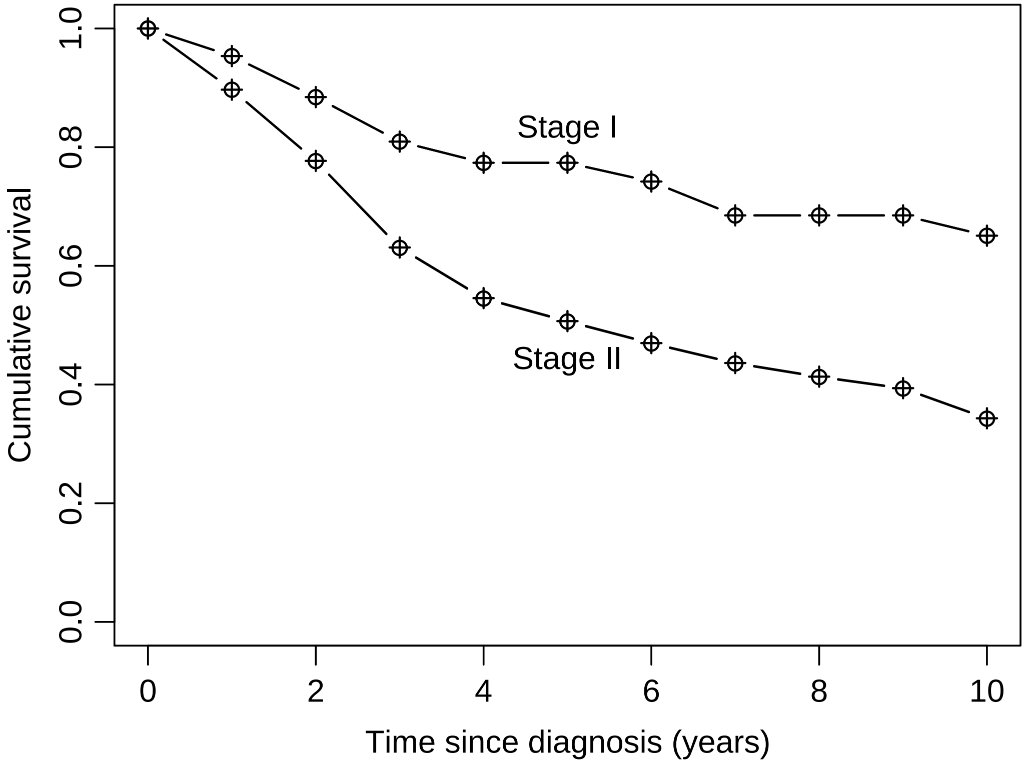
and thus the cumulative rate over the interval by

$$\hat{\Lambda} = \frac{Dy}{(N - D/2 - L/2)y} = \frac{D}{N - D/2 - L/2}$$

The survival probability for the interval is then estimated by:

$$\exp\left(-\frac{D}{N - D/2 - L/2}\right)$$

Multiplied together to a cumulative survival function, the resulting estimator is called the **modified lifetable estimator** (crosses)



Poisson likelihood

The log-likelihood for a rate, based on D deaths during Y risk time is:

$$D\log(\lambda) - \lambda Y$$

The log-likelihood for the parameter λ in a Poisson model with mean λY, based on an observation of D is:

$$D\log(\lambda Y) - \lambda Y = D\log(\lambda) + D\log(Y) - \lambda Y$$

Follow-up study observation: (D, Y)
Poisson observation: D, Y known constant.

Different models, same likelihood.

Precision of maximum likelihood estimates

General principle of likelihood theory for a parameter θ:

$$\mathrm{var}(\hat\theta) \approx -\left(\mathrm{D}^2_\theta\big[\ell(\theta)\big]\right)^{-1}\bigg|_{\theta=\hat\theta}$$

i.e. the variance of an estimate is minus the inverse of the 2nd derivative of the log-likelihood evaluated at the estimate (i.e. the maximum).

Precision of the rate estimate

The Poisson log likelihood, as a function of $\theta = \log(\lambda)$:

$$\ell(\theta) = D\theta - \mathrm{e}^{\theta} Y \qquad \mathrm{D}^2_{\theta}(D\theta - \mathrm{e}^{\theta} Y) = -\mathrm{e}^{\theta} Y = -\lambda Y$$

Inserting $\lambda = \hat{\lambda} = D/Y$ gives $-D$, so:

$$\mathrm{s.e.}(\log(\hat{\lambda})) = 1/\sqrt{D}$$

In practical terms, it means the confidence intervals has the form

$$\log(\hat{\lambda}) \pm 2/\sqrt{D} \qquad \Leftrightarrow \qquad \lambda \overset{\times}{\div} \mathrm{e}^{2/\sqrt{D}}$$

Late entries

The basic idea is to split time into intervals sufficiently small for the assumption of constant rates to hold.

The likelihood contribution for a small interval is **only** conditional on being alive (at risk of failure).

Thus late entries do not represent any problem for constructing the likelihood.

Late entries (left truncation) is a design / data acquisition problem.

Time-varying rates

If we assume rates vary by time, the log-likelihood just becomes a sum of terms with different λs for each interval with different rates:

$$\ell(\lambda_1, \lambda_2) = D_1 \log(\lambda_1) - \lambda_1 Y_1 + D_2 \log(\lambda_2) - \lambda_2 Y_2$$

In fact there is no need to aggregate the single empirical rates to (D, Y), even if they have the same rate.

Keep the little “atoms” of observation, and just assign the relevant time-period to each in the estimation.

Observational unit in survival studies

The central unit of observation is (d, y), small pieces of follow-up (many from each person).

Each one has covariates $(x_1, \dots, x_p)$ associated with it, e.g.:

- Sex
- Genotype
- Current age
- Parity
- Calendar time
- Blood pressure

. . .

Is time a response variable?

Yes and no.

Time comes in two guises:

1. The **risk time**, the small y-contributions to the empirical rates.

   This is (part of) the response variable.

2. The **time scale**, which may be time since entry, current age, time since hire, ...

   This is a covariate, whose effect we may take into the model.

Poisson-modelling

The Poisson-likelihood can accomodate the effect of covariates through a **regression model**:

$$\lambda(x_1, \ldots, x_p) = f(x_1, \ldots, x_p)$$

Thus, the problem remaining is a purely **technical** problem: Specification of a sensible form for f, that is

1. Interpretable — we should be able to formulate conclusions in plain language based on estimates from the model.

2. Practicable — we should be able to estimate in the model in finite time (i.e. use standard software).

The most practicable Poisson model for rates is the multiplicative:

$$\log\big(\lambda(x_1,\dots,x_p)\big) = \mu + \beta_1 x_1 + \cdots + \beta_p x_p = \eta, \text{ say.}$$

The log-likelihood then becomes:

$$\ell(\eta) = \sum_{\text{all } (d,y)\text{-pairs}} (d\eta - \mathrm{e}^{\eta} y)$$

This log-likelihood can be maximized by programs doing Poisson-likelihood estimation (generalized linear models):

- Pretend d is a Poisson observation.
- Specify the Poisson mean as:

$$\lambda y = \mathrm{e}^{\eta} y = e^{\eta + \log(y)}$$

The latter requires that the variable $\log(y)$ be included with the linear predictor with a fixed regression coefficient of 1.

This is in the GLM-jargon called an **offset**-variable.

Example: Diet data

A dietary survey from England. Used extensively as example in the book by Clayton & Hills[2]

Variables (amongst others):

`d` - indicator of coronary heart disease
`y` - years in the study
`height` - height in cm
`eng3` - energy intake group (1500/2500/3000)

Available as Stata dataset on the course homepage.

[2] David Clayton & Michael Hills: Statistical models in epidemiology. Oxford University Press, 1993

```
. use diet
. xi: glm d i.eng3 height, family(poisson) lnoffset(y)

i.eng3            _Ieng3_1-3        (_Ieng3_1 for eng3==1500 omitted)
...
Generalized linear models                  No. of obs      =       332
Optimization     : ML: Newton-Raphson      Residual df     =       328
                                           Scale param     =         1
Deviance         =  233.1298826            (1/df) Deviance =  .7107618
Pearson          =   1101.92786            (1/df) Pearson  =  3.359536
Variance function: V(u) = u                [Poisson]
Link function    : g(u) = ln(u)            [Log]
Standard errors  : OIM
Log likelihood   = -161.5649413            AIC             =  .9973792
BIC              =  209.9093428
------------------------------------------------------------------
       d |      Coef.  Std. Err.      z    P>|z|   [95% Cf. Interval]
---------+--------------------------------------------------------
_Ieng3_2 |   -.219921  .3434119   -0.64   0.522  -.8929959   .453154
_Ieng3_3 |  -.8956376  .4529728   -1.98   0.048  -1.783448 -.0078271
  height |  -.0802306  .0221933   -3.62   0.000  -.1237286 -.0367326
   _cons |   9.521837  3.719843    2.56   0.010   2.231078   16.8126
       y | (exposure)
------------------------------------------------------------------
```

```
> library( foreign )
> diet <- read.dta( "../data/diet.dta" )
> m1 <- glm( d ~ factor( eng3 ) + height + offset( log( y ) ),
+            family=poisson, data=diet, eps=1e-07 )
> summary( m1 )

Call:
glm(formula = d ~ factor(eng3) + height + offset(log(y)),
    family = poisson, data = diet, eps = 1e-07)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       9.52184    3.71968   2.560   0.0105
factor(eng3)2500 -0.21992    0.34339  -0.640   0.5219
factor(eng3)3000 -0.89564    0.45296  -1.977   0.0480
height           -0.08023    0.02219  -3.615   0.0003

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 255.01  on 331  degrees of freedom
Residual deviance: 233.13  on 328  degrees of freedom
AIC: 331.13

Number of Fisher Scoring iterations: 6
```

```
> names( summary( m1 ) )
 [1] "call"           "terms"      "family"         "deviance"
 [5] "aic"            "contrasts"  "df.residual"    "null.deviance"
 [9] "df.null"        "iter"       "deviance.resid" "aic"
[13] "coefficients"   "dispersion" "df"             "cov.unscaled"
[17] "cov.scaled"
> cf <- m1$coef
> cv <- summary( m1 )$cov.u
> ctr <- c(0,1,-1,0)
> ctr %*% cf
          [,1]
[1,] 0.6757167
> sqrt( t(ctr) %*% cv %*% ctr )
            [,1]
  [1,] 0.4147966
```

Survival analysis and medical demography

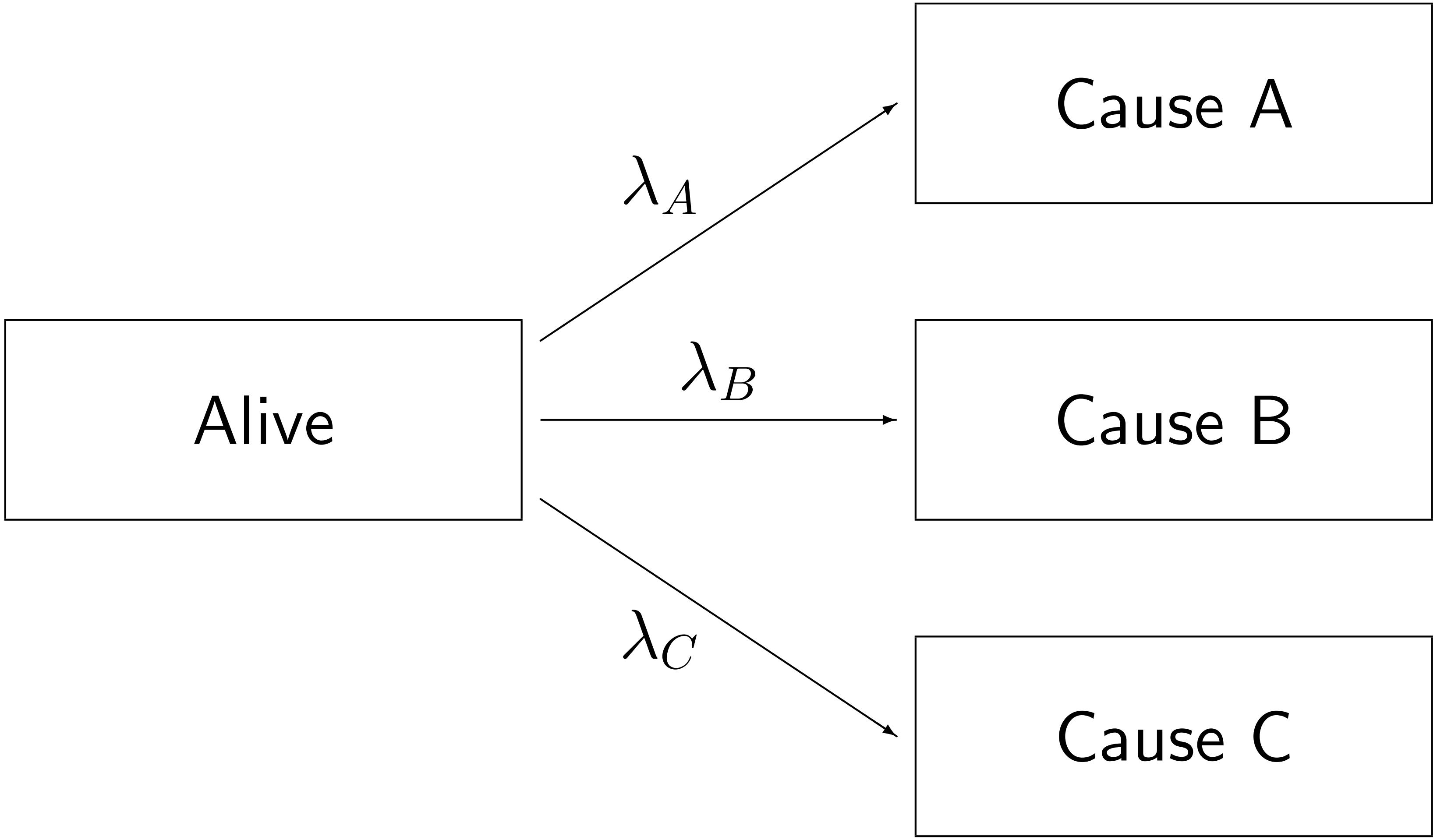
Competing risks

27 March 2003

Bendix Carstensen

Competing risks

You may die from more than one cause:



Cause-specific intensities

$$\begin{aligned}
\lambda_A(t) &= \lim_{h \to 0} \frac{\mathrm{P}\left\{\text{death from cause A in } (t, t+h] \mid \text{alive at } t\right\}}{h} \\
\lambda_B(t) &= \lim_{h \to 0} \frac{\mathrm{P}\left\{\text{death from cause B in } (t, t+h] \mid \text{alive at } t\right\}}{h} \\
\lambda_C(t) &= \lim_{h \to 0} \frac{\mathrm{P}\left\{\text{death from cause C in } (t, t+h] \mid \text{alive at } t\right\}}{h}
\end{aligned}$$

Total mortality rate:

$$\lambda_{\mathsf{Total}}(t) = \lim_{h \to 0} \frac{\mathrm{P}\left\{\text{death from any cause in } (t, t+h] \mid \text{alive at } t\right\}}{h}$$

$$\mathrm{P}\left\{\text{death from any cause in } (t, t+h] \mid \text{alive at } t\right\}$$

$$\begin{aligned} = \; & \mathrm{P}\left\{\text{death from cause A in } (t, t+h] \mid \text{alive at } t\right\} + \\ & \mathrm{P}\left\{\text{death from cause B in } (t, t+h] \mid \text{alive at } t\right\} + \\ & \mathrm{P}\left\{\text{death from cause C in } (t, t+h] \mid \text{alive at } t\right\} \end{aligned}$$

$$\Longrightarrow \qquad \lambda_{\text{Total}}(t) = \lambda_A(t) + \lambda_B(t) + \lambda_C(t)$$

Intensities are additive,
if they all refer to the same risk set, in this case "Alive".

Likelihood for competing risks

Data:

Y person years in “Alive”

D_A deaths from cause A

D_B deaths from cause B.

D_C deaths from cause C.

Assume for simplicity that rates are constant.

A survivor contributes to the log-likelohood:

$$\log(\mathrm{P}\left\{\text{Survival for a time of } y\right\}) = -(\lambda_A + \lambda_B + \lambda_C)y$$

A death from cause A contributes an additional $\log(\lambda_A)$, etc.

The total log-likelihood is then:

$$\begin{aligned}\ell(\lambda_A, \lambda_B, \lambda_C) &= D_A \log(\lambda_A) + D_B \log(\lambda_B) + D_C \log(\lambda_C) \\ &\quad -(\lambda_A + \lambda_B + \lambda_C)Y \\ &= [D_A \log(\lambda_A) - \lambda_A Y] + \\ &\quad [D_B \log(\lambda_B) - \lambda_B Y] + \\ &\quad [D_C \log(\lambda_C) - \lambda_C Y]\end{aligned}$$

The log-likelihood is made up of three contributions:
One for cause A, one for cause B and one for cause C.

Deaths are the cause-specific deaths, but the person-years are the same in all contributions.

Time varying rates:

This is the same business as with one rate; use time intervals sufficiently small to justify an assumtion of constant rate (intensity).

Practical implications

Analysis of the individual cause-specific rates effectively uses the same dataset for all causes, because the person-years are the same.

Thus the little “atoms” of data (the empirical rates (d, y) from each individual) will be the same for all analyses except for those where deaths occur.

Analysis of cause A: Contributions $(1, y)$ only for those intervals where a cause A death occurs.
Intervals with cause B or C deaths contribute only $(0, y)$
— for the analysis of cause A treated as censorings.

Competing risks are analysed by considering the cause specific rates separately. We shall return to the possibility of modelling the rates jointly.

Merely a technical issue.

Assumptions in competing risks

“Classical” way of looking at survival data:
description of the distribution of time to death.

For competing risks that would require three variables: T_A, T_B and T_C, representing times to death from each of the three causes.
But at most one of these is observed.

Often it is stated that these must be assumed independent in order to make the likelihoods machinery work.
If they are independent, it works, but is is not necessary.

An excellent account of these problems (and a counter example to the independence of survival times) are given in:

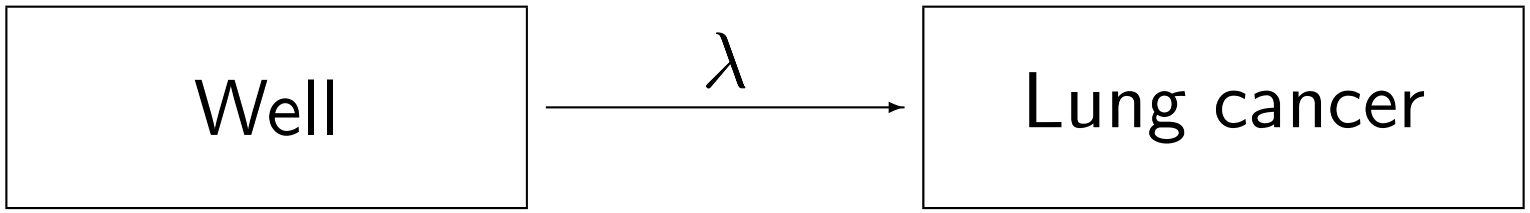
PK Andersen, SZ Abildstrøm & S Rosthøj:
Competing risks as a multistate model,
Research report 2001/12,
Department of Biostatistics, University of Copenhagen
Available as `.ps`-file at:
`http://www.biostat.ku.dk/publ-e.htm`

The paper also contains a guide for the practitioner.

Competing risk problems

The problems with competing risk models comes when estimated intensities (rates) are used to produce probability statements.

Classical set-up in cancer-registries:



Well

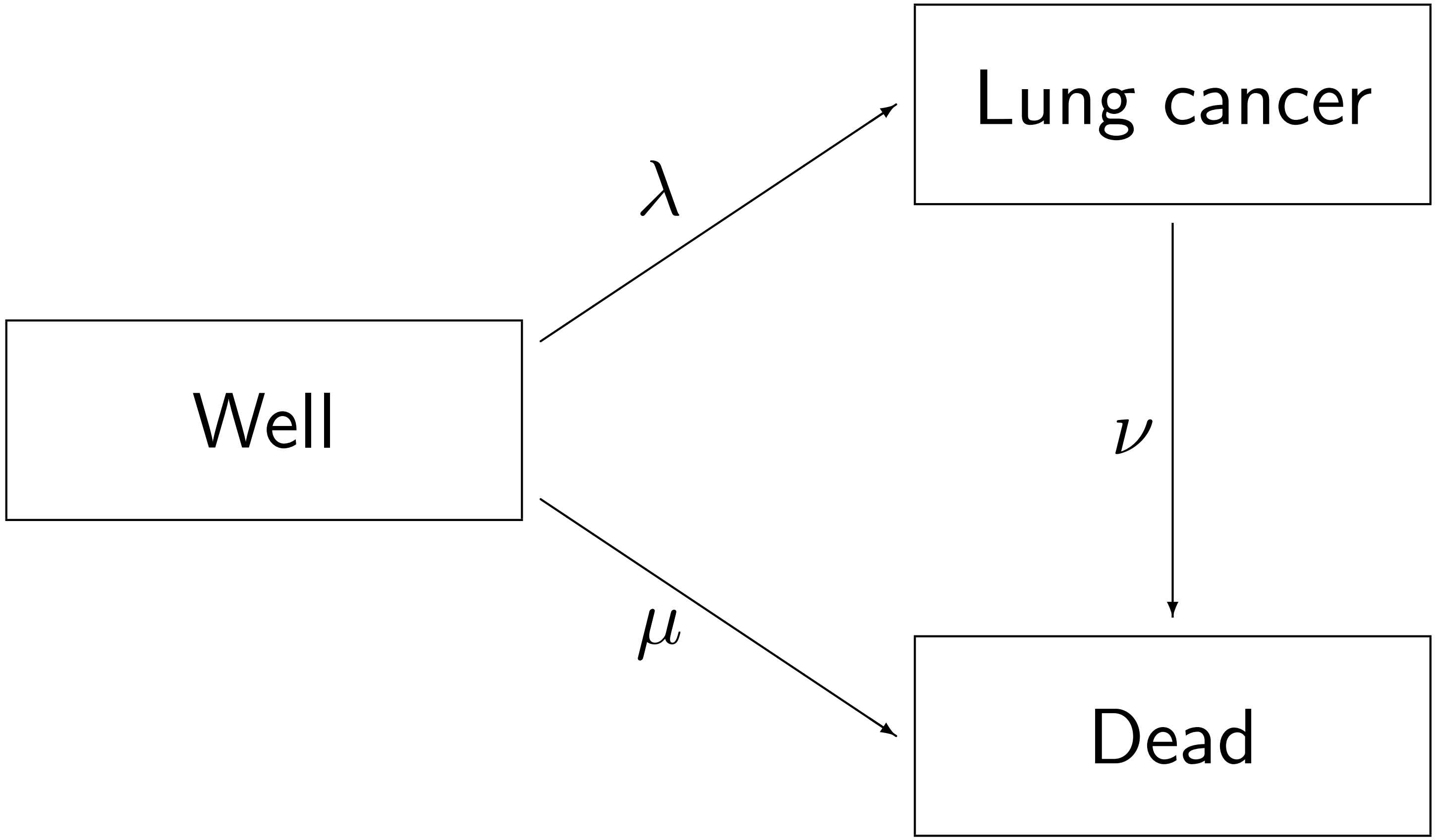
λ

Lung cancer

$$\mathrm{P}\left\{\text{Lung cancer before age 75}\right\} = 1 - \mathrm{e}^{-\Lambda(75)}$$

This is not quite right.

How the world really looks



Illness-death model. Little boxes with arrows.

How many get lung cancer before age a?

$$\mathrm{P}\left\{\text{Lung cancer before age 75}\right\} \neq 1 - \mathrm{e}^{-\Lambda(75)}$$

does not take the possibility of death prior to lung cancer into account.

$1 - \mathrm{e}^{-\Lambda(75)}$ often stated as the probaility of lung cancer before age 75, assuming all other acuses of death absent.

Lung cancer rates are however observed in a mortal population.

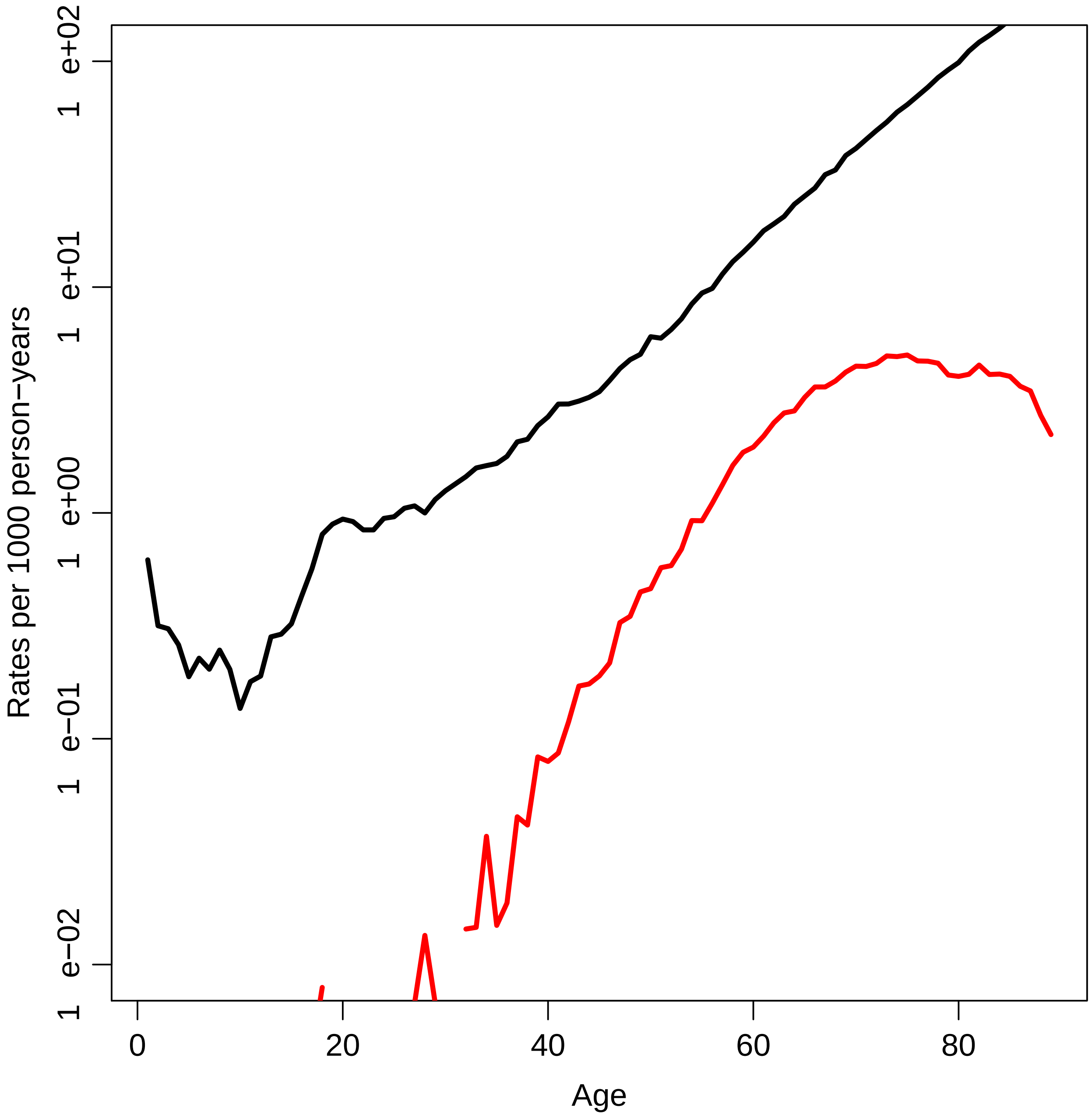
If all other causes of death were absent, this would assume that lung cancer rates remained the same.

$$\begin{aligned}
&\mathrm{P}\left\{\text{Lung cancer before age } a\right\}\\
&= \int_0^a \mathrm{P}\left\{\text{Lung cancer at age } u\right\}\mathrm{d}u\\
&= \int_0^a \mathrm{P}\left\{\text{Lung cancer in age } (u, u+\mathrm{d}u] \mid \text{alive at } u\right\}\\
&\qquad \times \mathrm{P}\left\{\text{alive at } u\right\}\mathrm{d}u\\
&= \int_0^a \lambda(u)\exp\left(-\int_0^u \mu(s)+\lambda(s)\mathrm{d}s\right)\mathrm{d}u
\end{aligned}$$

Probability of lungcancer

The rates are easily plotted for inspection in R:

```
matplot( age, 1000*cbind( D/Y, lung/Y ),
         log="y", type="l", lty=1, lwd=3,
         ylim=c(0.01,100), xlab="Age",
         ylab="Rates per 1000 person-years" )
```


Rates per 1000 person−years
1 e+02
1 e+01
1 e+00
1 e−01
1 e−02
0
20
40
60
80
Age

The probablility that a person contacts lung cancer before age a is (cf. the lecture notes):

$$\int_0^a \lambda(u) \exp\left(-\int_0^u \mu(s)+\lambda(s)\mathrm{d}s\right)\mathrm{d}u$$

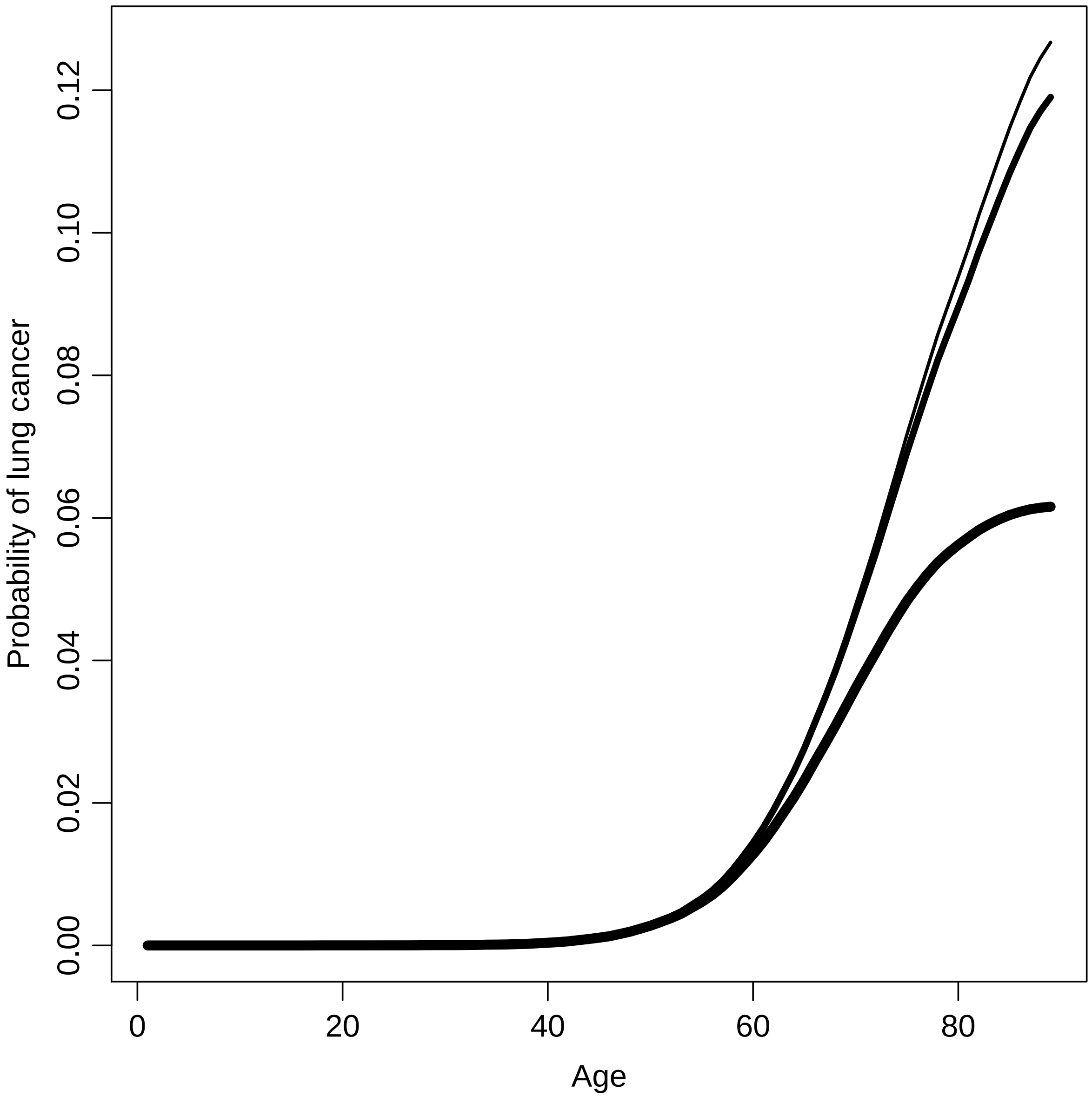
$$= \int_0^a \lambda(u) \exp\left(-\big(\mathrm{M}(u)+\Lambda(u)\big)\right)\mathrm{d}u$$

$\mathrm{M}(u)$ is the cumulative mortality rate

$\Lambda(u)$ is the cumulative lung cancer incidence rate.

R-commands needed to do the calculations:

```
cr.death <- cumsum( D/Y )
cr.lung <- cumsum( lung/Y )
p.simple <- 1 - exp( -cr.lung )
p.lung <- cumsum( lung/Y *
                  exp( -(cr.death+cr.lung) ) )
matplot( age, cbind( cr.lung, p.simple, p.lung ),
         type="l", lty=1, lwd=2*c(1,2,3),
         col="black", xlab="Age",
         ylab="Probability of lung cancer" )
```


Probability of lung cancer
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
20
40
60
80
Age

The assumption behind the calculation and the statement "6% of Danish males will get lung cancer" is that the lung cancer rates and the mortality rates in the file applies to a cohort of men. But they are cross-sectional rates, so the assumption is one of steady state of mortality rates (which is dubious) and lung cancer incidence rates (which is appaling).

Survival analysis and medical demography

Cohort studies and multiple timescales

3 April 2003

Bendix Carstensen

Likelihood for a rate

Empirical rate: (d, y) — (outcome,risk time) for a small piece of follow up.

Small enough to warrant assumption of constant rate.

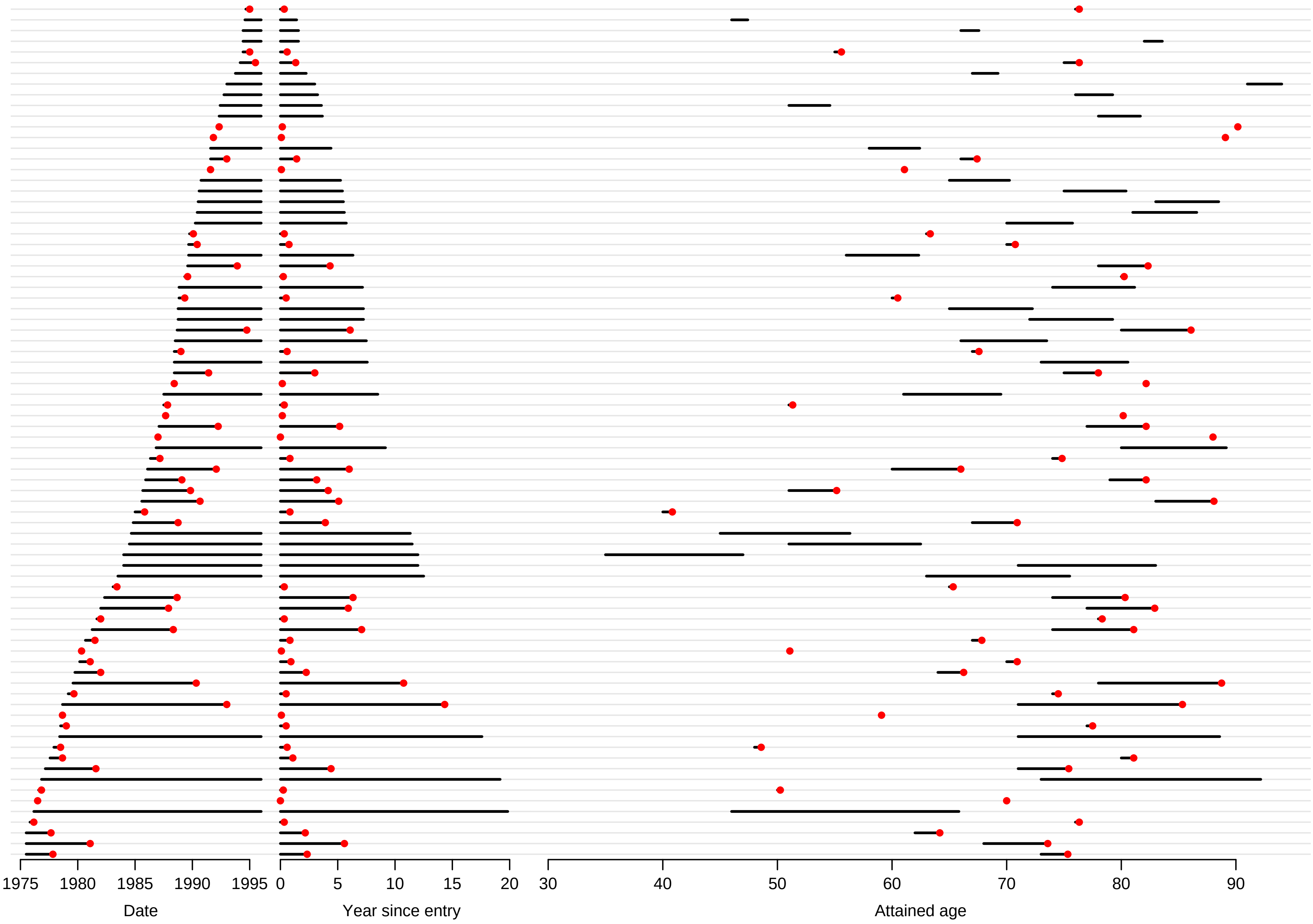
Log-likelihood:

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

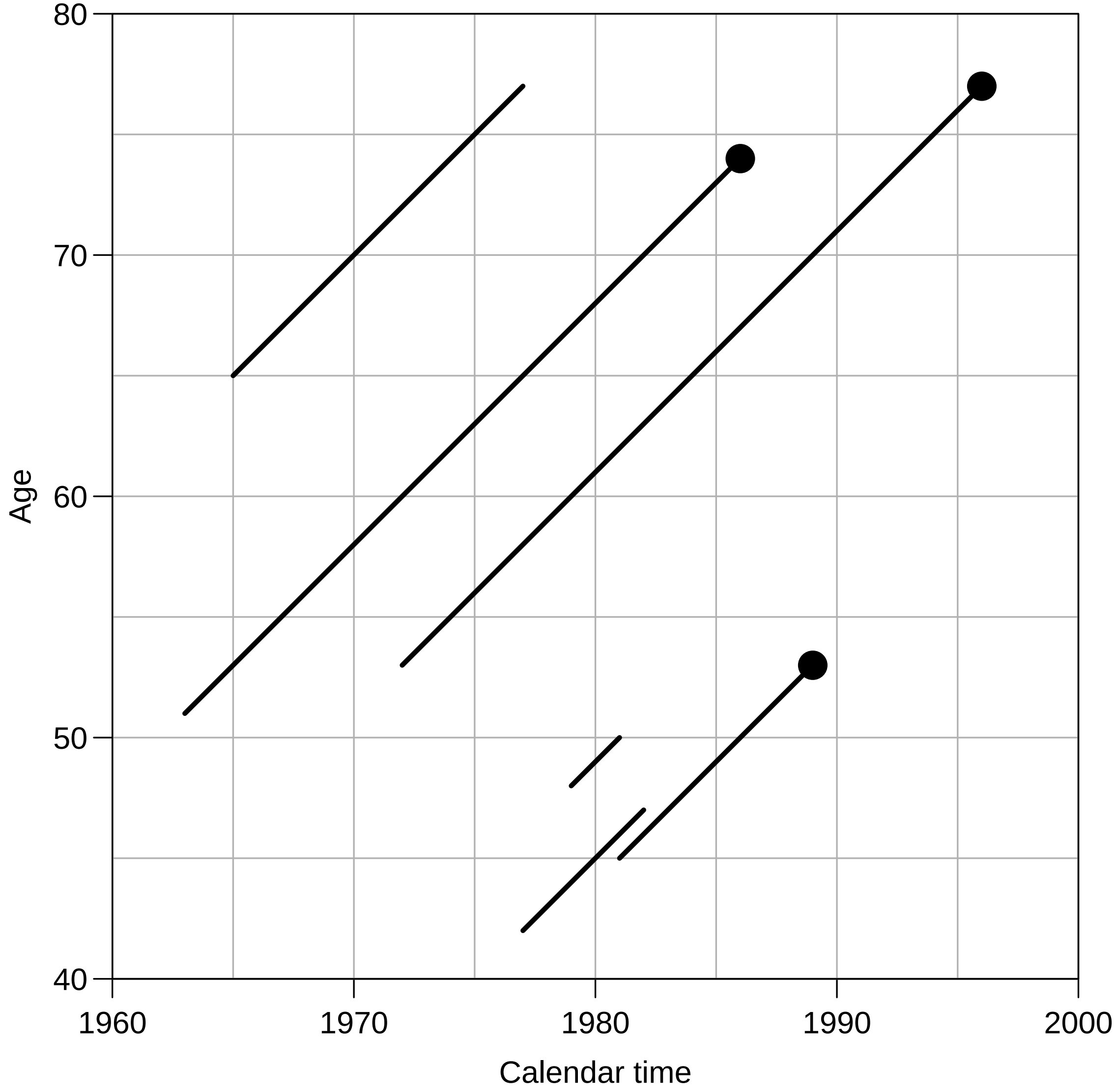
Each empirical rate has its own set of covariates attaced.

Fixed: Sex, Genotype

Varying: Parity, Medical history,
Calendar time, Time since entry, Attained age.


1975
1980
1985
1990
1995
Date
0
5
10
15
20
Year since entry
30
40
50
60
70
80
90
Attained age

Lexis-diagram



Named after the German statistician and economist **William Lexis** (1837–1914), who in the book:

“Einleitung in die Theorie der Bevölkerungsstatistik” (Karl J. Trübner, Strassburg, 1875)

developed the ideas of simultaneously classifying persons by age and calendar time, using this type of diagram.

Several timescales

Modelling the effect of several timescales:
Follow-up time must be subdivided by all of them:

FROM the registration of:
Entry, Exit times and Failure status
and the definition of the scales by:
Origin (when is time 0), Scale (units) and Cutpoints (subdivisions)

TO the set of empirical rates (d, y)
each with all covariates attached to it: $\big((d, y), \mathbf{z}\big)$

Cohort studies — medical demography

- Long follow up time. Constant rates not applicable.
- Several timescales are of interest:
  - Current age.
  - Time on study.
  - Time since exposure inception / cessation.
  - Cumulative exposure.
  - Calendar time.
- Delayed entry times due to data collection.

Example: Welsh Nickel Refinery study:

R. Doll, L.G. Morgan & F.E. Speizer:
Cancers of the lung and nasal sinuses in nickel workers.
Br.J.Cancer, **24**, 1970:

“Men employed in a nickel refinery in South Wales were investigated to determine whether the risk of developing carcinoma of the bronchi and nasal sinuses, which had been associated with the refining of nickel, are still present.

The data obtained were also used to compare the effect of age at exposure on susceptibility to cancer induction and to determine the rate of change of mortality after exposure to a carcinogenic agent has ceased”

Welsh nickel refinery workers

Data in Stata:

```
. list in 1/6, nodis

 id  icd  expos       dob       doe       dox     do1st
  3    0      5  1889.019  1934.246      1982    1906.5
  4  162      5  1885.978  1934.246  1949.249  1909.164
  6  163     10  1881.255  1934.246  1935.419    1906.5
  8  527      9   1886.34  1934.246  1956.019   1911.06
  9  150      0    1879.5  1934.246  1956.344  1909.458
 10  163      2  1889.915  1934.246  1952.456  1911.203
```

Follow-up started 1 Apr 1934 for all persons.

```
 id  icd  expos       dob       doe       dox     do1st
  3    0      5  1889.019  1934.246      1982    1906.5
  4  162      5  1885.978  1934.246  1949.249  1909.164
```

- Person no. 3, born 7 Jan 1889, and hired 3 Jul 1906 (when aged 17.48 years), he was 45.22 years at start of follow-up. He ceased to be at risk 31 Dec 1981, when he was 92.98 years old and still alive (`icd=0`).

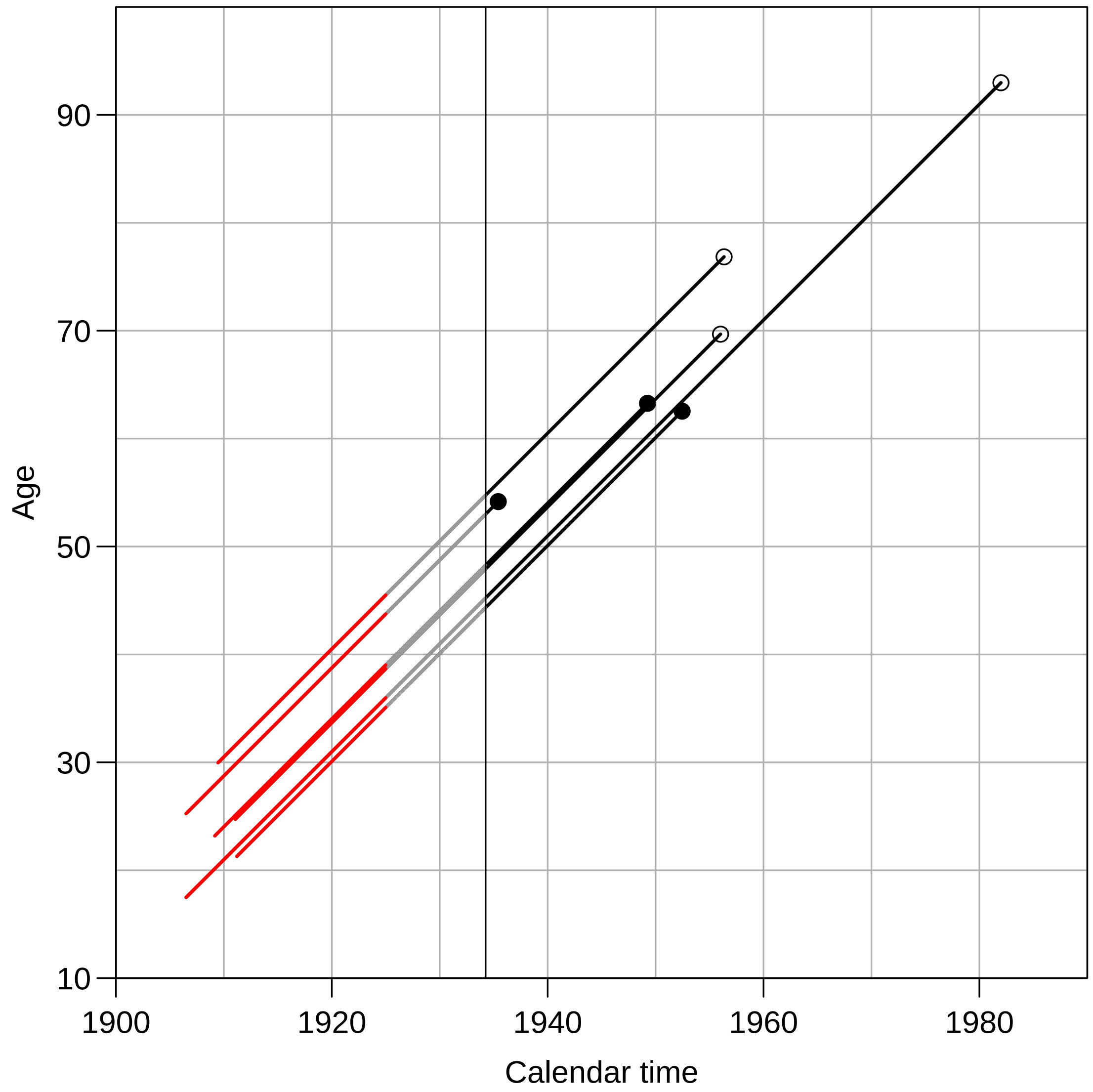
- Person no. 4, born 23 Dec 1885, and hired 1 Mar 1909 (when aged 23.19 years), he was 45.27 years at start of follow-up. He ceased to be at risk 1 Apr 1949, when he was 63.27 years old, because he died from lungcancer (`icd=162`).

Follow-up starts at 1 Apr 1934.

Some of the variables are constant throughout follow-up (e.g. sex), some vary deterministically (e.g. age and date) others require active registration to determine the values at a point in time.

In the Welsh nickel refinery the manufacturing process was changed in 1925, and relevant exposure is considered absent after this date. Cumulative exposure (`expos`) is therefore constant after 1925, and hence also throughout the follow-up.

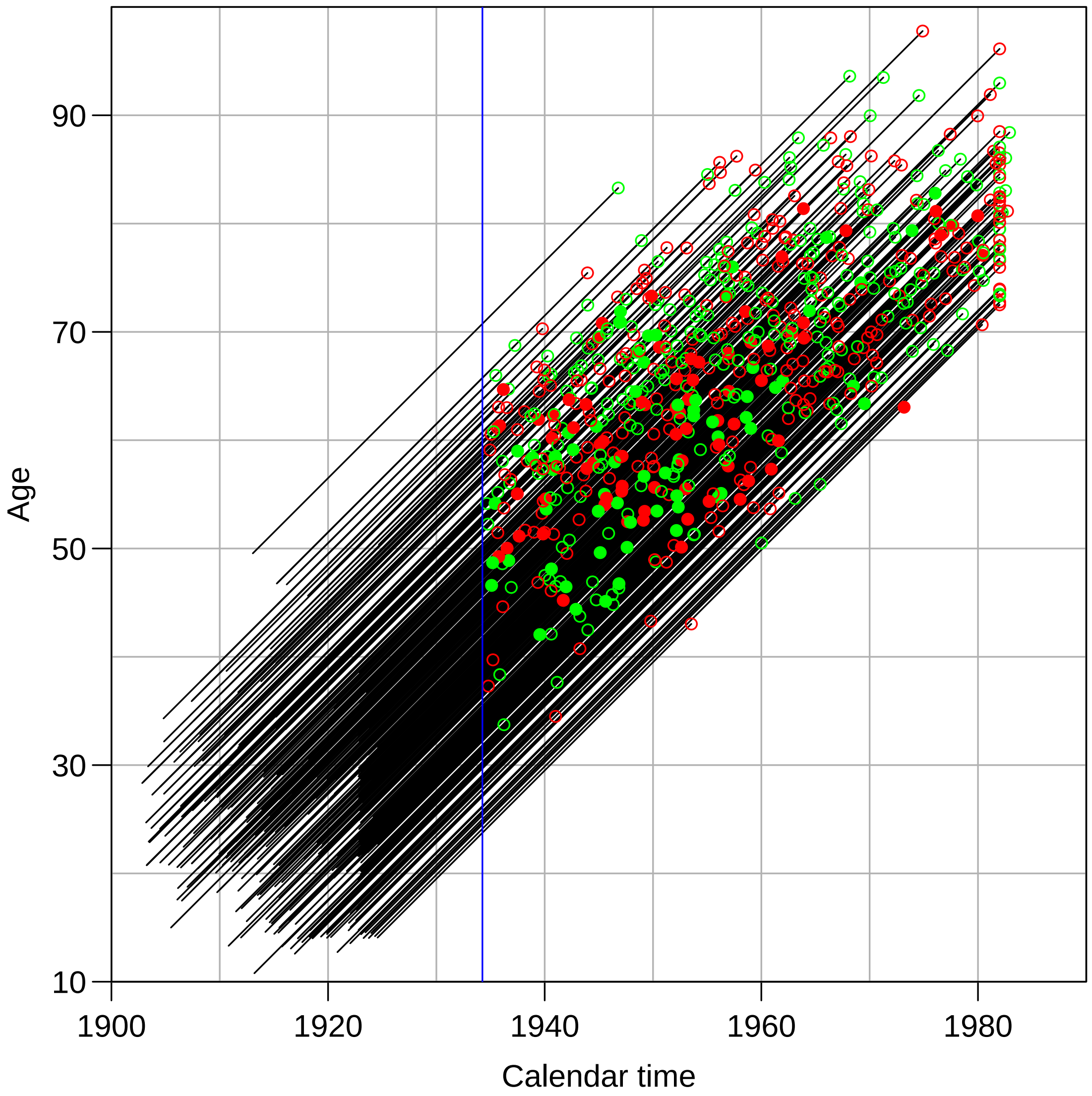
Follow-up for 6 persons: Lexis diagram



Red is time of exposure, black is the follow-up.

Persons died before 1 Apr 1934 are not included in the study.

Time



- Same scale for all individuals.
- Origin is different, whether we use age, time since hire, or time since exposure cessation.

Red: Death from lungcancer.

Green: Death from other causes or censoring.

Time scale exercise

- Name timescales that may be used in this study.
- List relevant covariates.
- Which covariates are constant and which vary with the follow-up of the individuals?

Cumulative exposure

Time in a job

Cumulative radiation dose

Cumulative no. cigarettes

Time employed in nickel refinery:

- Does not necessarily increase all the time.
- Does not increase in the same way for all individuals.

Cohort with 3 persons:

```
Id     Bdate     Entry      eXit St
 1 14/07/52 04/08/65 27/06/97  1
 2 01/04/54 08/09/72 23/05/95  0
 3 10/06/87 23/12/91 24/07/98  1
```

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

Splitting the follow up

	subj. 1	subj. 2	subj. 3
Age at **E**ntry:	13.06	18.44	4.54
Age at e**X**it:	44.95	41.14	11.12
Status at exit:	Dead	Alive	Dead
Y	31.89	22.70	6.58
D	1	0	1

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

Software for splitting the records

R: The function `Lexis` by David Clayton.

Stata: The function `stsplit`.
Originally written by David Clayton & Michael Hills, under the name `stlexis`.

SAS: The macro `%Lexis` by Bendix Carstensen

R-function and SAS-macro, as well as example programs `xLexis.xxx`, $\texttt{xxx} \in \{\texttt{sas}, \texttt{R}, \texttt{do}\}$ are available at `http://www.biostat.ku.dk/~bxc/Lexis`.

Splitting the follow-up with R:

Split by 10-year calendar time and 5-year age bands:

```
> source( "Lexis.R" )
...
> xcoh
  id      birth      entry       exit fail       bt      ent       ex
1  A 14/07/1952 04/08/1965 27/06/1997    1 1952.531 1965.588 1997.484
2  B 01/04/1954 08/09/1972 23/05/1995    0 1954.245 1972.685 1995.387
3  C 10/06/1987 23/12/1991 24/07/1998    1 1987.436 1991.973 1998.557
> x2 <-
+ Lexis( entry = ent,
+         exit = ex,
+         fail = fail,
+        scale = 1,
+       origin = list( per=0,                 age=bt          ),
+       breaks = list( per=seq(1900,2000,10), age=seq(0,80,5) ),
+      include = list( bt, en, ex, id ),
+         data = xcoh )
```

```
> x2
   Expand    Entry     Exit Fail  per age       bt       en       ex id
1       1 1965.588 1967.531    0 1960  10 1952.531 1965.588 1997.484  A
2       1 1967.531 1970.000    0 1960  15 1952.531 1965.588 1997.484  A
3       1 1970.000 1972.531    0 1970  15 1952.531 1965.588 1997.484  A
4       1 1972.531 1977.531    0 1970  20 1952.531 1965.588 1997.484  A
5       1 1977.531 1980.000    0 1970  25 1952.531 1965.588 1997.484  A
6       1 1980.000 1982.531    0 1980  25 1952.531 1965.588 1997.484  A
7       1 1982.531 1987.531    0 1980  30 1952.531 1965.588 1997.484  A
8       1 1987.531 1990.000    0 1980  35 1952.531 1965.588 1997.484  A
9       1 1990.000 1992.531    0 1990  35 1952.531 1965.588 1997.484  A
10      1 1992.531 1997.484    1 1990  40 1952.531 1965.588 1997.484  A
11      2 1972.685 1974.245    0 1970  15 1954.245 1972.685 1995.387  B
12      2 1974.245 1979.245    0 1970  20 1954.245 1972.685 1995.387  B
13      2 1979.245 1980.000    0 1970  25 1954.245 1972.685 1995.387  B
14      2 1980.000 1984.245    0 1980  25 1954.245 1972.685 1995.387  B
15      2 1984.245 1989.245    0 1980  30 1954.245 1972.685 1995.387  B
16      2 1989.245 1990.000    0 1980  35 1954.245 1972.685 1995.387  B
17      2 1990.000 1994.245    0 1990  35 1954.245 1972.685 1995.387  B
18      2 1994.245 1995.387    0 1990  40 1954.245 1972.685 1995.387  B
19      3 1991.973 1992.436    0 1990   0 1987.436 1991.973 1998.557  C
20      3 1992.436 1997.436    0 1990   5 1987.436 1991.973 1998.557  C
21      3 1997.436 1998.557    1 1990  10 1987.436 1991.973 1998.557  C
```

Splitting the follow-up with Stata:

Split by 10 year calendar time and 5-year age-bands:

```
. stset  ex,
        fail( fail )
       entry( ent )
      origin( time d(01jan1900) )
       scale( 365.25 )
          id( id )
. stsplit per, at( 0(10)100 )
. stsplit age, after( bth ) at( 0(5)90 )

id         bth        ent          ex  age  per     _t0      _t _d
 A  14jul1952  04aug1965  14jul1967   10   60  65.588  67.531  0
 A  14jul1952  04aug1965  01jan1970   15   60  67.531  70.000  0
 A  14jul1952  04aug1965  14jul1972   15   70  70.000  72.531  0
 A  14jul1952  04aug1965  14jul1977   20   70  72.531  77.531  0
 A  14jul1952  04aug1965  02jan1980   25   70  77.531  80.000  0
 A  14jul1952  04aug1965  14jul1982   25   80  80.000  82.531  0
 A  14jul1952  04aug1965  14jul1987   30   80  82.531  87.531  0
 A  14jul1952  04aug1965  01jan1990   35   80  87.531  90.000  0
 A  14jul1952  04aug1965  14jul1992   35   90  90.000  92.531  0
```

```
A  14jul1952  04aug1965  27jun1997   40   90  92.531  97.484  1
B  01apr1954  08sep1972  01apr1974   15   70  72.684  74.245  0
B  01apr1954  08sep1972  01apr1979   20   70  74.245  79.245  0
B  01apr1954  08sep1972  02jan1980   25   70  79.245  80.000  0
B  01apr1954  08sep1972  31mar1984   25   80  80.000  84.245  0
B  01apr1954  08sep1972  31mar1989   30   80  84.245  89.245  0
B  01apr1954  08sep1972  01jan1990   35   80  89.245  90.000  0
B  01apr1954  08sep1972  01apr1994   35   90  90.000  94.245  0
B  01apr1954  08sep1972  23may1995   40   90  94.245  95.387  0
C  10jun1987  23dec1991  09jun1992    0   90  91.973  92.436  0
C  10jun1987  23dec1991  09jun1997    5   90  92.436  97.436  0
C  10jun1987  23dec1991  24jul1998   10   90  97.436  98.557  1
```

Analysis of the Nickel refinery study

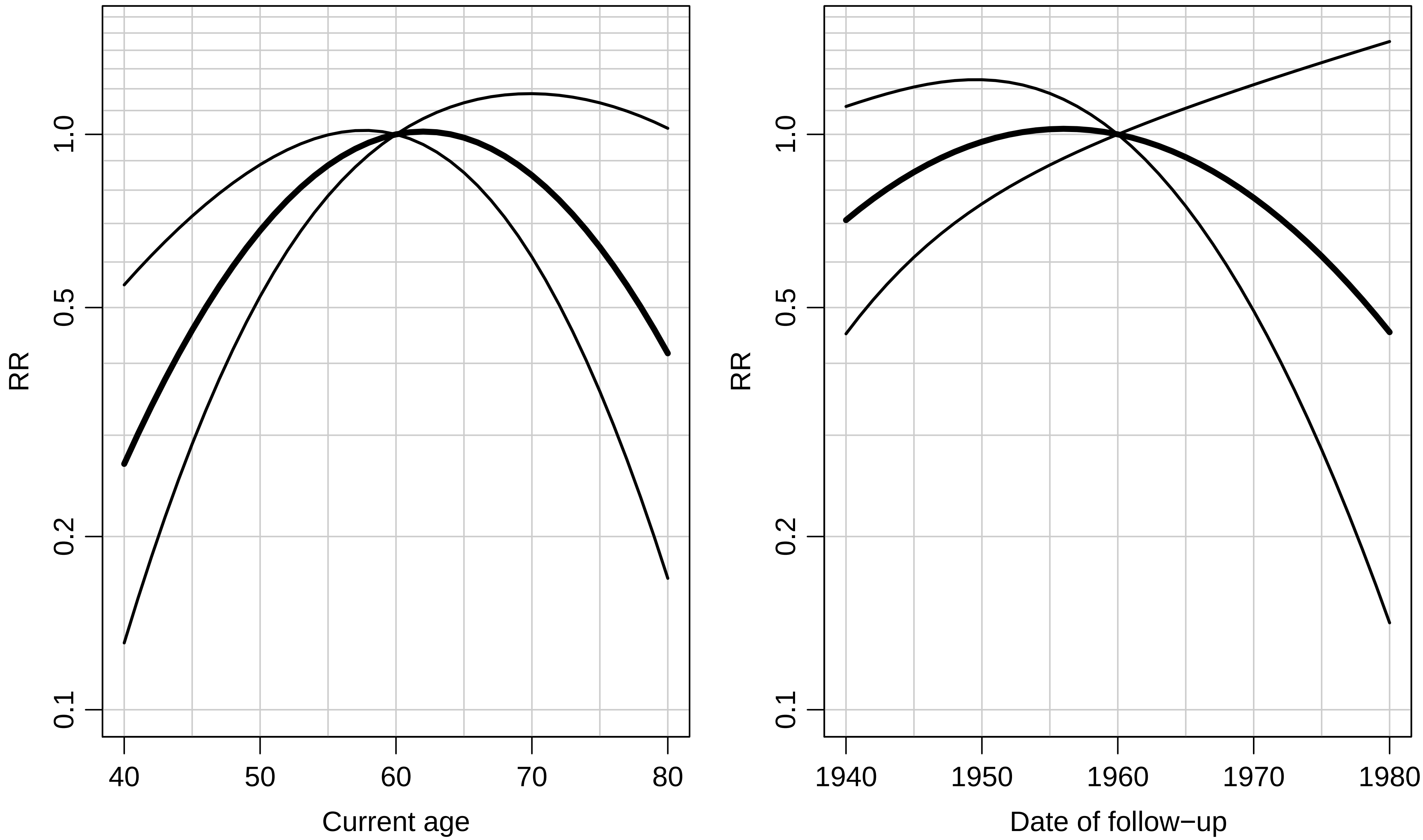
```
. use ../data/nickel, clear
. stset dox, entry(doe) origin(time 0) fail(icd==162,163) id(id)
                id:  id
     failure event:  icd == 162 163
obs. time interval:  (dox[_n-1], dox]
 enter on or after:  time doe
 exit on or before:  failure
--------------------------------------------------------------------------
      679  total obs.
        0  exclusions
--------------------------------------------------------------------------
      679  obs. remaining, representing
      679  subjects
      137  failures in single failure-per-subject data
 15348.06  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =  1934.246
                                  last observed exit t =  1982.912

. stsplit date, at(1900(5)2000)
             (3327 observations (episodes) created)
. stsplit age, after(dob) at(0(5)90)
             (3094 observations (episodes) created)
```

```
. list id dob doe dox date age _t0 _t _d in 1/20, nodis
id       dob       doe       dox  date  age     _t0      _t      _d
 3  1889.019  1934.246      1935  1930   45  1934.246  1935.000   0
 3  1889.019  1934.246  1939.019  1935   45  1935.000  1939.019   0
 3  1889.019  1934.246      1940  1935   50  1939.019  1940.000   0
 3  1889.019  1934.246  1944.019  1940   50  1940.000  1944.019   0
 3  1889.019  1934.246      1945  1940   55  1944.019  1945.000   0
 3  1889.019  1934.246  1949.019  1945   55  1945.000  1949.019   0
 3  1889.019  1934.246      1950  1945   60  1949.019  1950.000   0
```

```
. gen Y = _t-_t0
. gen a1 = ( age - 60 )
. gen a2 = ( age - 60 )^2
. gen d1 = ( date - 1960 )
. gen d2 = ( date - 1960 )^2

. glm _d a1 a2 d1 d2 expos, family(poisson)  lnoffset(Y)

Generalized linear models                   No. of obs      =      7100
Optimization     : ML: Newton-Raphson       Residual df     =      7094
------------------------------------------------------------------------
     _d |      Coef.   Std. Err.       z     P>|z|    [95% Cf. Interval]
-------+----------------------------------------------------------------
     a1 |   .0108726   .0127122     0.86    0.392  -.0140428  .0357881
     a2 |  -.0027659   .0008146    -3.40    0.001  -.0043625 -.0011694
     d1 |   -.011218   .0171422    -0.65    0.513   -.044816  .0223801
     d2 |  -.0014201   .0007319    -1.94    0.052  -.0028546  .0000144
  expos |   .0928812   .0205795     4.51    0.000   .0525461  .1332164
  _cons |  -4.383049   .1506798   -29.09    0.000  -4.678376 -4.087722
      Y | (exposure)
------------------------------------------------------------------------
```

What is the meaning of _cons   -4.383049?

Can any of the parameters be removed from the model?


RR
1.0
0.5
0.2
0.1
40
50
60
70
80
Current age
RR
1.0
0.5
0.2
0.1
1940
1950
1960
1970
1980
Date of follow−up

Time since entry

Implicit assumption in cohort studies:

All covariates are known at any point of follow up.
The precise history of the person is assumed known at any one timepoint.

In practise many covariates are only measured at entry, for example smoking status.

In analysis these are assumed constant throughout follow-up.

The quality of the used covariate (=the entry value) declines with time since follow-up:
Accuracy in covariates decrease by time since entry.

If time since entry is not in a model, follow-up with different quality of data is pooled.

Cohorts where all are exposed

Do mortality rates in cohort differ from those of an **external** population?

- Occupational cohorts
- Patient cohorts

Compared with reference rates obtained from:

- Population statistics (mortality rates)
- Disease registers (hospital discharges, cancer)

Accounting for age composition

- Compare rates in a study group with a standard set of age–specific rates

- Reference rates are normally based on large numbers of cases, so they can be assumed to be known

- Calculate “expected” number of cases, E, if the standard rates had applied in our study group, and compare this with the observed number of cases, D

$$\text{SMR} = D/E \qquad \text{s.e.}(\log[\text{SMR}]) = 1/\sqrt{D}$$

Statistical model for SMR

The rate in the cohort is assumed to be proportional to the rates in the population:

$$\lambda_{\textsf{coh}}(t) = \theta \lambda_{\textsf{pop}}(t)$$

The time t is normally the combination of current age, calendar time and sex.

The population rates are assumed known without error classified by age, calendar time and sex.

θ is the Rate-Ratio between the cohort and the general population.

Likelihood for SMR

For empirical rates (d, y) the log-likelihood is:

$$d\log(\lambda_{\text{coh}}) - \lambda_{\text{coh}} y = d\log(\theta\lambda_{\text{pop}}) - \theta\lambda_{\text{pop}} y = d\log(\theta) - \theta(\lambda_{\text{pop}} y)$$

omitting $d\log(\lambda_{\text{pop}})$ which does not depend on θ.

This is the likelihood for a theoretical rate θ, based on an empirical rate $(d, \lambda_{\text{pop}} y)$.

Thus under the assumption that θ is constant the maximum likelihood estimate is:

$$\hat{\theta} = \frac{D}{\lambda_{\text{pop}} Y} = \frac{\text{Observed}}{\text{Expected}} = \text{SMR}$$

SMR is the maximum likelihood estimator of the mortality rate ratio between the cohort and the population.

Expected numbers in practice

- From the file with expanded follow-up data:
  - y — The risk time in the record
  - age class — The ageclass of the record
  - period — The period of the record
  - sex — The sex of the record

- From the file with reference rates:
  - λ_R — The reference rate.
  - age class — The ageclass of the population rate
  - periode — The period of the population rate
  - sex — The sex of the population rate

- Population rates are matched up to the follow-up data, and expected numbers are computed as:

$$e = \lambda_{\text{pop}}(a, p, s) \times y$$

There are always two datasets in play with SMR:

1. The **cohort** dataset with follow-up information on a number of individuals.
   This is the dataset that must be split, to match with

2. The **rate** dataset with disease or death rates of a reference population.

SMR-calculations in Stata

1. Take a look at the population rate file. Determine the age and period classes.

2. `sort` the population rates by age, period and sex.

3. `clear` memory and get the cohort data.

4. `stset` the cohort data and `stsplit` them by age and period in classes as those of the available population rates.

5. `sort` the split data by age, period and sex (in the same way as the file of population rate was sorted)

6. `merge` the cohort data with the population rate file.

7. `generate` the number of expected cases by multiplying the cohort risk time with the rates.

The final result is a (very small) number of expected cases attached to every little piece of follow-up.

But when tabulated over the entire dataset it gives the number of deaths one would have expected to see in the cohort if it were subjected to population mortality rates.

SMR-calculations in Stata

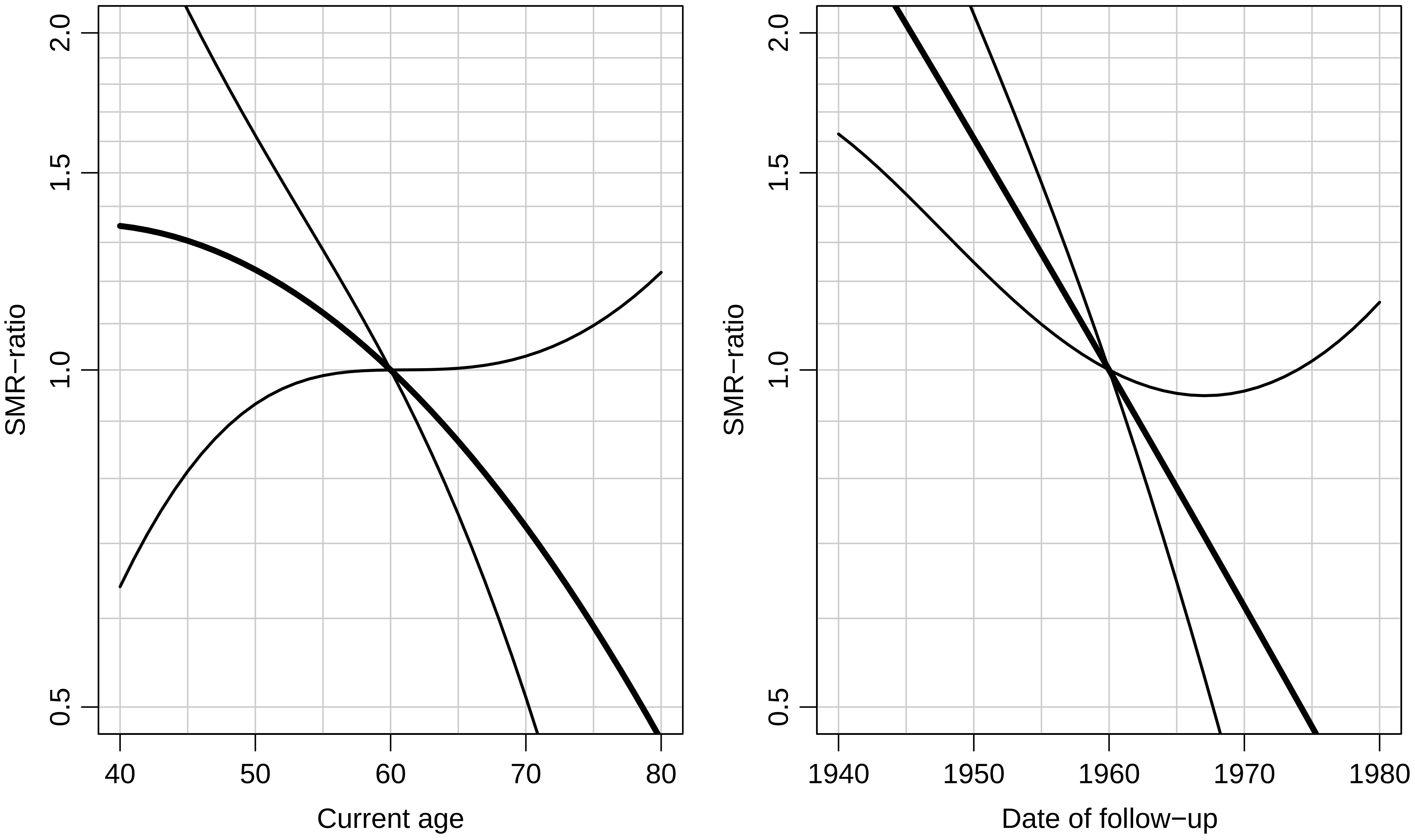
```
* Look at the population rate file and sort it
. use pop
. sort agr per sex
. save popsort

* Switch to cohort, split follow-up and sort it
. use cohort, clear
. stset exit, enter(entry) fail(st) id(id)
              origin(bdate) scale(365.25)
. stsplit agr, at(0(5)90)
. stsplit per, after(time=d(01jan1900)) at(0(5)100)
. sort agr per sex

* Merge the two files together
. merge agr per sex using popsort

* Keep only records with data from both sources
. keep if _merge==3
```

```
* Compute SMRs
. strate , smr(poprate)
. strate expos, smr(poprate)

* Generate expected numbers
. generate e_c = ( _t - _t0 ) * poprate

* Then we can do modelling
. glm _d expos sex, lnoffset( e_c )
```

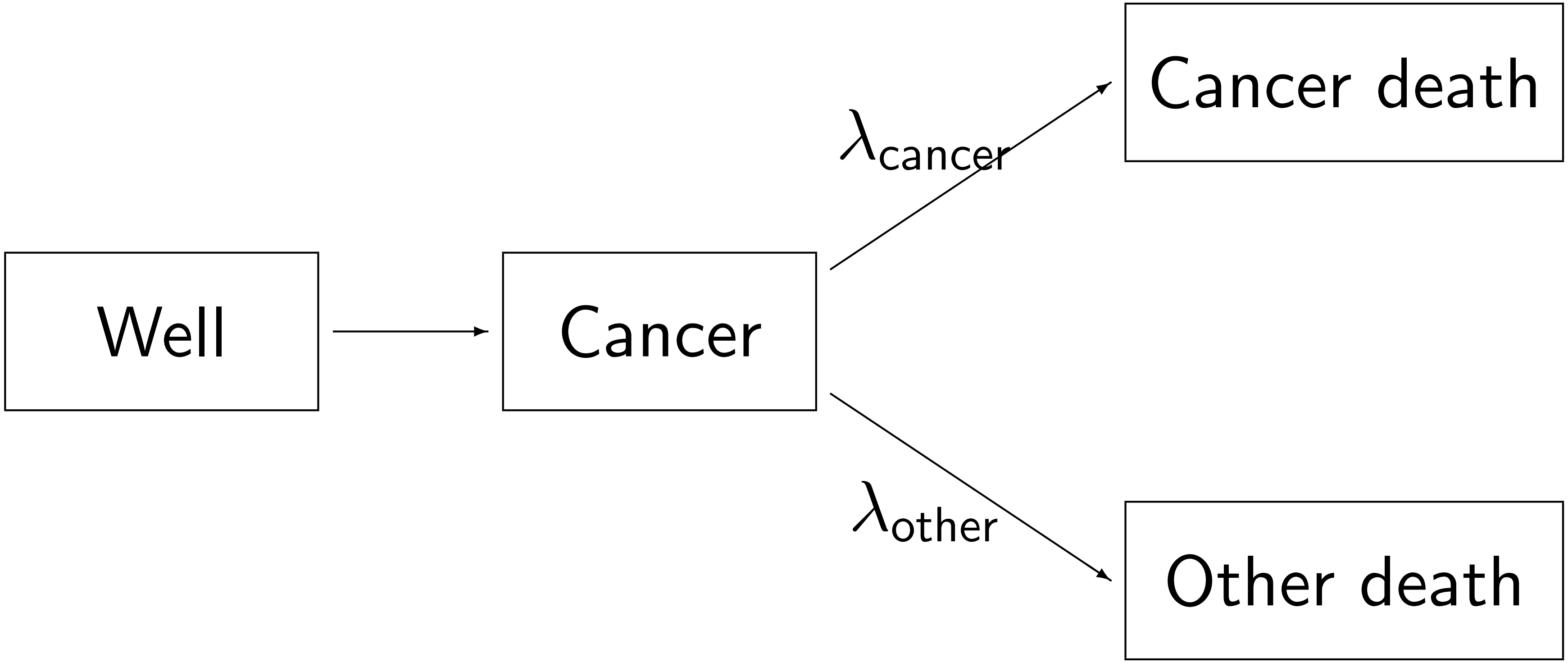
SMR-calculations in R

```
> Poprat <- read.dta( "../data/ewrates.dta" )
> Nickel <- read.dta( "../data/nickel.dta" )
>
> # Split time along two time-axes
> #
> Nsplit <-
+ Lexis( entry = doe, exit = dox, fail = (icd %in% c(162,163) ),
+       origin = list( year=0,                age=dob         ),
+       breaks = list( year=seq(1931,1981,5), age=seq(0,85,5) ),
+      include = list( expos ), data = Nickel )
> # Merge with the population rates
> #
> Nall <- merge( Nsplit, Poprat )
> M1 <- glm( Fail ~ I(age-60) + I((age-60)^2) + I(year-1960) + I((year-196
+                   expos + offset( log( (Exit-Entry)*lung/10^5 ) ),
+                   family=poisson, eps=10^-8, data=Nall )
round( summary( M1 )$coef, 6 )
                    Estimate Std. Error   z value Pr(>|z|)
(Intercept)        -1.095333   0.152962 -7.160808 0.000000
I(age - 60)        -0.026431   0.013639 -1.937944 0.052630
I((age - 60)^2)    -0.000581   0.000857 -0.677683 0.497972
I(year - 1960)     -0.048145   0.017156 -2.806348 0.005011
I((year - 1960)^2) -0.000050   0.000677 -0.074108 0.940925
expos               0.089864   0.020648  4.352281 0.000013
```

What is the meaning of: `(Intercept) -1.095333` ?


SMR−ratio
2.0
1.5
1.0
0.5
40
50
60
70
80
Current age
SMR−ratio
2.0
1.5
1.0
0.5
1940
1950
1960
1970
1980
Date of follow−up

Relative survival rates



What is the relative survival of cancer patients:
What is the the survival at t after the diagnosis, compared to what it would have been without cancer diagnosis?

Compare the survival for the group of cancer patients as it would have been had they died according to population rates:

Compute the expected survival for each patient from diagnosis and average these (on the probability scale). This is the expected survival for this group of patients.

Divide the observed survival with the expected survival. The observed survival will always be larger than the expected.

Expected survival

Assuming $\lambda_{\mathsf{other}}(t)$ is not affected by cancer diagnosis, the expected survival function is:

$$S_{\mathsf{E}}(t) = \exp\left(-\int_{\mathsf{diagnosis}}^{\mathsf{diagnosis}+t} \lambda_{\mathsf{other}}(u)\mathrm{d}u\right)$$

The observed survival function is:

$$S_{\mathsf{Obs}}(t) = \exp\left(-\int_{\mathsf{diagnosis}}^{\mathsf{diagnosis}+t} \lambda_{\mathsf{other}}(u) + \lambda_{\mathsf{cancer}}(u)\mathrm{d}u\right)$$

The relative survival rate, RSR (which is a proportion!):

$$\mathrm{RSR}(t) = \frac{S_{\mathsf{Obs}}(t)}{S_{\mathsf{E}}(t)} = \exp\left(- \int_{\text{diagnosis}}^{\text{diagnosis}+t} \lambda_{\text{cancer}}(u)\mathrm{d}u \right)$$

So RSR is based on the concept:
"what if the non-cancer mortality were absent".

If cause of death were easy to establish, we could just estimate $\lambda_{\text{cancer}}(t)$ and work it out.
For cancer patients the cause of death is not well recorded.

Take λ_{other} to be the population mortality, so the model becomes an **additive** hazards model for the total mortality of cancer patients:

$$\lambda(t) = \lambda_{\text{pop}} + \lambda_{\text{cancer}}$$

The likelihood for an empirical rate is then:

$$\ell(\lambda_{\text{cancer}}) = d\log(\lambda_{\text{pop}} + \lambda_{\text{cancer}}) - (\lambda_{\text{pop}} + \lambda_{\text{cancer}})y$$

This corresponds to a likelihood for a Poisson variate with mean $(\lambda_{\text{pop}} + \lambda_{\text{cancer}})y$.

Possible to model the λ_{cancer} as a function of covariates, if we can fit a Poisson model with identity link function.

These are known as **excess risk models**, and there is large literature on them. They are closely related to Aalen's additive hazards model.

Aalen O.O. (1989). A linear regression model for the analysis of lifetimes. *Statistics in medicine,* **8**, 907–925.

Aranda-Ordaz F.J. (1983). An extension of the proportional-hazards model for grouped data. *Biometrics,* **39**, 109–117

Ederer F., Axtell L.M. and Cutler S.J. (1961). The relative survival rate: a statistical methodology. *National Cancer Institute Monographs,* **6**, 101–121.

Hakulinen T. and Tenkanen L. (1987). Regression analysis of relative survival rates. *Applied Statistics,* **36**, 309–317.

Survival analysis and medical demography

The Cox model

11 April 2003

Bendix Carstensen

Modelling the hazard

The Cox-model allows the rate to depend not only on time, but also on covariates, as in regression analysis:

$$\lambda_i(t) = \lambda_0(t)\exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots)$$

If a person has covariate values 0, i.e. $x_{1i} = x_{2i} = \cdots = 0$, then this persons hazard (mortality rate) is $\lambda_0(t)$.

The function $\lambda_0(t)$ is called the baseline hazard, and not restricted to have any particular form. Hence the name "semiprametric" for this type model.

Note that this is where a choice of timescale is needed.

Components of the Cox-model

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots)$$

There are two components in a Cox-model:

1. The baseline hazard, $\lambda_0(t)$, a function of **time**. Hazard of a person with all covariates=0.

2. The relative risk (rate ratio) function,
   $\mathrm{RR} = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots) = \mathrm{e}^{\eta_i}$
   Usually summarized by the parameter estimates $\beta_1, \beta_2, \ldots$ and their standard errors.

Interpretation of the regression parameters

Two persons, A and B with covariate values

$$x_{1A}, x_{2A}, x_{3A}, \qquad x_{1B}, x_{2B}, x_{3B}, \ldots$$

Hazard ratio between A and B:

$$\begin{aligned}
\frac{\lambda_A(t)}{\lambda_B(t)} &= \frac{\lambda_0(t)\mathrm{RR}_A}{\lambda_0(t)\mathrm{RR}_B} = \frac{\mathrm{RR}_A}{\mathrm{RR}_B} \\
\\
&= \exp[(\beta_1 x_{1A} + \beta_2 x_{2A} + \cdots) - (\beta_1 x_{1B} + \beta_2 x_{2B} + \cdots)] \\
&= \exp[(\beta_1(x_{1A} - x_{1B}) + \beta_1(x_{1A} - x_{1B}) + \cdots]
\end{aligned}$$

Thus the rate ratio, or hazard ratio is **constant**, independent of time, i.e. the hazard rates for A and B are proportional. Hence the name **proportional hazards model**.

If A and B have identical values for all covariates except x_1, the the rate ratio is:

$$\mathrm{RR}_{A\ \text{vs.}\ B} = \exp[\beta_1(x_{1A} - x_{1B})]$$

Thus β_1 is the log rate ratio per unit change in variable x_1.

Exercise on Cox regression coefficients

A Cox mode is fitted and the cofficient to the variable holding the blood pressure in mmHg is:

$$\hat{\beta}_{\mathsf{BP}} = 0.0357$$

What is the rate ratio associated with an increase of 1 mmHg?

What is the rate ratio associated with an increase of 10 mmHg?

Exercise on Cox regression coefficients

Categorical covariates:

The variable `sex` is coded 1 for men and 0 for women.

A Cox mode is fitted and the cofficient to the variable is:

$$\hat{\beta}_{\textsf{sex}} = 0.712$$

What is the rate ratio between men and women?

What is the rate ratio between women and men?

The baseline survival function

Recall the relation between the hazard $\lambda(t)$ and the survival function $S(t)$:

$$S(t) = \exp\left(-\int_0^t \lambda(s)\,ds\right) = \exp\left(-\Lambda(t)\right)$$

The quantity $\Lambda(t) = \int_0^t \lambda(s)\,ds$ is the cumulative hazard.

An estimate of the cumulative baseline hazard, can be transformed to an estimate of the baseline survival function, i.e.
the expected survival for a person with all covariates $= 0$.

Estimation of the baseline

The baseline is usually estimated using the Breslow-estimator of the cumulative hazard:

$$\Lambda_0(t) = \sum_{j \leq t} \frac{1}{\sum_{i \in \mathcal{R}_j} \mathrm{e}^{\eta_i}}$$

Note that if $\eta_j = 0$ for all persons this is the Nelson-Aalen estimator of the cumulative hazard.

For a person with covariates $x_{1i}, x_{2i}, \ldots$ we have:

$$\Lambda_i(t) = \Lambda_0(t) \exp[\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots]$$

$$\begin{aligned} S_i(t) &= \exp[-\Lambda_i(t)] \\ &= \exp[-\Lambda_0(t)\mathrm{RR}_i] \\ &= S_0(t)^{\mathrm{RR}_i} \end{aligned}$$

The survival function for a person is the baseline survival raised to the power of RR_i, the rate-ratio function.

Baseline survival curve

If we get an estimate of the baseline survival function, $S_0(t) = \exp[-\Lambda_0(t)]$ this will refer to a person with all covariate values $= 0$.

This does not necessarily correspond to anything sensible.

Make sure that covariates are centered around sensible values, in order to make the baseline meaningful.

If we instead of age use age−60, then a person with a 0 value of the covariate will be aged 60 and not 0.

Estimation in Stata

As an example we use the Finnish colon cancer data, with survival in months since diagnosis in `survmm`, exit status in `status` (dead are coded 1 or 2), stage of disease in `stage` and age at diagnosis in `age`.

As for the Kaplan-Meier analysis, the first thing to do is to declare data as survival time data using `stset`:

```
. stset survmm, fail(status==1,2) scale(12)
```

The `scale(12)` transforms the survival time from months to years.

```
. stcox age

         failure _d:  status == 1 2
   analysis time _t:  survmm/12

No. of subjects =        14713   Number of obs    =     14713
No. of failures =        10067
Time at risk    =  57793.66667   LR chi2(1)       =   1084.33
Log likelihood  =   -89480.384   Prob > chi2      =    0.0000
 _t |
 _d | Haz. Ratio       z   P>|z|  [95% Conf. Interval]
----+------------------------------------------------------
age |   1.029548    31.47  0.000  1.027682    1.031416
-----------------------------------------------------------
```

The number 1.029 refers to the hazard ratio between two persons that differ one year in age at diagnosis.

The hazard ratio between two persons that differ 10 years in age at diagnosis would be $1.029^{10} = 1.338$. This can be directly computed by Stata by rescaling age:

```
. gen a10 =age/10
. stcox a10
------------------------------------------------------------
  _t |
  _d | Haz. Ratio  Std. Err.     z   P>|z|  [95% Conf. Int.]
----+-------------------------------------------------------
a10 |   1.338025  .0123806  31.47  0.000  1.313978  1.362512
------------------------------------------------------------
```

Thus, for evey 10 years older colon cancer patients are, their mortality increases by 34%.

Estimation of the baseline hazard

The baseline survival function, $S_0(t) = \exp[-\Lambda_0(t)]$ refers to a person with all covariate values $= 0$, but this does not necessarily correspond to anything sensible.

Only if we center all variables around a sensible value; e.g. recode age by:

```
. replace age = ( age - 60 ) / 10
```

This will make 60 the reference age and make the regression coefficient refer to an age-difference of 10 years.

Using the xi:-notation for generating dummy variables in Stata, the corresponding baseline hazard will be the one for the reference category.

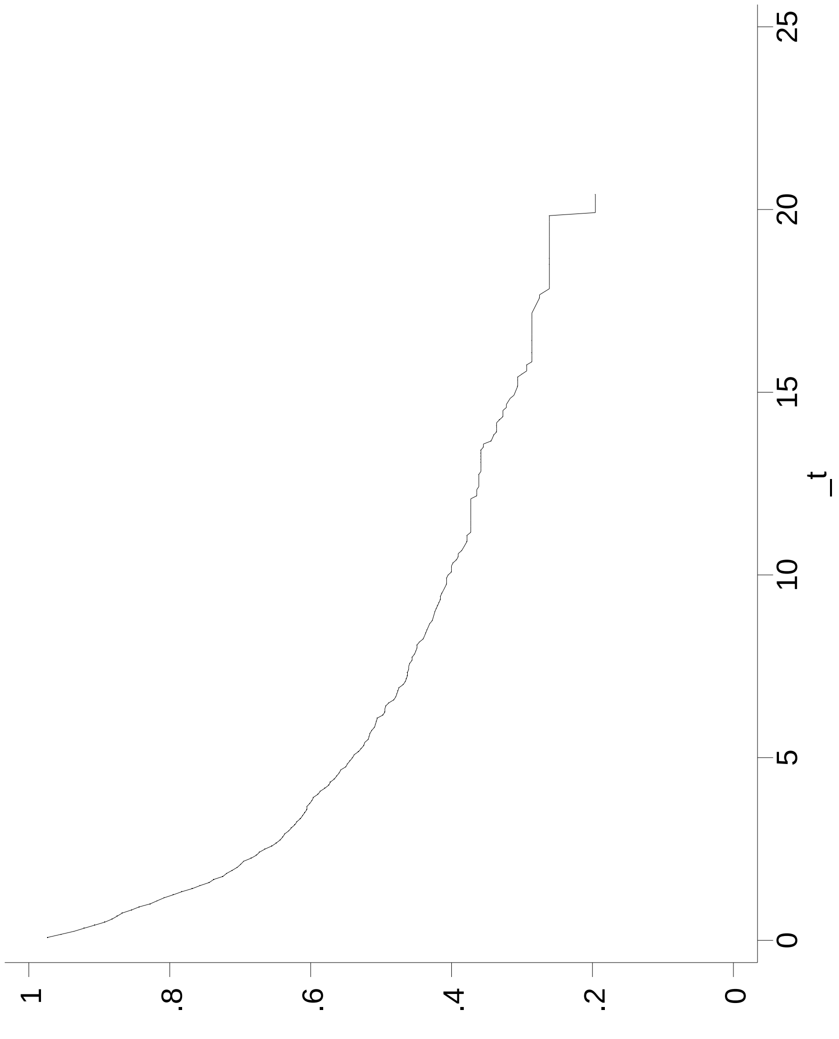
The way to get the baseline survival and cumulative hazard in Stata is:

```
. xi: stcox i.period i.stage, basech(L0) basesurv(S0)
```

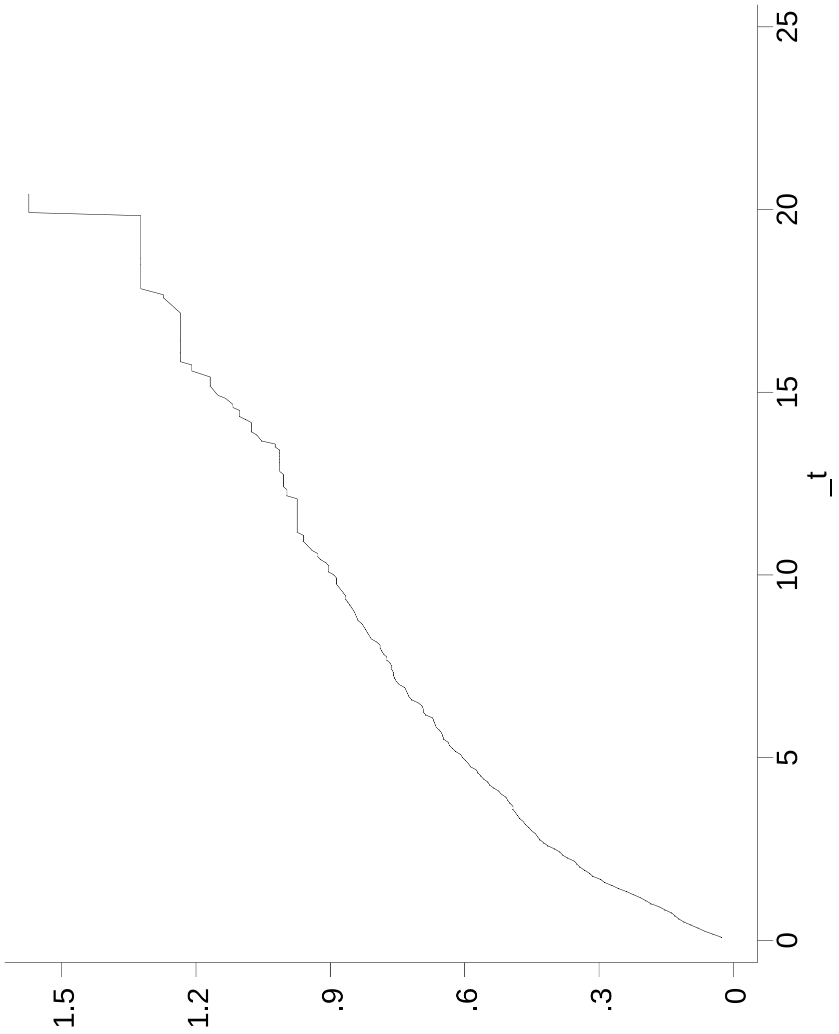
this will create two new variables S0 and L0 that contains the baseline survival and baseline cumulative hazard.

They can then be plotted against the survival time. The scaled version of the survival time is in _t:

graph S0 _t, xlabel(0(5)25) ylabel(0(0.2)1)



graph L0 _t, xlabel(0(5)25) ylabel(0(0.2)1)



Testing the assumption of proportionality

It is possible to relax the assumption about proportional hazards, by making a **stratified** model:

$$\lambda(t,x) = \lambda_s(t) \times \mathrm{RR}_x$$

Here we allow different baseline hazards for different levels of s, but maintain that the effect of other covariates (x) is the same across levels of s

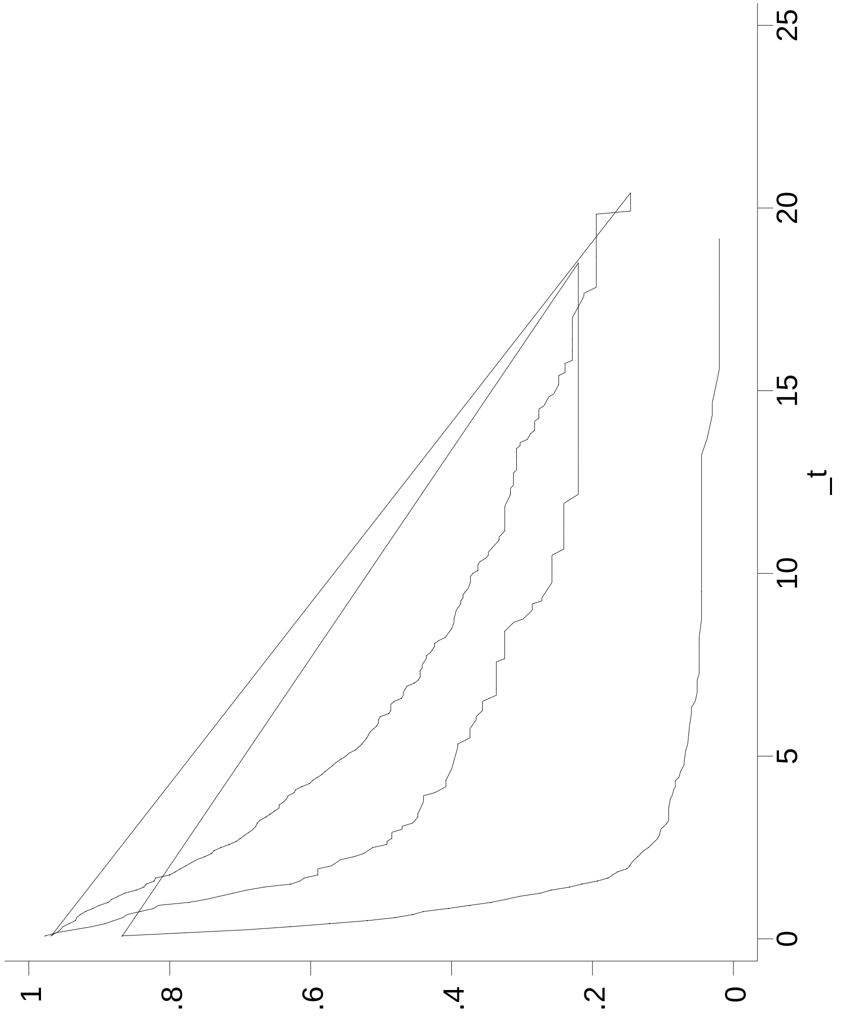
This is in other regression models called an **interaction** between time and the stratification variable.

```
. xi: stcox i.period, strata(stage) basech(sL0)
                       basesurv(sS0) sch(sr*)
. stphtest

Test of proportional hazards assumption
Time:  Time
            |          chi2       df       Prob>chi2
------------+---------------------------------------
global test |         11.98        3          0.0075
----------------------------------------------------
```

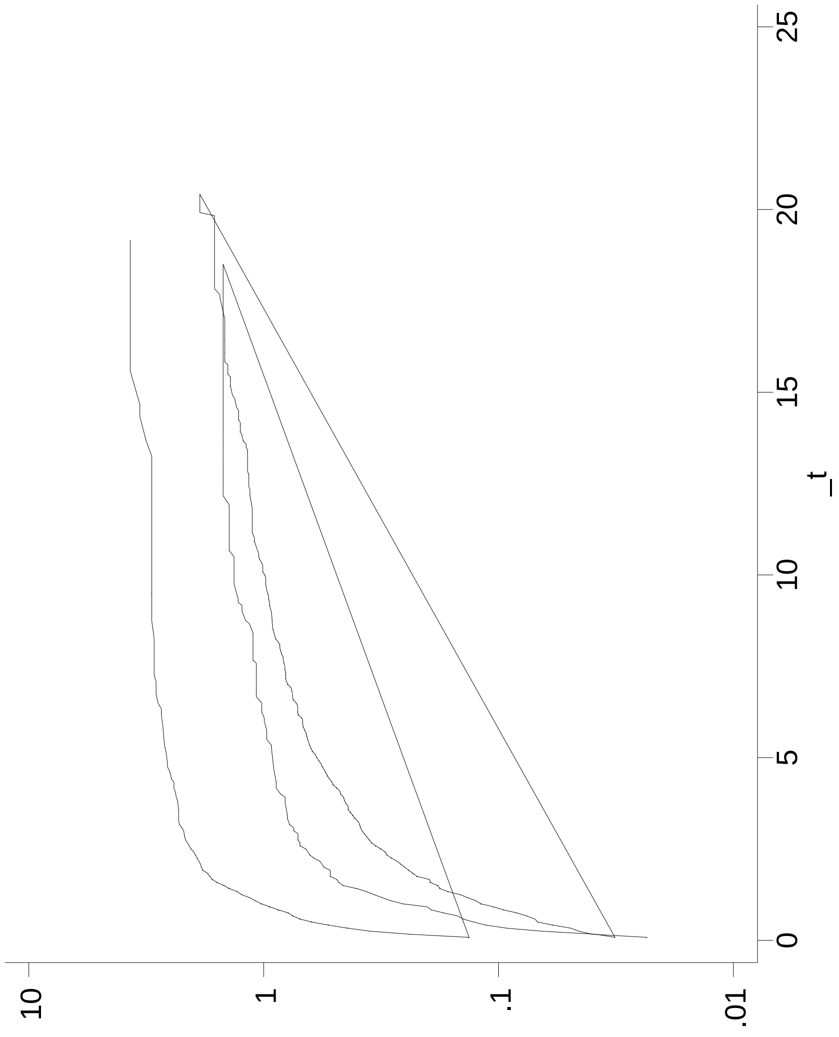
The `sch(sr*)` saves a set of residuals (Schoenfeld residuals) needed in order to make a test of the proportionality assumption, which is tested by the command `stphtest`.

So the proportionality assumtion does not hold in this case.

```
graph S0 _t, xlabel(0(5)25) ylabel(0(0.2)1)
```



graph L0 _t, xlabel(0(5)25) ylog



The Cox-likelihood

Cox devised the **partial** likelihood for the parameters $\beta = (\beta_1, \ldots, \beta_p)$ in the linear predictor $\eta_i = \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$:

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

where $\mathcal{R}_t$ is the risk set at time t,
i.e. the set of individuals at risk at time t.

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).

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

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

Assume w.l.o.g. the y for these emiprical rates are 1.

The log-likelihood contributions that contain information on a specific time-scale parameter α, relating to time t say, will be contributions from the empirical rate $(1, 1)$ with the death at time t, and all the empirical rates $(0, 1)$ from all the other individuals that were at risk at time t.

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

$$\ell_t(\alpha, \beta) = \sum_{i \in \mathcal{R}_t} \left\{ d_i(\alpha + \eta_i) - \mathrm{e}^{\alpha + \eta_i} \right\} = \alpha + \eta_{\textsf{death}} - \mathrm{e}^{\alpha} \sum_{i \in \mathcal{R}_t} \mathrm{e}^{\eta_i}$$

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

The derivative w.r.t. α is:

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

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

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

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

Implications for modelling

The model set up could have been formulated as one where there was a separate timescale parameter for each time-interval.

For those intervals on the time-scale where no deaths occur the estimate of the α will be $-\infty$, and so these intervals will not contribute to the log-likelihood.

The Cox-model can be estimated by standard Poisson-regression-software by splitting the data finely and specifying the model as having one rate parameter per time interval. The results will be the same, also for the $\mathrm{s.e.}$.

Example in Stata

```
. set matsize 150
. use ../data/ping-pong
(Written by R.              )
. stset exit, entry(entry) id(id) fail(event)

                id:  id
     failure event:  event ~= 0 & event ~= .
obs. time interval:  (exit[_n-1], exit]
 enter on or after:  time entry
 exit on or before:  failure

-------------------------------------------------------------------
       60  total obs.
        0  exclusions
-------------------------------------------------------------------
       60  obs. remaining, representing
       60  subjects
       37  failures in single failure-per-subject data
     2960  total analysis time at risk, at risk from t =          0
                             earliest observed entry t =          0
                                  last observed exit t =        122
```

```
. stcox ping pong, nohr
   _d |      Coef.   Std. Err.    z    P>|z|   [95% Conf. Interval]
------+------------------------------------------------------------
 ping |  -.5161977      .3892  -1.33   0.185  -1.279016    .2466203
 pong |  -.6945726     .17471  -3.98   0.000  -1.036998   -.3521473
-------------------------------------------------------------------
```

```
. stsplit time, at(1(1)125)
(2741 observations (episodes) created)
. gen risk = _t-_t0
. xi: glm _d i.time ping pong, family(Poisson) lnoffset(risk)

        _d |      Coef.  Std. Err.      z   P>|z|  [95% Cf. Interval]
-----------+-------------------------------------------------------
  _Itime_2 |  -.9301846  12246.85  -0.00   1.000 -24004.32  24002.46
  _Itime_3 |     -.8332  11546.38  -0.00   1.000 -22631.31  22629.65
  ...
_Itime_121 |   18.37402  9999.272   0.00   0.999 -19579.84  19616.59
_Itime_122 |  -1.098276  14141.62  -0.00   1.000 -27718.17  27715.97
      ping |  -.5161977  .3891999  -1.33   0.185 -1.279015  .2466201
      pong |  -.6945726    .17471  -3.98   0.000 -1.036998 -.3521474
     _cons |  -18.40874  9999.272  -0.00   0.999 -19616.62   19579.8
      risk | (exposure)
-------------------------------------------------------------------
```

Example in R

Artificial data `ping-pong.dta`

```
> library( survival )
> library( splines )
> ds <- read.dta( file="../data/ping-pong.dta" )
> dx <- Lexis( entry=entry, exit=exit, fail=event,
+              breaks=sort( unique( c( ds$entry, ds$exit ) ) ),
+              data=ds, include=list( id, ping, pong ) )
> c.res <- coxph( Surv( entry, exit, event ) ~ ping + pong, data=ds,
+                 method="breslow", eps=10^-8, iter.max=25 )
> p.res <- glm( Fail ~ factor( Time ) - 1 + ping + pong +
                      offset( log(Exit-Entry) ),
+               family=poisson, data=dx,
+               eps=10^-8, maxit=25 )
> s.res <- glm( Fail ~ bs( Time, df=5 ) + ping + pong + offset( log(Exit-E
+               family=poisson, data=dx,
+               eps=10^-8, maxit=25 ) )
```

```
> cr <- ci.lin( c.res )[,1:2]
> pr <- ci.lin( p.res, subset=length( coef( p.res ) )-1:0 )[,1:2]
> sr <- ci.lin( s.res, subset=length( coef( s.res ) )-1:0 )[,1:2]
> all <- cbind(
+        rbind( cr[1,], pr[1,], sr[1,], (cr/pr)[1,], (cr/sr)[1,] ),
+        rbind( cr[2,], pr[2,], sr[2,], (cr/pr)[2,], (cr/sr)[2,] ) )
> rownames( all ) <- c("Cox","Poisson","Spline","C/P","C/S")
> colnames( all ) <- paste( rep( rownames( pr ), rep( 2, 2 ) ),
+                           rep( c("Est","SE"), 2 ) )
> print( round( all, 5 ) )
        ping Est ping SE pong Est pong SE
Cox     -0.51620 0.38920 -0.69457 0.17471
Poisson -0.51620 0.38920 -0.69457 0.17471
Spline  -0.47706 0.38693 -0.65151 0.16702
C/P      1.00000 1.00000  1.00000 1.00000
C/S      1.08203 1.00587  1.06610 1.04601
```

Summary of methods

- Likelihood is a product of contributions from single empirical rates (d, y), each represnenting the data from a small interval of follow-up.

- The correspondng data-layout requires split of follow-up for each person.

- Modelling is specification of sensible (interpretable, practical) parametric models for covariates associated with each empirical rate.

- The effect of timescales should not in principle be more complicated to model that that of e.g. blood pressure or height.

- Time lends itself to detailed modelling because of the structure of the likelihood function as a product at separate terms for each time.

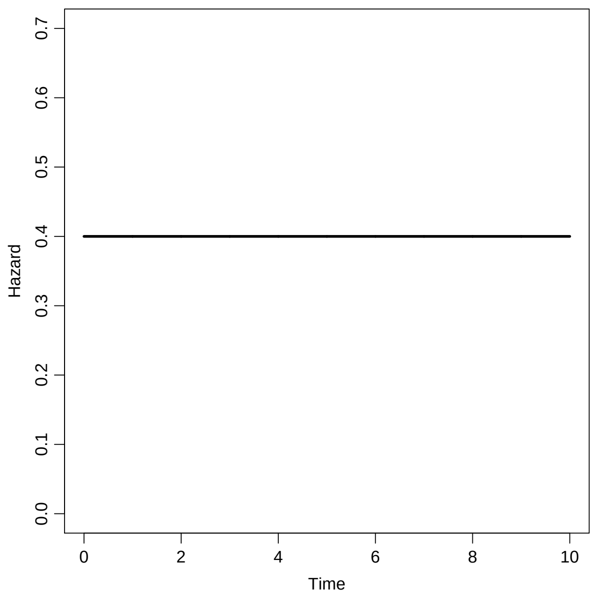
  This is not the case for any other covariate. But it does not necessarily make the exploitation of the fact sensible let alone desirable in practical biostatistics.

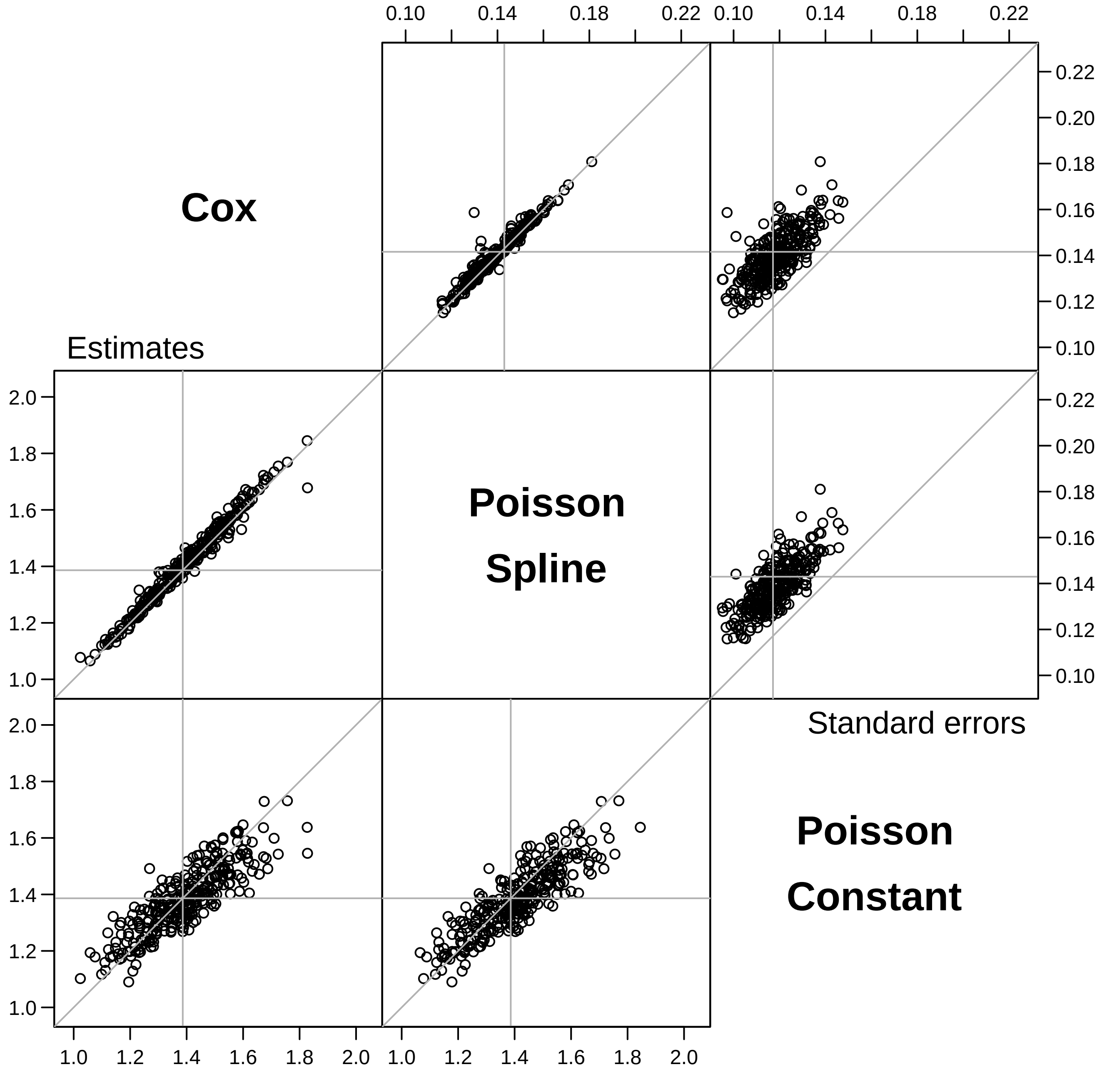
Practical comparison of Cox and Poisson

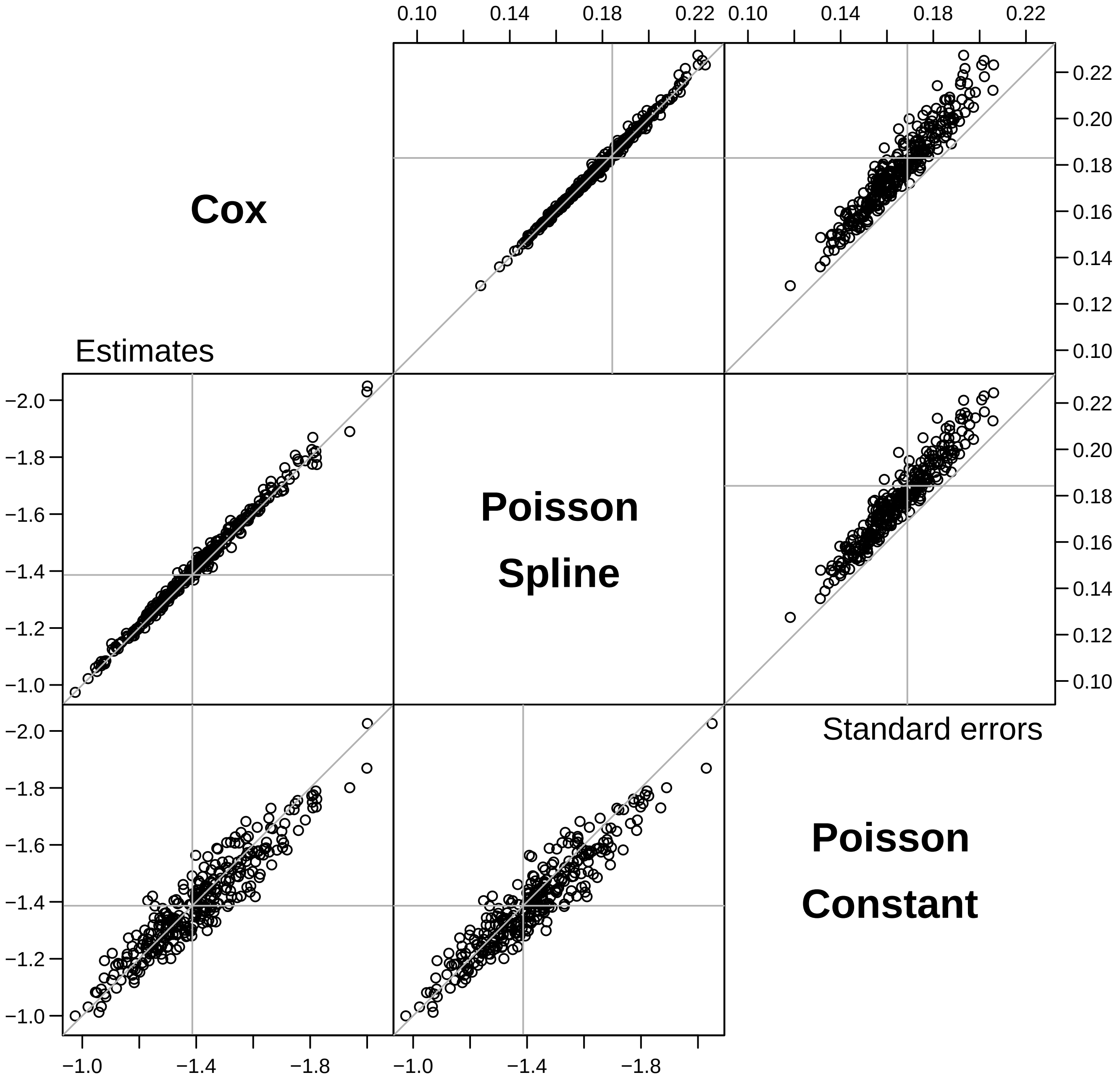
Simulate 300 datasets of 200 persons, followed for 10 years with a baseline hazard which constant or neither constant nor monotone, and two covariates `ping` and `pong` with relative risks of 4 and 0.25, respectively.

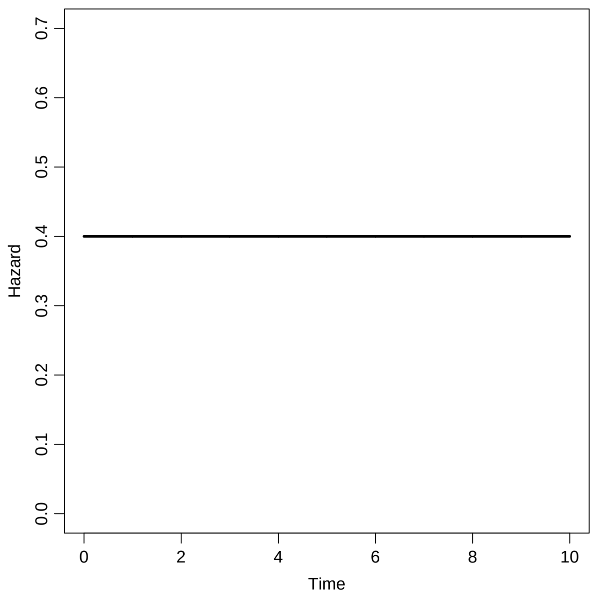
For each dataset split data in 20 intervals over the follow-up time and fit:

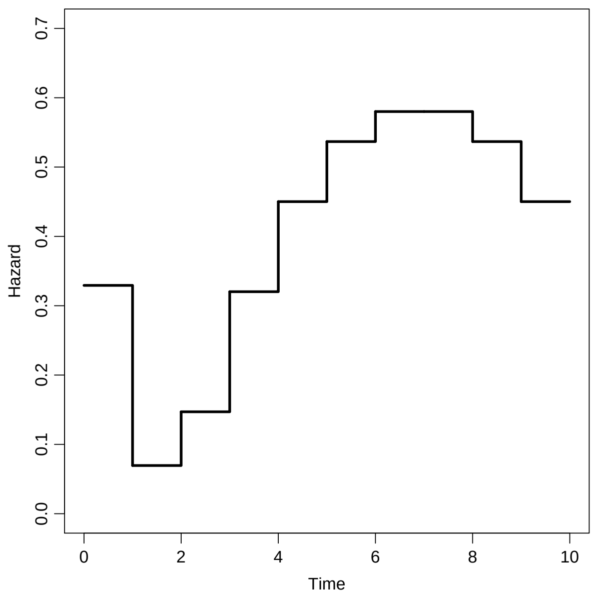
- Cox-model (to the original 300 obs dataset)
- Poisson model with a 5-parameter parametric basline.
- Poisson model assuming constant baseline.

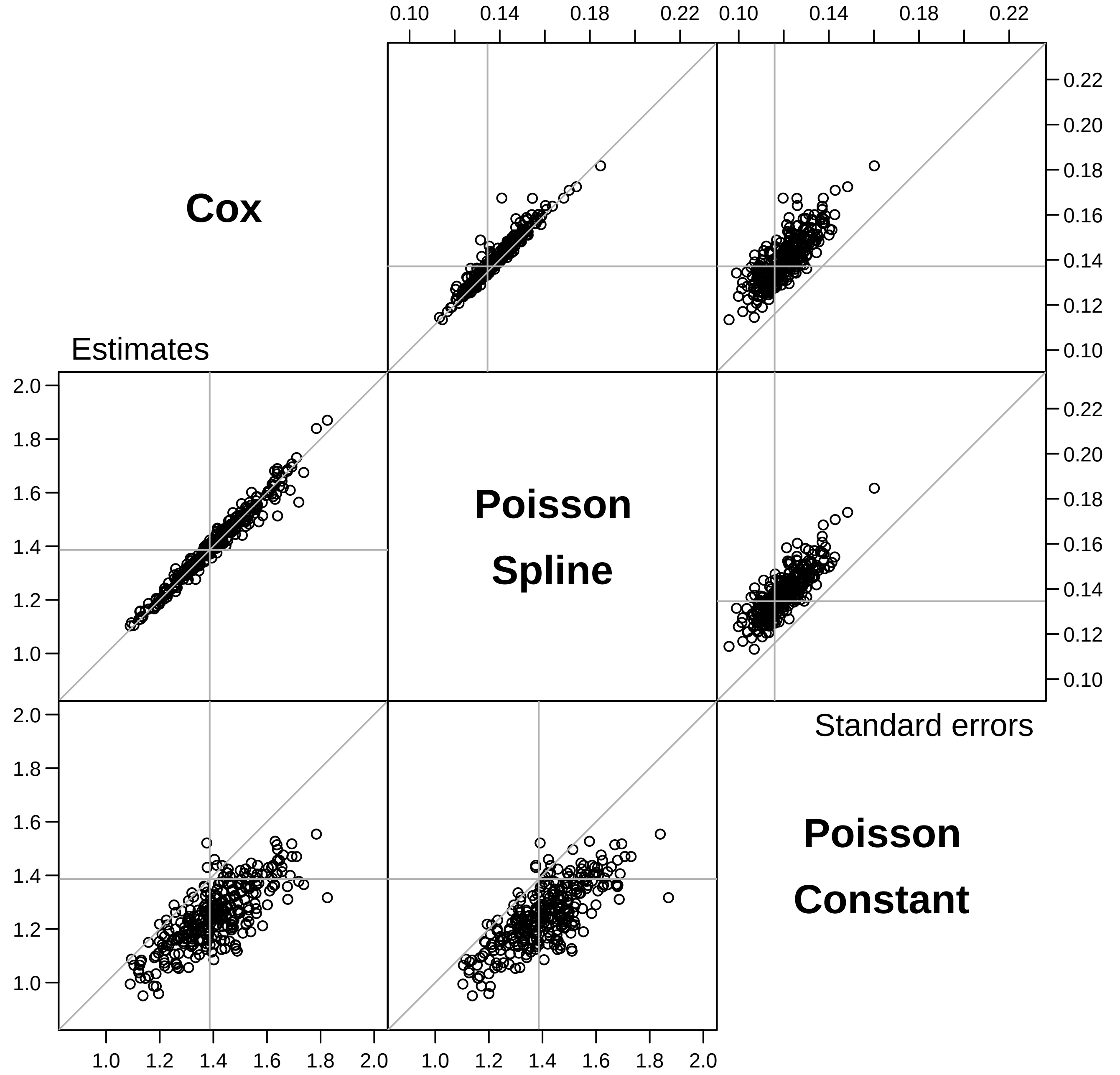

Hazard
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Time
0
2
4
6
8
10

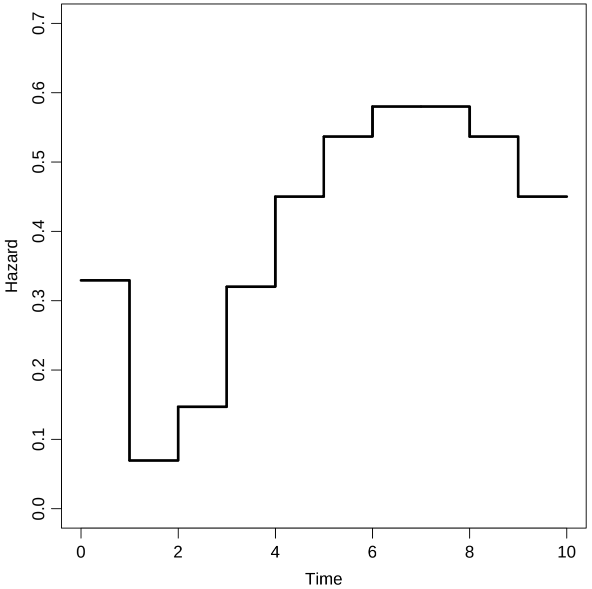

Cox
Poisson
Spline
Poisson
Constant
Estimates
Standard errors
1.0
1.2
1.4
1.6
1.8
2.0
0.10
0.12
0.14
0.16
0.18
0.20
0.22

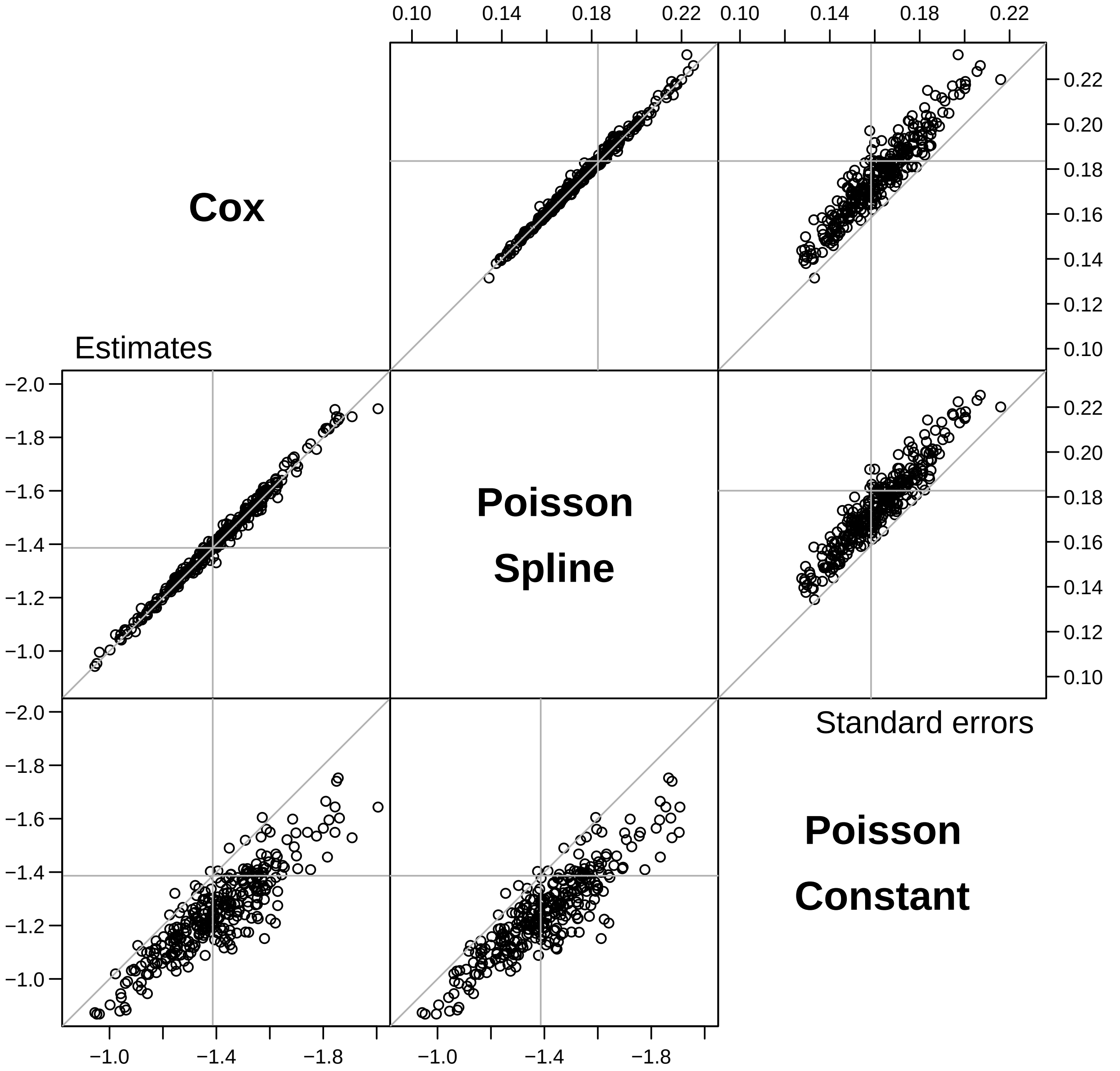

Cox
Poisson
Spline
Poisson
Constant
Estimates
Standard errors
0.10
0.14
0.18
0.22
0.12
0.16
0.20
−1.0
−1.2
−1.4
−1.6
−1.8
−2.0


Hazard
Time
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8
10


Hazard
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Time
0
2
4
6
8
10


Cox
Poisson
Spline
Poisson
Constant
Estimates
Standard errors
0.10
0.14
0.18
0.22
0.10
0.12
0.14
0.16
0.18
0.20
0.22
1.0
1.2
1.4
1.6
1.8
2.0


Hazard
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Time
0
2
4
6
8
10


Cox
Poisson Spline
Poisson Constant
Estimates
Standard errors
0.10
0.14
0.18
0.22
0.12
0.16
0.20
−2.0
−1.8
−1.6
−1.4
−1.2
−1.0

Conslusion

- Cox-model and Poisson model gives the same results.

- Even with very few parameters to accomodate the time effect.

- Modelling of the effect of time must be proper.

Types of estimates: Survival function

The Cox-model allows estimation of a survival function for a person with a given set of covariates.
Confidence intervals well described.

Usually calculated as c.i. for

$$\Lambda(t) = -\log(S(t))$$

or

$$\log(\Lambda(t)) = \log(-\log(S(t)))$$

and then backtransformed.

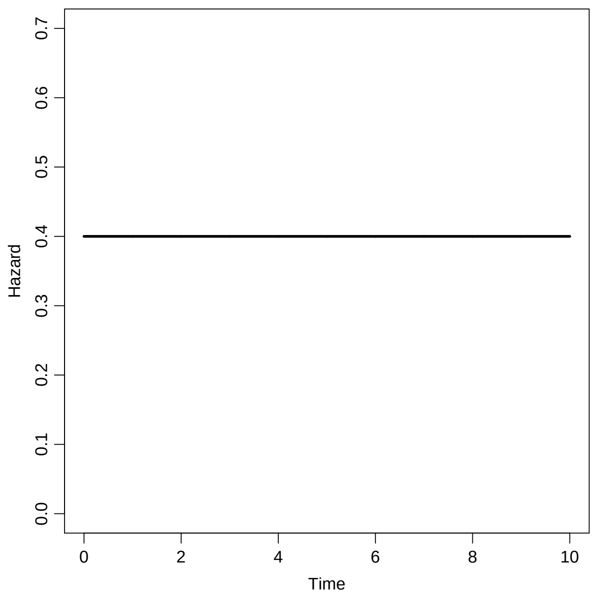
With a parametric model we estimate $\log(\lambda(t))$ as a linear function of parameters.

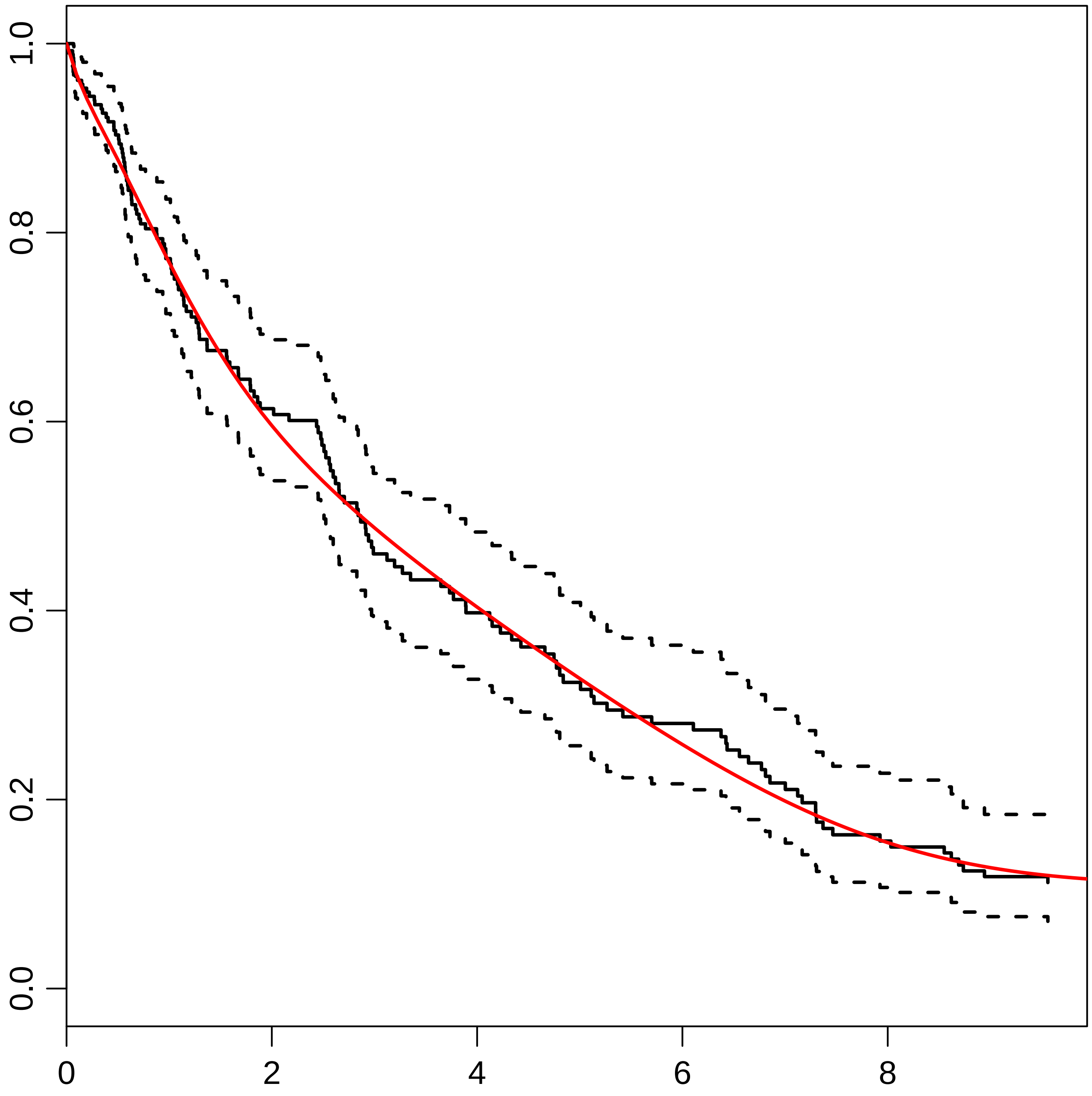
Hence we can estimate $\lambda(t)$ and the cumulative sum (integral), as e.g.:

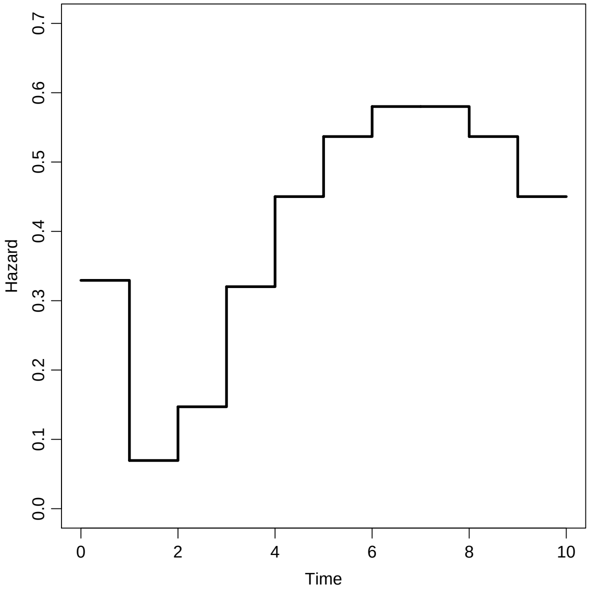
$$\hat{\Lambda}(t) = \sum \exp(\hat{\alpha}_t) \times l$$

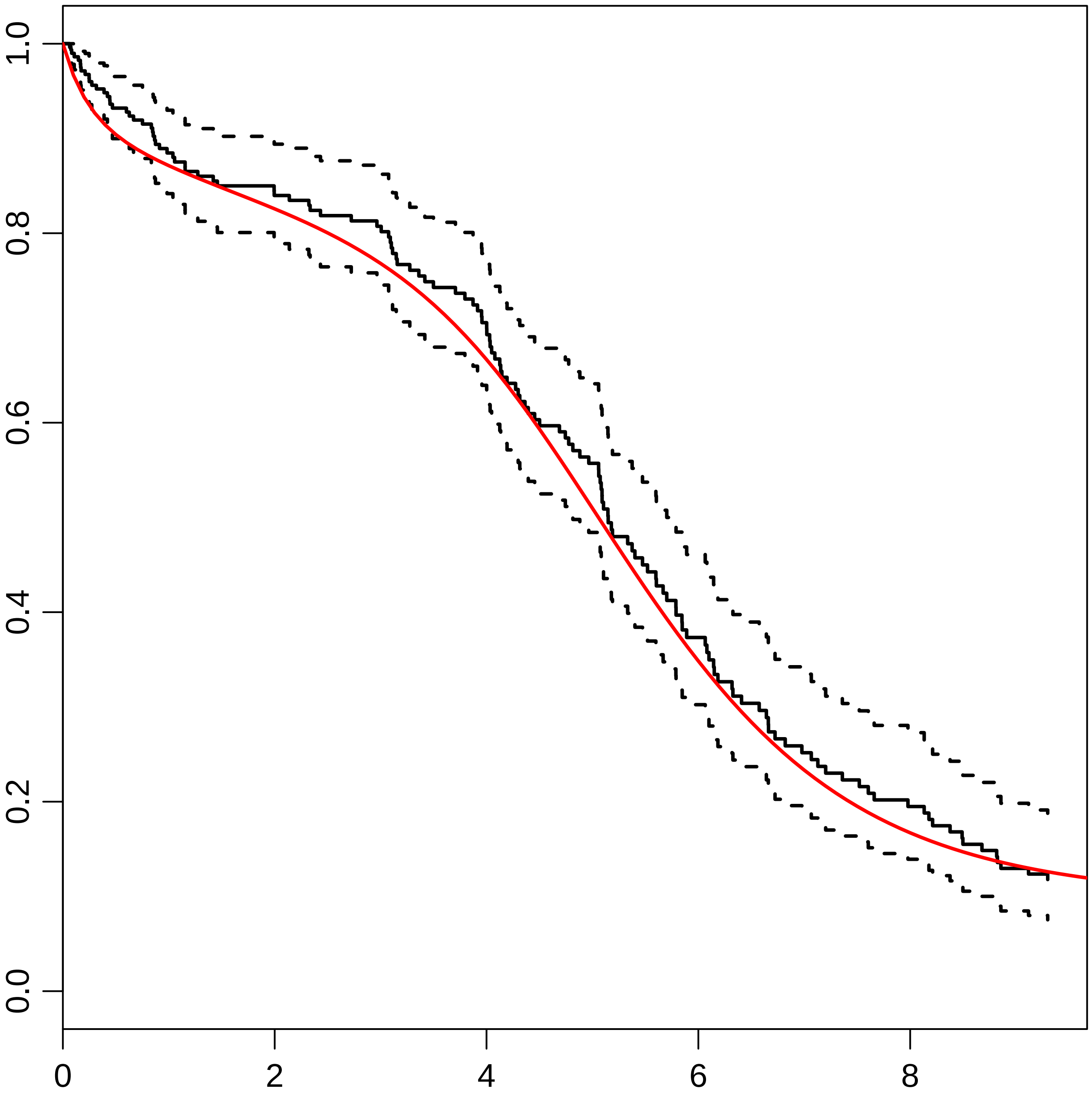
where α_t is the constant log-rate in an interval around t.

But standard errors of Λ in this formulation is a major headache...


Hazard
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8
10
Time


0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8


Hazard
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8
10
Time

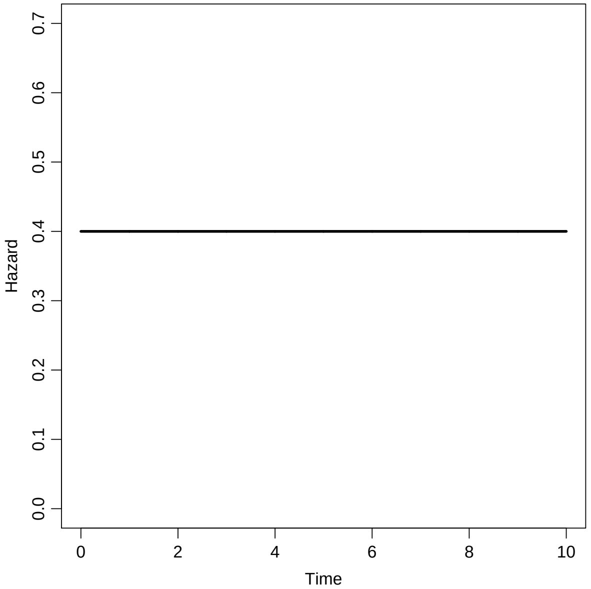

0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8

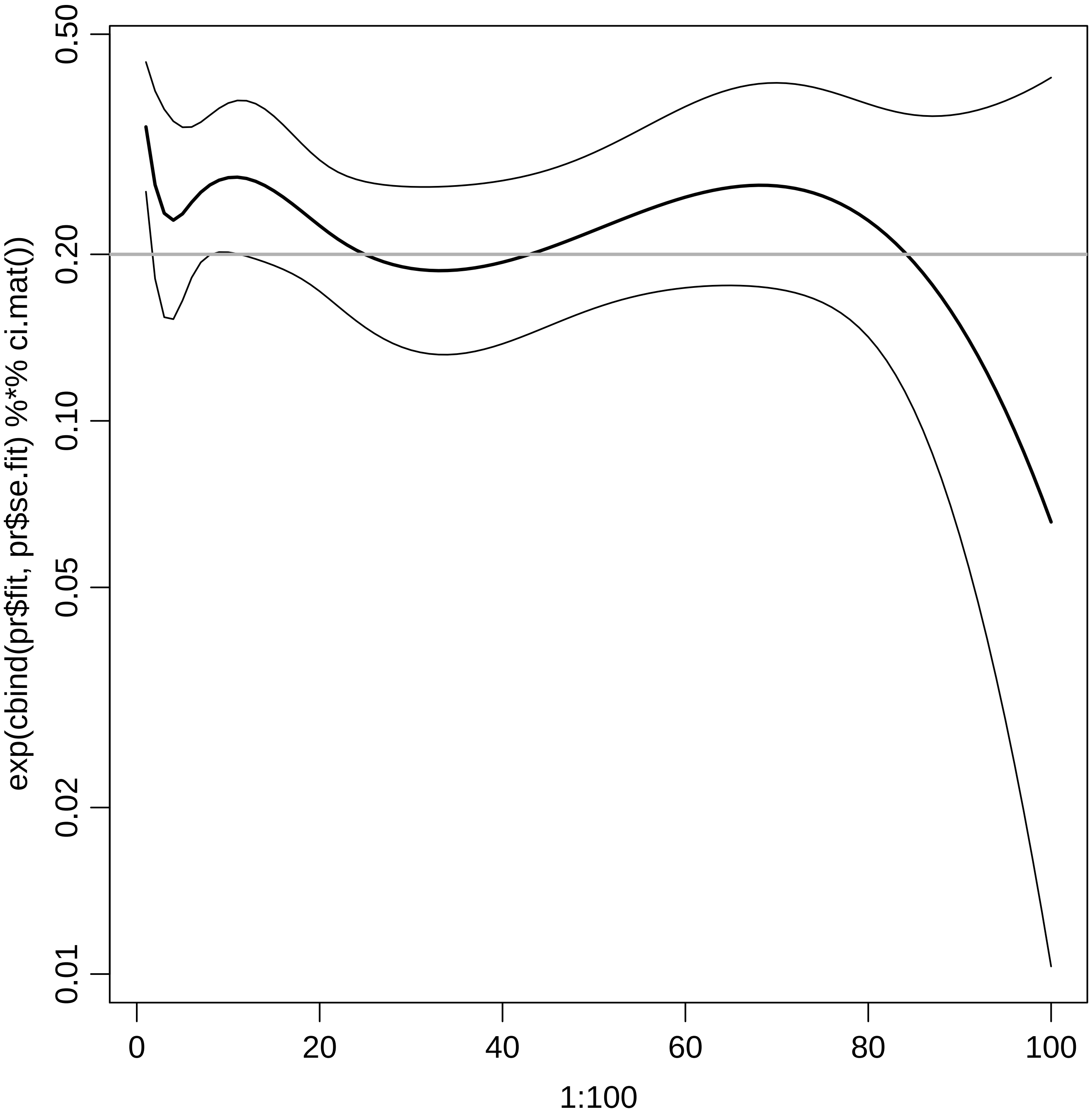
The underlying rate

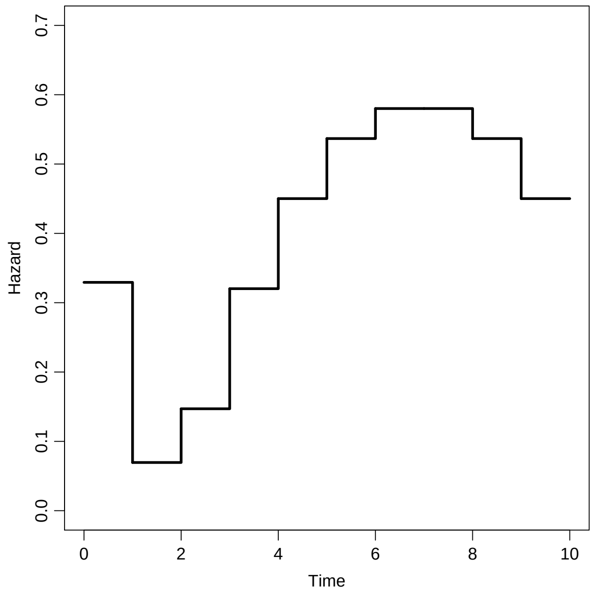
A large litterature is concerned with smoothing survival/cumulative hazards in order to produce estimates of the underlying hazard function.

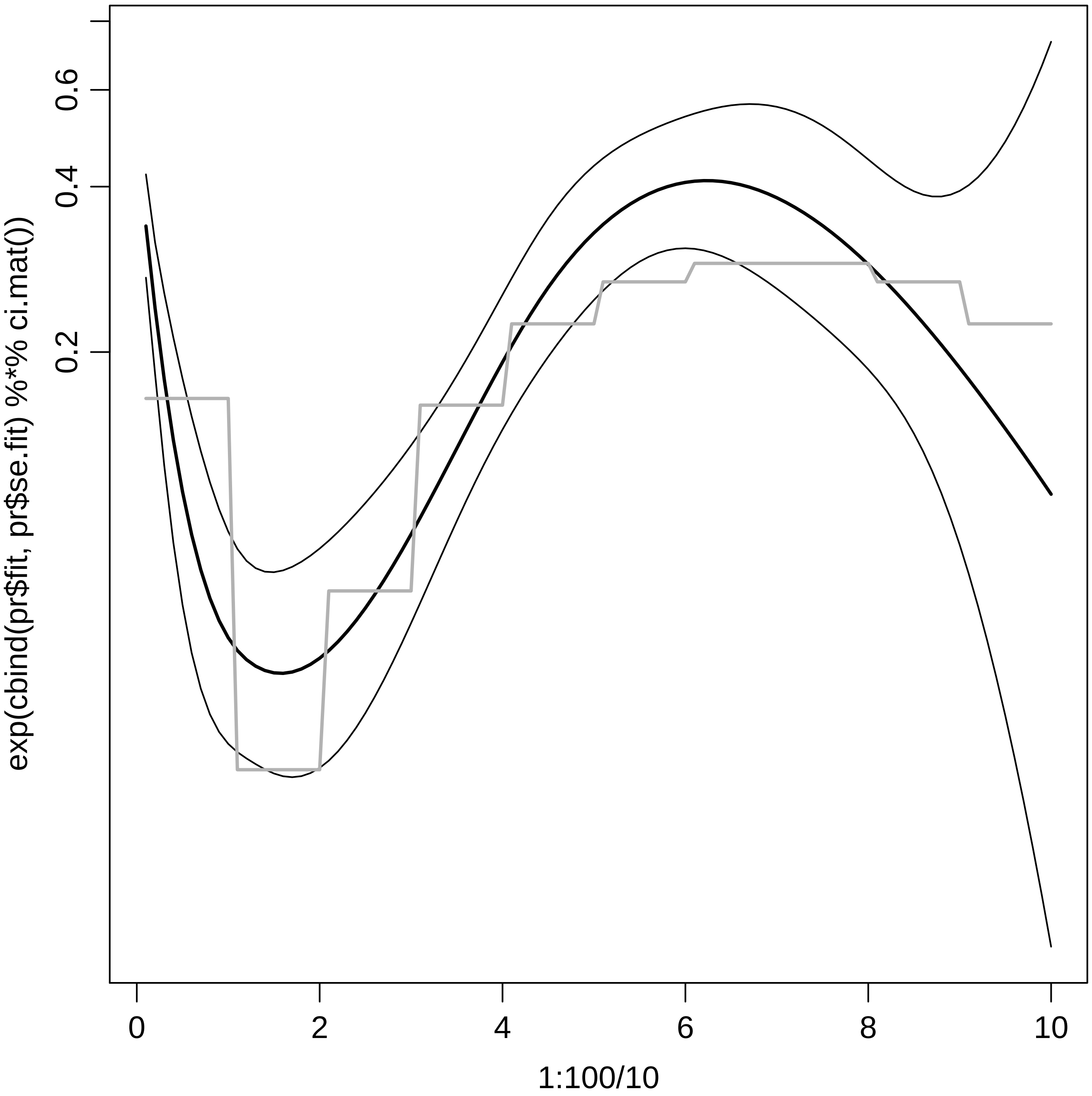
This is like crossing the river to water the horses.

The underlying hazard is best estimated directly by a parametric model jointly with the relative risks associated with covariates.

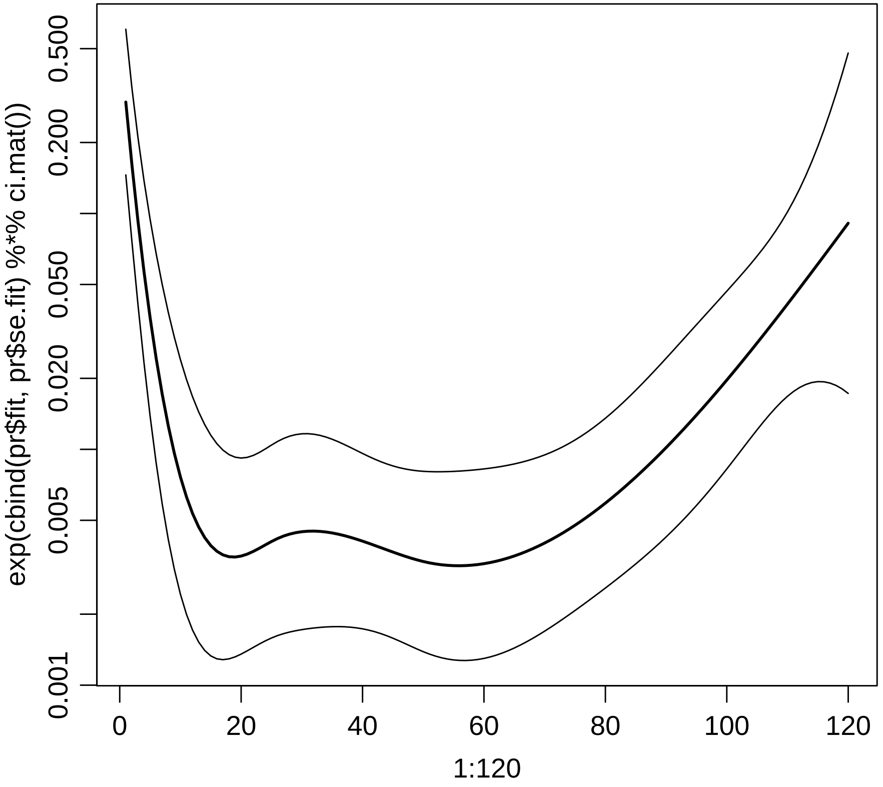

Hazard
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Time
0
2
4
6
8
10

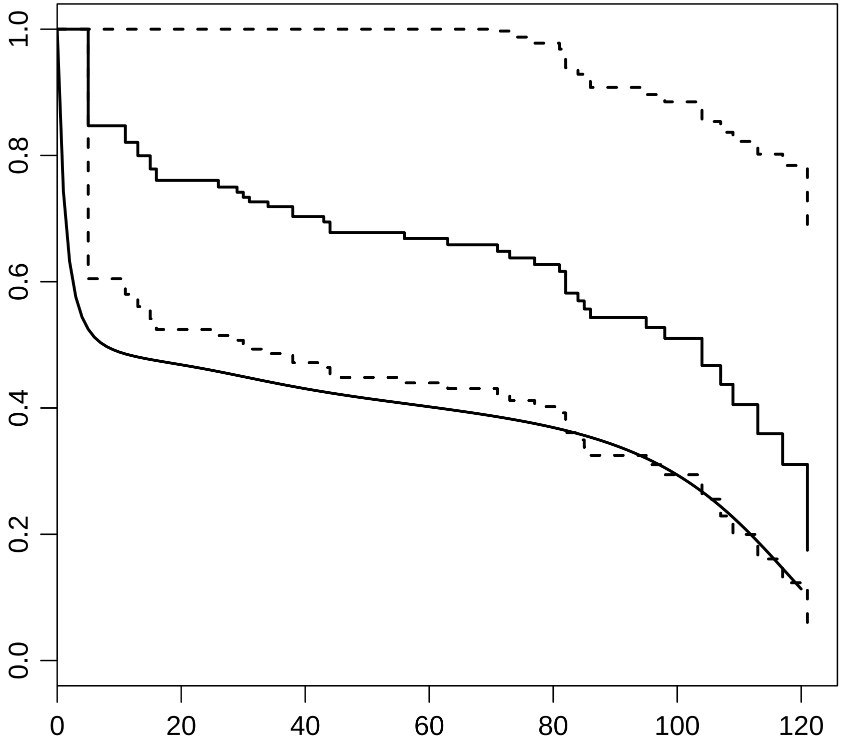

exp(cbind(prfit, prse.fit) %*% ci.mat())
0.01
0.02
0.05
0.10
0.20
0.50
1:100
0
20
40
60
80
100


Hazard
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8
10
Time


exp(cbind(prfit, prse.fit) %*% ci.mat())
0.2
0.4
0.6
0
2
4
6
8
10
1:100/10

Example of baseline problems





Why are the two survival estimates so different?

Which model

- Cox-model:
  - Survival function estimates desired. (No late entries).
  - One well-defined (major) time scale.
  - Religiously founded obsession with overparametrization.

- Poisson-model:
  - Hazard function estimates desired.
  - Any number of time scales.
  - Sound judgement of parametrization.

Survival function for whom?

Estimating a survival function from a Cox-model requires a specification of a set of covariates:

- Stata: All set to 0.
- R/ Splus: The population mean of parameters.
- SAS: The population mean of parameters.

All choices may be problematic.

Relative survival revisited

Compute the expected survival for each patient from diagnosis and average these (on the probability scale). This is the expected survival for this group of patients:

$$\frac{1}{n}\sum_i S_{i\textsf{pop}}(t) = \frac{1}{n}\sum_i \exp(-\Lambda_{i\textsf{pop}}(t))$$

Observed survival:

$$\frac{1}{n}\sum_i \hat{S}_i(t) = \frac{1}{n}\sum_i \exp(-\hat{\Lambda}_i(t))$$

Relative survival

The classical relative survival does:

$$\text{RSR} = \frac{\frac{1}{n}\sum_i \exp(-\hat{\Lambda}_i(t))}{\frac{1}{n}\sum_i \exp(-\Lambda_{i\mathsf{pop}}(t))}$$

The excess risk modelling does:

$$\frac{\exp(-\frac{1}{n}\sum_i \hat{\Lambda}_i(t))}{\exp(-\frac{1}{n}\sum_i \Lambda_{i\mathsf{pop}}(t))}$$

Not the same thing. But the modelling possibilities in the latter may be preferable.

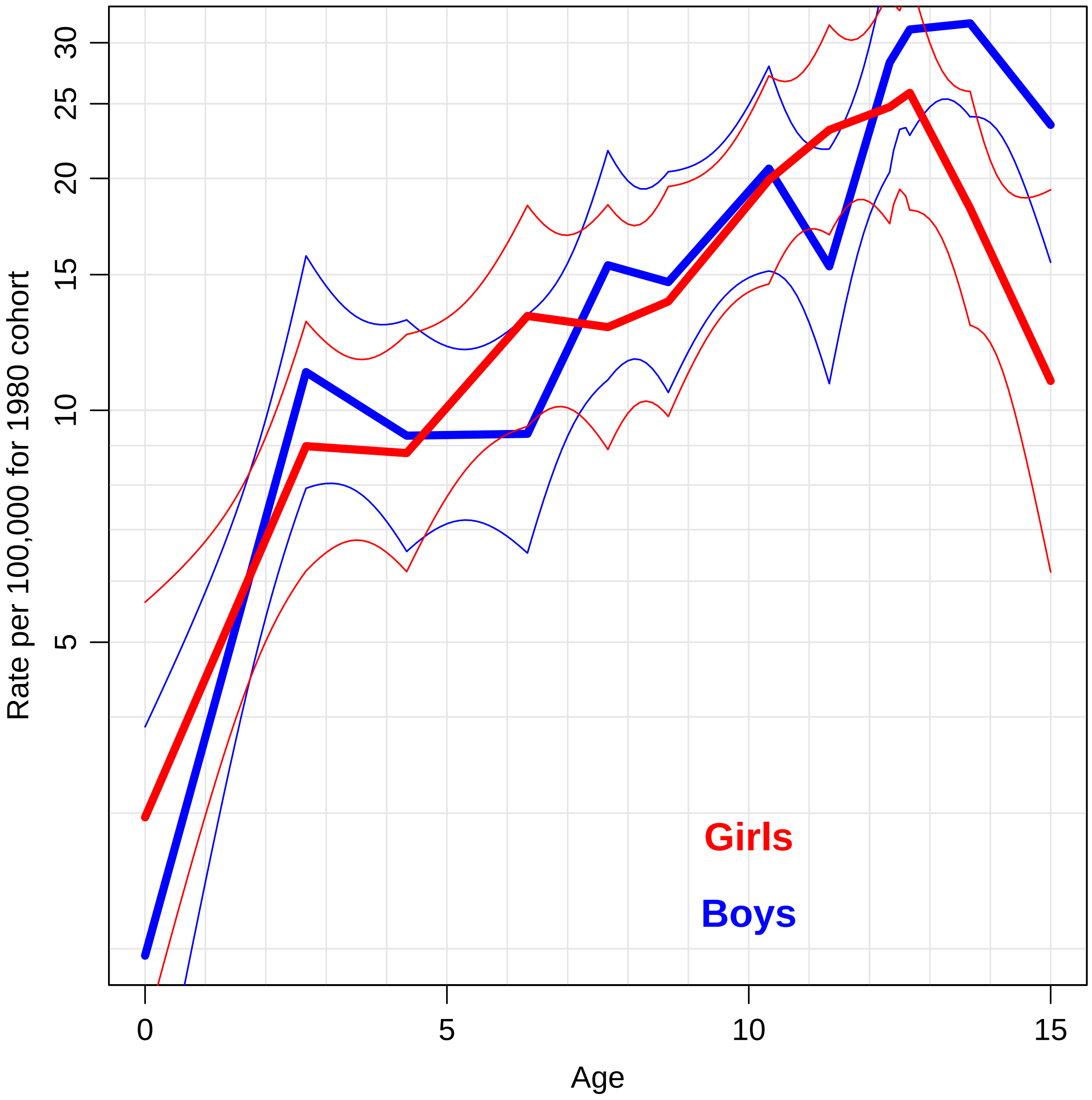
Survival analysis and medical demography

Age-Period-Cohort Models

1 May 2003

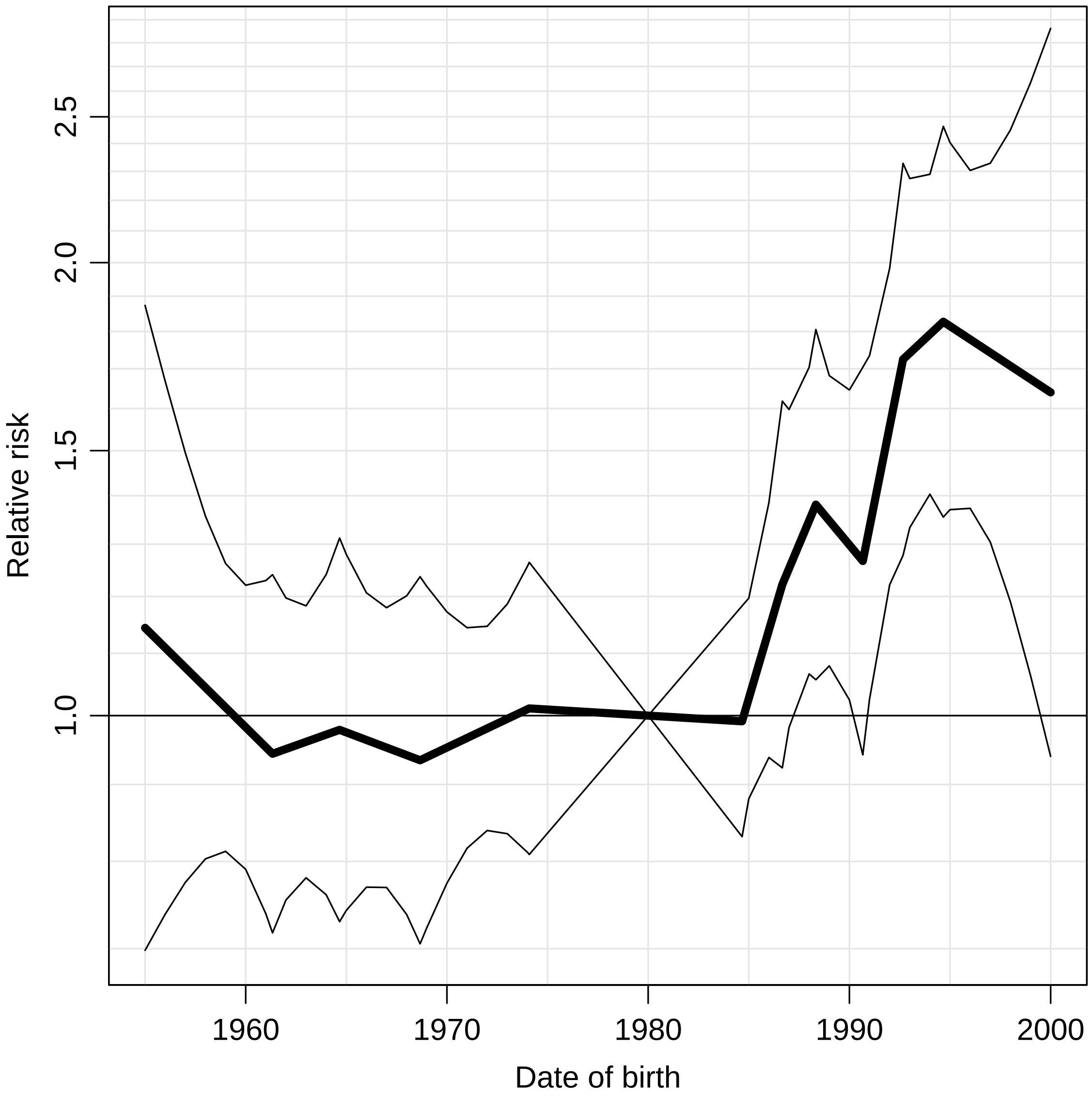
Bendix Carstensen

Incidence rates of IDDM in Denmark



The best fitting model is one with separate age-effects for males and females,

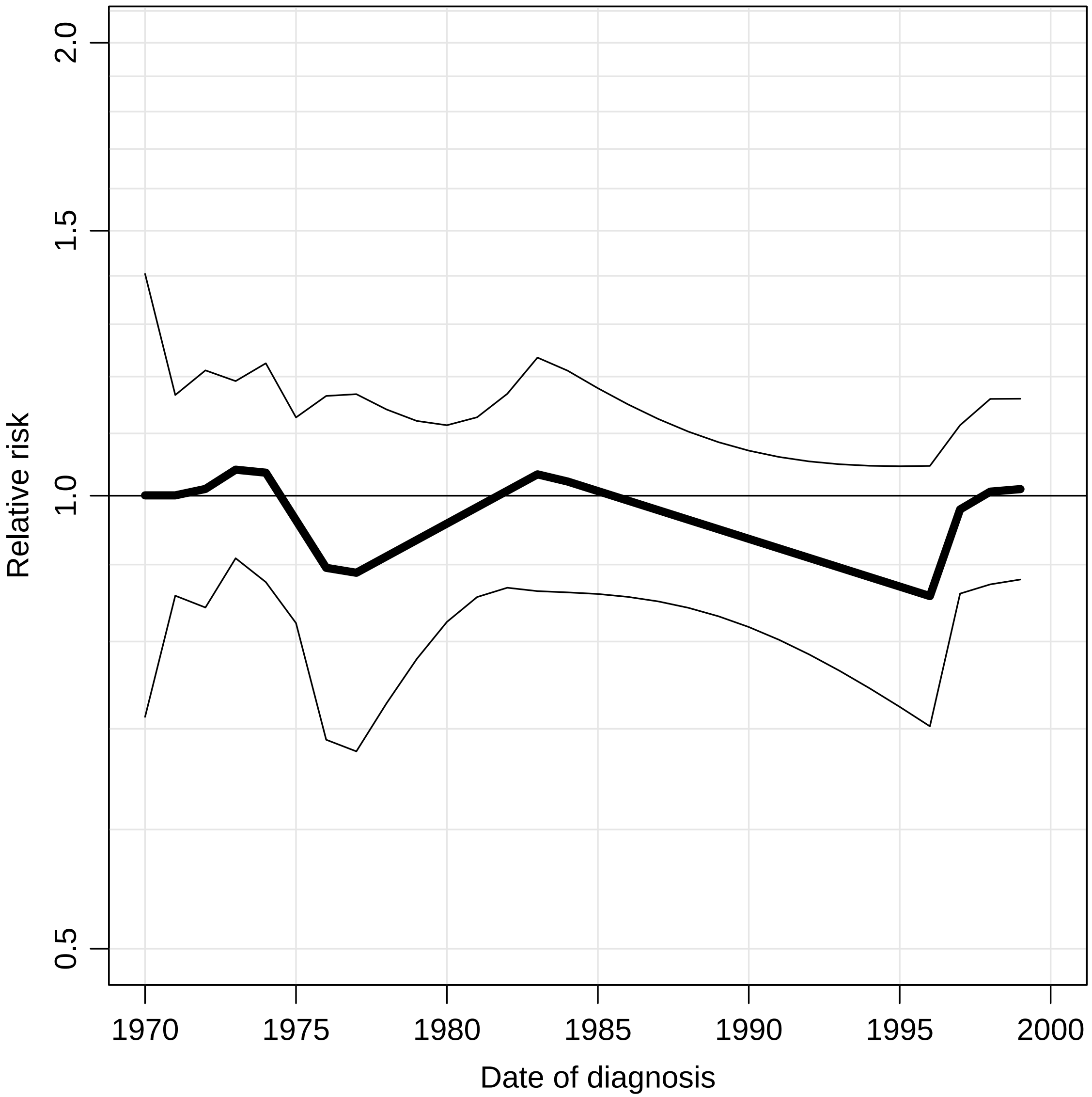
Incidence rates of IDDM in Denmark



with a common effect of birth cohort:

$$\log(\lambda_{aps}) = f_s(a)+g(p-a)$$

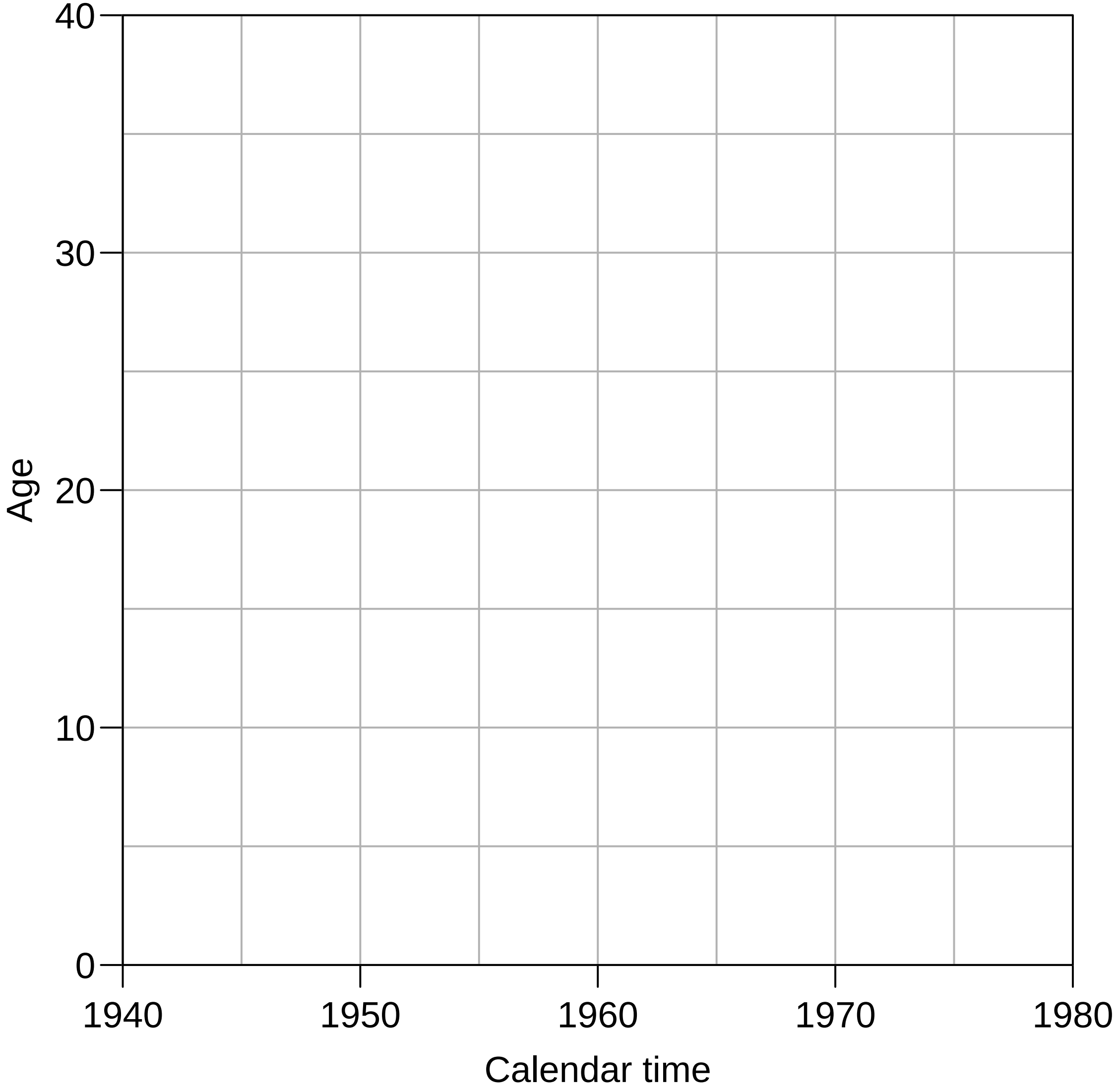
Incidence rates of IDDM in Denmark



and virtually no period effect:

$$\begin{aligned} h(p) &= \log(\lambda_{aps}) \\ &- (\hat{f}_s(a) + \hat{g}(p - a \end{aligned}$$

Lexis diagram [1]



Disease registers record events.

Official statistics collect population data.

[1] Named after the German statistician and economist **William Lexis** (1837–1914), who devised this diagram in the book "Einleitung in die Theorie der Bevölkerungsstatistik" (Karl J. Trübner, Strassburg, 1875).

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.:

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

Register data - rates

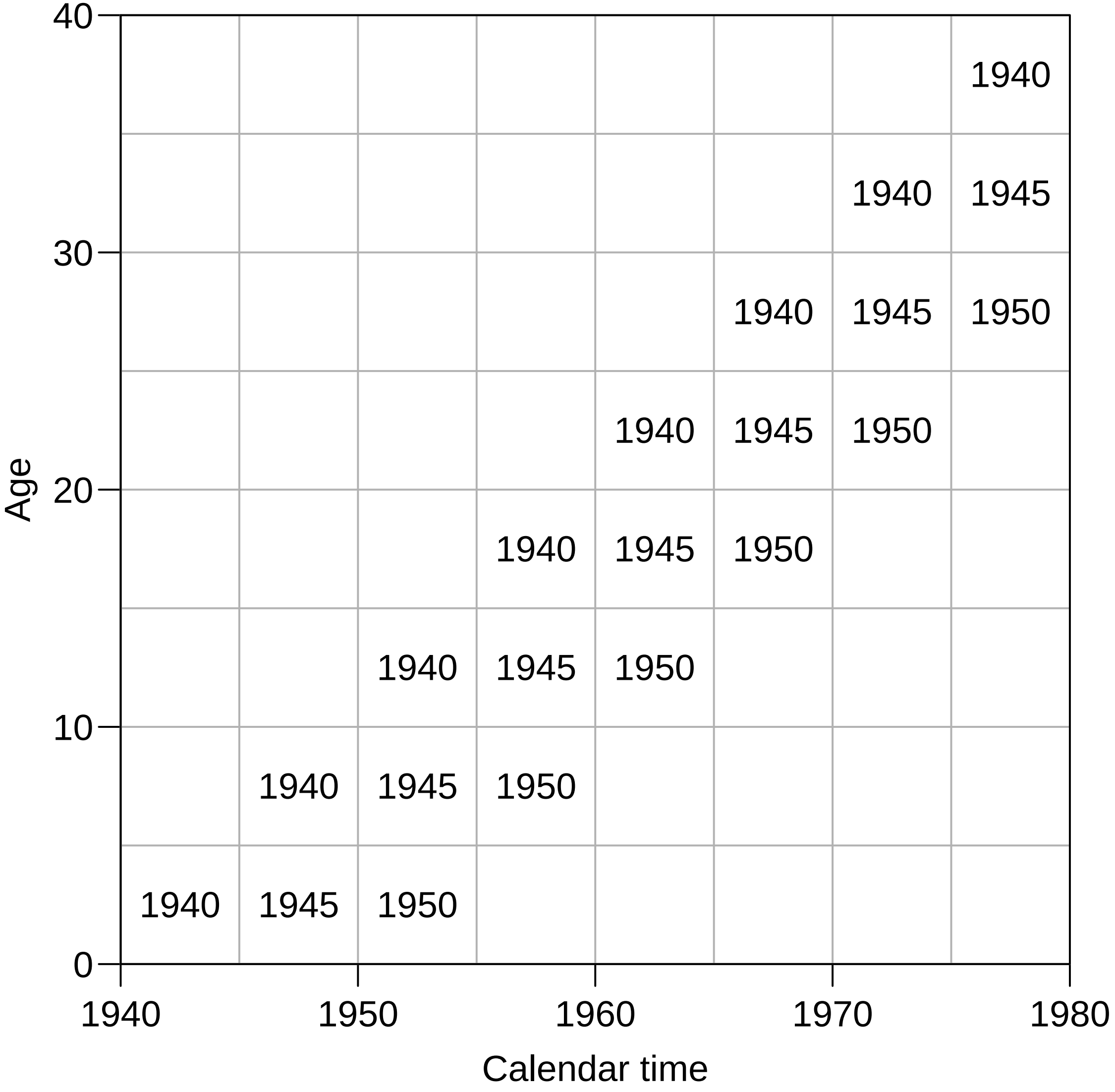
Thus we have access to rates in “tiles” of the Lexis daigram:

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

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

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

Synthetic cohorts



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

Successively overlapping 10-year periods.

Disease rates in a Lexis diagram

Three variables (factors) involved:

- Age at diagnosis, A (Age) — Current age.
- Date of diagnosis, P (Period) — Current date.
- Date of birth, C (Cohort)

$c = p - a$ produce an identifiability / parametrization problem.

APC-models based on factors from tables.

A multiplicative model for rates has the same likelihood as a Poisson model for the mean of disease counts D_{ap}:

$$\log[\mathrm{E}(D_{ap})] = \log(Y_{ap}) + \alpha_a + \beta_p + \gamma_c$$

The linear relationship between A, P and C induce a linear constraint, so the model has dimension:

$$1 + (A-1) + (P-1) + (C-1) - 1 = A + P + C - 3$$

$C = A + P - 1$, so the dimension is $2(A+P) - 4$.

Factor modelling

Only second order differences (curvature) are identifiable:

$$\alpha_a - 2\alpha_{a+1} + \alpha_{a+2}$$

In the APC-factor model there are

$$(A-2)+(P-2)+(C-2)$$

such invariants, leaving 3 dimensions unaccounted for.

Parametrizing by curvature

Age factor, parametrized by invariants:

age	contrast matrix					linear predictor
	μ	δ	ζ_2	ζ_3	ζ_4	
1	1	0	0	0	0	$\eta_1 = \mu$
2	1	1	0	0	0	$\eta_2 = \mu + 1\delta$
3	1	2	1	0	0	$\eta_3 = \mu + 2\delta \ + 1\zeta_2$
4	1	3	2	1	0	$\eta_4 = \mu + 3\delta \ + 2\zeta_2 + 1\zeta_3$
5	1	4	3	2	1	$\eta_5 = \mu + 4\delta \ + 3\zeta_2 + 2\zeta_3 + 1\zeta_4$

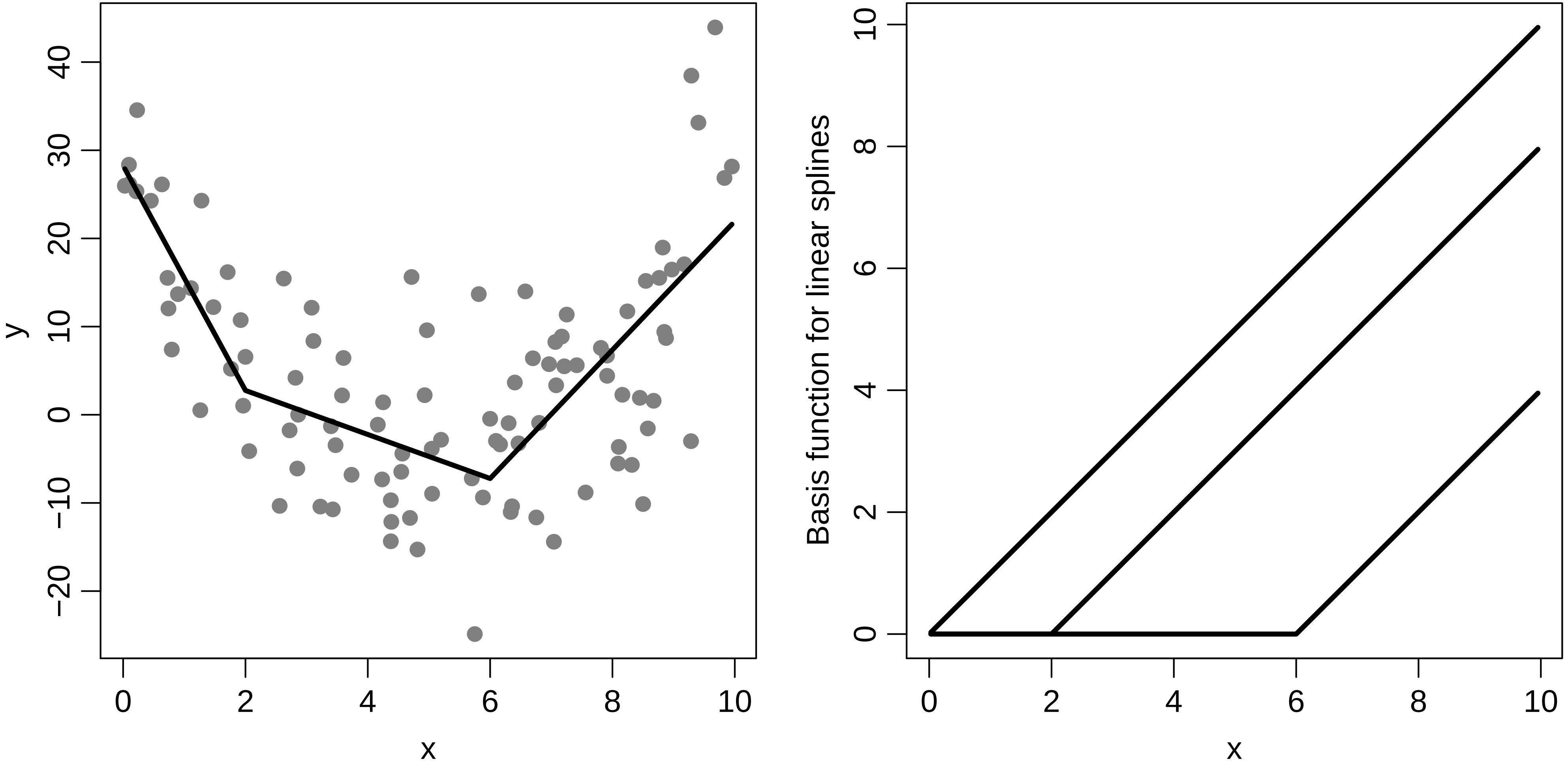
$$\Longrightarrow \quad \eta_3 - 2\eta_4 + \eta_5 = \zeta_5, \quad \text{etc.}$$

Digression: Linear splines

Linear regression with a piecewise linar function. If the curve breaks at “knots” k_1 and k_2, say, use covariates:

$$x \qquad \max(0, x - k_1) \qquad \max(0, x - k_2)$$

```
   x      max(0,x-2.5)      max(0,x-6.0)
 1.0               0.0               0.0
 2.0               0.0               0.0
 3.0               0.5               0.0
 4.0               1.5               0.0
 5.0               2.5               0.0
 6.0               3.5               0.0
 7.0               4.5               1.0
 8.0               5.5               2.0
 9.0               6.5               3.0
10.0               7.5               4.0
```



Coefficients are **changes** in slope at each knot.

Linear splines and 2nd order differences

age	contrast matrix					linear predictor
	μ	δ	ζ_2	ζ_3	ζ_4	
1	1	0	0	0	0	$\eta_1 = \mu$
2	1	1	0	0	0	$\eta_2 = \mu + 1\delta$
3	1	2	1	0	0	$\eta_3 = \mu + 2\delta\ + 1\zeta_2$
4	1	3	2	1	0	$\eta_4 = \mu + 3\delta\ + 2\zeta_2 + 1\zeta_3$
5	1	4	3	2	1	$\eta_5 = \mu + 4\delta\ + 3\zeta_2 + 2\zeta_3 + 1\zeta_4$

This is just a design matrix for linear splines, with knots at the middle of the tabulation intervals.

Modelling

Common approach:
One parameter per age / period / cohort.
Only feasible with coarse tables (many cases per cell).

Alternative: fine tabulation, e.g. 1-year classes, with the model describing rates as a "smooth" function of
mean age (a),
mean date of diagnosis (p) and
mean date of birth (c).

No need to tabulate only by age and period in equidistant classes. Any subdivision of the Lexis diagram will work.

APC-model in general

A model describing incidence rates by three functions:

$$\begin{aligned}\log(\lambda_{ap}) &= f(a) + g(p) + h(c) \\ &= [f(a) + \delta a] + \\ &\quad [g(p) - \delta p] + \\ &\quad [h(c) + \delta c]\end{aligned}$$

Any function of the form $f(a) + \delta a$ can be used for age-effect, and still give the same model.

The linear component of f, g and h cannot be determined.

Data and the model

An APC-model describes how rates vary by age, period and cohort.

The description cannot be more detailed than data.

A fine tabulation of data (cases and population) will still allow a rather coarse model. (=few parameters), but not vice versa.

The amount of data (no. of cases) should guide the number of parameters, but not the tabulation. And the number of tabulation intervals should not guide the number of parameters.

The common tabulation curse

Much confusion has arisen from the tight marriage between tabulation and modelling — one parameter per interval.

On top of this age classes has traditionally been numbered $a = 1, \ldots, A$, periods $p = 1, \ldots, P$ and cohorts $c = 1, \ldots, C$, causing funny relations like $c = p - a + A$.

Better to label classes by means of age, period and cohort.

Information in factor level parametrization.

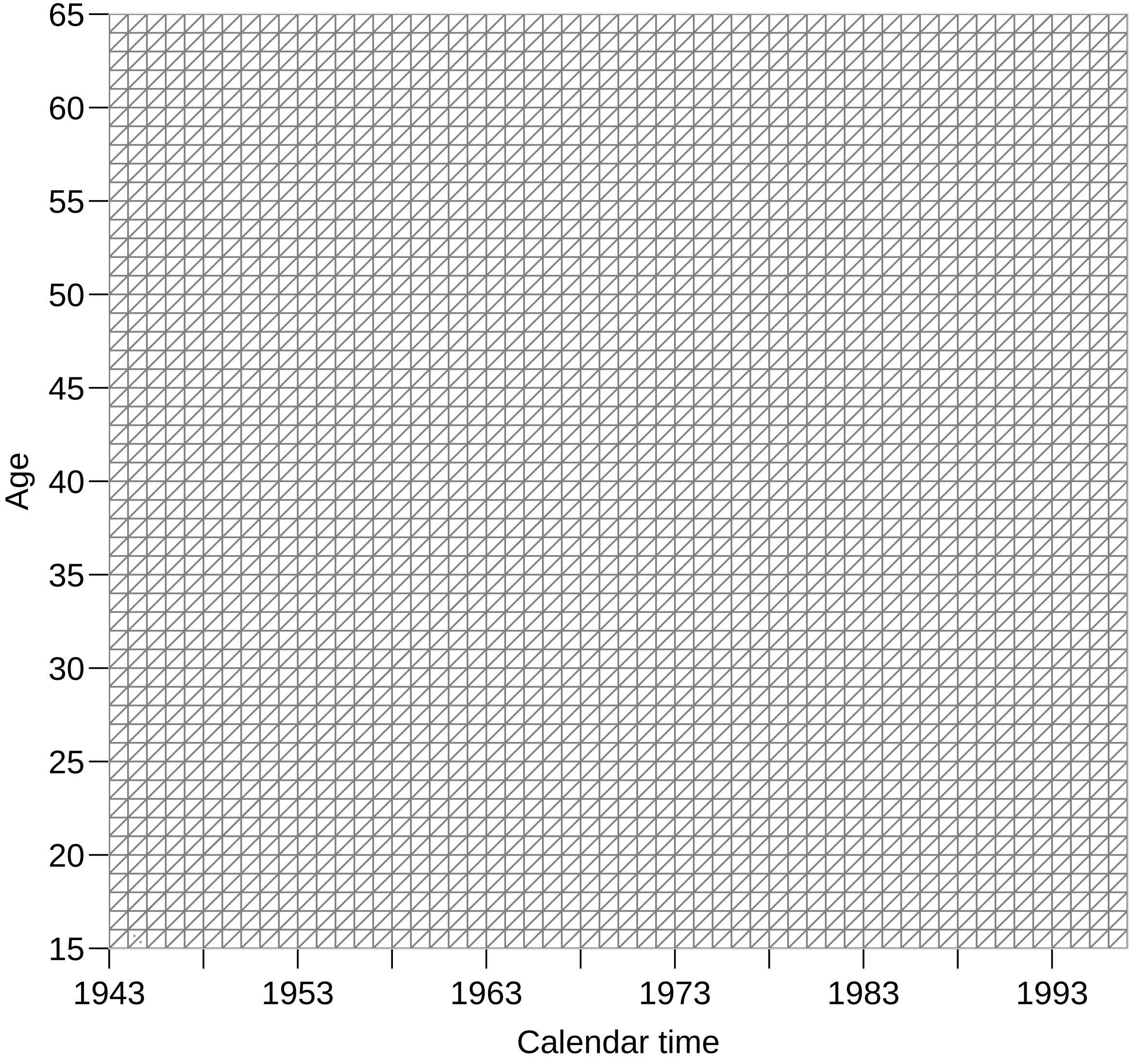
Age class information vary by the disease rate variation by age (usually a lot).

Period classes are usually quite balanced in size.

Cohort classes vary both by age-variation in disease rates and by the sampling frame for the data. The youngest and the oldest cohort are only represented by one cell each.

Not reasonable to parametrise by factor levels: knots for linear splines are put in places whith little information.

Tabulation of Seminomas of testis in Denmark



Tabulation by age, period and date of birth in 1-year classes.

5400 cells.

4461 cases.

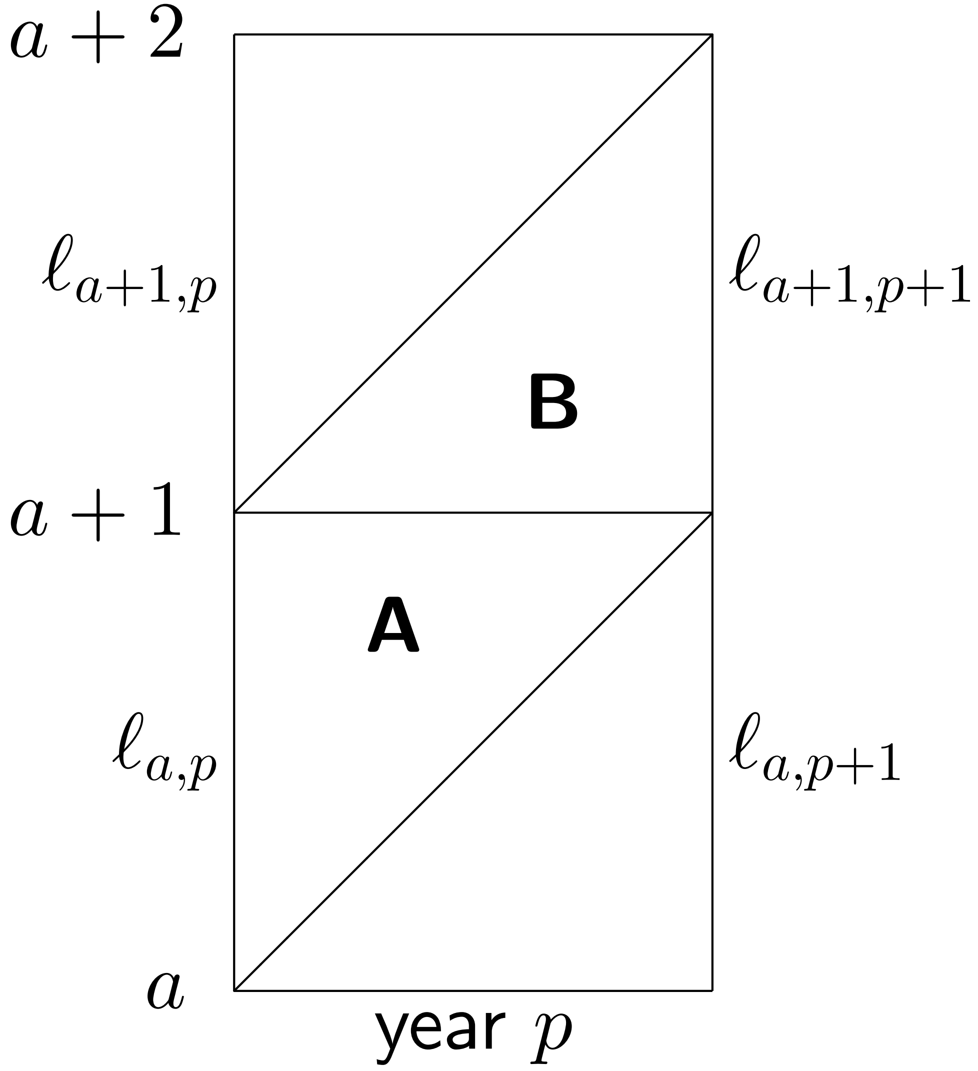
Population risk time in triangles

Upper triangle (**A**):

$\frac{1}{3}\ell_{ay} + \frac{1}{6}\ell_{a+1,y+1}$

Lower triangle (**B**):

$\frac{1}{6}\ell_{a-1,y} + \frac{1}{3}\ell_{a,y+1}$.



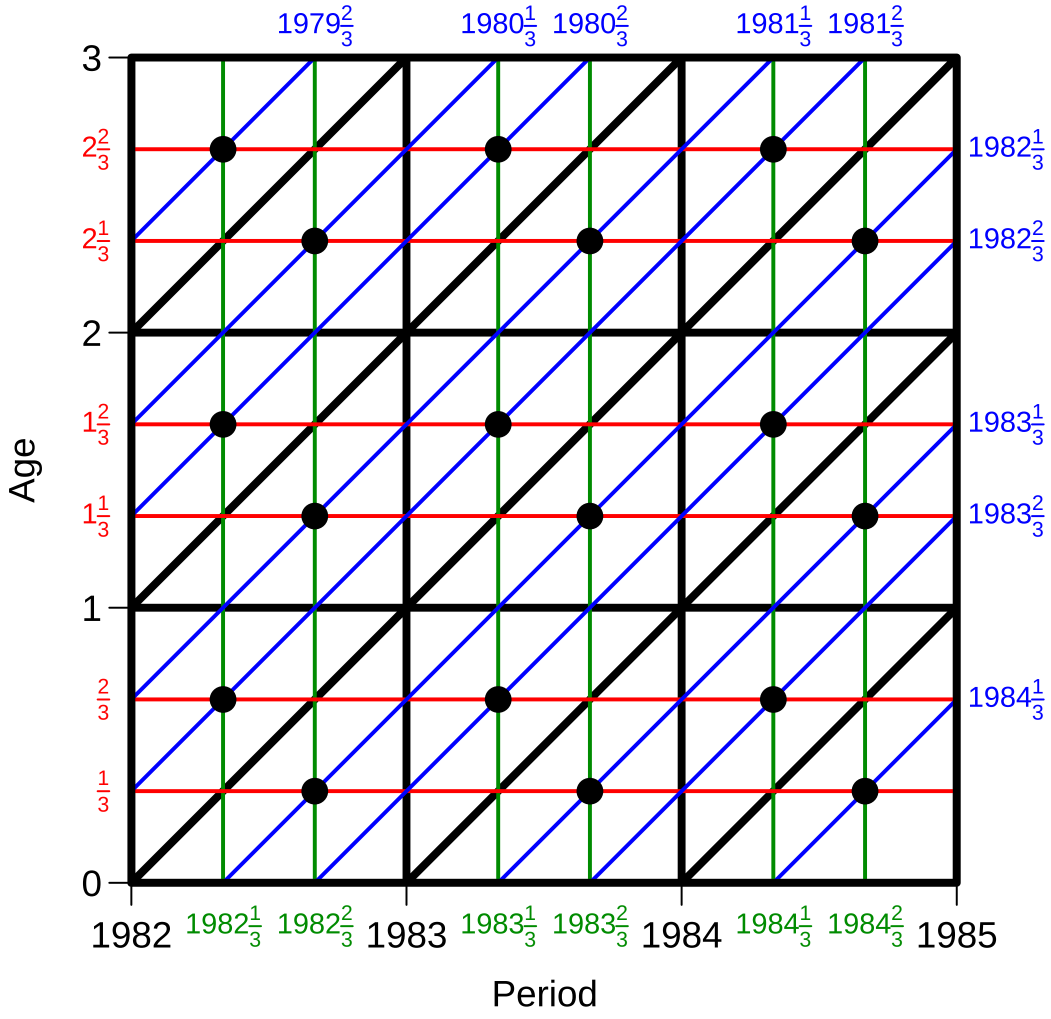
Average age, period and cohort in triangles

$$\mathrm{E}_{\mathsf{A}}(a) = \int_{p=0}^{p=1} \int_{a=p}^{a=1} 2a \, da \, dp = \int_{p=0}^{p=1} 1 - p^2 \, dp = \frac{2}{3}$$

$$\mathrm{E}_{\mathsf{A}}(p) = \int_{a=0}^{a=1} \int_{p=0}^{p=a} 2p \, dp \, da = \int_{a=0}^{a=1} a^2 \, dp = \frac{1}{3}$$

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

Correct coding of age, period and cohort



If tabulation is in triangles, use:
– mean age,
– mean date of diagnosis
– mean date of birth.

Using the class midpoints give an erroneous model, see e.g. Clayton & Schifflers.

Factor modelling for triangles

Tabulation by triangles of the Lexis diagram gives $2A$ different age-means, $2P$ different period means and $2C-2$ cohort means.

Setting up a model with these as factor actually results in a likelihood which is a product two likelihoods:

One for the upper triangles and one for the lower. Impossible to report.

Suggestion by Holford

We can write the model:

$$\begin{aligned} f(a)+g(p)+h(c) &= \tilde{f}(a)+\hat{\mu}_a+\hat{\delta}_a a+ \\ &\quad \tilde{h}(c)+\hat{\mu}_c+\hat{\delta}_c c+ \\ &\quad \tilde{g}(p)+\hat{\mu}_p+\hat{\delta}_p p \end{aligned}$$

i.e. extract any linear function from f, h and g.

Holfords point was to let $f(a)=\alpha_a$ and $\hat{\mu}_a, \hat{\delta}_a$ be the regression of α_a on a; similarly for p and c.

That is, extract the linear trends. But not very helpful in describing the rates.

Putting things back together

1. Intercept (which carries the rate-dimension) plus the age-terms, with reference to a specific cohort, c_0.

   Age-specific incidence rates in cohort c_0.

2. Linear effect of cohort (or period).

   Usually termed “drift”.

3. Non-linear effect of cohort.

4. Non-linear effect of period.

$$\begin{aligned} f(a) + g(p) + h(c) &= \tilde{f}(a) + \hat{\mu}_a + \hat{\delta}_a a + \\ &\quad \tilde{h}(c) + \hat{\mu}_c + \hat{\delta}_c c + \\ &\quad \tilde{g}(p) + \hat{\mu}_p + \hat{\delta}_p p \end{aligned}$$

Put them back together in different order:

$$\begin{array}{rl} \text{Age:} & \tilde{f}(a) + \hat{\mu}_a + \hat{\mu}_c + \hat{\mu}_p + (\hat{\delta}_a + \hat{\delta}_p)a + (\hat{\delta}_c + \hat{\delta}_p)c_0 \\ \text{Drift:} & (\hat{\delta}_c + \hat{\delta}_p)(c - c_0) \\ \text{Cohort:} & \tilde{h}(c) \\ \text{Period:} & \tilde{g}(p) \end{array}$$

But no confidence intervals available (the parametrization is data-driven), and a hazzle to program.

A practical suggestion

The primary time scale in any descripive epidemiological study based on a disease register is age: Age-specific rates.

The second substantial question asked is what the time trend is. Addressed by fitting an age-drift-model:

$$\log(\lambda_{ap}) = f(a) + \delta(c - c_0) \qquad c = p - a$$

with $f(a)$ chosen as a linear spline function.
$f(a)$ is the age-specific incidence rates in the c_0 birth cohort.
δ is the "average annual change in disease rates".

Non-linear cohort effect

The non-linear (identifiable) effects of cohort, h, can be estimated as residuals:

$$\log(\lambda_{ap}) = \hat{f}(a) + \hat{\delta}(c - c_0) + h(c)$$

Fitted by taking $\log(Y) + \hat{f}(a) + \hat{\delta}(c - c_0)$ $(\mathrm{link}($fitted values$))$ as offset in a model for $h(c)$.

Not maximum likelihood, but cohort effects conditional on estimated age-effect and drift.
Same procedure for the period variable.

Gives standard errors (albeit **conditional**).

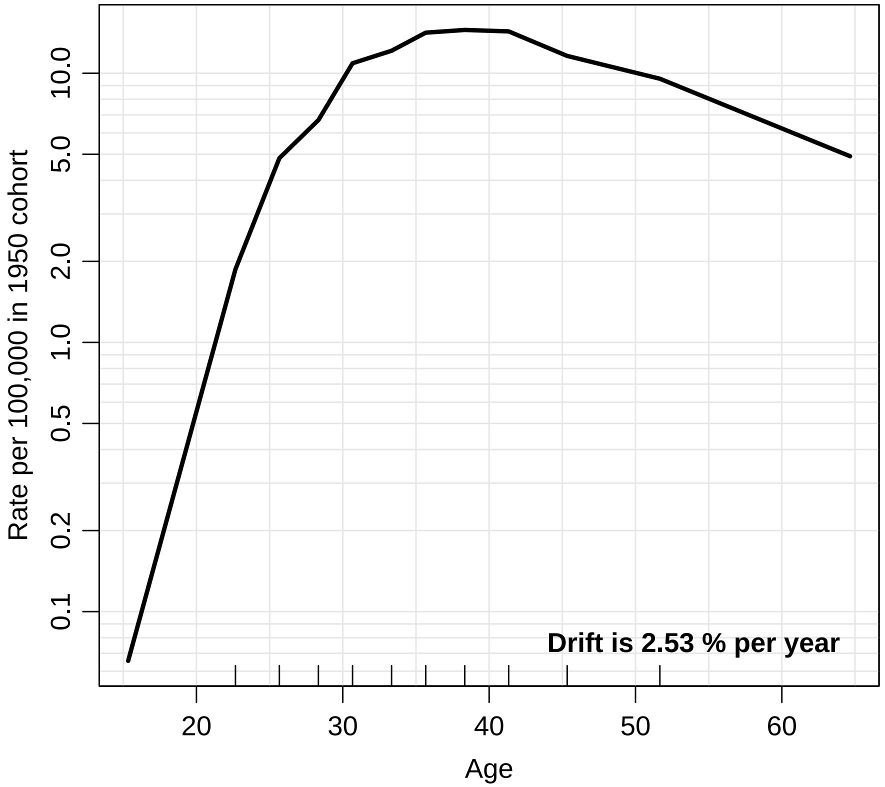
How to in R

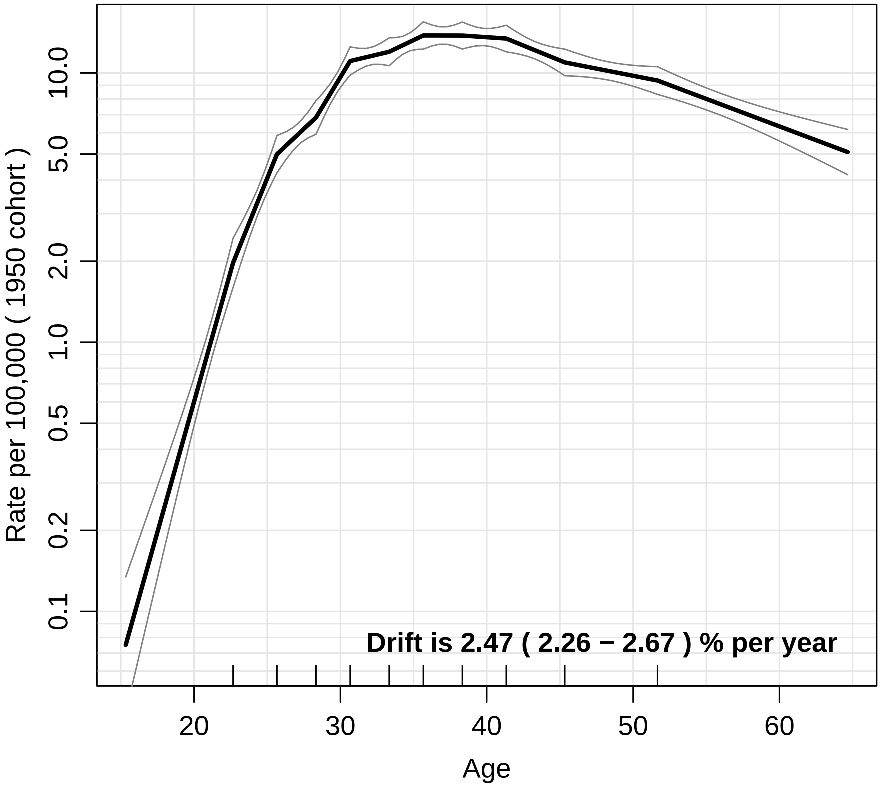
Assume `spl` is a function that generates a spline basis for a variable:

```
m.drift <- glm( D ~ spl(A) + I(C-c0) + offset(log(Y)),
                family=poisson )

m.coh   <- glm( D ~ spl(C) + offset(log(fitted(m.drift))),
                family=poisson )

m.per   <- glm( D ~ spl(P) + offset(log(fitted(m.coh))),
                family=poisson )
```

The rest is plotting of the estimates from these models.

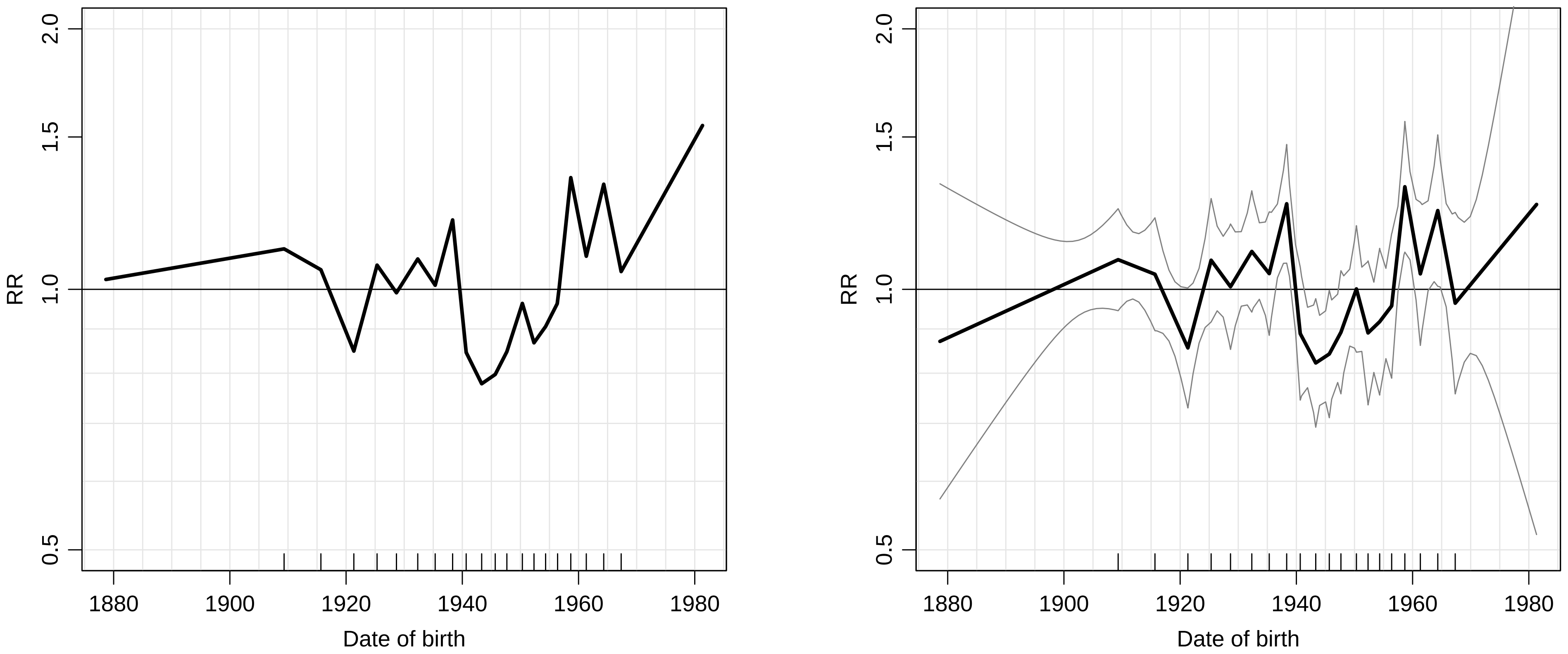
Maximum likelihood vs. conditional residuals





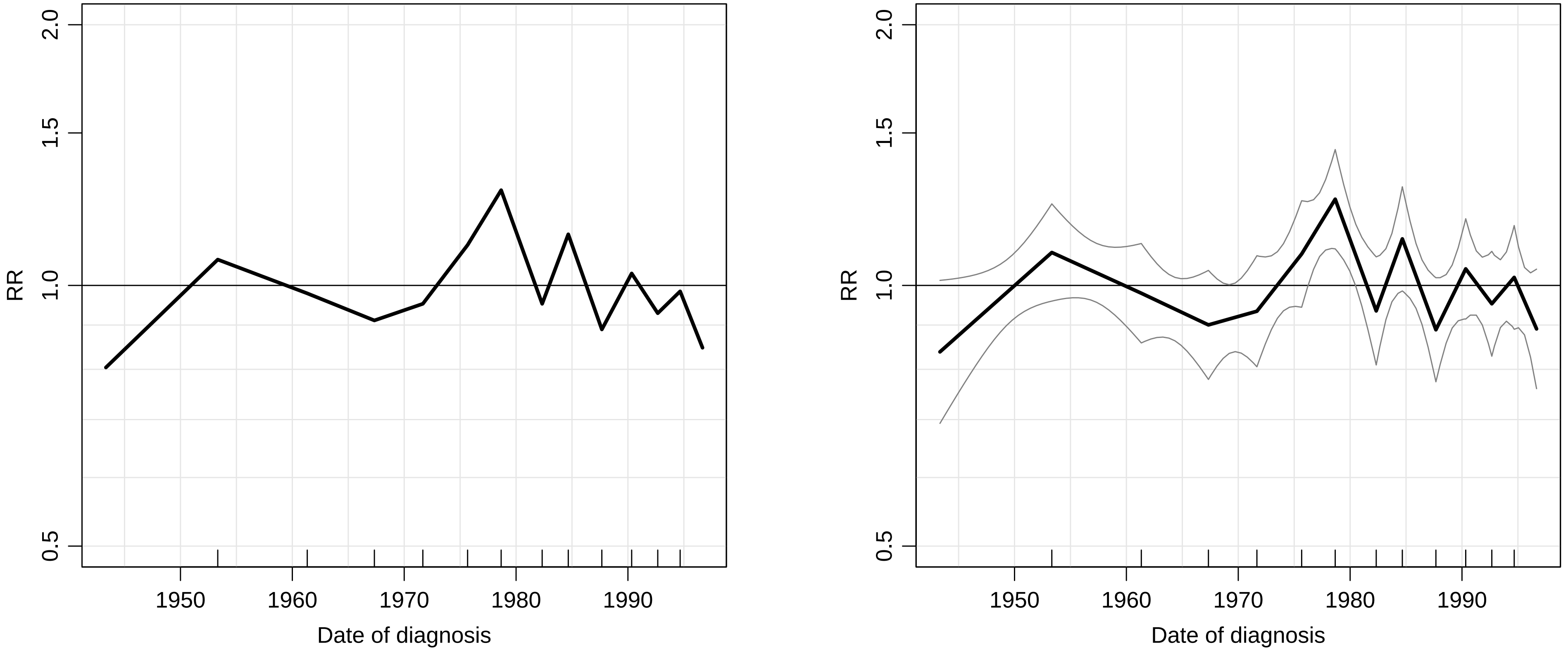
Maximum likelihood estimates where period and cohort effects constrained to be flat on average.

Maximum likelihood vs. conditional residuals



Maximum likelihood estimates where period and cohort effects constrained to be flat on average.

Maximum likelihood vs. conditional residuals



Maximum likelihood estimates where period and cohort effects constrained to be flat on average.

Parametrization recommendation:

- Tabulate data as finely as possible.

- Use parametric models for age, period and cohort effects. Splines are simple to implement in standard software. But, anything goes (C. Porter).

- Use an epidemiologically sensible sequence:
  - Age, drift
  - Cohort $|$ age, drift
  - Period $|$ age, drift, cohort

- Optionally, include drift in the cohort term.

Age-period-cohort models: Summary

There is a huge, largely confusing litterature on this.

Confusion mainly from the failure to recognize the inherently continuous nature of the problem — inference in the Lexis diagram.

Separate the modelling from the tabulation of data:
Tabulate data as finely as possible. **Then** model.
Reporting only possible with assumptions (decisions!) about which timescale is the more relevant.

A discussion in more depth with the central references is in:
`www.biostat.ku.dk/~bxc/Lexis/Lexis.pdf`

Survival analysis and medical demography

Interval Censoring

8 May 2003

Bendix Carstensen

Panel studies

A panel study is one where a set of people (the panel) is examined for some condition at specific points in time.

If it is an asymptomatic condition only detected by a test, the we only know that the person has acquired it between two testing times.

If data are complete we put:

$$\mathrm{P}\left\{\text{Event in interval } i\right\} = p_i$$

So the likelihood is a simple binomial likelihood for observations for each interval.

Assumptions

- Condition asymptomatic.

- Condition irreversible.

- Everyone is on the same timescale, calendar time.

  Other timesacales can be accomodated by introducing covariates in a logistic regression model for the p_is.

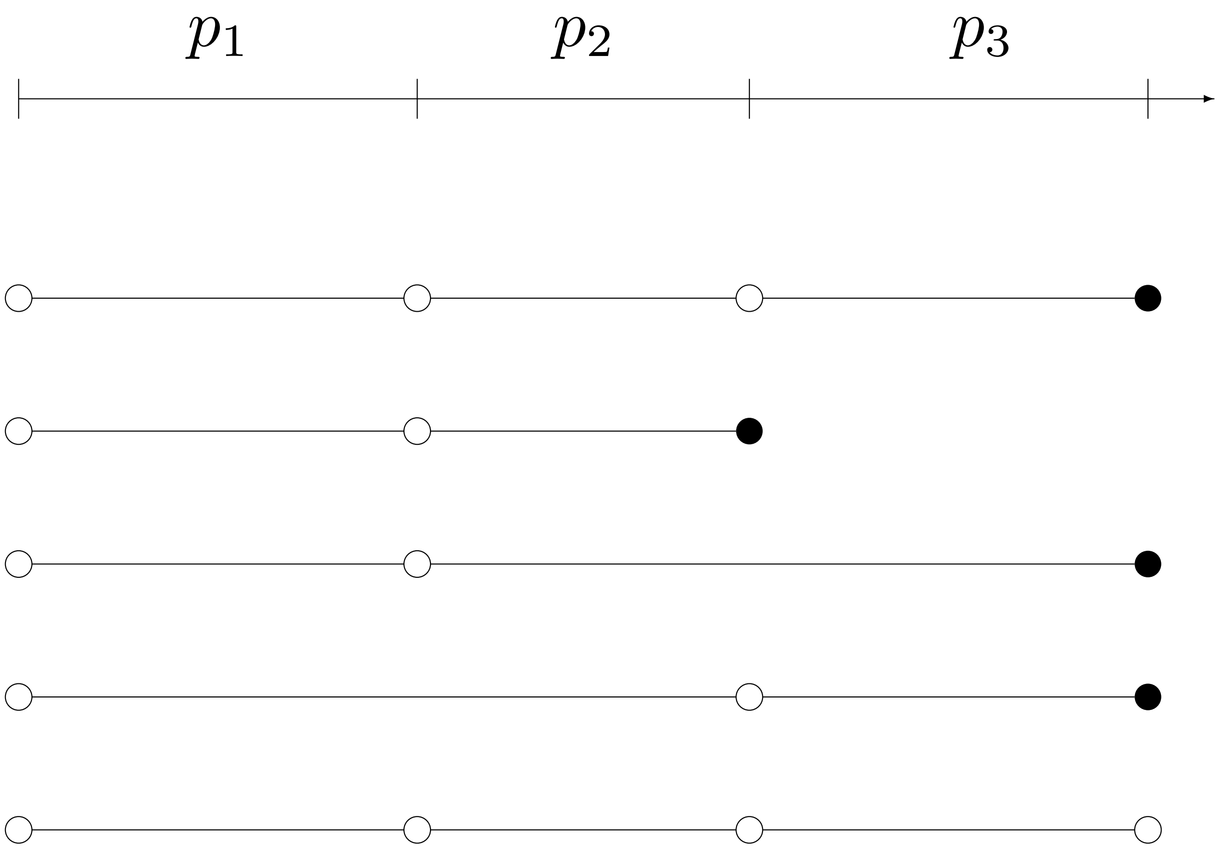
- Everyone appears every time.

If people fail to show up, there will be likelihood problems.

Panel study example

Survival probabilities

Likelihood contribution



$p_1p_2(1-p_3)$

$p_1(1-p_2)$

$p_1(1-p_2p_3)$

$p_1p_2(1-p_3)$

$p_1p_2p_3$

Peto's approach

The likelihood terms of the form p_i, ad $(1 - p_i)$ are just usual binomial likelihood terms.

The not-so-nice terms, $(1 - p_i p_{i+1})$, appear whenever a positive test appears after a missed visit.

Peto [**?**] essentially proposed just to maximize this resulting likelihood by standard methods for function optimization.

If the number of intervals is fairly small, this is feasible.

Becker's approach

Becker [?] instead parametrized by the cumulative hazards over the intervals:

$$p_i = \exp(-\Lambda_i) \qquad \Leftrightarrow \quad \Lambda_i = \log(p_i)$$

The likelihood would then be made up of terms of the form:

$$\exp(-\Lambda_i), \quad 1 - \exp(-\Lambda_i), \quad 1 - \exp\big(-(\Lambda_i + \Lambda_{i+1})\big)$$

These are just terms from a binomial likelihood for a model with log-link.

Becker's implementation

The likelihood contribution from person p will be the same as a likelihood from one or two Bernoulli trials, y_p, with success probability $\exp(-\sum \Lambda_i) = \exp(\eta)$, $\eta = \sum \beta_i x_{pi}$:

- Intervals survived: $y_p = 1$, $x_{pi} = -1$ for intervals survived.
- Intervals with event: $y_p = 0$, $x_{pi} = -1$ for intervals since last seen disease-free:

$$1 - \exp(\beta_2 x_2 + \beta_3 x_3) = 1 - \exp(-\beta_2)\exp(-\beta_3) = 1 - p_2 p_3$$

Generalized linear model: binomial error and log-link.

Interval censoring in general

Turnbull [?] gives an estimator for any pattern in the interval censoring, not just the panel study situation.

Estimator in the Kaplan-Meier / Nelson-Aalen tradition:

Essentially one parameter per distinct observation time.

Not easy to implement, let alone extend with covariates.

Becker's approach again

The likelihood that Becker set up was with the cumulative hazards over an interval as parameters.

If we instead assume that hazards are constant in each interval, we can estimate the hazards directly by replacing

$$x_i \quad \text{by} \quad x_i l_i$$

where l_i is the length of interval i.

Parametrization by $\Lambda_i = \lambda_i l_i$ or λ_i is just covariate scaling.

Makes more general parametric models for interval censored data possible.

The general formulation

Time of event is only known to be in an interval (t_w, t_d) — the person is last seen well at t_w and diseased at t_d.

Thus for person p we have three time-points:

t_{pe} — time of entry

t_{pw} — time last seen well

t_{pd} — time first seen diseased

Typical of asymptomatic conditions like carrier status where persons are tested at regular intervals.

Likelihood with constant rate

If the rate is constant, λ, the likelihood-contribution for person p is:

$$\begin{aligned} & \mathrm{P}\left\{\text{no event from } t_{pe} \text{ to } t_{pw} \ \right\} \times \\ & \left(1 - \mathrm{P}\left\{\text{no event from } t_{pw} \text{ to } t_{pd} \middle| \text{ no event till } t_{pw}\right\}\right) \ = \\ & \exp(-\lambda(t_{pw} - t_{pe})) \times \{1 - \exp(-\lambda(t_{pd} - t_{pw}))\} \end{aligned}$$

Binomial likelihood with $p = \exp(\lambda x)$ for two observations:

y	x
1	$-(t_{pw} - t_{pe})$
0	$-(t_{pd} - t_{pw})$

Likelihood for piecewise constant rates

If the rates are assumed constant in intervals, the likelihood from an interval where event for person p has occurred is:

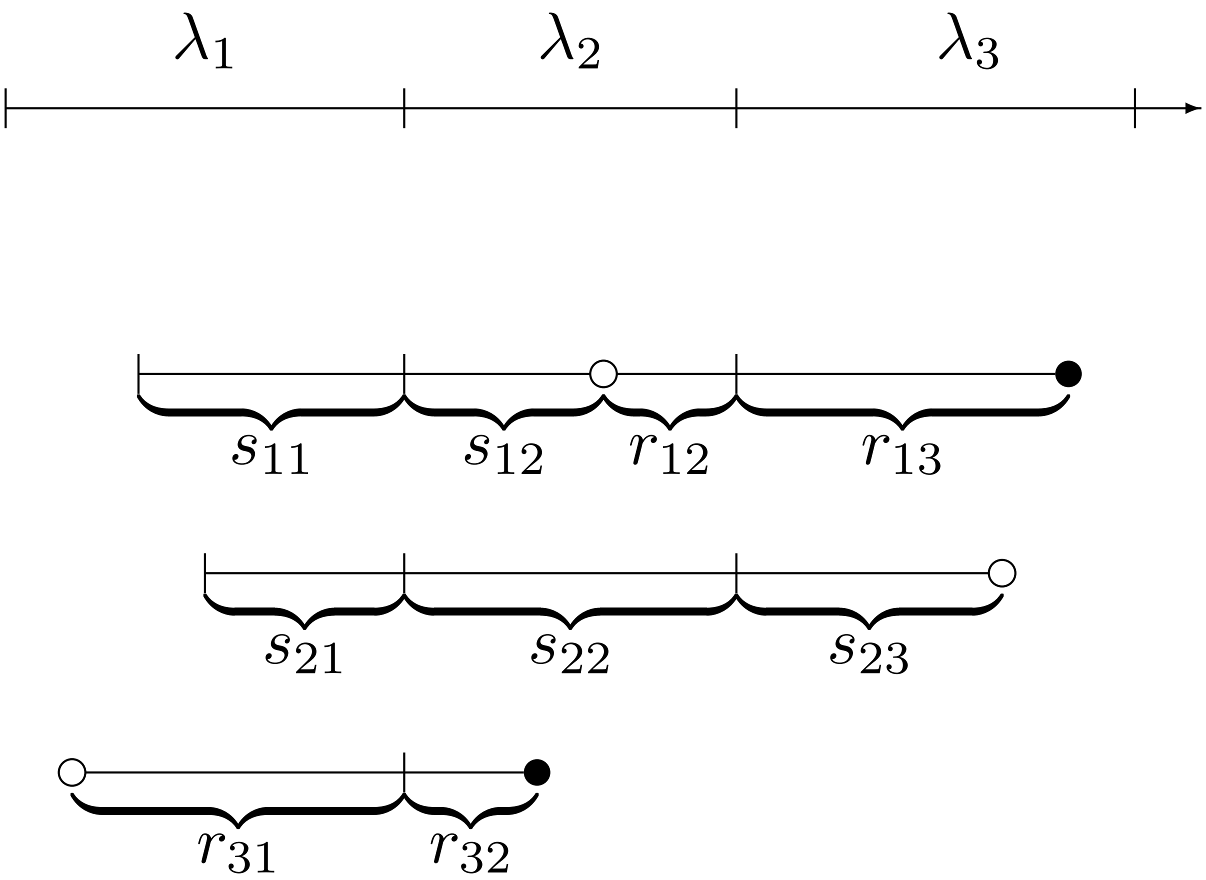
$$1 - \exp\left(-\int \lambda_p(s)\mathrm{d}s\right) = 1 - \exp\left(\sum \lambda_i x_{pi}\right)$$

where x_{pi} is minus the length of the **part of** interval i where the event may have occurred.

Lifting the assumtion about simultaneous testing takes the problem out of the panel-study setup, and allows any relevant time scale to be used.

Coding setup for intensity covariates

Intensities



Variables

p	y_{pj}	x_{1pj}	x_{2pj}	x_{3pj}
1	1	$-s_{11}$	$-s_{12}$	0
1	0	0	$-r_{12}$	$-r_{13}$
2	1	$-s_{21}$	$-s_{22}$	$-s_{23}$
3	0	$-r_{31}$	$-r_{32}$	0

Model vs. likelihood

The likelihood derived is like a binomial likelihood.

But the **model** is not a binomial model; hence the $\mathrm{s.e.}$s will be wrong if the program used computes the expected rather than the observed information.

The normal approximation to the distribution of estimates in this model is likely to be dubious. Intervals based on profile likelihood would probably be preferable.

Regression models

Two identical papers by Farrington [?] and Carstensen [?].

The structure of the likelihood induces a model which is additive on the intensity scale.

The simplest regression model is the (additive) excess risk model:

$$\lambda_i(\mathbf{z}) = \lambda_i(\mathbf{0}) + \sum_k \beta_k z_k.$$

where the zs are covariates for each person.

Replace the terms λ_i in the likelihood for the simple case by the terms $\lambda_i + \sum_k \beta_k z_k$:

$$\exp\left(\sum_i \lambda_i x_{pi}\right) = \exp\left(\sum_i (\lambda_i + \sum_k \beta_k z_k) x_{pi}\right)$$

i.e. a binomial likelihood with success probability:

$$\mu_p(\mathbf{z}) = \exp\Big\{\sum_i \lambda_i x_{ip} + \sum_k \beta_k (z_{kp} \sum_i x_{ip})\Big\}$$

Inclusion of the **covariates** $z_1, \ldots, z_K$ in the **model** for the intensities means that

variables $z_k \sum_i x_i$ should be added in the **estimation**.

Variables in the regression model

- Response variable:

  y_p: 1 for intervals survived, 0 for event intervals

- Covariates:
  - Baseline hazard:

    $x_1, x_2, \ldots$: Minus the time at risk in interval i.
  - Covariates:

    $z_{pk} \times \sum_i x_i$: The actual covariates should be multiplied by the sum of the $x$es before used in the estimation.

Multiplicative model — proportional hazards

Replace the terms λ_i in the likelihood for the simple case above by the terms $\lambda_i \exp(\sum_k \beta_k z_k)$.

The likelihood is then as a binomial likelihood with means:

$$\begin{aligned} \mu_p(\mathbf{z}) &= \exp\Big\{ \textstyle\sum_i \lambda_i x_{ip} \exp(\sum_k \beta_k z_{kp}) \Big\} \\ &= \exp\Big\{ -\exp(\ln[-\textstyle\sum_i \lambda_i x_{ip}] + \sum_k \beta_k z_{kp}) \Big\} \end{aligned}$$

For fixed λs this is a generalized linear model
For fixed βs this is a generalized linear model.

$$\exp\Big\{\sum_i \lambda_i x_{ip} \exp(\sum_k \beta_k z_{kp})\Big\}$$

For fixed βs it is a generalised linear model:

- parameters: λ_i
- covariates: $x_i \exp(\sum_k \beta_k z_k)$,
- error: Binomial
- link: logarithmic

$$\exp\left\{-\exp(\ln[-\sum_i \lambda_i x_{ip}] + \sum_k \beta_k z_{kp})\right\}$$

For fixed λs it is a generalised linear model:

- parameters: β_k
- covariates: z_k
- error: Binomial
- link: log$-$log
- offset: $\ln(-\sum_i \lambda_i x_i)$

Algorithm for fitting the model.

1. Fit a model without covariates, to obtain initial estimates of the λs.

2. Fix the λs, and fit a model with covariates z_k, log$-$log-link and offset $\ln(-\sum_i \lambda_i x_i)$ to obtain estimates of the βs.

3. Fix the βs, form the covariates $x_i \exp(\sum_k \beta_k z_k)$, and fit a model with these covariates and log-link, to obtain estimates of the λs.

4. Repeat 2. and 3. until convergence.