PhD course 2023

Epidemiological methods in medical research

The Cox regression model

Clayton & Hills, Ch. 30

2 March 2023

Per Kragh Andersen

Analysis of cohort studies

In cohort studies:

- Subjects are followed over *time*
- Occurrences of *events* of interest are observed
- The frequency measure typically used is the *rate*

$$\lambda(t) = P(\text{event in interval } (t, t+dt) \mid \text{no event before } t)/dt$$

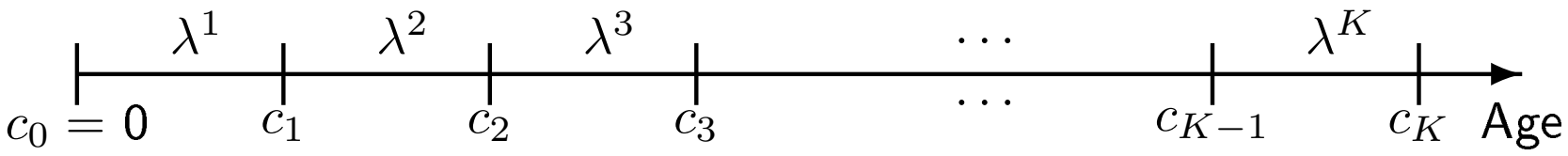
- Data are often analyzed using *Poisson regression*, e.g.

$$\log(\text{Rate}) = \text{ Corner + Exposure + Time} \quad (*)$$

where Time is often *categorized* age.

Poisson regression: piecewise constant rate

The time variable (age) is divided into K intervals and the rate in each of the intervals (λ^k) is assumed constant (but possibly different among intervals).



Thus

$$\lambda(t) = \lambda^k \text{ for } t \text{ between } c_{k-1} \text{ and } c_k, \quad k = 1, \ldots, K$$

The intervals need not have the same length.

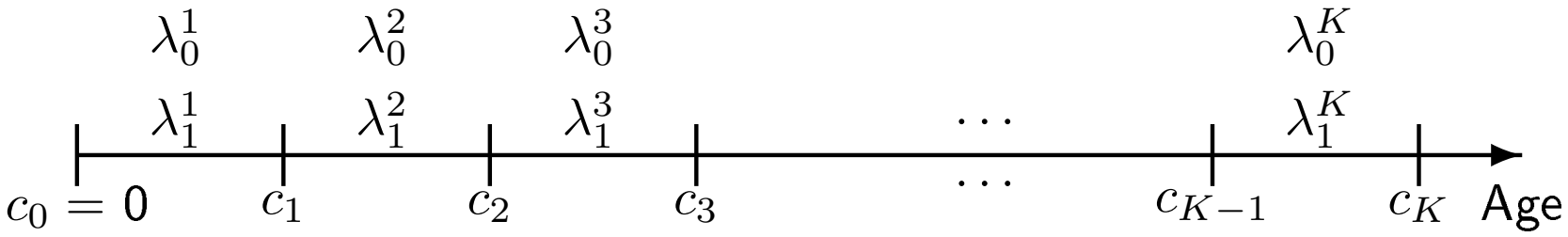
Poisson regression: piecewise constant rate

For estimation, we only need to keep record of the total number of events and the time at risk in each interval.

For example, the diet data split by age bands:

Age band	IHD cases	P-yrs. at risk
40-49	6	919.8
50-59	17	2149.2
60-69	22	1556.4

Typically, we further relate the rate to covariates, e.g. exposure (E=0, unexposed, E=1, exposed):



- Not part of standard SAS to split the time variable, but user-written SAS-macros exist.
- R – packages exist (e.g., Epi Package)

  More later in the course.
- Often an assumption of no interaction is imposed as in (*):
  $\log(\lambda_1^j) = \log(\theta_E) + \log(\lambda_0^j), \quad j = 1, 2, ..., K$

Poisson regression: piecewise constant rate

For estimation, we need to keep record of the total number of events and the time at risk in each interval combined with each exposure group.

Example, the diet data:

	Exposed		Unxposed	
Age band	IHD cases	P-yrs. at risk	IHD cases	P-yrs. at risk
40-49	2	311.9	4	607.9
50-59	12	878.1	5	1271.1
60-69	14	667.5	8	888.9

Similarly when there are more (categorical) covariates.

Survival data

Time (T) to *death* (or other *event* of interest), measured from a well-defined starting time point (time-origin, time=0):

- Time from start of randomized clinical trial to death
- Time from first employment to pension
- Time from filling of a tooth to filling falls out

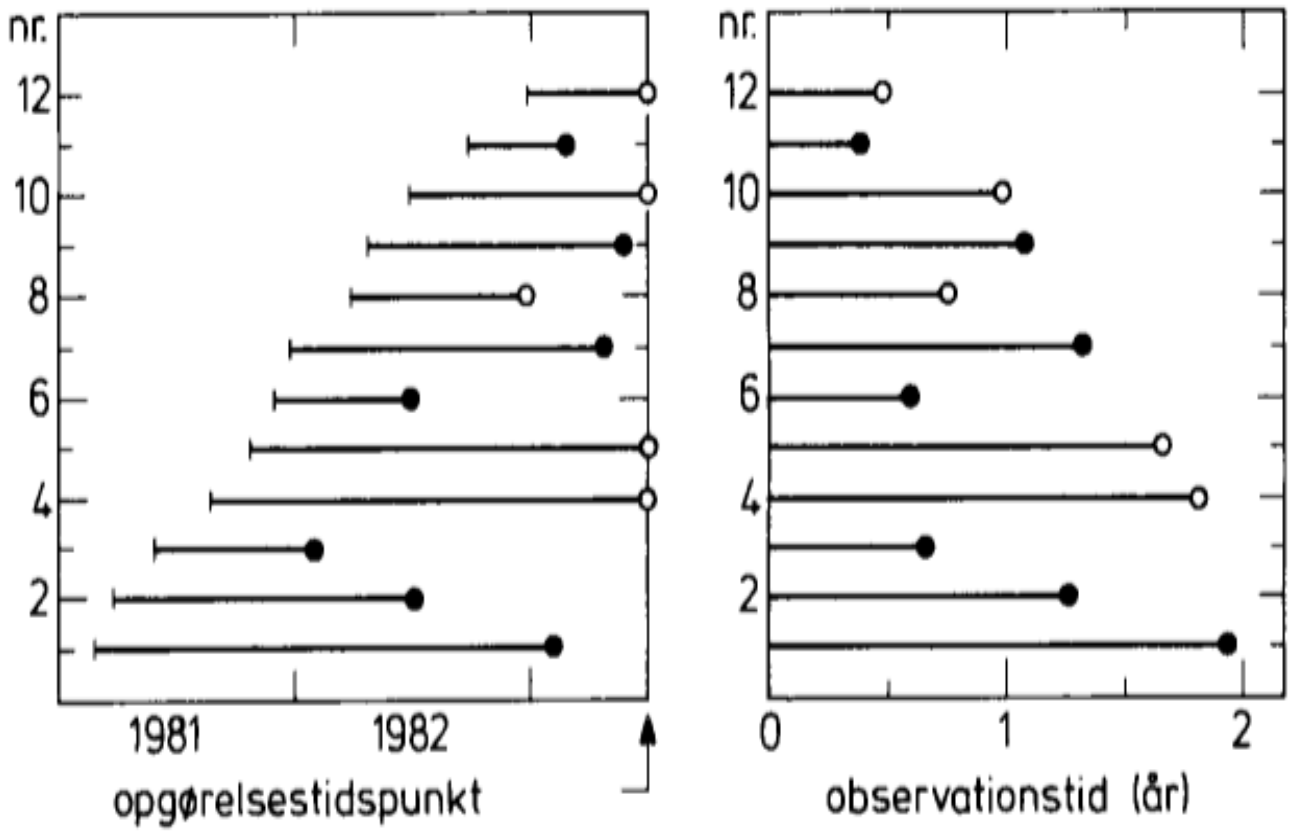
In clinical applications, choice of time 0 is often 'obvious' (randomization, diagnosis, start of treatment).

What is special about survival data?

- *Censoring*: For some, we will only know a period in which the event was not observed, but not when (and sometimes: if) the event will happen

Survival data studies are special cases of cohort studies.

A small data set



Ordered times: 5, 6*, 7 , 8, 9*, 12*, 13, 15, 16, 20*, 22*, 23

* indicates censored observations.

A small data set

Ordered times: 5, 6*, 7 , 8, 9*, 12*, 13, 15, 16, 20*, 22*, 23

* indicates censored observations.

How to estimate the mean survival time?

$$\frac{5+6+7+8+9+12+13+15+16+20+22+23}{12} = \frac{156}{12} = 13.0?$$

$$\frac{5+7+8+13+15+16+23}{7} = \frac{87}{7} = 12.4?$$

Which fraction of patients survives past 12 months? $\frac{6}{12} = 0.5?$

Exercise: Why are these estimates biased? And in which direction?

We need methods that are able to account for censoring.

This leads to a focus on other parameters than the mean value.

Survival and hazard function

Let T be the time (from time 0) to the event of interest:

$$\begin{aligned} S(t) &= P(T > t) \\ &= \text{probability of survival to time } t \end{aligned}$$

$$\begin{aligned} \lambda(t) &= \text{rate or hazard} \\ \lambda(t)dt &\approx P(T \leq t + dt \mid T > t) \\ &= \text{probability of failure before } t + dt \text{ given survival beyond } t \end{aligned}$$

Relationship (when there are *no competing risks*):

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

(Here, $\Lambda(t)$ is the *integrated (or cumulative) hazard function*.)

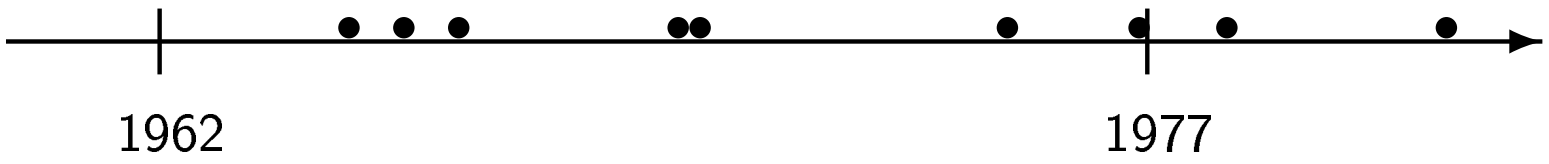
$F(t) = 1 - S(t) = P(T \leq t)$ is the *failure risk* before time t.

Malignant melanoma data

In the period 1962-77, 205 patients had their tumor removed and were followed until 1977. At the end of 1977:

- 57 had died of malignant melanoma (status=1)
- 134 were still alive (status=2)
- 14 had died from other causes (status=3)

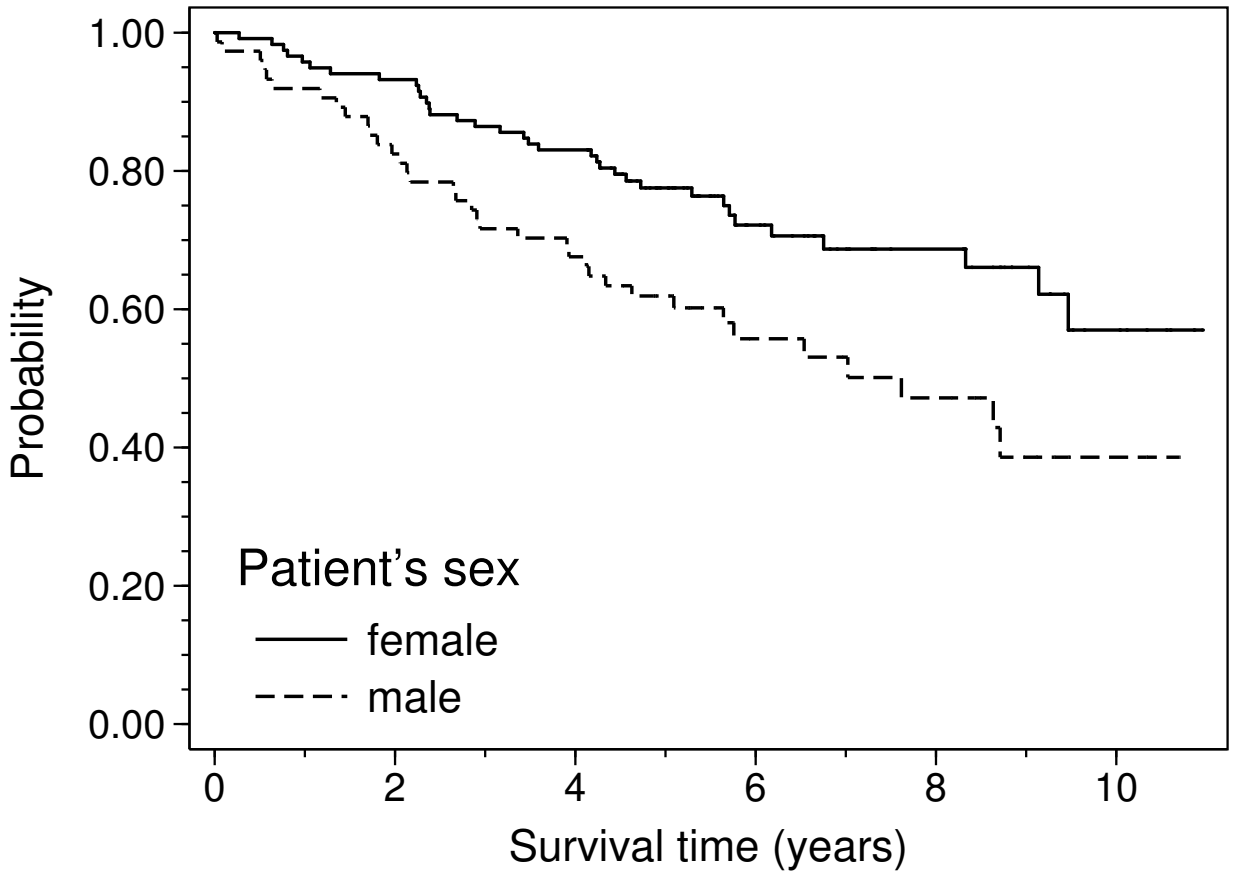
Purpose: Study effect on survival of sex, age, thickness of tumor, ulceration, etc. (no treatment factor)



Malignant melanoma data

id	days	status	sex	age	year	thick	ulc
1	10	3	1	76	1972	6.76	1
2	30	3	1	56	1968	0.65	0
3	35	2	1	41	1977	1.34	0
4	99	3	0	71	1968	2.90	0
5	185	1	1	52	1965	12.08	1
6	204	1	1	28	1971	4.84	1
7	210	1	1	77	1972	5.16	1
8	232	3	0	60	1974	3.22	1
9	232	1	1	49	1968	12.88	1
10	279	1	0	68	1971	7.41	1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
203	4688	2	0	42	1965	0.48	0
204	4926	2	0	50	1964	2.26	0
205	5565	2	0	41	1962	2.90	0

Kaplan-Meier survival curves



How to quantify the difference between males and females?

The Cox Model

The Cox model assumes that the rate for the ith individual is

$$\lambda_i(t) = \lambda_0(t) \exp(\beta x_{i1})$$

where

$$x_{i1} = \begin{cases} 0 & \text{if individual } i \text{ is a female} \\ 1 & \text{if individual } i \text{ is a male} \end{cases}$$

That is,

$$\lambda_i(t) = \begin{cases} \lambda_0(t) & \text{if individual } i \text{ is a female} \\ \lambda_0(t) \exp(\beta) & \text{if individual } i \text{ is a male,} \end{cases}$$

so, $\lambda_0(t)$ (the "baseline hazard") is the hazard rate for females. This is completely *unspecified*.

Time t is the chosen time variable, e.g. time since randomization, age, or (like here) time since operation.

Hazard ratio

If

$$\lambda_i(t) = \begin{cases} \lambda_0(t) & \text{if individual } i \text{ is a female} \\ \lambda_0(t)\exp(\beta) & \text{if individual } i \text{ is a male} \end{cases}$$

then the *rate ratio*, *RR* (or *hazard ratio*, *HR*) between males and females is

$$\text{RR} = \frac{\lambda_0(t)\exp(\beta)}{\lambda_0(t)} = \exp(\beta) = \theta.$$

This ratio is independent of time, i.e. we have *proportional hazards*. The Cox model is also called the *proportional hazards model*.

Note that proportional hazards is a *modeling assumption* that may or may nor describe the data well.

Estimation

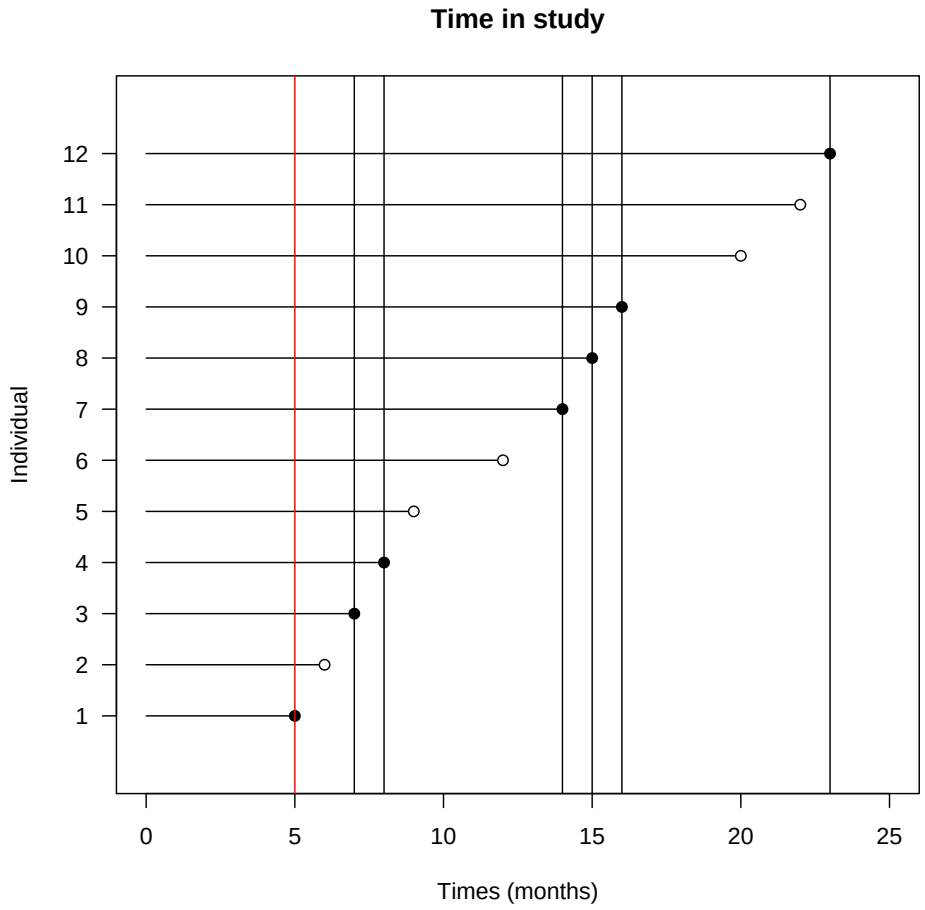
The parameters: β and the baseline hazard $\lambda_0(t)$, may be estimated based on the *likelihood principle*. The maximum likelihood estimate is denoted $\widehat{\beta}$.

When deriving the likelihood function ("Cox's partial likelihood"), the concept of a *risk set* is crucial. The risk set, $R(t_i)$ at death time t_i is the set of individuals being at risk of dying (alive and uncensored) just before time t_i.

In Cox's partial likelihood the covariates for the individual failing are "compared to" the covariates for patients who could have failed at that time, i.e. patients from the risk set:

$$\sum_{\text{failures}} \log\left(\frac{\theta_{(\text{for case})}}{\sum_{\text{Risk set}}\theta}\right) \quad (\text{with } \theta_i = \exp(\beta x_{i1}))$$

Risk set (at time t): $R(t) =$ set of subjects who could have been the case (at time t). Depends on how *time* is defined (age, time on study, calendar time,...). May induce *delayed entry* (more later):



The Cox model

$$\lambda_i(t) = \lambda_0(t) \exp(\beta x_{i1})$$

can also be written on log-scale

$$\begin{aligned} \log(\lambda_i(t)) &= \log(\lambda_0(t) \exp(\beta x_{i1})) \\ &= \log(\lambda_0(t)) + \beta x_{i1}. \end{aligned}$$

In 'Clayton-Hills notation':

$$\log(\text{Rate}) = \text{Corner(Time)} + x_1.$$

Compare with the Poisson regression model: later!

For the melanoma data, we get $\widehat{\beta} = 0.656, \text{SD} = 0.238$, i.e., the *hazard ratio* is $\exp(0.656) = 1.93$ with 95% confidence limits from $1.93/\exp(1.96 \cdot 0.238) = 1.21$ to $1.93 \times \exp(1.96 \cdot 0.238) = 3.07$.

The Cox model in SAS and R

In SAS, `proc phreg` can be used for estimation in the Cox model.

```
proc phreg data=melanom;
  class sex (ref="0");
  model days*status(2) = sex / rl;
run;

/* The 'rl' option (or 'risklimits') adds confidence limits
   for hazard ratios to the output */
```

Part of output from `proc phreg`:

.
.
.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
sex	1	1	0.65586	0.23761	7.6190	0.0058	1.927	1.209	3.070

The column `Parameter Estimate` is $\widehat{\beta}$.

```
mel <-
read.table("melanom-surv.txt", header = TRUE)


library(survival)
fit <- coxph(Surv(days, status != 2) ~ factor(sex), data = mel)
summary(fit)

Call:
coxph(formula = Surv(days, status != 2) ~ factor(sex), data = mel)

               coef exp(coef) se(coef)     z Pr(>|z|)
factor(sex)1 0.6559    1.9269   0.2376 2.761  0.00577 **

             exp(coef) exp(-coef) lower .95 upper .95
factor(sex)1     1.927      0.519      1.21      3.07
```

Adjustment for confounding

Add covariates $x_{i2}, \ldots, x_{ip}$ to the model including only 'exposure' x_{i1}:

$$\begin{aligned} \lambda_i(t) &= \lambda_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}), \text{ resp.} \\ \log(\text{Rate}) &= \text{Corner(Time)} + x_1 + x_2 + \cdots + x_p. \end{aligned}$$

The Cox model assumes that:

1. the hazards for all individuals are proportional,
2. the effects of covariates are additive and linear on the log rate scale, (the 'linear predictor')

Items 1. and 2. should be checked as part of the data analysis.

Results for melanoma data

Adding age (per 10 years), `ulceration` (yes vs. no), and `tumor thickness` (in *mm*) to the model including `sex`, we get the results

Covariate	$\widehat{\beta}$	SD	HR
Sex	0.413	0.240	1.51
Age	0.218	0.078	1.24
Ulceration	0.952	0.268	2.59
Thickness	0.099	0.035	1.10

Note the reduced log(hazard ratio) for `sex` after adjusting for (primarily) thickness.

Exercise: Calculate a 95% confidence interval for HR for `sex`.

Solution: From
$\exp(0.413 - 1.96 \cdot 0.240) = 0.94$ to $\exp(0.413 + 1.96 \cdot 0.240) = 2.42$.

Checking model assumptions

Linearity or interactions may be tested as for other models with a linear predictor, e.g. using splines to test linearity.

Proportional hazards may be checked by adding *interactions with time* - this is possible in both `SAS` and `R`.

There are many other ways of checking the assumptions, including some based on *cumulative martingale residuals* or *cumulative score ('Schoenfeld') residuals*.

These methods are available in `SAS PROC PHREG` via the `ASSESS` statements. In `R`, the method is implemented in various packages, e.g. `timereg`.

What to do if proportional hazards fail?

Then, in principle, the model is *incorrect* and results should be interpreted with caution.

In a classical 'exposure-confounder' situation, if proportional hazards fail for a confounder then *sensitivity analyses* may be a way forward: Estimate the exposure effect with and without allowing non-proportional hazards for the confounder and compare. The Cox model tends to be remarkably robust towards this kind of model mis-specification!

If the assumption fails for the exposure then a single hazard ratio is misleading and something like survival curves could be presented instead of a single parameter. This does complicate matters!

Independent censoring

The Cox model (and other methods for survival data, including Kaplan-Meier) rely on an assumption of *independent censoring*.

This means that individuals censored at any given time t should not be a biased sample of those who are *at risk* at time t.

Stated in other words: the hazard $\lambda(t)$ gives the event rate at time t, i.e. the failure rate given that the subject is still alive ($T > t$).

Independent censoring then means that the extra information that the subject is not only alive, but also uncensored at time t does not change the failure rate.

Choice of time variable

A study is conducted over calendar time but the natural time variable may be time since treatment, e.g. the melanoma study.

Cohort studies are often conducted by recruiting a random sample of the population at the start of the study and then these subjects are followed for a number of years.

In such a study, age may be a more natural time variable than time on study.

Vaccinations in Guinea-Bissau 1990-96

Rural Guinea-Bissau: 5274 children under 7 months of age were visited two times at home, with an interval of $\approx$ six months. Information about vaccination (BCG, DTP, measles vaccine) was collected at each visit, and at second visit death during follow-up was registered. Children were censored if they moved away during follow-up or survived until second visit.

Variables in the Bissau data set (`bissau.txt`):

`id`	Child id
`fuptime`	Follow-up time in days
`dead`	0 = censored, 1 = dead
`bcg`	1 = Yes, 2 = No
`dtp`	Number of doses of DTP (0, 1, 2 or 3)
`age`	Age at first visit in days
`agem`	Age at first visit in (whole) months

5275 lines, first line contains variable names.

Is the risk of dying associated with vaccination?

	Outcome		
Exposure	Died	Survived	Total
BCG vaccinated	125 (3.8%)	3176	3301
not BCG vaccinated	97 (4.9%)	1876	1973
Total	222 (4.2%)	5052	5274

Risk ratio, $RR = 0.77$, odds ratio $OR = 0.76$.

NB: This table (and the RR and OR estimates) *ignore* the varying follow-up times among children.

```
proc phreg data=bissau;
  class bcg;
  model fuptime*dead(0)=bcg / rl ;
run;
```

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.2824	1	0.0385
Score	4.3761	1	0.0364
Wald	4.3474	1	0.0371

Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
bcg	1	4.3474	0.0371

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
bcg	1	-0.28214	0.13532	4.3474	0.0371	0.754	0.578	0.983

```
proc phreg data=bissau;
  class bcg agem;
  model fuptime*dead(0)=bcg agem / rl ;
run;
```

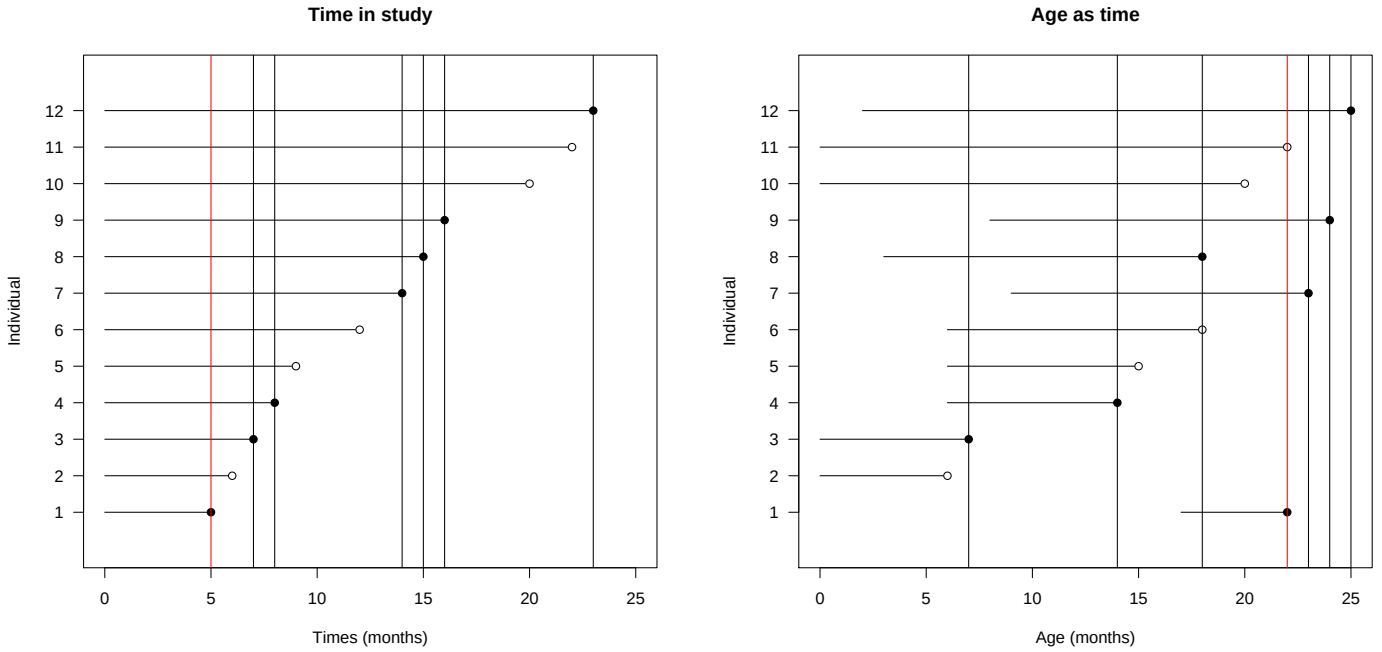
Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
bcg	1	5.6510	0.0174
agem	6	7.7246	0.2590

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
bcg	1	1	-0.34720	0.14605	5.6510	0.0174	0.707	0.531	0.941
agem	0	1	0.01053	0.35339	0.0009	0.9762	1.011	0.506	2.020
agem	1	1	0.12553	0.34494	0.1324	0.7159	1.134	0.577	2.229
agem	2	1	-0.24631	0.35903	0.4707	0.4927	0.782	0.387	1.580
agem	3	1	0.20946	0.34502	0.3686	0.5438	1.233	0.627	2.425
agem	4	1	0.34300	0.34265	1.0020	0.3168	1.409	0.720	2.758
agem	5	1	0.34118	0.34699	0.9668	0.3255	1.407	0.713	2.777

Age as time variable: Delayed entry (small data set)



Subjects are only at risk from the age at entry and onwards: Had they died before the age of entry, we would not have observed that.

Handling delayed entry is quite easily done by careful control of the *risk set* $R(t_i)$ at death time t_i.

```
data bissau2;
  set bissau;
  outage=age+fuptime; /* age is in days */
run;
proc phreg data=bissau2;
  class bcg;
  model (age,outage)*dead(0)= bcg / rl;
/* Alternatively:  model outage*dead(0)= bcg / rl, entry = age */;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
bcg	1	-0.35542	0.14065	6.3854	0.0115	0.701	0.532	0.923

R code and (edited) output

```
# load data
bissau <- read.table("bissau.txt", header = TRUE)

library(survival)
fit <- coxph(Surv(fuptime, dead != 0) ~ (bcg == 1) + factor(agem), data = bissau)
summary(fit)

bissau$outage <- bissau$age + bissau$fuptime

fitage <- coxph(Surv(age,outage, dead != 0) ~ (bcg == 1), data = bissau)
summary(fitage)

Call:
coxph(formula = Surv(age, outage, dead != 0) ~ (bcg == 1), data = bissau)

                coef exp(coef) se(coef)      z Pr(>|z|)
bcg == 1TRUE -0.3552    0.7011   0.1407 -2.525   0.0116 *

             exp(coef) exp(-coef) lower .95 upper .95
bcg == 1TRUE    0.7011      1.426    0.5321    0.9236
```

Comparison of Poisson and Cox models (alternative) diet data

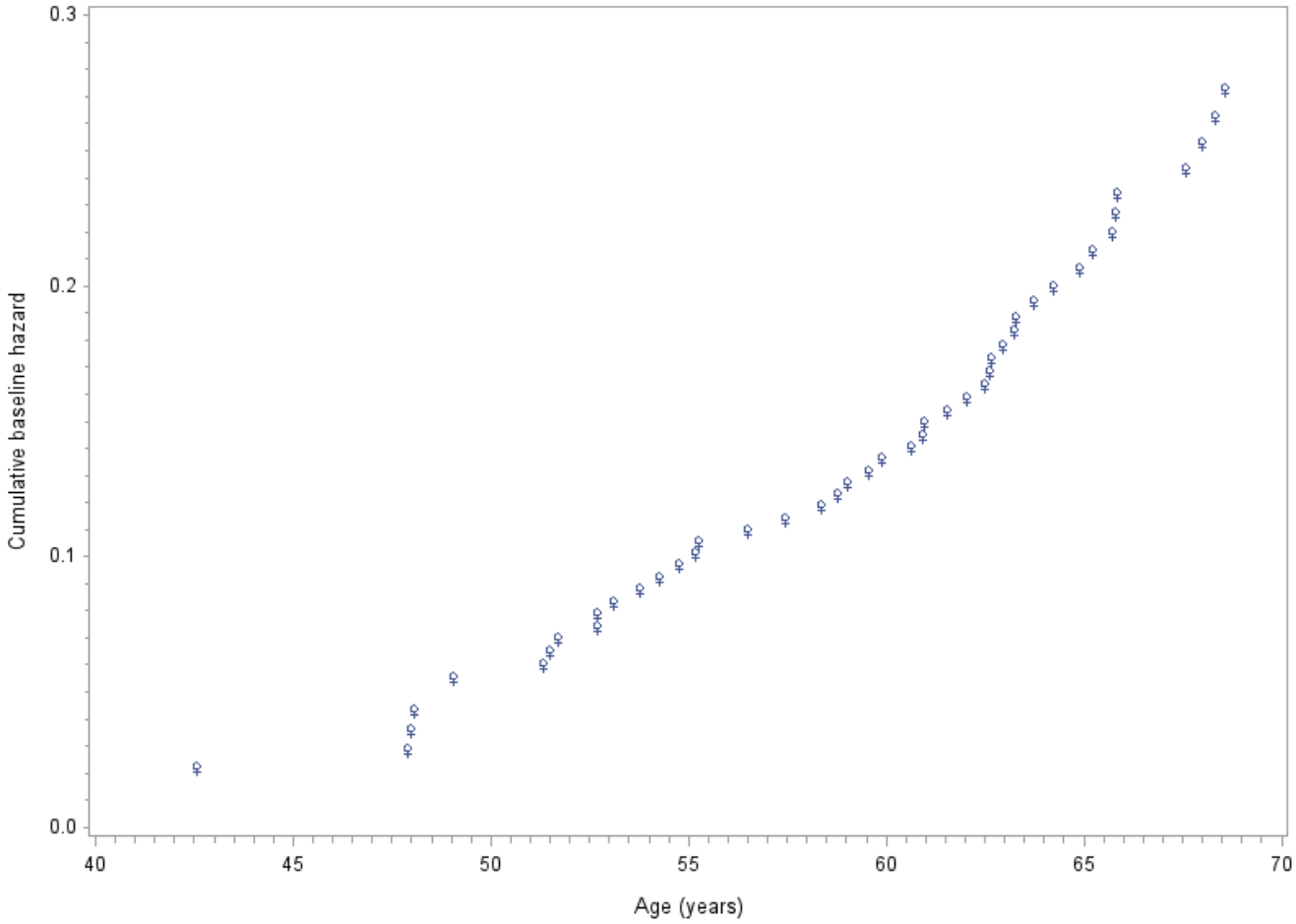
Poisson			
Parameter	Estimate	SD	W
Exposure (1)	0.622	0.303	4.22
Age (1)	0.204	0.472	0.19
Age (2)	0.747	0.461	2.62

Cox			
Parameter	Estimate ($\widehat{\beta}$)	SD	W
Exposure (1)	0.611	0.303	4.08

Estimation of (cumulative) baseline hazard

In addition, the Cox model contains the baseline rate (the effect of age) which may, optionally, be estimated (or rather the cumulative baseline rate) and presented graphically, the so-called 'Breslow' estimator.

Diet data: estimated cumulative baseline hazard



Results for melanoma data ($K = 3$ intervals)

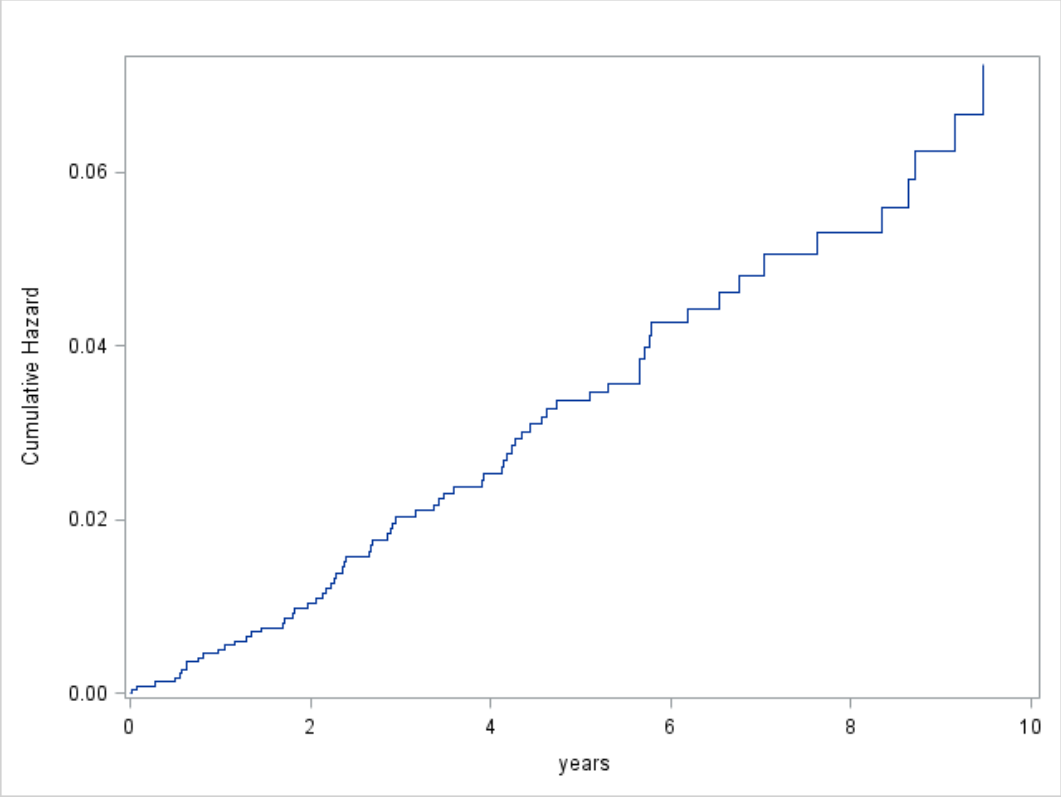
	Cox		Poisson	
Covariate	$\widehat{\beta}$	SD	$\widehat{\beta}$	SD
Sex	0.413	0.240	0.396	0.240
Age	0.218	0.078	0.222	0.076
Ulceration	0.952	0.268	0.960	0.269
Thickness	0.099	0.035	0.096	0.035

The Poisson model, additionally, provides estimates for the baseline hazard in each interval (0-2.5, 2.5-5, 5+ years).

For the Cox model, as above, the (cumulative) baseline hazard may be extracted from the output and plotted.

Note again the similarity between the two sets of estimates.

Baseline hazards



Poisson baseline rates: 0-2.5 years 614, 2.5-5 years 719, 5years- 699 per 100000 years.

Bissau data: Cox, Poisson, or logistic regression?

Analysing Guinea-Bissau data using three different regression models all adjusting for age in months as a categorical variable. In the Cox model, follow-up time was used as the time variable. In the Poisson model, the follow-up time was used as time at risk. The logistic regression did not take the follow-up time into account.

Cox RR (95%CI)	Poisson RR (95%CI)	Logistic OR (95%CI)
0.71 (0.53-0.94)	0.71 (0.53-0.94)	0.71 (0.53-0.96)

Follow-up time does not seem important.

OR and RR are close since the mortality rate is (rather) low.

Cox or Poisson?

Items to consider:

- sample size: Cox needs individual records, Poisson can use tabulated data (for categorical covariates)
- parametric/non-parametric time effects; "strong" effects
- which effects are of interest?
- Choice of "basic" time scale: only for Cox
- Poisson can handle several time variables 'in parallel'
- Examination of proportional hazards

Take-home message:

Cox and Poisson tend to give *very similar results*