Biostatistics III: Survival analysis for epidemiologists

Therese Andersson, Anna Johansson and Mark Clements
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden

http://www.biostat3.net/

Karolinska Institutet
Feb 10–19, 2020

http://kiwas.ki.se/katalog/katalog/kurs/3244

Friendly faces

- Primary teachers:
  Therese Andersson (therese.m-l.andersson@ki.se),
  Anna Johansson (anna.johansson@ki.se)
- Course director: Mark Clements (mark.clements@ki.se)
- Course administrator: Gunilla Nilsson Roos (gunilla.nilsson.roos@ki.se)
- Teaching assistants: Mark Clements, Alessandro Gasparini, Yuliya Leontyeva.

Overview of the course

- Central concepts in survival analysis: censoring, truncation, survivor function, hazard function.
- Estimating survival non-parametrically.
- Non-parametric method for testing differences in survival between groups.
- Estimating rates and modelling them using Poisson regression.
- Effect modification and confounding when modelling rates.
- Modelling rates using Cox proportional hazards model.
- The proportional hazards assumption.
- Assessing the proportional hazards assumption (for both Cox and Poisson regression).

- Modelling non-proportional hazards (for both Cox and Poisson regression).
- Comparison of the Cox and Poisson regression models (illustration that they are very similar).
- Stratified Cox models.
- Flexible parametric survival models.
- More on censoring and truncation, including informative censoring.
- Competing risks analysis (limited coverage)
- Standardised mortality/incidence ratios
- Some biases in survival analysis/cohort studies.
- Risk set sampling (e.g., the nested case-control design and case-cohort design)

Teaching format

- Generally lectures in the morning followed by exercises in the afternoon.
- We have constructed exercises and provided solutions to most exercises. We will suggest appropriate exercises for each afternoon.
- There are a lot of exercises, don't worry if you don't have time to finish all of them. We will let you know which exercises that are the key exercises.
- Course participants have a wide range of backgrounds and diverse interests. It is hoped that the lab sessions will provide time for you to study or ask questions about topics of special interest.
- The lecture notes are very comprehensive, and contain topics which we will only cover briefly. Some slides will not be covered during the lectures, but are included for completeness.

Group work

- Group work on day 5, each group will be assigned an article to read and discuss.
- You can see which group you belong to in the list of participants.
- The papers can be found on the course web site.
- Read your allocated paper in advance, and discuss the paper in your group on day 5 of the course.
- More information will be given later during the course.

Computing

- We will be using Stata.
- We have created a zip-file, containing all files needed for the labs, that can be downloaded from the course website. All Stata datasets can also be accessed directly from the web, for example

```
. use http://www.biostat3.net/download/colon.dta
```

- Further details will come on working in the MEB computer lab.
- See A brief introduction to Stata (`http://www.biostat3.net/download/stataintro.pdf`) for
  - A general introduction to Stata; and
  - Notes on using Stata for survival analysis.
- We encourage you to work in pairs/small groups and discuss the computer exercises.

For SAS users

- Although we will be using Stata during the course, we have made the data sets available in SAS format and you are welcome to attempt the exercises using SAS. See `http://www.biostat3.net/download/index.php?dir=sas/`.
- We have written brief notes on how the methods described during the course can be implemented in SAS (`http://www.biostat3.net/download/sas/notes_for_sas_users.pdf`).
- We have provided SAS code for reproducing some of the key exercises.
- Please keep in mind that if you choose to work with SAS we expect you to bring your own laptop (with SAS installed) and to be familiar with SAS so that the teaching assistants can focus on helping with the statistical concepts rather than statistical programming.

For R users

- Moreover, you are welcome to attempt the exercises using R. See `http://www.biostat3.net/download/index.php?dir=R/` for an example using the melanoma dataset.
- Please keep in mind that if you choose to work with R we expect you to bring your own laptop (with R installed) and to be familiar with R so that the teaching assistants can focus on helping with the statistical concepts rather than statistical programming.

Textbook

- We have not assigned any compulsory texts since experience has shown that course participants have widely varying preferences.
- We will provide extensive course notes and many participants do not find a great need for additional texts. A large number of textbooks are available and we suggest students interested in additional reading identify a textbook at a technical level suitable for them. Many general textbooks in medical statistics contain a chapter on survival analysis.
- Very few books are targeted at epidemiologists (e.g., you won't find Poisson regression mentioned in many books).
- The definitive text for epidemiologists is 'Statistical Methods in Cancer Research: Volume II - The Design and Analysis of Cohort Studies' by Breslow and Day [4] although it is rather advanced. [full text available free from `http://www.iarc.fr/en/publications/pdfs-online/stat/sp82/SP82.pdf`]

- 'An Introduction to Survival Analysis Using Stata' [9] is highly recommended for Stata users. Many parts, however, assume a solid grasp of mathematical statistics.
- The SAS 'books by user' [6, 1] are recommended for SAS users.

Learning outcomes

– The course plan is the formal document upon which the course (and the examination) is based (see `http://kiwas.ki.se/katalog/katalog/kurs/3244`).
– KI uses outcome-based learning. The learning outcomes are listed in the course plan and reproduced below.

After successfully completing this course you should be able to:

– Propose a suitable statistical model for assessing a specific research hypothesis using data from a cohort study, fit the model using standard statistical software, evaluate the fit of the model and interpret the results.
– Explain the similarities and differences between Cox regression and Poisson regression.
– Discuss the concept of timescales in statistical models for time-to-event data, control for different timescales using standard statistical software, and argue for an appropriate timescale for a given research hypothesis.
– Discuss the concept of confounding in epidemiological studies and control/adjust for confounding using statistical models.

– Apply and interpret appropriate statistical models for studying effect modification and be able to reparameterise a statistical model to estimate appropriate contrasts.
– Critically evaluate the methodological aspects (design and analysis) of a scientific article reporting a cohort study.

Examination

The course grade is based solely on a take-home examination. The exam is the only mandatory part of the course.

– The examination is individual-based: **you are not allowed to cooperate with anyone**, although you are encouraged to consult the available literature. The teachers will use Urkund in order to assess potential plagiarism (`http://ki.se/sites/default/files/cheating_is_forbidden_2013.pdf`)
– The examination will be made available at 12:00 on Wednesday 19 February 2020 and the examination is due by 17:00 on Wednesday 26 February 2020.
– Students who do not obtain a passing grade in the first examination will be offered a second examination within 2 months of the final day of the course.
– Do not write answers by hand: please use Word, LaTeX or a similar format for your examination report.
– Motivate all answers and show all calculations in your examination report, but write as brief an answer as possible without loss of clarity. Define any notation that you use for equations. The examination report should be written in English.
– You are expected to write computer code to read the data and for your

analysis. Include your computer code in your report. You are encouraged to use Stata, R or SAS for your analysis; if you wish to use other software, please contact Mark Clements `mark.clements@ki.se`.
– Email the examination report containing the answers **as a pdf file** to `gunilla.nilsson.roos@ki.se`. **Write your name in the email, but do not write your name in the document containing the answers.**

Topics for Day 1

- Central concepts in survival analysis: censoring, survivor function, hazard function.
- Estimating survival non-parametrically, using the Kaplan-Meier and the life table methods.
- Non-parametric methods for testing differences in survival between groups (log-rank and Wilcoxon tests).

Analysis of Time-to-Event Data (survival analysis)

- Survival analysis is used for e.g. cohort studies and randomized clinical trials (RCTs), where study particpants are followed from a start time to an endpoint (failure or event).
- Survival analysis is also known as failure time analysis (primarily in engineering), lifetime analysis, and time-to-event analysis.
- Survival analysis concerns analysing the time to the occurence of an event, e.g. time until a cancer patient dies.
- The event is not necessarily death, despite the name survival analysis. It can also be occurrence of disease, or any other event.
- Survival data are generated when the response measurement of interest is the time from a well-defined origin of measurement to occurrence of an event of interest.

- The outcome can be thought of as comprising two dimensions
  a. an event indicator (0/1), and
  b. a time at risk (continuous) (a.k.a. survival time, risktime, follow-up time, persontime)
- In some studies, the event of interest (e.g. death) is bound to occur if we are able to follow-up each individual for a sufficient length of time. However, whether or not the event of interest is inevitable generally has no consequence for the design, analysis, or interpretation of the study[1].
- The basic statistical methodology is similar for randomised and observational studies, although some methods are more appropriate for some designs than others (e.g. need to control for confounding in observational studies).
- The characteristic that complicates the use of standard statistical methods is *censoring* — unobserved values of the response measurement of interest. Censoring leads to differences in follow-up time between individuals.

[1] For completeness, an exception is when we are interesting in estimating the proportion "cured".

Formal requirements of time-to-event data

- Three basic requirements define time-to-event measurements

  a. precise definition of the start and end of follow-up time
  b. unambiguous origin for the measurement of 'time'; scale of time (e.g. time since diagnosis, attained age)
  c. precise definition of 'response,' or occurrence of the event of interest

- We will discuss the concept of timescales (b) and how to choose an appropriate timescale later in the course.

- On the upcoming slide we will see how (a) and (c) are not always perfectly satisfied in practice.

Examples of time-to-event measurements

- Time from diagnosis of cancer to death due to the cancer
- Time from an exposure to cancer diagnosis
- Time from HIV infection to AIDS
- Time from diagnosis of localised cancer to metastases
- Time from randomisation to recurrence in a cancer clinical trial
- Time from remission to relapse of leukemia
- Time between two attempts to donate a unit of blood for transfusion purposes
- Time to the first goal (or next goal) in a hockey game

- Epidemiological cohort studies are time-to-event studies and are analysed in the framework of survival analysis.

- Examples of time-to-event data can be found in almost every discipline.

- In each of these examples what is the start and end of follow-up, and event?

Sample data sets

- The following data sets will be used during the course:

  colon : colon carcinoma diagnosed during 1975–1994 with follow-up to 31 December 1995.
  melanoma : skin melanoma diagnosed during 1975–1994 with follow-up to 31 December 1995.
  colon_sample : a random sample of 35 patients from the colon data.
  diet : data from a pilot study evaluating the use of a weighed diet over 7 days in epidemiological studies. The primary hypothesis is the relation between dietary energy intake and incidence of coronary heart disease (CHD).

- The diet data are analysed extensively by David Clayton and Michael Hills in their textbook [8]. These data are also used in examples in the Stata manual (for example, `stsplit`, `strate`, and `stptime`).

Variables in the colon carcinoma data set

```
obs: 15,564
-------------------------------------------------------
value
variable name label      variable label
-------------------------------------------------------
sex            sex        Sex
age                       Age at diagnosis
stage          stage      Clinical stage at diagnosis
mmdx                      Month of diagnosis
yydx                      Year of diagnosis
surv_mm                   Survival time in months
surv_yy                   Survival time in years
status         status     Vital status at exit
subsite        colonsub   Anatomical subsite of tumour
year8594       year8594   Indicator for diagnosed during 1985-94
agegrp         agegrp     Age in 4 categories
dx                        Date of diagnosis
exit                      Date of exit
id Unique patient ID
```

ID	Sex	Age at dx	Clinical stage	Date of dx. mmyy	Surv. time mm	yy	Status
1	male	72	Localised	2.89	2	0	Dead - other
2	female	82	Distant	12.91	2	0	Dead - cancer
3	male	73	Distant	11.93	3	0	Dead - cancer
4	male	63	Distant	6.88	5	0	Dead - cancer
5	male	67	Localised	5.89	7	0	Dead - cancer
6	male	74	Regional	7.92	8	0	Dead - cancer
7	female	56	Distant	1.86	9	0	Dead - cancer
8	female	52	Distant	5.86	11	0	Dead - cancer
9	male	64	Localised	11.94	13	1	Alive
10	female	70	Localised	10.94	14	1	Alive
11	female	83	Localised	7.90	19	1	Dead - other
12	male	64	Distant	8.89	22	1	Dead - cancer
13	female	79	Localised	11.93	25	2	Alive
14	female	70	Distant	6.88	27	2	Dead - cancer
15	male	70	Regional	9.93	27	2	Alive
16	female	68	Distant	9.91	28	2	Dead - cancer
17	male	58	Localised	11.90	32	2	Dead - cancer
18	male	54	Distant	4.90	32	2	Dead - cancer
19	female	86	Localised	4.93	32	2	Alive
20	male	31	Localised	1.90	33	2	Dead - cancer
21	female	75	Localised	1.93	35	2	Alive
22	female	85	Localised	11.92	37	3	Alive
23	female	68	Distant	7.86	43	3	Dead - cancer
24	male	54	Regional	6.85	46	3	Dead - cancer
25	male	80	Localised	6.91	54	4	Alive
26	female	52	Localised	7.89	77	6	Alive
27	male	52	Localised	6.89	78	6	Alive
28	male	65	Localised	1.89	83	6	Alive
29	male	60	Localised	11.88	85	7	Alive
30	female	71	Localised	11.87	97	8	Alive
31	male	58	Localised	8.87	100	8	Alive
32	female	80	Localised	5.87	102	8	Dead - cancer
33	male	66	Localised	1.86	103	8	Dead - other
34	male	67	Localised	3.87	105	8	Alive
35	female	56	Distant	12.86	108	9	Alive

Variables in the skin melanoma data set

```
obs: 7,775
-------------------------------------------------
variable name   variable label
-------------------------------------------------
sex             Sex
age             Age at diagnosis
stage           Clinical stage at diagnosis
mmdx            Month of diagnosis
yydx            Year of diagnosis
surv_mm         Survival time in months
surv_yy         Survival time in years
status          Vital status at exit
subsite         Anatomical subsite of tumour
year8594        Indicator for diagnosed during 1985-94
agegrp          Age in 4 categories
dx              Date of diagnosis
exit            Date of exit
id Unique patient ID
-------------------------------------------------
```

Vital status in colon and melanoma data sets

Table 1: Codes for vital status

Code and description
0 Alive
1 Dead: melanoma was the cause
2 Dead: other cause of death
4 Lost to follow-up

Variables in the diet data set

```
. describe
obs: 337       vars: 12
-------------------------------------------------------------

variable    label      variable label
-------------------------------------------------------------
id                     Subject identity number
chd                    Failure: 1=chd, 0 otherwise
y                      Time in study (years)
hieng       hieng      Indicator for high energy
energy                 Total energy (kcals per day)
job         job        Occupation
month                  Month of survey
height                 Height (cm)
weight                 Weight (kg)
doe                    Date of entry
dox                    Date of exit
dob                    Date of birth
-------------------------------------------------------------
```


What can we estimate from time-to-event data?

- Survival probability (survivor function), i.e. the proportion who have not experienced the event at a given time point during follow-up
- Mean survival time
- Median survival time
- Event rates (hazard rates, incidence rates), i.e. instantaneous risk that the event will occur at a given time point (hazard function)
- Hazard ratios, i.e. ratios of event rates between different groups (e.g., exposed vs. unexposed) while adjusting for confounders
- In some studies the time-to-event (or survival probability) is of primary interest whereas in many epidemiological cohort studies we are primarily interested in comparing the event rates between the exposed and unexposed.

Censoring and follow-up

- Censoring refers to the situation where the individual can no longer be followed up and event of interest has not occurred during the observed follow-up.
- We will not be able to observe the event if it happens after the censoring event.
- In studying the survival of cancer patients, for example, patients enter the study at the time of diagnosis (or the time of treatment in randomised trials) and are followed up until the event of interest is observed. Censoring may occur in one of the following forms:
  - Termination of the study before the event occurs (administrative censoring);
  - Loss to follow-up, for example, if the patient emigrates; and
  - Death due to a cause not considered to be the event of interest (in cause-specific survival analyses).

- We say that the survival time is censored.
- These are examples of right censoring, which is the most common form of censoring in medical studies.
- With right censoring, we know that the event has not occurred during follow-up, but we are unable to follow-up the patient further. We know only that the true survival time of the patient is greater than a given value.
- In other words, follow-up time (time at risk) may differ between individuals.
- If we do not account for these differences (by using survival analysis) then results may be biased.

Examples of events and censorings

Table 2: Examples of some common events and censorings

Events	Censorings
Death	Emigration
	End-of-study (e.g. 2006-12-31)
Cancer death	Death due to other causes than cancer
	Emigration
	End-of-study (e.g. 2006-12-31)
Breast cancer incidence	Death
	Emigration
	End-of-study (e.g. 2006-12-31)
	Mastectomy

Why do we need survival analysis?

- In biostat I and biostat II we covered statistical methods for comparing means and proportions (e.g., logistic regression). What happens if we apply these methods now?
- Let's assume a new treatment was introduced in late 1992 and we are interested in studying whether patient survival has improved for patients diagnosed 1993–94 compared to those diagnosed earlier.
- Let's compare the proportion of patients who die between the two diagnosis periods, using the colon_sample data. The patients were followed until end of 1995.
- This means that patients who were diagnosed 1993-1994 only had follow-up for at most 36 months (3 years) due to administrative censoring.
- Whereas, patients diagnosed 1985-1992 had follow-up for at most 11 years.


```
. tab dx93 dead, row chi2

           |         dead
      dx93 |     alive       dead |     Total
-----------+----------------------+----------
dx 1985-92 |        10         18 |        28
           |     35.71      64.29 |    100.00
-----------+----------------------+----------
dx 1993-94 |         6          1 |         7
           |     85.71      14.29 |    100.00
-----------+----------------------+----------
     Total |        16         19 |        35
           |     45.71      54.29 |    100.00

          Pearson chi2(1) =   5.6414   Pr = 0.018
```

- We see that only 1 of the 7 (14%) patients diagnosed in the recent period died compared to 18 of 28 (64%) in the early period and this difference is statistically significant.

- It is not surprising that the proportion of deaths was lower among patients diagnosed more recently since these patients had a shorter follow-up time; they did not have the same opportunity to die.

- Let's instead compare the average 'survival time' (the lengths of the lines) between the two groups while ignoring whether or not the patient died.

```
. ttest surv_mm, by(dx93)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group | Obs       Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 1985-92 |  28   48.39286    7.067202    37.39612    33.89216    62.89356
 1993-94 |   7   21.28571     4.37914    11.58612    10.57034    32.00108
---------+--------------------------------------------------------------------
combined |  35   42.97143    5.988713     35.4297     30.8009    55.14196
---------+--------------------------------------------------------------------
    diff |       27.10714    14.44577               -2.282995    56.49728
------------------------------------------------------------------------------
```

- Patients diagnosed in 1985-92 'survived' on average for 48 months compared to 21 months for patients diagnosed 1993-94.

- Restricting this analysis to patients who died (i.e., mean survival time among those who died) is not appropriate either. By definition, the maximum survival time for patients diagnosed 1993-1994 is 36 months.

```
. ttest surv_mm if dead, by(dx93)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 1985-92 |      18        29.5     7.03783    29.85898    14.65148    44.34852
 1993-94 |       1           3           .           .           .           .
---------+--------------------------------------------------------------------
combined |      19    28.10526           .           .           .           .
---------+--------------------------------------------------------------------
    diff |                26.5           .                       .           .
------------------------------------------------------------------------------
```

- What we would like is some measure of the risk of death adjusted for the fact that individuals were at risk for different lengths of time.

- Methods used for making inference about proportions (e.g., logistic regression) are only appropriate when all individuals have the same time at risk. This is typically not the case when we have survival data.

- There may, however, be situations where everyone has the same potential follow-up.

- That is, when we have a binary outcome and all individuals are at risk for the same length of time the proportion is an appropriate outcome measure.

$$\text{proportion who experience the event} = \frac{\text{number of events}}{\text{number of individuals}}$$

- Every individual contributes the same amount to the denominator.

- If, however, individuals are at risk for differing lengths of time we use 'person-time' as the denominator and estimate the event rate (a mortality rate in this example).

$$\text{event rate} = \frac{\text{number of events}}{\text{person-time at risk}}$$


```
. stset surv_mm, fail(dead) scale(12)
. strate dx93

Estimated rates and lower/upper bounds of 95% confidence intervals
(35 records included in the analysis)
  +-------------------------------------------------------------+
  |    dx93    Events   p-time       Rate      Lower      Upper |
  |-------------------------------------------------------------|
  | dx 1985-92   18   112.9167   0.159410   0.100435   0.253014 |
  | dx 1993-94    1    12.4167   0.080537   0.011345   0.571737 |
  +-------------------------------------------------------------+
```

- The main message is that, in survival analysis, the outcome has two dimensions – the event indicator and the time at risk.
- The event rate is not the only appropriate outcome measure; it is also possible to estimate the proportion surviving (or proportion dying) while controlling for the fact that individuals are at risk for different lengths of time. This, in fact, will be the focus for today's lectures.

Terminology

- In the strictest sense, a *ratio* is the result of dividing one quantity by another. In the sciences, however, it is mostly used in a more specific sense, that is, when the numerator and the denominator are two separate and distinct quantities [14].
- A *proportion* is a type of ratio in which the numerator is included in the denominator, e.g. the incidence proportion (aka cumulative incidence).
- A *rate* is a measure of change in one quantity per unit of another quantity. In epidemiology, rates typically have units events per unit time.
- We will be estimating both proportions (e.g., survival proportions) and rates (e.g., mortality rates) and should recognise that these are conceptually different.

The survivor function

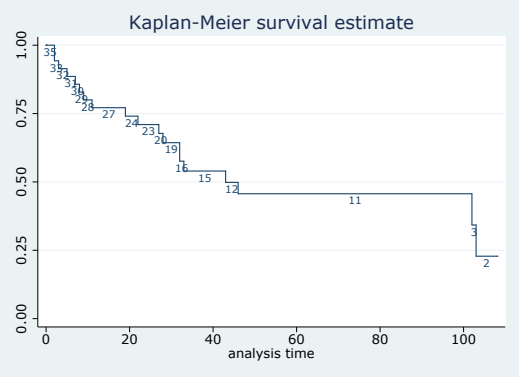


Figure 1: Estimates of $S(t)$ for the 35 patients diagnosed with colon carcinoma. All deaths are considered events ($S(t)$ is called the observed survivor function).

- The survivor function, $S(t)$, gives the probability of surviving until at least time t.
- $S(t)$ is a nonincreasing function with a value 1 at the time origin and a value 0 as t approaches infinity.
- Note that $S(t)$ is a function (the survivor function) which depends on t and should not be referred to as the survival rate.
- The survivor function evaluated at a specific value of t is often referred to as the 'survival rate', for example, the '5-year survival rate'.
- We prefer to use the term 'survival proportion', for example, the '5-year survival proportion'.
- For example, the 5-year survival proportion for the data presented in Figure 1 is 45%.

- Nonparametric methods for estimating $S(t)$ (described later) generally involve estimating the survival proportion at discrete values of t and then interpolating these to obtain an estimate of $S(t)$.

Interpreting $S(t)$ and comparing estimates of $S(t)$ between groups

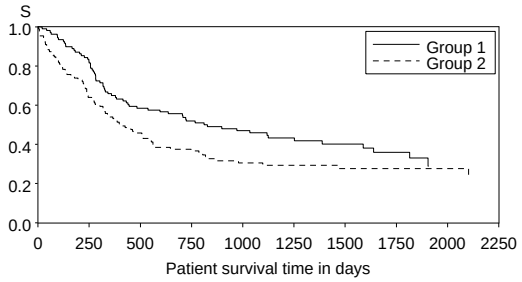


Figure 2: Estimated survivor function (S) for two groups of patients

- Individuals in group 1 experience superior survival compared to individuals in group 2 (even if the long-term survival proportions are similar).
- The gap between the survival curves is decreasing after approximately 850 days.
- It is, however, difficult to determine the essence of the failure pattern, and even more difficult to compare it between groups, simply by studying plots of the survivor function.
- The rate of decline of the survivor function, in survival analysis called the hazard function, $\lambda(t)$, can be thought of as "the speed with which a population is dying".[2]
- When the survival difference is first increasing and then decreasing, is an example of non-proportional hazards, a concept we will return to later.

[2] strictly, the hazard is the rate of change (and the derivative of the negative logarithm) of the survivor function, such that $\lambda(t) = \left(\frac{d}{dt}S(t)\right)/S(t) = -\frac{d}{dt}\ln[S(t)]$.

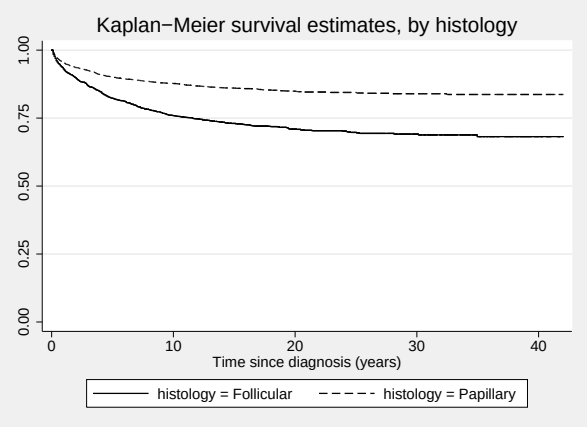
- The survival experience of a cohort can be expressed in terms of the survival proportion or the hazard rate.
- In epidemiological cohort studies where the incidence of a *disease* is the outcome (rather than *death*), we often present the failure proportion, given by $1 - S(t)$, rather than $S(t)$.
- We can model the hazard function (the incidence rate) and estimate the hazard ratio (incidence rate ratio) for the exposed compared to the unexposed.
- Often it is the hazard ratio, rather than the survivor function, which is of primary interest.

The hazard function, $\lambda(t)$

- The term 'hazard rate' is the generic term used in survival analysis to describe the 'event rate'. If, for example, the event of interest is disease incidence then the hazard represents the incidence rate.
- The hazard function, $\lambda(t)$, is the instantaneous event rate at time t, conditional on survival up to time t. The units are events per unit time.
- In contrast to the survivor function, which describes the probability of *not* failing before time t, the hazard function focuses on the failure rate at time t among those individuals who are alive at time t.
- That is, a lower value for $\lambda(t)$ implies a higher value for $S(t)$ and vice-versa.
- Note that the hazard is a rate, not a proportion or probability, so $\lambda(t)$ can take on any value between zero and infinity, as opposed to $S(t)$ which is restricted to the interval $[0, 1]$.

Survival of Swedes with differentiated thyroid cancer

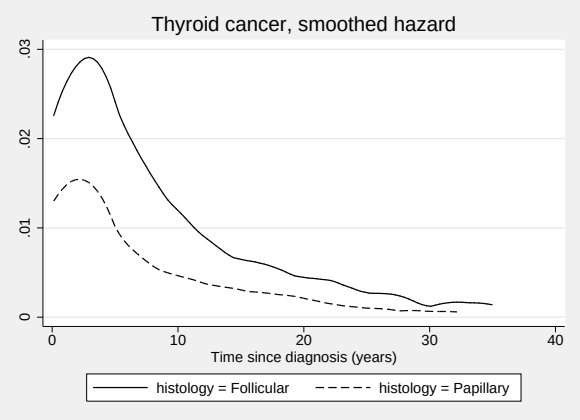
```
sts graph, by(histology)
```



- What do we see? Consider the questions on the following page.

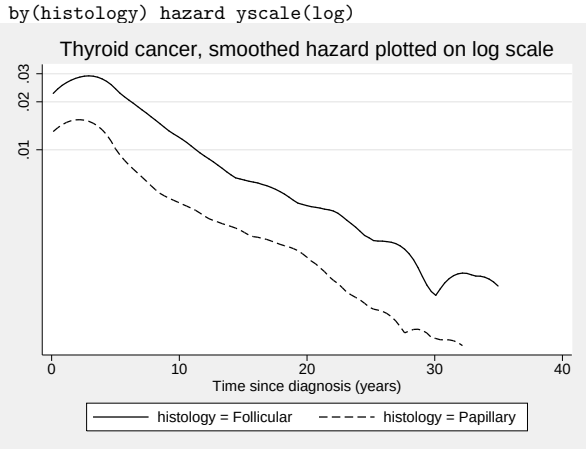
- Which group (histological type) experiences the best survival?
- Does the group with best survival experience lower mortality throughout the follow-up?
- At what point in the follow-up is mortality the highest?

```
sts graph, by(histology) hazard
```



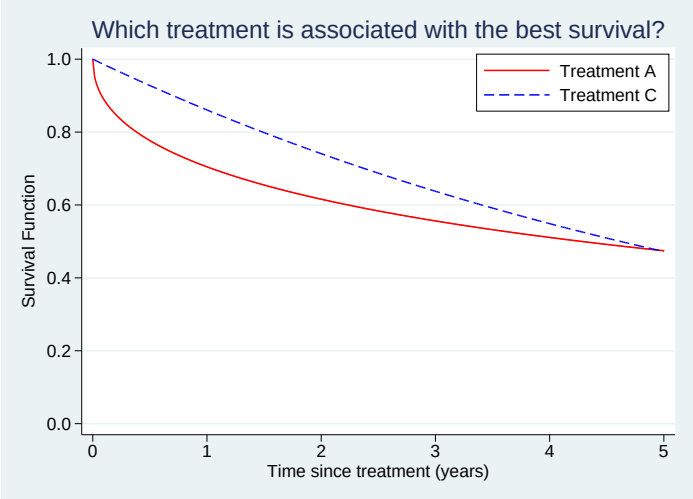
- Is an assumption of proportional hazards appropriate?


```
sts graph, by(histology) hazard yscale(log)
```



- I would be very happy with an assumption of proportional hazards (it doesn't get much better than this).

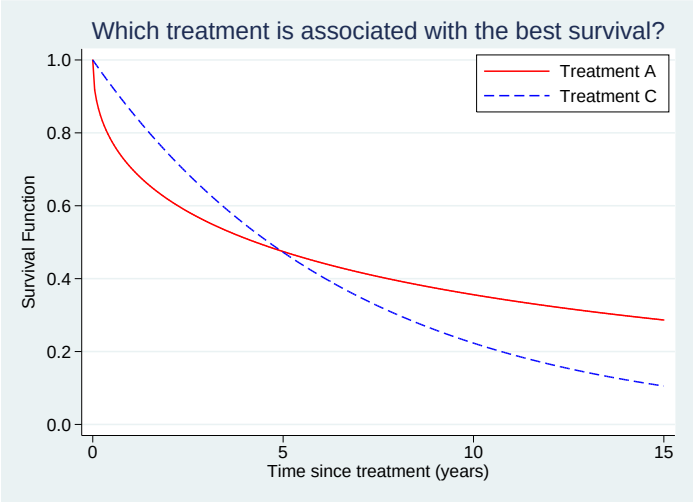
Which group (A or C) has the best survival?



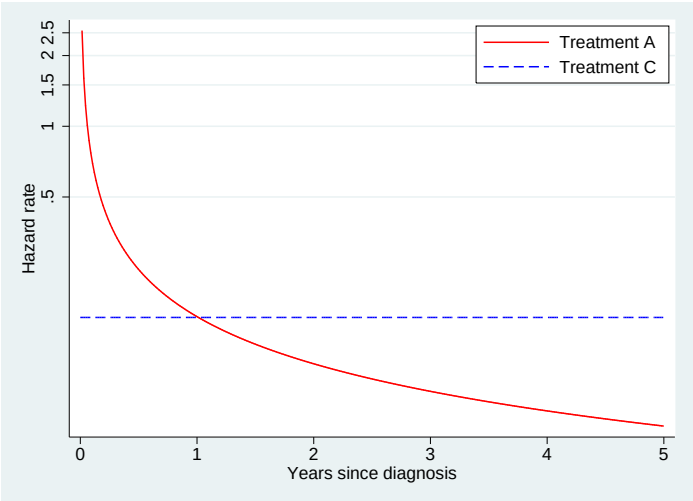
- If we were to base inference solely on the data shown in the graph (i.e., make no assumptions about what happens after 5 years) we would conclude that group C experiences superior survival compared to group A (even if the 5-year survival proportions are similar).

- Patients in group C have lower mortality for the interval up to approximately 15 months following diagnosis but then have higher mortality than group A after 15 months. (Non-proportional hazards)

What about if we extend the follow-up?

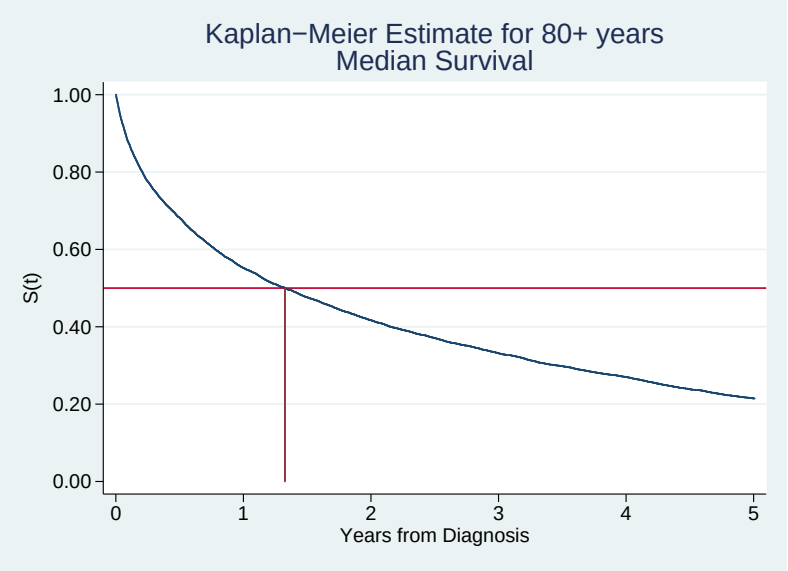


Now plot the two hazard functions



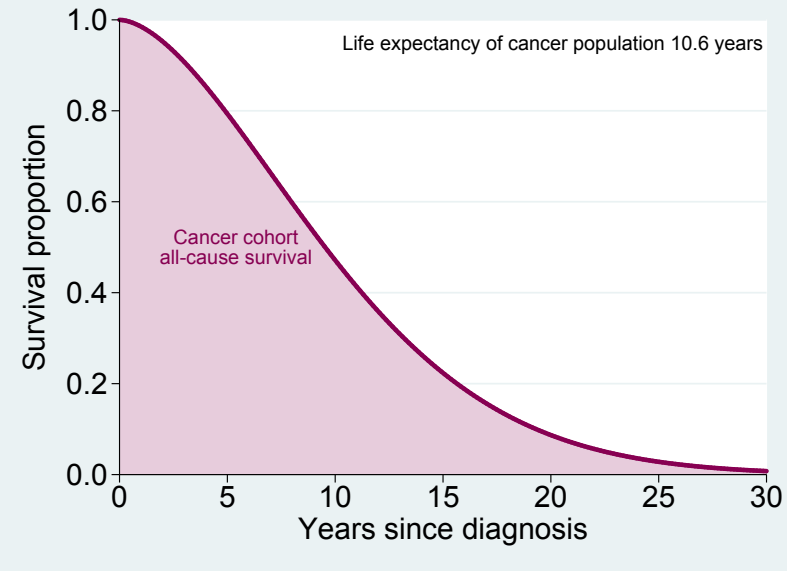
Other measures of survival: Median survival time

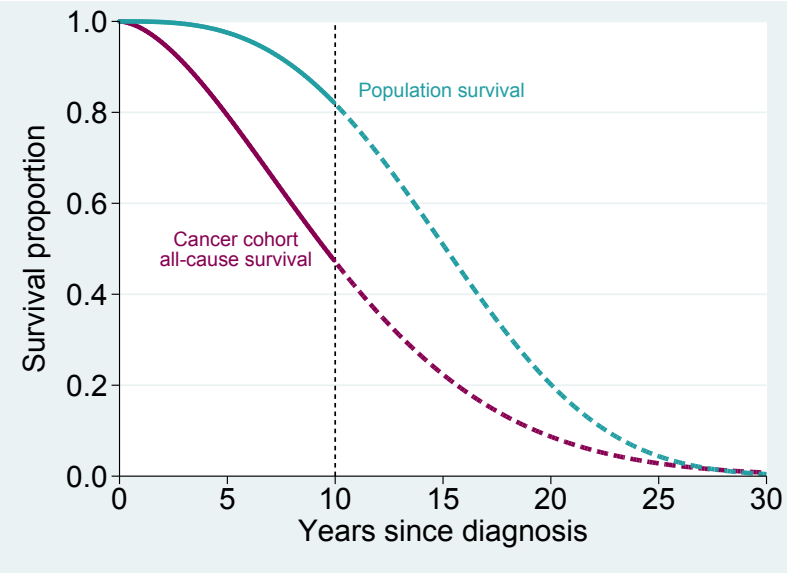
- The median survival time is another measure used to summarise the survival experience of the patients.
- The median survival time is the time at which $S(t) = 0.5$. That is, the time beyond which 50% of the individuals in the population are expected to survive.
- It is estimated by the time at which the estimate of $S(t)$ falls below 0.5.
- The median survival time for the example shown on the next slide is approximately 1.3 years.
- The median can be estimated by extrapolation if the survivor function does not sink below 0.5 during the period the patients are under follow-up.



Other measures of survival: Mean survival time

- The mean survival, i.e. average survival time, is the area under the survival curve (the integrated survival function).
- Not the same as mean follow-up time, i.e. taking the mean of all follow-up times.
- When the survival function doesn't reach 0, the restricted mean survival time can be estimated. Otherwise the survival function has to be extrapolated.





Estimating the survivor function, $S(t)$

- There are two main methods to estimate $S(t)$: The Kaplan-Meier method and the life table method.
- Consider the sample data for the 35 colon cancer patients introduced on slide 61.
- We may be interested in estimating $S(t)$ where the event of interest is death due to any cause.
- An estimate of $S(t)$ could be obtained by simply calculating the proportion of individuals still alive at selected values of t, such as completed years.
- We had 35 patients alive at start. Eight of the 35 patients died during the first year of follow-up so the estimate for $S(1)$ is $\hat{S(1)} = (35 - 8)/35 = 27/35 = 0.771$.

- We encounter problems when attempting to estimate $S(2)$. Ten patients died within two years of follow-up, but 2 patients (patients 9 and 10) could not be followed-up for a full 2 years.

- We could exclude these two patients from the analysis alltogether and let $\hat{S}(2) = (33 - 10)/33$, but this will underestimate the true survival proportion since it ignores the fact that each of these two patients were at risk of death for between one and two years but did not die while under observation.

- If we instead use $\hat{S}(2) = (35 - 10)/35$ then we will overestimate the true survival proportion, since we are assuming that each of these two patients survived for a full two years.

- Two common (and similar) methods for estimating $S(t)$ in the presence of censoring are the lifetable (actuarial) method and the Kaplan-Meier (product-limit) method.

ID	Sex	Age at dx	Clinical stage	dx date (mmyy)	Survt time (mm)	yy	Status
1	male	72	Localised	2.89	2	0	Dead - other
2	male	72	Distant	12.91	2	0	Dead - cancer
3	male	73	Distant	11.93	3	0	Dead - cancer
4	male	63	Distant	6.88	5	0	Dead - cancer
5	male	67	Localised	5.89	7	0	Dead - cancer
6	male	74	Regional	7.92	8	0	Dead - cancer
7	female	56	Distant	1.86	9	0	Dead - cancer
8	female	52	Distant	5.86	11	0	Dead - cancer
9	male	64	Localised	11.94	13	1	Alive
10	female	70	Localised	10.94	14	1	Alive
11	female	83	Localised	7.90	19	1	Dead - other
12	male	64	Distant	8.89	22	1	Dead - cancer
13	female	79	Localised	11.93	25	2	Alive
14	female	70	Distant	6.88	27	2	Dead - cancer
15	male	70	Regional	9.93	27	2	Alive
16	female	68	Distant	9.91	28	2	Dead - cancer
17	male	58	Localised	11.90	32	2	Dead - cancer
18	male	54	Distant	4.90	32	2	Dead - cancer
19	female	86	Localised	4.93	32	2	Alive
20	male	31	Localised	1.90	33	2	Dead - cancer
21	female	75	Localised	1.93	35	2	Alive
22	female	85	Localised	11.92	37	3	Alive
23	female	68	Distant	7.86	43	3	Dead - cancer
24	male	54	Regional	6.85	46	3	Dead - cancer
25	male	80	Localised	6.91	54	4	Alive
26	female	52	Localised	7.89	77	6	Alive
27	male	52	Localised	6.89	78	6	Alive
28	male	65	Localised	1.89	83	6	Alive
29	male	60	Localised	11.88	85	7	Alive
30	female	71	Localised	11.87	97	8	Alive
31	male	58	Localised	8.87	100	8	Alive
32	female	80	Localised	5.87	102	8	Dead - cancer
33	male	66	Localised	1.86	103	8	Dead - other
34	male	67	Localised	3.87	105	8	Alive
35	female	56	Distant	12.86	108	9	Alive

Summary of possible approaches to estimating $S(2)$

- We've now seen one approach that leads to an overestimate and one that leads to an underestimate.

$$\frac{35-10}{35} = 0.714 \text{ is an overestimate.}$$

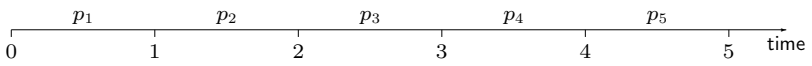
$$\frac{34-10}{34} = 0.706 \text{ reasonable estimate?}$$

$$\frac{33-10}{33} = 0.670 \text{ is an underestimate.}$$

- We don't actually use $\frac{34-10}{34}$ as an estimate of $S(2)$ but we do make a similar type of adjustment.

The general approach to nonparametric estimation of $S(t)$

- Assume we wish to estimate five year survival, $S(5)$.



- We start by estimating the following conditional survival probabilities:
  p_1, the probability of surviving at least 1 year from time 0
  p_2, the probability of surviving at least 2 years conditional on surviving 1 year
  p_3, the probability of surviving at least 3 years conditional on surviving 2 years
  p_4, the probability of surviving at least 4 years conditional on surviving 3 years
  p_5, the probability of surviving at least 5 years conditional on surviving 4 years

- The probability of surviving at least 5 years (from time zero) is then given by the product of these conditional survival probabilities.

$$S(5) = \prod_{i=1}^{5} p_i$$

- That is, to survive five years one must survive year 1 and year 2 and year 3, and year 4, and year 5.

- The advantage of this approach is that we can appropriately account for censoring when estimating the probability of surviving a small time interval (i.e., when estimating the conditional survival probabilities).

- The cumulative survival is estimated as the product of conditional survival proportions, where the estimate of each conditional survival proportion is based upon only those individuals under follow-up.

- That is, the individuals who are censored are assumed to have the same prognosis as those individuals who could be followed up.

- This requires the assumption that censoring is *non-informative*.

- That is, we make the assumption that, conditional on the values of any explanatory variables, censoring is unrelated to prognosis (the probable course and outcome of the disease).

- If censoring was informative, for example if censored were more likely to die, then we would be left with healthier patients in the study, showing a better survival than the true survival of the patients.

- More on informative censoring later during the course.

- This approach is employed by both the lifetable (actuarial) method and the Kaplan-Meier (product-limit) method.
- We chose, arbitrarily, to estimate conditional probabilities for one year intervals (time-bands) but the intervals may be any width.
- The primary differences between the lifetable and Kaplan-Meier methods is the manner in which the intervals are chosen (not really a difference in theory) and the method for dealing with ties.
- If two individuals have the same survival time (time to event or time to censoring), we say that the survival times are 'tied'.
- Many of the standard methods for survival analysis, such as the Kaplan-Meier method and the Cox proportional hazards model, assume that survival time is measured on a continuous scale and that ties are therefore rare.
- In population-based survival analysis, however, ties are common.

Life table method for estimating $S(t)$

- Also known as the 'actuarial method'. The approach is to divide the period of observation into a series of time intervals and estimate the conditional (interval-specific) survival proportion for each interval.
- The cumulative survivor function, $S(t)$, at the end of a specified interval is then given by the product of the interval-specific survival proportions for all intervals up to and including the specified interval.
- In the absence of censoring, the interval-specific survival proportion is $p = (l - d)/l$, where d is the number of events (deaths) observed during the interval and l is the number of patients alive at the start of the interval.
- In the presence of censoring, it is assumed that censoring occurs uniformly throughout the interval such that each individual with a censored survival time is at risk for, on average, half of the interval. This assumption is known as the actuarial assumption.

- The effective number of patients at risk during the interval is given by $l' = l - \frac{1}{2}w$ where l is the number of patients alive at the start of the interval and w is the number of censorings during the interval.
- The estimated interval-specific survival proportion is then given by $p = (l' - d)/l'$.
- The cumulative survival is estimated as the product of conditional survival proportions, where the estimate of each conditional survival proportion is based upon only those individuals under follow-up.

$$S(t_k) = \prod_{i=1}^{k} p_i$$

Table 3: Life table with annual interval for the 35 patients.

time	l	d	w	l'	p	$S(t)$
[0-1)	35	8	0	35.0	0.77143	0.77143
[1-2)	27	2	2	26.0	0.92308	0.71209
[2-3)	23	5	4	21.0	0.76190	0.54254
[3-4)	14	2	1	13.5	0.85185	0.46217
[4-5)	11	0	1	10.5	1.00000	0.46217
[5-6)	10	0	0	10.0	1.00000	0.46217
[6-7)	10	0	3	8.5	1.00000	0.46217
[7-8)	7	0	1	6.5	1.00000	0.46217
[8-9)	6	2	3	4.5	0.55556	0.25676
[9-10)	1	0	1	0.5	1.00000	0.25676

- l is the number alive at the start of the interval
- d is the number of events (deaths) during the interval
- w is the number of censorings (withdrawals) during the interval
- l' is the effective number at risk for the interval
- p is the interval-specific survival proportion
- $S(t)$ is the estimated cumulative survivor function (proportion) at the end of the interval

The Kaplan-Meier method for estimating $S(t)$

- Also known as the product-limit method but is more commonly known as the Kaplan-Meier method, after the two researchers who first published the method in English in 1958 [19]. The method was published earlier (1912) in German [19, 3].
- Rather than using pre-specified time intervals, interval-specific survival is estimated at each event time.
- To obtain Kaplan-Meier estimates of survival, the patient survival times are first ranked in increasing order.
- The times where events (deaths) occur are denoted by t_i, where $t_1 < t_2 < t_3 < \ldots$.
- The number of deaths occurring at t_i is denoted by d_i.

- If both censoring(s) and death(s) occur at the same time, then the censoring(s) are assumed to occur immediately after the death time.
- That is, individuals with survival times censored at t_i are assumed to be at risk at t_i.
- In essence, the Kaplan-Meier method uses an interval size decreased towards zero so that the number of intervals tends to infinity.
- The Kaplan-Meier method was developed for applications where survival time is measured on a continuous scale.
- In practice, survival time is measured on a discrete scale (e.g. minutes, hours, days, months, or years) so the interval length is limited by the accuracy to which survival time is measured.
- We should therefore use as accurate a time scale as possible. That is, don't base the estimate on time in days if time in minutes is also known.

- In practice, only those intervals containing an event contribute to the estimate, so we can ignore all other intervals.

- The Kaplan-Meier estimate of the cumulative survivor function at time t is given by

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t}(1 - \frac{d_i}{l_i}) & \text{if } t \geq t_1 \end{cases} \quad (1)$$

  where l_i is the number of persons at risk.

- A plot of the Kaplan-Meier estimate of the survivor function (slide 39) takes the form of a step function, in which the survival probabilities decrease at each death time and are constant between adjacent deaths times.

- Censorings do not affect the estimate of $S(t)$, but contribute in Equation 1 by decreasing l_i at the next death time.

- If the largest observed survival time (which we will call t_z) is a censored survival time, then $\hat{S}(t)$ is undefined for $t > t_z$, otherwise $\hat{S}(t) = 0$ for $t > t_z$.

- The standard error of the estimate can be obtained using Greenwood's method [16] (slide 97).

- Confidence intervals an be obtained as described on slide 99.

- As for the life table method, non-informative censoring is assumed.

K-M estimates for the sample data (up to 25 months)

t	at risk	observed deaths	p_i	$S(t)$	SE
0	35	0	1.0000	1.0000	–
2	35	2	0.9429	0.9429	0.0392
3	33	1	0.9697	0.9143	0.0473
5	32	1	0.9688	0.8857	0.0538
7	31	1	0.9677	0.8571	0.0591
8	30	1	0.9667	0.8286	0.0637
9	29	1	0.9655	0.8000	0.0676
11	28	1	0.9643	0.7714	0.0710
13+	27	0			
14+	26	0			
19	25	1	0.9600	0.7406	0.0745
22	24	1	0.9583	0.7097	0.0776
25+	23	0			
		...			

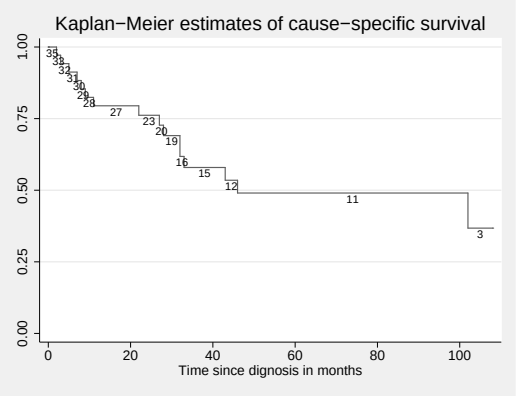


Figure 3: Estimates of $S(t)$ for the 35 patients diagnosed with colon carcinoma. All deaths are considered events ($S(t)$ is called the observed survivor function).

Summary: nonparametric estimation of $S(t)$

1. Split follow-up into intervals (timebands). If there are both deaths and censorings within an interval then

K-M: Assume the events precede the censorings, that is, everyone is at risk when the events occur.

Actuarial: Assume half of the censored individuals are at risk when the events occur.

2. Estimate conditional probabilities of surviving each interval

$$p_i = 1 - d_i/n_i$$

where d_i is the number of events and n_i number at risk for interval i.

3. $S(t)$ is the product of the conditional probabilities up to time t.

$$S(t_k) = \prod_{i=1}^{k} p_i$$

- The only difference between the Kaplan-Meier method and the actuarial method is the approach to dealing with ties (which affects the value of n_i in estimating the conditional probabilities), and how the intervals are chosen.
- The Kaplan-Meier approach is slightly biased in the presence of ties so one should define time as accurately as possible (e.g., don't use time in months if you have time in days) in order to minimise the number of ties.
- If survival times are generated on a truly discrete scale (e.g., patients are contacted annually to ascertain vital status) and ties are common then the actuarial approach is preferable.
- The actuarial method can, however, also be used with many small intervals.

Testing for differences in survival between groups

- Comparing survival at a fixed time point (e.g. five years) wastes available information.
- It is invalid to compare the proportion surviving at a given time, based on the comparison of two binomial proportions, where the time point for comparison is chosen after viewing the estimated survivor functions (e.g. testing for a difference at the point where the Kaplan-Meier curves show the largest difference).
- Various tests are available (parametric and non-parametric) for testing equality of survival curves. The most common is the log rank test, which is non-parametric.
- Start by tabulating the number at risk in each exposure group and the total number of events (deaths) at every time point when one of more deaths occur.

- Under the null hypothesis that the two survival curves are the same, the expected number of deaths in each group will be proportional to the number at risk in each group.
- For example (see slide 81), at $t = 2$ months we observed 2 deaths (one male and one female). Conditional on 2 deaths being observed, we would expect $2 \times 19/35 = 1.086$ deaths among the 19 males at risk and $2 \times 16/35 = 0.914$ deaths among the 16 females at risk.
- Now calculate the totals of the observed and expected number of deaths for each group (1=males, 2=females), calling them O_1, O_2, E_1, and E_2, and calculate the following test statistic

$$\theta = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}. \quad (2)$$

- Under the null hypothesis, θ will approximately follow a χ^2 distribution with 1 degree of freedom. That is, if θ is greater than 3.84 then we reject the null hypothesis and conclude that there is a statistically significant difference between the two survival curves.

Log rank test for comparing survival of males and females

event	males			females		
time	at risk	obs	exp	at risk	obs	exp
2	19	1	1.086	16	1	0.914
3	18	1	0.545	15	0	0.455
5	17	1	0.531	15	0	0.469
7	16	1	0.516	15	0	0.484
8	15	1	0.500	15	0	0.500
9	14	0	0.483	15	1	0.517
11	14	0	0.500	14	1	0.500
19	13	0	0.520	12	1	0.480
22	13	1	0.542	11	0	0.458
27	12	0	0.545	10	1	0.455
28	11	0	0.550	9	1	0.450
32	11	2	1.158	8	0	0.842
33	9	1	0.563	7	0	0.438
43	8	0	0.615	5	1	0.385
46	8	1	0.667	4	0	0.333
102	2	0	0.500	2	1	0.500
103	2	1	0.667	1	0	0.333

Totals: $O_1 = 11$, $E_1 = 10.488$, $O_2 = 8$, $E_2 = 8.512$

- The test statistic is $\theta = (O_1 - E_1)^2/E_1 + (O_2 - E_2)^2/E_2 = 0.056$, which is less than 3.84 implying no evidence of a difference in survival between males and females.
- For k groups, the log rank test statistic is

$$\theta = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

  which has an approximate χ^2_{k-1} distribution under the null hypothesis.
- The log rank test is designed to be sensitive to departures from the null hypothesis in which the two hazards (instantaneous death rates) are proportional over time. It is very insensitive to situations in which the hazard functions cross.
- The log rank test puts equal weight on every failure (irrespective of the number at risk at the time of the failure).

- An alternative test, the generalised Wilcoxon test, is constructed by weighting the contribution of each failure time by the total number of individuals at risk and is consequently more sensitive to differences early in the follow-up period (when the number at risk is larger).
- The Wilcoxon test is more powerful than the log rank test if the proportional hazards assumption does not hold.
- It is difficult to apply the log rank test while simultaneously controlling for potential confounding variables (a regression approach is preferable).
- In a randomised clinical trial, however, potential confounders are controlled for in the randomisation, so we can use the log rank test to compare survival curves for the different treatment groups.
- The log rank test provides nothing more than a test of statistical significance for the difference between the survival curves, it tells us nothing about the size of the difference. A regression approach allows us to both determine statistical significance and to estimate the size of the effect.

Testing for differences in survival – Summary of key points

- Various tests are available for testing equality of survival curves, the most well-known being the log rank test.
- These tests are rarely used in observational epidemiology; we prefer to use modelling since it:
  1. provides estimates of the size of the effect (i.e., rate ratios); the log-rank test just gives a p-value;
  2. provides greater possibilities for confounder control and effect modification.
- The log-rank test assumes proportional hazards.
- Consider the situation where we have two groups; a Cox model with one explanatory variable gives us everything the log-rank test does (a p-value). It also gives us the estimated hazard ratio and CI but, more importantly, it is simple to extend the model to compare survival between the two groups while controlling for potential confounders.

Survival analysis using Stata

- In order to analyse survival data it is necessary to specify (at a minimum) a variable representing survival time and a variable specifying whether or not the event of interest was observed (called the failure variable).
- Instead of specifying a variable representing survival time we can specify the entry and exit dates (this is necessary if subjects enter the study at different times).
- In many statistical software programs (such as SAS), these variables must be specified every time a new analysis is performed.
- In Stata, these variables are specified once using the `stset` command and then used for all subsequent survival analysis (`st`) commands (until the next `stset` command).

- For example

```
. use melanoma
. stset surv_mm, failure(status==1)
```

- The above code shows how we would `stset` the skin melanoma data in order to analyse cause-specific survival with survival time in completed months (`surv_mm`) as the time variable.
- Of the four possible values of `status`, we have specified that only code 1 indicates an event (death due to melanoma).
- If we wanted to analyse observed survival (where all deaths are considered to be events) we could use the following command

```
. stset surv_mm, failure(status==1,2)
```

- Some of the Stata survival analysis (st) commands relevant to this course are given below. Further details can be found in the manuals or online help.

```
stset       Declare data to be survival-time data
stsplit     Split time-span records
stdes       Describe survival-time data
stsum       Summarize survival-time data
sts         Generate, graph, list, and test the survivor
                and cumulative hazard functions
strate      Tabulate failure rate
stptime     Calculate person-time at risk and failure rates
stcox       Estimate Cox proportional hazards model
stphtest    Test of Cox proportional hazards assumption
stphplot    Graphical assessment of the Cox
              proportional hazards assumption
stcoxkm     Graphical assessment of the Cox
              proportional hazards assumption
streg       Estimate parametric survival models
```


- Once the data have been stset we can use any of these commands without having to specify the survival time or failure time variables.

- For example, to plot Kaplan-Meier estimates of the cause-specific survivor function by sex and then fit a Cox proportional hazards model with sex and calendar period as covariates

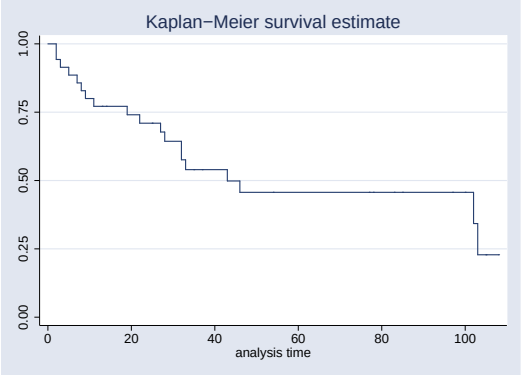
```
. sts graph, by(sex)
. stcox sex year8594
```


Kaplan-Meier estimates in Stata

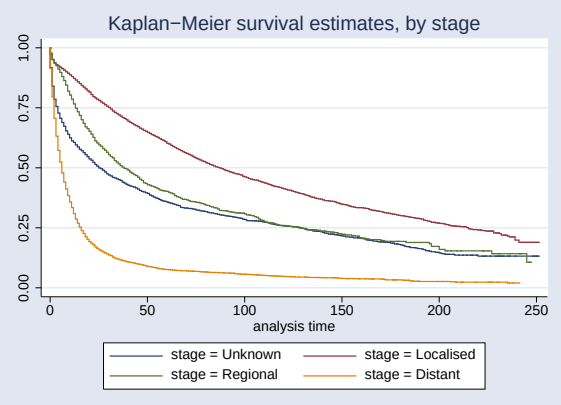
```
. stset surv_mm, failure(status==1,2)
. sts list
       Beg.          Net     Survivor    Std.
Time  Total   Fail   Lost    Function   Error    [95% CI]
-------------------------------------------------------------
   2     35      2      0      0.9429  0.0392  0.7903  0.9854
   3     33      1      0      0.9143  0.0473  0.7573  0.9715
   5     32      1      0      0.8857  0.0538  0.7236  0.9555
   7     31      1      0      0.8571  0.0591  0.6903  0.9379
   8     30      1      0      0.8286  0.0637  0.6577  0.9191
   9     29      1      0      0.8000  0.0676  0.6258  0.8992
  11     28      1      0      0.7714  0.0710  0.5946  0.8785
  13     27      0      1      0.7714  0.0710  0.5946  0.8785
  14     26      0      1      0.7714  0.0710  0.5946  0.8785
  19     25      1      0      0.7406  0.0745  0.5603  0.8558
  22     24      1      0      0.7097  0.0776  0.5271  0.8323
  25     23      0      1      0.7097  0.0776  0.5271  0.8323
-------------------------------------------------------------
```

Plotting Kaplan-Meier estimates of $S(t)$ using Stata

```
. use http://www.biostat3.net/download/colon_sample, clear
. stset surv_mm, failure(status==1,2)
. sts graph
```




```
. use http://www.biostat3.net/download/colon
. stset surv_mm, failure(status==1,2)
. sts graph, by(stage)
```



Log rank test in Stata

```
. use colon_sample
. stset surv_mm, failure(status==1,2)
. sts test sex
Log-rank test for equality of survivor functions
------------------------------------------------
       |  Events
sex    |  observed       expected
-------+-------------------------
Male   |        11          10.49
Female |         8           8.51
-------+-------------------------
Total  |        19          19.00
             chi2(1) =        0.06
             Pr>chi2 =      0.8113
```

- The log rank test is non-significant indicating no difference in survival between males and females (if the assumptions hold –e.g. no uncontrolled confounding, proportional hazards and non-informative censoring).

The same test as a Cox model

```
. use colon_sample
. stset surv_mm, failure(status==1,2)
. stcox sex

No. of subjects =          35        Number of obs  =       35
No. of failures =          19
Time at risk    =        1504
                                     LR chi2(1)     =     0.06
Log likelihood  =  -56.259206        Prob > chi2    =   0.8118
------------------------------------------------------------------
 _t |  HR       Std. Err.     z    P>|z|  [95% Conf. Interval]
----+-------------------------------------------------------------
sex | 0.89501   .4179592   -0.24   0.812  .3583709    2.235266
------------------------------------------------------------------

. di (-0.24)^2
.0576
```


Summary of Day 1

- Time-to-event analysis (survival analysis) is necessary when
  - We are interested in studying the time to an event, e.g. time to diagnosis, time to death
  - Individuals in a study are followed for different lengths of time, and therefore are 'at risk' for different amounts of time, e.g. in cohort studies.
- The outcome in survival analysis consists of both an event indicator (0/1) and a time dimension (continuous).
- The outcome can be expressed as either a survival proportion or an event rate (hazard). Comparison between groups are primarily made using hazard ratios.
- The survivor function (survival proportion) can be estimated using several alternative methods, e.g. the Kaplan-Meier method and the life table method.
- The log rank test can be used to test for differences in survival, but is rarely used in observational epidemiology.

- In observational epidemiology we prefer modelling since it:
  - enables us to compare survival between exposure categories while controlling for confounding (although we can also perform an adjusted log rank test).
  - places a focus on estimation rather than testing (i.e., we obtain estimated hazard ratios and CIs).
  - enables us to study effect modification.
  - is extentable in other useful ways.

Exercises for Day 1

100. Hand calculation: Kaplan-Meier estimates of cause-specific survival (35 patients)

101. Kaplan-Meier estimates of cause-specific survival using Stata (35 patients)

102. Localised melanoma: Comparing various approaches to estimating the 10-year survival proportion

103. Melanoma: Comparing survival proportions and mortality rates according to stage

104. Localised melanoma: Comparing estimates of cause-specific survival between periods; first graphically and then using the log rank test

Appendix day 1: Estimating the standard error and confidence intervals for estimated survival proportions

- The most widely used method for estimating the standard error of the estimated survival proportion is the method described by Greenwood (1926) [13, 16].

- Appropriate for both the actuarial and Kaplan-Meier methods.

- Appropriate for both observed and cause-specific survival.

- Known as Greenwood's method or Greenwood's formula. The formula,

$$\mathrm{SE}({}_1p_i) = {}_1p_i \left[\sum_{j=1}^{i} \frac{d_j}{l'_j(l'_j - d_j)} \right]^{\frac{1}{2}}, \qquad (4)$$

(where l is the number of patients alive at the start of the interval, w is the number of censorings during the interval, and $l' = l - \frac{1}{2}w$) is slightly laborious for hand calculation, but readily available in many computer programs.

- This is the default method for the software used in this course.

- Non-integer values for l'_i, e.g. $l'_i = 20.5$, do not cause any problems in practical use.

- For a single interval, Equation 4 reduces to

$$\mathrm{SE}(p_i) = p_i \left\{ \frac{d_i}{l'_i(l'_i - d_i)} \right\}^{\frac{1}{2}} = \sqrt{p_i(1-p_i)/l'_i},$$

  which is the familiar binomial formula for the standard error of the observed interval-specific survival proportion based on l'_i trials.

- It can also be shown for the general case that Equation 4 reduces to the binomial standard error in the absence of censoring.

- Confidence intervals can be calculated for any estimated survival proportion in order to provide a measure of uncertainty associated with the point estimate.
- A 95% confidence interval (CI) is an interval, i.e. a range of values, such that under repeated sampling, the true survival proportion will be contained in the interval 95% of the time (if the model is correct).
- The CI is often called an interval estimate for the true survival proportion, while the estimated survival proportion is called the point estimate.
- A confidence interval for the true survival proportion can be obtained by assuming that the estimated survival proportion is normally distributed around the true value with estimated variance given by the square of the standard error.
- A two-sided $100(1-\alpha)\%$ confidence interval ranges from $p - z_{\alpha/2}\text{SE}(p)$ to $p + z_{\alpha/2}\text{SE}(p)$, where p is the estimated survival proportion (which can be an

interval-specific or cumulative), $\text{SE}(p)$ the associated standard error, and $z_{\alpha/2}$ the upper $\alpha/2$ percentage point of the standard normal distribution.

- For a 95% confidence interval, $z_{\alpha/2} = 1.96$, and for a 99% confidence interval, $z_{\alpha/2} = 2.58$.
- The standard error of the observed and cause-specific survival proportion can be obtained using Greenwood's method (slide 97).
- As a rule of thumb, the normal approximation for a single interval i is usually appropriate when both $l_i' p_i$ and $l_i'(1-p_i)$ are greater than or equal to 5 [2].
- Confidence intervals obtained in this way are symmetric about the point estimate and can sometimes contain implausible values for the survival proportion, i.e., values less than zero or greater than one.
- One method of obtaining confidence intervals for the observed survival proportion in the range [0,1] is to transform the estimate to a value in the

range $[-\infty, \infty]$, obtain a confidence interval for the transformed value, and then back-transform the confidence interval to [0,1].

- One such transformation is the complementary log-log transformation, $\ln[-\ln(p)]$, which is equivalent to constructing the confidence intervals on the log cumulative hazard scale.
- To estimate confidence intervals for the survival proportion using this method, we first transform the estimated cumulative observed survival rate (OSR).
- We will write this transformation as $g(\text{OSR}) = \ln[-\ln(\text{OSR})]$, where g is the complementary log-log transformation.
- We also require an estimate of the variance of the OSR on the log hazard scale.
- Using a Taylor series approximation[3], the variance of a function, g, of a

[3] In this setting, this is called the *delta method*.

random variable, X, can be approximated by

$$\text{var}\{g(X)\} \approx \left\{\frac{\text{d}g(X)}{\text{d}X}\right\}^2 \text{var}(X)$$

- If we denote the cumulative observed survival proportion by X then, noting that
$$\frac{\text{d}\ln[f(X)]}{\text{d}X} = \frac{1}{f(X)}\frac{\text{d}f(X)}{\text{d}X},$$
we have
$$\text{var}\{g(X)\} = \text{var}\{\ln[-\ln(X)]\} \approx \frac{1}{[X\ln(X)]^2}\text{var}(X).$$

- An estimated 95% confidence interval on the log hazard scale is therefore given by $g(\text{OSR}) \pm 1.96\sqrt{\text{var}\{g(OSR)\}}$, which is then back-transformed to give a 95% confidence interval for the OSR.

Topics for Day 2

- Rates and person-time
- Hazard rates and hazard function
- Time scales
- Estimating and modelling constant rates, using Poisson regression
- Main effect models
- Interaction models and parameterisation
- Confounding by time scale and time-varying rates
- Effect modification by time scale (see day 3)

Rates and person-time

- A cohort study is characterized by persons being followed until either an event or censoring.
- The rate is a measure of event occurrence in the cohort.
- Because persons in a cohort are followed for different lengths of time due to censoring, we cannot calculate risks as "number of cases" divided by "number of persons".
$$\text{risk} = \frac{\text{events}}{\text{persons at risk}}.$$
- We must use a denominator which takes different lengths of follow-up into account.
$$\text{rate} = \frac{\text{events}}{\text{time at risk}}.$$
- Persons followed for a longer time have a larger chance of having the event, since they are under observation for a longer time.

- If the cohort was followed for a longer time, then we would expect more events to occur than if the cohort was followed for a shorter time (given that the same number of persons were at risk).
- Time-at-risk or person-time is measured in units of person-years, person-months or similar.
- Person-time is a method of measurement combining persons and time; it is used to aggregate the total population at risk assuming that 10 people at risk for one year is equivalent to 1 person at risk for 10 years.
- If five people are followed for one year, they are followed for 5 person-years.
- If two persons are followed for 2.5 years, they are followed for 5 person-years.
- A *rate* is a measure of change in one quantity per unit of another quantity. In epidemiology, rates typically have units 'events per unit time'.
  - Mortality rate: 0.5 deaths per 1,000 person-years

  - Incidence rate: 14 cancers per 100,000 person-years
- Mortality rates and incidence rates are *event rates*.
- The term 'hazard rate' (or 'hazard') is the generic term used in survival analysis to describe the 'event rate'. If, for example, the event of interest is disease incidence then the hazard rate represents the incidence rate.

$$\text{hazard rate} = \frac{\text{events}}{\text{time at risk}}.$$

- If five people are followed for one year, and one experience a cancer, then the incidence rate is $1/5 = 0.2$ cases per person-year.
- If two persons are followed for 2.5 years, and one experience a cancer, then the incidence rate is $1/5 = 0.2$ cases per person-year.

- Often disease incidences are reported per 100,000 person-years. For example, an incidence rate of 4 per 100,000 person-years is equivalent to 0.04 per 1,000 person-years and 0.00004 per person-year.

Hazard rates and the hazard function, $\lambda(t)$

- In contrast to the survivor function, which describes the probability of *not* failing before time t, the hazard function focuses on the failure rate at time t among those individuals who are alive at time t. So, the survival function is formally defined for a random time variable T by

$$S(t) = \Pr(T > t) = 1 - F(t). \tag{5}$$

  where $F(t)$ is the failure proportion (aka the cumulative density function).

- The hazard function is formally defined for a random time variable T by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{6}$$

- The hazard function shows how the hazard rate varies over time.

- The hazard function, $\lambda(t)$, is the instantaneous event rate at time t, conditional on survival up to time t.

- It can be thought of the 'speed with which the cohort experiences the event over time' or an 'instantaneous risk of the event over time'.

- From Equation 6, one can see that $\lambda(t)\Delta t$ may be viewed as the 'approximate' probability of an individual who is alive at time t experiencing the event in the next small time interval Δt.

- The units are events per unit time.

- Note that the hazard is a rate, not a probability, so $\lambda(t)$ can take on any value between zero and infinity, as opposed to $S(t)$ which is restricted to the interval $[0, 1]$.

- A lower value for $\lambda(t)$ implies a higher value for $S(t)$ and vice-versa.

- One relationship of particular importance is

$$\begin{aligned} S(t) &= \exp\left[-\int_0^t \lambda(s)\,\mathrm{d}s\right] \\ &= \exp(-\Lambda(t)), \end{aligned} \tag{7}$$

  where $\Lambda(t)$ is called the cumulative hazard (or integrated hazard) at time t. The cumulative hazard has no simple interpretation and is rarely used or reported for epidemiological purposes.

Constant and time-varying rates

- Rates can be constant over time, or they can vary over time. [Here: time means 'time scale', e.g. age, time-since-diagnosis, etc.]
- If the rate is constant, then it can easily be estimated as

$$\text{hazard rate} = \frac{\text{events}}{\text{time at risk}}.$$

- This is an overall rate or average rate which is assumed to be the same (i.e. constant) across time.
- If the rate is time-varying, however, then we must account for the time scale when we estimate the rate.
- Rates may be constant on one time scale, but vary across another time scale.

- It is also important to separate between 'time at risk' (i.e. amount of risk time in the denominator) and 'time scale' (i.e. on which scale is the risk time measured).

Choice of time scale

- There are several time scales along which rates might vary. These differ from one another only in the choice of *time origin*, the point at which time is zero.
- Consider the following questions?
  - What is the time?
  - How old are you?
  - For how long have you lived at your current address?
- What is the time origin for each? When was time zero? When did the clock start?
- In which units did you specify time? Could different units have been used?
- Time progresses in the same manner but, in answering these questions, we have applied a different time origin and used different units.

Change in time scale - change in alignment of risk times

- Same cohort, same amount of risk time, but different time scales.

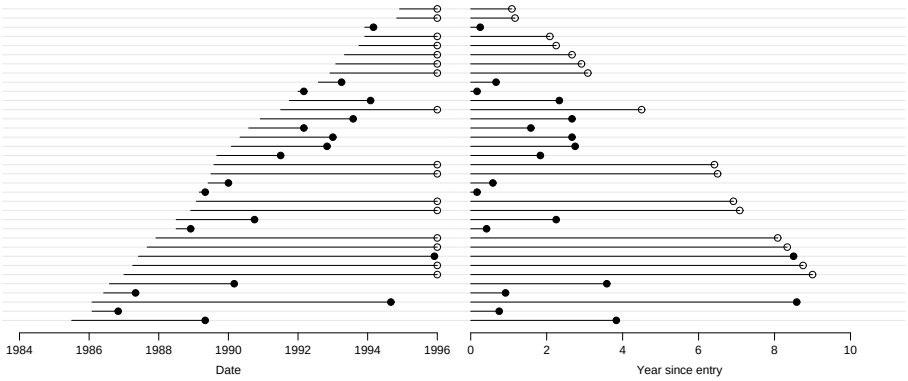


Figure 4: Calendar time (left) and time from entry in years (right)

- Same constant (average, overall) rate for both time scales, but different time-varying rates across the time scales.

- The time-varying rates depend on where the events occur and where the risk time is distributed along the time scale.

Common time scales in epidemiology

Origin	Time scale
Birth	Age
A fixed date	Calendar time
First exposure	Time exposed
Entry into study	Time in study
Disease onset	Time since onset
Diagnosis	Time since diagnosis
Start of treatment	Time on treatment

- In many of the methods used in survival analysis, effects are adjusted for the underlying time scale. Choice of time scale therefore has important implications.

- On many time scales, subjects do not enter follow-up at the time origin, $t = 0$.

- To deal with these issues `stset` has two additional options, 'origin' to specify the origin of time, and 'enter' to specify the time of entry to the study (when a person starts being at risk).

Time scales in the diet data

- We will stset the diet data using three different time scales.
- In the diet data we have the following variables
  - date of entry = `doe`
  - date of exit = `dox`
  - event indicator = `chd`
- Stset will generate time variables (start and end) needed for the analysis, and also set the time scale for the analysis.
- To stset the time scale as *time since entry*, we specify `doe` as origin:

```
. stset dox, fail(chd) enter(doe) origin(doe) scale(365.24)
```

- Each individual enters the study (becomes 'at risk') at the date specified by `doe`.

- The date of entry is also the time origin (time zero).
- By specifying `scale(365.24)` we are scaling the time unit from days to years.
- To use *attained age* as the time scale we specify

```
. stset dox, fail(chd) enter(doe) origin(dob) scale(365.24)
```

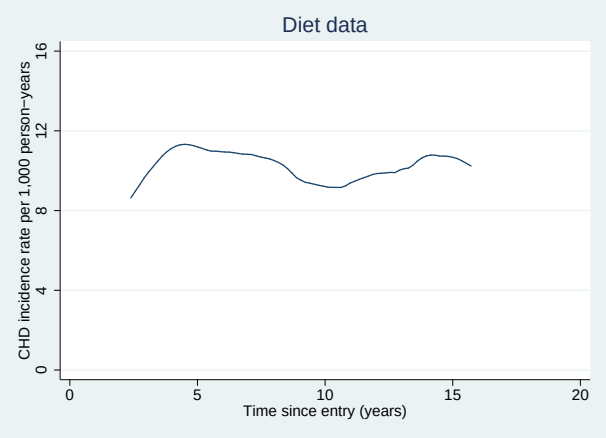
- Individuals enter the study at `doe` (as before) but the time origin is now the date of birth.
- To use *calendar time* as the time scale we specify a fixed date as the time origin. For example

```
. stset dox, fail(chd) enter(doe) origin(d(1/1/1900))
```

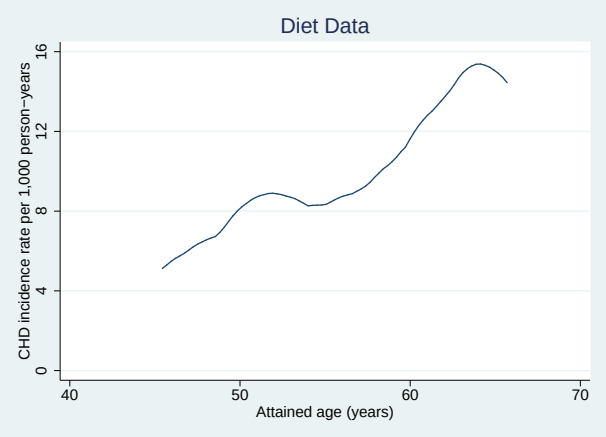
- Rates may be constant over one time scale, while they may vary over another time scale.

- Time varying rates can be estimated as average rates (events over person-time) within segments of time. If we put a smoother across those segments, we may see the following graphs.

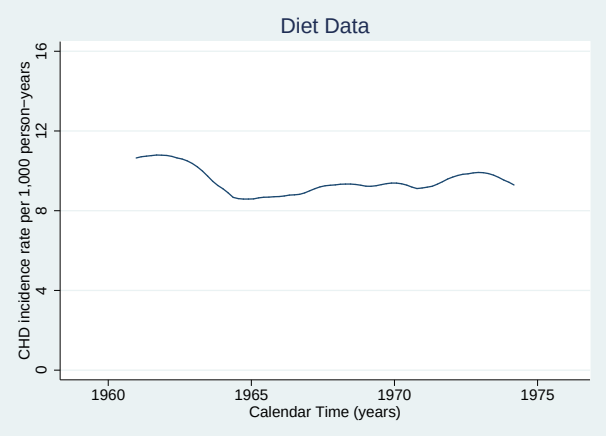
CHD rate with time-since-entry as the time scale



CHD rate with attained age as the time scale



CHD rate with calendar time as the time scale



Estimating CHD rates in the diet data

- We first `stset` the data using time since entry as the timescale.

```
. stset dox, fail(chd) origin(doe) enter(doe) scale(365.24) id(id)
     failure event:  chd != 0 & chd < .
obs. time interval:  (dox[_n-1], dox]
 exit on or before:  failure
    t for analysis:  (time-origin)/365.24
            origin:  time doe
--------------------------------------------------------------
      337  total obs.
        0  exclusions
--------------------------------------------------------------
      337  obs. remaining, representing
      337  subjects
       46  failures in single failure-per-subject data
 4603.669  total analysis time at risk, at risk from t = 0
```


- To estimate the overall rate of CHD

```
. strate, per(1000)
         failure _d:  chd
   analysis time _t:  (dox-origin)/365.24
             origin:  time doe
Estimated rates (per 1000) and lower/upper bounds of 95% CI
(337 records included in the analysis)

  +----------------------------------------+
  |  D        Y     Rate    Lower     Upper |
  |----------------------------------------|
  | 46   4.6038   9.9918   7.4841   13.3397 |
  +----------------------------------------+
```

- `D` is number of events; `Y` is person-time at risk (in units of 1000 years).
- The overall (average) rate is D/Y, i.e. 9.99 events per 1000 person-years. It will be the same, regardless of which time scale we stset the data on (try this!).

Estimating CHD rates by exposure (energy intake)

- The `stptime` command tabulates the number of events and person-time at risk and calculates event rates.

```
. stptime, by(hieng) per(1000)
         failure _d:  chd
   analysis time _t:  (dox-origin)/365.24
             origin:  time doe
hieng | person-time  failures     rate   [95% Conf. Interval]
------+-------------------------------------------------------
  low |   2059.4305     28   13.595992   9.387478    19.69123
 high |   2544.2382     18   7.0748093   4.457431     11.2291
------+-------------------------------------------------------
total |   4603.6687     46   9.9920309   7.484296    13.34002
```

- Note that person-time is in years but the rates are per 1000 years.
- The rates represent the *overall rates* of CHD in each group during follow-up.

- The `strate` command performs similar calculations.

```
. strate hieng, per(1000)
        failure _d:  chd
  analysis time _t:  (dox-origin)/365.24
            origin:  time doe
Estimated rates (per 1000) and lower/upper bounds of 95% CI
(337 records included in the analysis)
  +-------------------------------------------------+
  | hieng    D        Y      Rate    Lower     Upper |
  |-------------------------------------------------|
  |   low   28   2.0594   13.5960   9.3875   19.6912 |
  |  high   18   2.5442    7.0748   4.4574   11.2291 |
  +-------------------------------------------------+
```

- `D` is number of events; `Y` is person-time at risk (in units of 1000 years).

- The incidence rate ratio (IRR) for individuals with a high compared to low energy intake is $7.1/13.6 = 0.52$.
- That is, without controlling for any possible confounding factors, we estimate that individuals with a high energy intake have a CHD risk that is approximately half that of individuals with a low energy intake.
- This is sometimes called a 'crude estimate'; it is not adjusted for potential confounders.
- Is this a true effect? What important confounder might we need to consider?

A model for the rate

- When working with rates, we believe that effects are most likely to be multiplicative.
- That is, we believe that the rate in the high energy group (λ_1) is likely to be a multiple of the rate in the low energy group (λ_0). The multiplication factor is the incidence rate ratio, θ.

$$\lambda_1 = \lambda_0 \times \theta, \text{ for example, } 7.1 = 13.6 \times 0.52$$

$$\text{IRR} = \frac{\lambda_1}{\lambda_0} = \theta, \text{ for example, } 0.52 = 7.1/13.6$$

- If the explanatory variable X is equal to 1 for individuals with a high energy intake and 0 for individuals with a low energy intake then we can write

$$\lambda(X) = \lambda_0 \times \theta^X$$

- So for each increase of one unit in X the rate increases with a multiple of θ, i.e. the effects are multiplicative (we multiply the constant).

- That is,

$$\lambda = \lambda_0 \text{ when } X = 0$$
$$\lambda = \lambda_0\theta \text{ when } X = 1$$

- For instance, the rate λ_1 among the individuals with high energy intake is

$$\lambda_1 = \lambda(1) = \lambda_0 \times \theta^1 = 13.6 \times 0.52 = 7.1$$

- In practice, it is more convenient to work on a logarithmic scale.

$$\begin{aligned} \lambda &= \lambda_0 \times \theta^X \\ \ln(\lambda) &= \ln(\lambda_0 \times \theta^X) \\ &= \ln(\lambda_0) + \ln(\theta^X) \\ &= \ln(\lambda_0) + \ln(\theta)X \\ \ln(\lambda) &= \beta_0 + \beta_1 X \end{aligned}$$

  where $\beta_0 = \ln(\lambda_0)$ is the log baseline rate and $\beta_1 = \ln(\theta)$ is the log IRR, or log rate ratio. [This is a key result!]

- On the log scale, the effects are additive. For an increase of one unit in X, the log rate increases with an constant $\ln(\theta)$, or β_1 (we add the constant).

- $\ln(\lambda) = \beta_0 + \beta_1 X$ is a Poisson regression model with one binary explanatory variable, X.

- Exercise: What are the estimates of β_0 and β_1?

- The estimate of β_0 is the log of the rate at baseline, ln(13.6)=2.61

- The estimate of β_1 is the log of the IRR comparing group 1 to group 0, ln(0.52)=-0.65

Three regression models commonly applied in epidemiology

- Linear regression
$$\mu = \beta_0 + \beta_1 X$$
- Logistic regression
$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$
- Poisson regression
$$\ln(\lambda) = \beta_0 + \beta_1 X$$
- In each case β_1 is the effect per unit of X, measured as a change in the mean (linear regression); the change in the log odds (logistic regression); the change in the log rate (Poisson regression).

The effect of high energy, using Poisson regression

hieng	X	D	Y	Rate per pyr
low	0	28	2059.4	0.01360
high	1	18	2544.2	0.00707

- If we assume a Poisson regression model

$$\begin{aligned}
\ln(\lambda) &= \beta_0 + \beta_1 X \\
X = 0 : \ln(28/2059.4) &= \beta_0 = -4.3 \\
X = 1 : \ln(18/2544.2) &= \beta_0 + \beta_1 \\
\ln(IRR) = \ln\left(\frac{18/2544.2}{28/2059.4}\right) &= \beta_1 \\
-0.6532 &= \beta_1 = \ln(\text{IRR}) \\
0.52 &= \exp(\beta_1) = \text{IRR}
\end{aligned}$$

Poisson regression in Stata

```
. poisson chd hieng, expos(y)
  chd |      Coef.   [95% Conf. Interval]
------+------------------------------------
hieng |  -.6532341  -1.245357   -.0611114
_cons |   -4.29798  -4.668379   -3.927582
---------------------------------------
on a log scale

. poisson chd hieng, expos(y) irr
    chd |       IRR   [95% Conf. Interval]
--------+------------------------------------
  hieng |  .5203602   .2878382    .9407184
----------------------------------------
on a ratio scale
```

- This is not a st command, survival time must be specified in a separate variable, y.

- The model is estimated using the method of maximum likelihood.
- Confidence intervals are constructed by assuming the estimated regression parameters are normally distributed. 95% CI :$\beta_1 \pm 1.96$ stderr(β_1)
- That is, confidence intervals are constructed on the log scale, as is standard for ratio measures.
- As such, the CI for the IRR is not symmetric around the point estimate.
- We see that the confidence limits for the IRR are simply the exponentiated limits of the ln IRR, and turned around.

$$\begin{aligned} \text{upper 95\% CI IRR: } \exp(\text{lower limit for } \beta_1) &= \exp(\beta_1 - 1.96 \text{ stderr}(\beta_1)) \\ \text{lower 95\% CI IRR: } \exp(\text{upper limit for } \beta_1) &= \exp(\beta_1 + 1.96 \text{ stderr}(\beta_1)) \end{aligned}$$

- To fit a Poisson regression model, we can also use the `streg` (which fits the model in the framework of parametric survival models) or `glm` (generalised linear model) commands.

What happened to the time scale?

- The Poisson model we just fitted did not take into account that rates may vary over follow-up time.
- The data were stset using time since entry as time scale, but the rate we estimated was the 'overall rate' (constant rate, average rate) of CHD throughout the follow-up, i.e. simply all events of CHD divided by total persontime at risk.
- When we estimate the *overall rate*, we assume that the rates (13.6 per 1,000 person-years among low energy group, and 7.1 among high energy group) are constant throughout the follow-up time.
- We will get back to how to model rates which vary over time later. But first we will have a look at how to model main effects and interactions, in general, in Poisson regression, while assuming constant rates over follow-up.

Main effects model for constant rates

- The energy intake variable `hieng` has two levels.
- The variable eng3, created below, has 3 levels. Here are the crude rates and IRRs.

```
. egen eng3=cut(energy),at(1500,2500,3000,4500)
```

```
     energy    eng3       Rate
------------------------------
   1500-2499   1500  |   16.90
   2500-2999   2500  |   10.91    IRR (2500 vs 1500) = 10.91/16.90 = 0.65
   3000-4500   3000  |    4.88    IRR (3000 vs 1500) =  4.88/16.90 = 0.29
------------------------------
```


- To include eng3 in regression commands we need to use indicator variables for the 3 levels.

eng3	$X1$	$X2$	$X3$
1500	1	0	0
2500	0	1	0
3000	0	0	1

- We create dummy variables X1, X2, X3 for exposure eng3.

```
. tabulate eng3, generate(X)
```

- We set exposure level 1500 as reference by omitting X1 from the model.


```
. poisson chd X2 X3, e(y) irr

  chd |      IRR   [95% Conf. Interval]
------+------------------------------
   X2 |   .6452   .3388815    1.228561
   X3 |   .2886   .1235342    .6744495
_cons |   .0169   .0103547    .0275892
```

- The variable $(X1)$ that indicates the category with the lowest energy intake (1500) is omitted, meaning this is the reference category.

- In terms of the parameters

$$\ln(\lambda) \quad = \quad \beta_0 + \beta_2 X_2 + \beta_3 X_3$$

energy	X	ln(rate)	rate
1500	X_1	β_0	$\exp(\beta_0)$
2500	X_2	$\beta_0 + \beta_2$	$\exp(\beta_0 + \beta_2) = \exp(\beta_0)\exp(\beta_2)$
3000	X_3	$\beta_0 + \beta_3$	$\exp(\beta_0 + \beta_3) = \exp(\beta_0)\exp(\beta_3)$

$$\text{IRR}_{X_2/X_1} \quad = \quad \exp(\beta_2) = \frac{\exp(\beta_0)\exp(\beta_2)}{\exp(\beta_0)}$$

$$\text{IRR}_{X_3/X_1} \quad = \quad \exp(\beta_3) = \frac{\exp(\beta_0)\exp(\beta_3)}{\exp(\beta_0)}$$

Automatic generation of indicators using factor variables

```
. poisson chd i.eng3, e(y) irr
------------------------------------------
     chd |      IRR    [95% Conf. Interval]
---------+--------------------------------
    eng3 |
   2500  |   .6452416   .3388815    1.228561
   3000  |   .2886479   .1235342    .6744495
```

- i. tells Stata that eng3 should be treated as a categorical variable
- The baseline is, by default, the first level (1500-2499), but this can be changed to (say) the third level (3000–4500) with

```
. poisson chd ib3000.eng3, e(y) irr
```


- If you want to change the baseline for all analysis you can use

```
. fvset base 3000 eng3
```

- There are more choices to `fvset base`, for example

```
. fvset base last eng3
. fvset base first eng3
. fvset base frequent eng3
```

- To re-set the default, use `fvset clear eng3`

Metric (continuous) exposure variables

- The effect of energy on failure, when energy is measured as a continuous variable

```
. poisson chd energy , e(y) irr
--------------------------------------
   chd |        IRR  [95% Conf. Interval]
-------+------------------------------
energy |     .99885  .9981367    .9995637
```

- For each 1 unit increase in energy intake, the CHD rate is reduced by 0.1%. The units of energy are kcals per day.

- However, a 1 unit increase may not be clinically relevant. We may want to rescale the continuous variable.

```
. summarize energy
Variable |  Obs     Mean    Std. Dev.     Min      Max
---------+-----------------------------------------------
  energy |  337    2828.9    441.8      1748.4   4395.8
```

- To get the IRR for an increase of, say, 100 units

```
. gen energy100=energy/100
. poisson chd energy100, e(y) irr
------------------------------------------
      chd |        IRR  [95% Conf. Interval]
----------+-------------------------------
energy100 |   .8913034  .8298593    .9572968
```

- The estimated IRR is $0.99885^{100} = 0.8913$. That is, for each 100 unit increase in energy intake, we estimate that the CHD rate is reduced by 11%.

The main effects model — constant effect over strata

- If the true effect of exposure does not vary across strata of another variable we can use a main effects model.
- For example, if the effect of high energy intake is the same in all occupations, we can estimate an energy effect that is the same for all occupations.
- Thus, if the estimates of high energy differ only randomly over occupation level, we can consider a model in which the true effect of high energy is constant over occupation level, i.e. no interaction.
- This allows us to combine the information from different strata to yield a single estimate of exposure effect.
- This combined estimate of the effect we call *the main effect*, which is then *controlled for* the stratifying (confounding) variable(s).

- The effect of exposure is assumed to be the same in all levels of other variables in a main effects model.
- Statistical tests for the presence of confounding are not available, although there are statistical tests for effect modification.

Main effects model using Poisson regression

```
. poisson chd i.hieng i.job, e(y) irr
------------------------------------------------
         chd |        IRR   [95% Conf. Interval]
-------------+----------------------------------
     1.hieng |   .5247666   .290225    .9488499
             |
         job |
           2 |   1.358442   .6282879    2.937133
           3 |   .8843023   .4322823    1.808981
             |
       _cons |   .0132337   .0071823    .0243837
```

- Occupation is measured in job variable (1=driver, 2=conductor, 3=banker).
- The Stata Poisson regression command makes no distinction between the exposure variable and the control variable.

- The first number reported is the effect of `hieng` controlled for `job`, and the next two are the effects of `job` controlled for `hieng`.

Models and parameters in Poisson regression

- In the Poisson regression model we estimated 4 parameters. One parameter (the intercept) is a log rate and the other three are log incidence rate ratios.
- The model is $\ln(\lambda) = \beta_0 + \beta_1 \text{hieng} + \beta_2 \text{cond} + \beta_3 \text{bank}$
- The parameters are

$\exp(\beta_0)$ = rate at reference level of all covariates, i.e. driver with low energy
$\exp(\beta_1)$ = rate ratio (comparing high vs low energy)
$\exp(\beta_2)$ = rate ratio (comparing conductors vs drivers)
$\exp(\beta_3)$ = rate ratio (comparing bank vs drivers)

Parameters estimates with and without the `irr` option

```
. poisson chd i.hieng i.job, e(y)
        chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+----------------------------------------------------------------
     1.hieng |  -.6448017   .3021979    -2.13   0.033    -1.237099   -.0525046
         job |
          2  |   .3063385   .3934232     0.78   0.436    -.4647568    1.077434
          3  |  -.1229563     .36517    -0.34   0.736    -.8386764    .5927639
       _cons |  -4.324988   .3118157   -13.87   0.000    -4.936136   -3.713841

. poisson chd i.hieng i.job, e(y) irr
        chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+----------------------------------------------------------------
     1.hieng |   .5247666   .1585834    -2.13   0.033      .290225    .9488499
         job |
          2  |   1.358442   .5344426     0.78   0.436     .6282879    2.937133
          3  |   .8843023   .3229207    -0.34   0.736     .4322823    1.808981
       _cons |   .0132337   .0041265   -13.87   0.000     .0071823    .0243837
```

- The irr option exponentiates the coefficients.

- From the model, the estimated rate for each combination of explanatory variables can be formulated as a function of the baseline rate λ and the three incidence rate ratios. The baseline is the reference group of all variables (drivers with low energy).
- The model is $\ln(\lambda) = \beta_0 + \beta_1 \text{hieng} + \beta_2 \text{cond} + \beta_3 \text{bank}$
- Which on the rate scale is $\lambda = \exp(\beta_0 + \beta_1 \text{hieng} + \beta_2 \text{cond} + \beta_3 \text{bank})$
- These are the rates, λ:

job	hieng=0	hieng=1
1=driv	$\exp(\beta_0)$	$\exp(\beta_0 + \beta_1)$
2=cond	$\exp(\beta_0 + \beta_2)$	$\exp(\beta_0 + \beta_1 + \beta_2)$
3=bank	$\exp(\beta_0 + \beta_3)$	$\exp(\beta_0 + \beta_1 + \beta_3)$

- The estimated incidence rate for drivers with a low energy intake (ref group) is $\exp(-4.325) = 0.0132$ events/person-year.
- The estimated incidence rate for conductors with a high energy intake is $\exp(\beta_0 + \beta_1 + \beta_2) = \exp(\beta_0) \times \exp(\beta_1) \times \exp(\beta_2) = 0.0132 \times 1.358 \times 0.525 = 0.0094$ events/person-year.

- To get the IRRs (with reference group 'drivers with low energy intake') we simply divide all cells with the baseline rate $\lambda = \exp(\beta_0)$

job	hieng=0	hieng=1
1=driv	$\exp(\beta_0)/\exp(\beta_0) = 1$	$\exp(\beta_0 + \beta_1)/\exp(\beta_0) = \exp(\beta_1)$
2=cond	$\exp(\beta_0 + \beta_2)/\exp(\beta_0) = \exp(\beta_2)$	$\exp(\beta_0 + \beta_1 + \beta_2)/\exp(\beta_0) = \exp(\beta_1 + \beta_2)$
3=bank	$\exp(\beta_0 + \beta_3)/\exp(\beta_0) = \exp(\beta_3)$	$\exp(\beta_0 + \beta_1 + \beta_3)/\exp(\beta_0) = \exp(\beta_1 + \beta_3)$

- If we put the estimates from the Stata output into our table of parameters, we get

job	hieng=0	hieng=1
driv	1.0	0.52
cond	1.36	$0.52 \times 1.36 = 0.71$
bank	0.88	$0.52 \times 0.88 = 0.46$

- Compared to drivers with low energy intake, the conductors with high energy have a 29% lower incidence of CHD (IRR=0.71).
- Effect of being high vs low energy is 0.52 (48% lower).
- Effect of being conductor vs driver is 1.36 (36% higher).
- Effect of being banker vs driver is 0.88 (12% lower).

- If we want to get the effect of energy separately in all levels of occupation, we simply change the reference rate to be low energy in all levels.

job	hieng=0	hieng=1
1=driv	$\exp(\beta_0)/\exp(\beta_0) = 1$	$\exp(\beta_0 + \beta_1)/\exp(\beta_0) = \exp(\beta_1)$
2=cond	$\exp(\beta_0 + \beta_2)/\exp(\beta_0 + \beta_2) = 1$	$\exp(\beta_0 + \beta_1 + \beta_2)/\exp(\beta_0 + \beta_2) = \exp(\beta_1)$
3=bank	$\exp(\beta_0 + \beta_3)/\exp(\beta_0 + \beta_3) = 1$	$\exp(\beta_0 + \beta_1 + \beta_3)/\exp(\beta_0 + \beta_3) = \exp(\beta_1)$

- We see that the main effect models give the same effect of energy, $\exp(\beta_1)$, regardless of job level.

job	hieng=0	hieng=1
1=driv	1.0	$\exp(\beta_1)$
2=cond	1.0	$\exp(\beta_1)$
3=bank	1.0	$\exp(\beta_1)$

- This is what a main effect model assumes, i.e. the effect of exposure is the same in all levels of another variable.

- Similarly, if we want to get the effect of job in all levels of energy intake, we change the reference rate to the driver in both levels of energy intake.

job	hieng=0	hieng=1
1=driv	$\exp(\beta_0)/\exp(\beta_0) = 1$	$\exp(\beta_0 + \beta_1)/\exp(\beta_0 + \beta_1) = 1$
2=cond	$\exp(\beta_0 + \beta_2)/\exp(\beta_0) = \exp(\beta_2)$	$\exp(\beta_0 + \beta_1 + \beta_2)/\exp(\beta_0 + \beta_1) = \exp(\beta_2)$
3=bank	$\exp(\beta_0 + \beta_3)/\exp(\beta_0) = \exp(\beta_3)$	$\exp(\beta_0 + \beta_1 + \beta_3)/\exp(\beta_0 + \beta_1) = \exp(\beta_3)$

- Similarly, the main effects of job are the same, $\exp(\beta_2)$ and $\exp(\beta_3)$, regardless of energy level.

job	hieng=0	hieng=1
1=driv	1.0	1.0
2=cond	$\exp(\beta_2)$	$\exp(\beta_2)$
3=bank	$\exp(\beta_3)$	$\exp(\beta_3)$

Effect modification

- If the true effect of exposure is not the same, but varies across strata of another variable there is said to be 'effect modification' (interaction) — the effect of exposure cannot then be represented by one IRR for all levels.
- For example, the effect of high energy may depend on occupation, so that drivers have a greater effect of high energy than conductors.
- Then we say that the effect is *modified by occupation*. There is an interaction between energy intake and occupation.
- Does `job` modify the effect of `hieng`? If we estimate the IRR of high vs low energy separately in all three job groups we get


```
job       Effect of hieng
driver        0.41
conductor     0.66
bank          0.52
```

- The numbers represent the incidence rate ratios (comparing high to low energy intake) within each job category.
- If the effect of high energy is not modified by job then we would expect these to be similar.
- In the previous main effect model, we assumed (and estimated) the effect to be 0.52 in all job levels.

Interaction model using Poisson regression

- If we want to estimate separate effects of hieng in job levels, then four parameters are not enough.
- We have six combinations of energy and job level, hence we need six parameters to estimate six different rates.
- This is easily done by including interaction terms in the Poisson model.
- The model is
  $\ln(\lambda) = \beta_0 + \beta_1 \text{hieng} + \beta_2 \text{cond} + \beta_3 \text{bank} + \beta_4 \text{hieng x cond} + \beta_5 \text{hieng x bank}$

- The parameters are

$\exp(\beta_0)$ = rate at reference level of all covariates, i.e. driver with low energy
$\exp(\beta_1)$ = rate ratio (comparing high vs low energy among drivers)
$\exp(\beta_2)$ = rate ratio (comparing conductors vs drivers among low energy)
$\exp(\beta_3)$ = rate ratio (comparing bank vs drivers among low energy)
$\exp(\beta_4)$ = interaction term (excess rate for cond vs driv among high energy)
$\exp(\beta_5)$ = interaction term (excess rate for bank vs driv among high energy)


```
. poisson chd i.hieng##i.job, e(y) irr
         chd |        IRR   [95% Conf Int]
-------------+------------------------
     1.hieng |   .4102648   .1235412    1.362438
             |
         job |
          2  |   1.136857   .4266828    3.029051
          3  |    .813427   .3325064    1.989927
             |
   hieng#job |
        1 2  |   1.596755   .3222813    7.911183
        1 3  |   1.261973   .2824452    5.638532
             |
       _cons |   .0144648   .0072338     .028924
```

- 0.41 is the effect of `hieng` when `job` is at its first level.
- 1.14 and 0.81 are the effects of `job` when `hieng` is at its first level.
- 1.60 and 1.26 are the interactions between `hieng` and `job`.

- 0.014 is the baseline rate in the reference level of both hieng and job.

Parameters for the interaction model

- The *fitted rates* (λ) from the interaction model can be expressed in terms of the model parameters as follows

job	hieng=0	hieng=1
1=driv	$\exp(\beta_0)$	$\exp(\beta_0 + \beta_1)$
2=cond	$\exp(\beta_0 + \beta_2)$	$\exp(\beta_0 + \beta_1 + \beta_2 + \beta_4)$
3=bank	$\exp(\beta_0 + \beta_3)$	$\exp(\beta_0 + \beta_1 + \beta_3 + \beta_5)$

- $\exp(\beta_4)$ and $\exp(\beta_5)$ are the interaction parameters. They measure deviations from the hypothesis of main effect of `hieng` in all `job` categories.

- The model is
$\ln(\lambda) = \beta_0 + \beta_1$hieng $+ \beta_2$cond $+ \beta_3$bank $+ \beta_4$hieng x cond $+ \beta_5$hieng x bank

- If we wish to tabulate the IRR, using the drivers with low energy intake as reference group, then we simply divided all cells with the reference rate.

job	hieng=0	hieng=1
1=driv	$\exp(\beta_0)/\exp(\beta_0) = 1$	$\exp(\beta_0 + \beta_1)/\exp(\beta_0) = \exp(\beta_1)$
2=cond	$\exp(\beta_0 + \beta_2)/\exp(\beta_0) = \exp(\beta_2)$	$\exp(\beta_0 + \beta_1 + \beta_2 + \beta_4)/\exp(\beta_0) = \exp(\beta_1 + \beta_2 + \beta_4)$
3=bank	$\exp(\beta_0 + \beta_3)/\exp(\beta_0) = \exp(\beta_3)$	$\exp(\beta_0 + \beta_1 + \beta_3 + \beta_5)/\exp(\beta_0) = \exp(\beta_1 + \beta_3 + \beta_5)$

- If we add the Stata output into the table of IRRs we get

job	hieng=0	hieng=1
driv	1.0	0.41
cond	1.14	$0.41 \times 1.14 \times 1.60 = 0.75$
bank	0.81	$0.41 \times 0.81 \times 1.26 = 0.42$

- Compare to the estimates from main effects model

job	hieng=0	hieng=1
driv	1.0	0.52
cond	1.36	$0.52 \times 1.36 = 0.71$
bank	0.88	$0.52 \times 0.88 = 0.46$

Testing for interaction using Stata

- A test of interaction is simply to test if the excess (1.26 and 1.60) is equal to 1 (or 0 on the log scale).

```
. poisson chd i.hieng i.job i.hieng##i.job, e(y) irr
------------------------------------------------------------------------------
         chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.hieng |   .4102648   .2512349    -1.45   0.146     .1235412    1.362438
             |
         job |
          2  |   1.136857   .5684285     0.26   0.798     .4266828    3.029051
          3  |    .813427   .3712769    -0.45   0.651     .3325064    1.989927
             |
   hieng#job |
        1 2  |   1.596755   1.303745     0.57   0.567     .3222813    7.911183
        1 3  |   1.261973   .9638479     0.30   0.761     .2824452    5.638532
             |
       _cons |   .0144648   .0051141   -11.98   0.000     .0072338     .028924
------------------------------------------------------------------------------
```



```
. testparm 1.hieng#2.job 1.hieng#3.job
  chi2(  2) =    0.33
Prob > chi2 =    0.8475
```

- No evidence of a statistically significant interaction.
- This is a so-called Wald test, which approximates the likelihood ratio test. We could also use a likelihood ratio test, where we compare the log-likelihoods from the main effects model and the interaction model.

Reparameterising the model to directly estimate the effect of exposure in each stratum

- We are often interested in the effect of the exposure (comparison between high and low energy intake) for each level of the modifier (job).
- We can divide the rates with different reference rates to obtain the effect of high energy in each level of job.

job	hieng=0	hieng=1
1=driv	$\exp(\beta_0)/\exp(\beta_0) = 1$	$\exp(\beta_0+\beta_1)/\exp(\beta_0) = \exp(\beta_1)$
2=cond	$\exp(\beta_0+\beta_2)/\exp(\beta_0+\beta_2) = 1$	$\exp(\beta_0+\beta_1+\beta_2+\beta_4)/\exp(\beta_0+\beta_2) = \exp(\beta_1+\beta_4)$
3=bank	$\exp(\beta_0+\beta_3)/\exp(\beta_0+\beta_3) = 1$	$\exp(\beta_0+\beta_1+\beta_3+\beta_5)/\exp(\beta_0+\beta_3) = \exp(\beta_1+\beta_5)$

- This yields the following IRRs:

job	hieng=0	hieng=1
driv	1.0	$\exp(\beta_1)$
cond	1.0	$\exp(\beta_1 + \beta_4)$
bank	1.0	$\exp(\beta_1 + \beta_5)$

- We can reparameterise the model in Stata to directly estimate parameters of high energy, one for each job level.

How to make Stata produce stratified effects

- Instead of just one baseline rate, you need three baseline rates, λ (one for each job level). In addition, for each job level, you need an IRR for the energy effect (three rate ratios).

```
. poisson chd ibn.job i.hieng#i.job, e(y) irr nocons
         chd |        IRR   [95% Conf. Interval]
-------------+----------------------------------
         job |
           1 |   .0144648   .0072338    .028924
           2 |   .0164445   .0082238   .0328825
           3 |   .0117661   .0066821   .0207183
             |
   hieng#job |
         1 1 |   .4102648   .1235412   1.362438
         1 2 |   .6550924   .2273009   1.888008
         1 3 |   .5177431    .211639   1.266581
```


- Note that this is the same model; there are still 6 parameters and the fitted rates are identical. It's just that the 6 parameters in this model have a different interpretation.

- The log-likelihood for this model is the same as the previous interaction model. This is because we are fitting the exact same model, but with different parameterisation.

Effects of exposure within each stratum of the modifier

- If we insert the Stata output in our previous table, we get

job	hieng=0	hieng=1
driv	1.0	0.41
cond	1.0	0.66
bank	1.0	0.52

- The stratum-specific IRRs are similar, there is no evidence of interaction.
- Effect of being high vs low energy is 0.41 for drivers, 0.41x1.60=0.66 for conductors, and 0.41x1.26=0.52 for bankers.
- Compare this to the main effects model: 0.52 for all occupation levels.

- Similarly, we can reparameterise the models to show effect of job in levels of energy.
- This yields the following IRRs

job	hieng=0	hieng=1
driv	1.0	1.0
cond	$\exp(\beta_2)$	$\exp(\beta_2 + \beta_4)$
bank	$\exp(\beta_3)$	$\exp(\beta_3 + \beta_5)$


```
. poisson chd ibn.hieng i.job#i.hieng, e(y) irr nocons
         chd |        IRR   [95% Conf. Interval]
-------------+----------------------------------
       hieng |
          0  |   .0144648   .0072338    .028924
          1  |   .0059344   .0022273    .0158117
             |
   job#hieng |
        2 0  |   1.136857   .4266828    3.029051
        2 1  |   1.815282   .5122665    6.432687
        3 0  |   .813427    .3325064    1.989927
        3 1  |   1.026523   .3091123    3.408953
```

Linear combinations of parameters

- As an alternative to reparameterising the interaction model we can use the Stata `lincom()` command to estimate the effect of exposure within each level of the modifier together with confidence intervals. Here again is the interaction model with the default parameterisation.

```
.poisson chd i.hieng##i.job, exp(y) irr
         chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.hieng |   .4102648   .2512349    -1.45   0.146     .1235412    1.362438
             |
         job |
           2 |   1.136857   .5684285     0.26   0.798     .4266828    3.029051
           3 |    .813427   .3712769    -0.45   0.651     .3325064    1.989927
             |
   hieng#job |
         1 2 |   1.596755   1.303745     0.57   0.567     .3222813    7.911183
         1 3 |   1.261973   .9638479     0.30   0.761     .2824452    5.638532
             |
       _cons |   .0144648   .0051141   -11.98   0.000     .0072338     .028924
------------------------------------------------------------------------------
```


- The effect of `hieng` for job=1 is 0.41. We now estimate the effect of `hieng` for the other two categories of job.

```
. lincom 1.hieng + 1.hieng#2.job, irr
------------------------------------------------------------------------------
chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
----+-------------------------------------------------------------------------
(1) |   .6550924   .3537903    -0.78   0.434     .2273009    1.888008
------------------------------------------------------------------------------

. lincom 1.hieng + 1.hieng#3.job, irr
------------------------------------------------------------------------------
chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
----+-------------------------------------------------------------------------
(1) |   .5177431   .2363163    -1.44   0.149      .211639    1.266581
------------------------------------------------------------------------------
```

- The calculation $0.410 \times 1.596 = 0.655$ isn't difficult but calculating the standard error and CI is non-trivial (a combination of variances and covariances). Lincom is useful for obtaining CI's.

Review: Parameterisations of the interaction model

- Interaction models can be parameterised in different ways to show effects using different reference groups.
- Interaction models measure deviations from the hypothesis of main effect of `hieng` in all `job` categories.
- If we wish to change reference group, simply divide the cells with corresponding reference group.
- E.g. if we wish to tabulate the IRR, using the drivers with low energy intake as reference group, then we simply divided all cells with the rate for drivers with low energy intake.
- Interactions are symmetrical, meaning that we can choose either variable job or hieng as the effect modifier of the other.

How the two parameterisations are related

job	hieng=0	hieng=1
1=driv	$\exp(\beta_0) = \exp(\gamma_0)$	$\exp(\beta_0 + \beta_1) = \exp(\gamma_0 + \gamma_3)$
2=cond	$\exp(\beta_0 + \beta_2) = \exp(\gamma_1)$	$\exp(\beta_0 + \beta_1 + \beta_2 + \beta_4) = \exp(\gamma_0 + \gamma_4)$
3=bank	$\exp(\beta_0 + \beta_3) = \exp(\gamma_2)$	$\exp(\beta_0 + \beta_1 + \beta_3 + \beta_5) = \exp(\gamma_0 + \gamma_5)$

- The model is $\ln(\lambda) =$
$\gamma_0$driv $+ \gamma_1$cond $+ \gamma_2$bank $+ \gamma_3$hieng x driv $+ \gamma_4$hieng x cond $+ \gamma_5$hieng x bank

Time varying rates and confounding by time

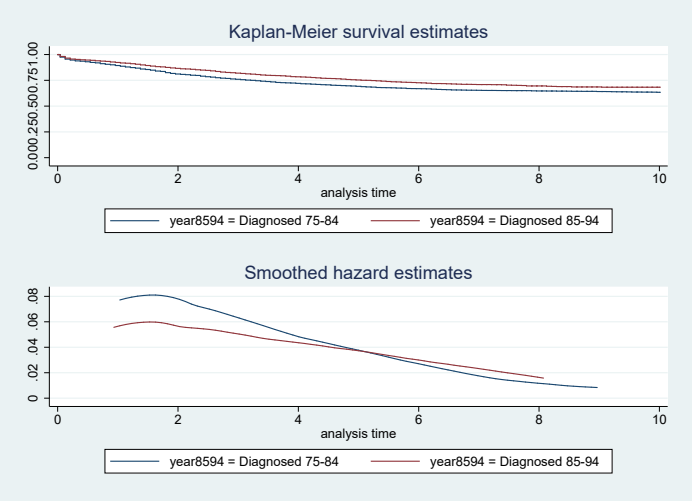
- So far, we have modelled the *overall rate*, i.e. a constant rate throughout the follow-up.
- We have modelled how this overall rate could vary according to exposure variables using main effects models and interaction models.
- These models are general for many kinds of exposure variables.
- In survival analysis, time (i.e. time scale) is a special variable (exposure).
- Now, we will look at how to model and adjust for time (time scale) when it confounds the effect of interest.
- The elegant way we can model time (time scale) is one of the beauties of survival analysis.

- We will look at time as a confounder of the rates and time as an effect-modifier of other variables (later on).
- Important to remember, risk time (amount of time at risk) is different from time scale (where on a scale is the risk time distributed).

Confounding

- Confounding occurs when the association between an exposure and an outcome is induced by a third factor associated with both the exposure and outcome, but not in the causal path.
- For example, if the exposure is smoking and the outcome is death, and smoking is more common among elderly, the association between smoking and death may simply be an effect of age. The age distribution among smokers differs from the age distribution among non-smokers.
- We can adjust for confounding by conditioning on the confounder via stratification or adjusting in a regression model.
- Confounding by time (i.e. time scale) is similar, i.e. where is the person-time distributed along the time scale for different exposure groups.
- We shall use the colon cancer data, where patients were diagnosed 1975-1984 and 1985-1994, with follow-up for death until 1995.

For which calendar period is mortality lowest?



For which calendar period is mortality lowest?

```
. use colon, clear
. stset surv_mm if stage==1, failure(status==1) scale(12) exit(120)
. strate year8594, per(1000)

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(6274 records included in the analysis)
  +-------------------------------------------------------------+
  |        year8594     D         Y     Rate    Lower    Upper |
  |-------------------------------------------------------------|
  | Diagnosed 75-84   862   15.7531   54.719   51.186   58.497 |
  | Diagnosed 85-94   825   15.2024   54.268   50.688   58.100 |
  +-------------------------------------------------------------+
```

- The graphs suggest that patients diagnosed in the recent period have lower mortality (better survival) but the estimated (overall) rates suggest that they are similar.

- The end of follow-up is 1995. We have restricted to 10 years of follow-up (120 months).
- Those diagnosed 1975-84 are all followed for up to 10 years, whereas those diagnosed 1985-94 are followed for at most 10 years (and many will be followed for less than 10 years due to end of study in 1995).
- So, those diagnosed 1985-94 have shorter follow-up. Their person-time will be distributed close to diagnosis date, and the overall (average) rate will be weighted towards the higher early mortality.
- The rates are confounded by follow-up time.
- Hence, the overall rates look very similar, instead of a lower rate in 1985-94 that we would expect.

- If we restrict the calculation to first five years (60 months) of follow-up, the rates are more what we would expect with higher rate in the early period (1975-84) as indicated in the graph.

```
. stset surv_mm if stage==1, failure(status==1) scale(12) exit(60)
. strate year8594, per(1000)

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(6274 records included in the analysis)
  +-------------------------------------------------------------+
  |        year8594     D          Y     Rate    Lower    Upper |
  |-------------------------------------------------------------|
  | Diagnosed 75-84   748     9.5836   78.050   72.652   83.848 |
  | Diagnosed 85-94   745    12.1193   61.472   57.213   66.049 |
  +-------------------------------------------------------------+
```

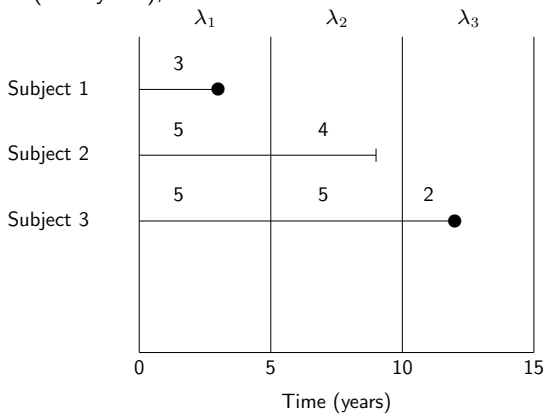
- This indicates that it is important to adjust for follow-up time when estimating rates and rate ratios.

- Because different exposure groups have different distributions of person-time along the time scale, the overall rate may be biased (over- or under-estimated).

Time as a confounder

- If the rate is constant over time, then time will not confound the overall estimate of the rate. (A constant rate means that time is not associated with the rate.)
- When the rate changes with time then time may confound the effect of exposure.
- We will, for the moment, assume that the rates are constant within broad time bands but can change from band to band.
- This approach (categorising a metric variable and assuming the effect is constant within each category) is standard in epidemiology.
- We often categorise metric variables — the only difference here is that the variable is 'time'.

- Consider a group of subjects with rates λ_1 during band 1 (0-5 years), λ_2 during band 2 (5-10 years), etc.



- What are the estimated failure rates, λ_1, λ_2, λ_3, for each of the bands?

Splitting the records by follow-up time

- A convenient way to fit these models using a computer is to replace the single record for this subject by three new records, one for each band of observation.
- The new subject–band records can be treated as independent records.

subject	timeband	follow-up	failure
1	0-5	3	1
2	0-5	5	0
2	5-10	4	0
3	0-5	5	0
3	5-10	5	0
3	10-15	2	1

- The rate for timeband 0-5 is then $1/(3+5+5)$, and so on for other timebands.
- This method can be used whether rates are varying simply as a function of time or in response to some time–varying exposure.

System variables created by stset

_t0	time at entry
_t	time at exit
_d	failure indicator
_st	inclusion indicator

- For example, to stset the Diet data with time since entry as the time scale.

```
. use http://www.biostat3.net/download/diet
. stset dox, id(id) fail(chd) origin(doe) enter(doe) sc(365.24)
. list id  _t0 _t _d _st doe dox in 1/5, clean
 id   _t0          _t   _d   _st         doe         dox
127     0   16.791239    0     1   16Feb1960   01Dec1976
200     0   19.958932    0     1   16Dec1956   01Dec1976
198     0   19.958932    0     1   16Dec1956   01Dec1976
222     0   15.394935    0     1   16Feb1957   10Jul1972
305     0   1.4948665    1     1   16Jan1960   15Jul1961
```


Splitting on 'time in study' (time since entry)

```
. use http://www.biostat3.net/download/diet
. stset dox, id(id) failure(chd) origin(doe) ent(doe) sc(365.24)
```

- It is good to check what the data looks like BEFORE splitting!

```
. list id  _t0 _t _d _st if id==78, clean
       id   _t0          _t   _d   _st
 28.   78     0   5.6180698    1     1
```

- Split the data using the `stsplit` command, which will also generate a `timeband` variable

```
. stsplit timeband, at(0(2)20) trim
(0 + 4 obs. trimmed due to lower and upper bounds)
(2122 observations (episodes) created)
```

- It is good to check what the data looks like AFTER splitting!

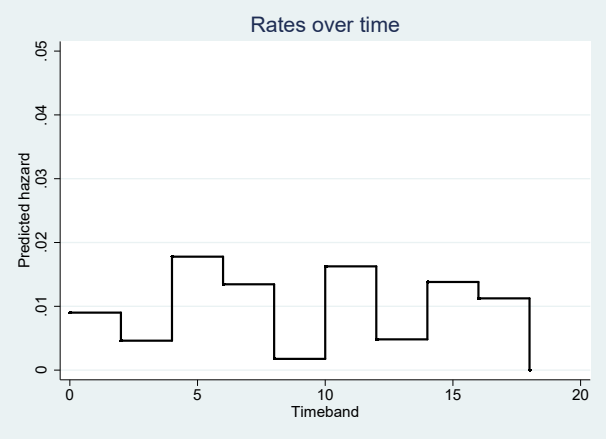

```
. list id timeband _t0 _t _d _st if id==78, clean
       id   timeband   _t0          _t   _d   _st
 189.   78          0    0           2    0    1
 190.   78          2    2           4    0    1
 191.   78          4    4   5.6180698    1    1
```

- Person ID=78 was followed up for 5.618 years, and when we split the record we got three rows of data, one for each time band 0-2, 2-4, 4-6 years where this person contributes risk time.

Rates for different time bands

```
. strate timeband, per(1000)
  +-------------------------------------------------------+
  | timeband     D        Y       Rate     Lower      Upper |
  |-------------------------------------------------------|
  |        0     6   0.6658    9.01205   4.04876   20.05973 |
  |        2     3   0.6499    4.61589   1.48872   14.31189 |
  |        4    11   0.6187   17.77860   9.84579   32.10291 |
  |        6     8   0.5947   13.45180   6.72721   26.89835 |
  |        8     1   0.5670    1.76370   0.24844   12.52060 |
  |-------------------------------------------------------|
  |       10     8   0.4919   16.26292   8.13305   32.51949 |
  |       12     2   0.4148    4.82158   1.20586   19.27877 |
  |       14     5   0.3619   13.81571   5.75048   33.19266 |
  |       16     2   0.1778   11.24988   2.81357   44.98197 |
  |       18     0   0.0610    0.00000         .          . |
  +-------------------------------------------------------+
```


- We can plot the rates over timebands. This produces a step function.



- Poisson regression can also be performed using the `streg` command. This is preferable when the data have been 'stsplit' since `streg` respects the internal variables (`_d`, `_t0`, `_t`, and `_st`) created by `stset` and `stsplit`.

```
. streg hieng, dist(exp)
Exponential regression -- log relative-hazard form
No. of subjects =          337         Number of obs   =      2455
No. of failures =           46
Time at risk    =  4603.504449
                                       LR chi2(1)      =      4.82
Log likelihood  =   -175.00017         Prob > chi2     =    0.0282
------------------------------------------------------------------
   _t | Haz. Ratio   Std. Err.    z    P>|z|   [95% Conf. Interval]
------+-----------------------------------------------------------
hieng |   .5203748   .1572099  -2.16   0.031   .2878463    .9407449
 cons |   .0135959   .0025694   -22.74   0.000     .0093874    .0196911
------------------------------------------------------------------
```

- This rate ratio 0.520 is not adjusted for time.

- The effect of `hieng` controlled for `timeband` is found with:

```
. streg hieng i.timeband, dist(exp)
          _t | Haz. Ratio   Std. Err.      z  [95% Conf. Interval]
-------------+----------------------------------------------------
       hieng |   .5192324   .1568783    -2.17  .2871997    .9387276
             |
    timeband |
           2 |   .5135604    .363132    -0.94  .1284451    2.053361
           4 |   1.994108   1.012072     1.36  .7374578    5.392127
           6 |   1.509821   .8154207     0.76  .5238543    4.351515
           8 |    .197761   .2136135    -1.50  .0238071     1.64276
          10 |   1.808417   .9766387     1.10  .6274883    5.211844
          12 |   .5339264   .4359361    -0.77  .1077703    2.645232
          14 |   1.536019   .9300917     0.71  .468788     5.032884
          16 |   1.261454   1.029979     0.28  .2546036    6.249978
          18 |   1.29e-06   .0015518    -0.01       0           .
       _cons |    .012205   .0051785   -10.38  0.000      .0053135    .0280349
```

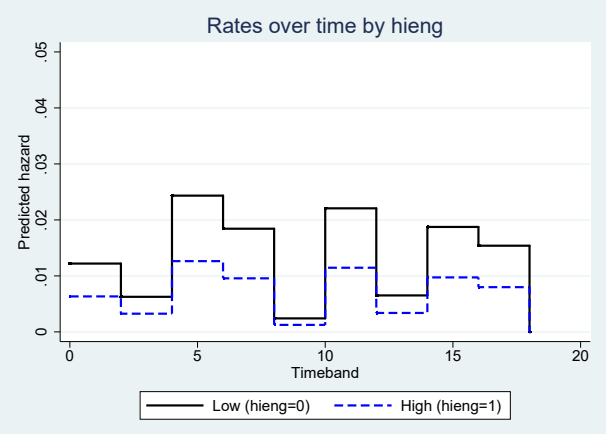
- The estimated rate ratio adjusted for time is 0.519.

- There is no reason to believe that time-on-study would be a confounder for these data. This would, however, be of interest in the cancer examples.

- Because this is a main effects model, the effect of hieng is assumed to the same (0.519) across all timebands. (If we believed the effect of hieng was different over time, then we would need to include interaction between hieng and timeband.)

- The ratio for hieng is adjusted for timeband. Meaning that we are comparing persons within the same timeband with respect to energy intake.

- Again, we can plot the rates over timebands for high and low energy intake. The rate ratio (ratio between curves) will be the same (0.519) for all timebands, since we have assumed a main effect model.



- We fitted a main effects model, and can calculate the rate ratios using the same technique as we did earlier.

timeband	hieng=0	hieng=1
0	1.0	0.52
2	0.51	0.52×0.51
4	1.99	0.52×1.99
...	...	...

Splitting the follow–up on the age scale

- Attained age is a possible confounder for the diet study, young and old people may differ in both energy intake and risk for CHD. Attained age is more interesting as a potential confounder than age at entry.

```
. stset dox, fail(chd) enter(doe) origin(dob) sc(365.24) id(id)
. list id _t0 _t _d _st if id==163
 id         _t0          _t   _d   _st
163    47.55373   60.922656    1     1

. stsplit ageband, at(30,40,50,60,70) trim
. list id ageband _t0 _t _d _st if id==163
id   ageband       _t0          _t   _d   _st
163       40  47.55373          50    0     1
163       50        50          60    0     1
163       60        60   60.922656    1     1
```


- We see that, as expected, the CHD incidence rate depends on attained age.

```
. strate ageband, per(1000)
 ageband    _D         _Y     _Rate
      30     0     0.0963    0.0000
      40     6     0.9070    6.6152
      50    18     2.1070    8.5428
      60    22     1.4933   14.7325
```

The effect of hieng controlled for attained age

```
. streg hieng i.ageband, dist(exp)
------------------------------------------------
          _t | Haz. Ratio   [95% Conf. Interval]
-------------+----------------------------------
       hieng |   .5370252   .2967504    .9718473
             |
     ageband |
          40 |    4865216    0           .
          50 |    5963551    0           .
          60 |   1.03e+07    0           .
```

- Poor choice of baseline for ageband!

- Let's use a different reference category.

```
. fvset base 40 ageband
. streg hieng i.ageband, dist(exp)
------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|
-------------+----------------------------------------
       hieng |   .5370252   .1625226    -2.05   0.040
             |
     ageband |
          30 |   2.06e-07   .0005733    -0.01   0.996
          50 |   1.225752   .5786603     0.43   0.666
          60 |   2.108791   .9728264     1.62   0.106
```

- Is there evidence that the effect of hieng is confounded by attained age?
- Minor indication of confounding, the crude hieng estimate is 0.52, as we saw in previous slides.

Summary of Day 2

- A *rate* is defined as *events* divided by *total time-at-risk*, where time at risk is usually measured in person-years, person-months etc.
- The rates can vary across various *time scales*, e.g. time since entry, attained age, calendar period.
- Rates can be modelled using Poisson regression, which estimates the baseline hazard rate and the rate ratios for different exposure levels.
- Main effects models estimate rate ratios for exposures while adjusting for other covariates (confounders).
- Interaction models estimate separate rate ratios for different exposure levels (effect modification).
- Interactions can be re-parameterised in various ways to show the different aspects of effect modification.

- The estimates of rates and rate ratios can be confounded by time (i.e. time scale).

- Assessing goodness-of-fit in Poisson regression is only possible on collapsed data using the deviance and the Pearson chi-square statistic (see appendix).

Exercises for Day 2

110. Diet data: tabulating incidence rates and modelling with Poisson regression. [working through what was presented in the lectures].

111. Localised melanoma: model cause-specific mortality with Poisson regression. [this is a key exercise, next time we will fit a Cox model to the same data and compare the results]

112. Diet data: Using Poisson regression to study the effect of energy intake adjusting for confounders. [Something for you to do if you've finished the other two]

Appendix Day 2: Statistical models

- Multiple regression models are important in that they allow simultaneous estimation and testing of the effect of many prognostic factors on survival.

- The aim of statistical modelling is to derive a mathematical representation of the relationship between an observed response variable and a number of explanatory variables, together with a measure of the uncertainty of any such relationship.

- The uses of a statistical model can be classified into the following three areas:

  1. Descriptive: To describe any structure in the data and quantify the effect of explanatory variables, and to study the pattern of any such associations;
  2. Hypothesis testing: To statistically test whether an observed response variable is associated with one or more explanatory variables; and
  3. Prediction: For example, predicting excess mortality for a future time period, or predicting the way in which the outcome may change if certain explanatory variables changed in value.

- Note that a statistical model is never true, but may be useful.
- When making inference based on the model we assume that the model is true.
- If the model is badly misspecified then inference will be erroneous.
- It is therefore important to consider the validity of any assumptions (e.g. proportional hazards) underlying the model and to check for evidence of lack-of-fit.

An introduction to generalised linear models, GLM

- A simple linear model (i.e. least squares regression) can be written as

$$y_i = \mathbf{x}\beta + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2). \quad (8)$$

- For a generalised linear model (GLM), it is assumed that the probability distribution function of the outcome, y_i, belongs to the exponential family (which includes the normal, binomial, and Poisson distributions), and that the relationship between the expectation of y_i and its linear predictor is given by the link function g. That is,

$$g(u_i) = \mathbf{x}\beta, \quad (9)$$

  where $u_i = E(y_i)$ and g is the link function (which is monotonic and differentiable).
- Many widely used models can be fitted in the framework of generalised linear models. For example:

- Linear regression – link: identity, error: normal

$$u_i = \mathbf{x}\beta.$$

- Poisson regression – link: log, error: Poisson

$$\ln(u_i) = \mathbf{x}\beta.$$

  When modelling event rates, the outcome is y_i/n_i, where n_i is person-time at risk. The model can then be rewritten as

$$\ln(u_i) = \ln(n_i) + \mathbf{x}\beta, \text{ where } \ln(n_i) \text{ is known as an offset term.}$$

- logistic regression – link: logit, error: binomial

$$\ln(\frac{\pi_i}{1-\pi_i}) = \mathbf{x}\beta,$$

  where $\pi_i = E(y_i/n_i)$ is the outcome.

- To define a generalised linear model, it is therefore necessary to specify
  - the error distribution
  - the link function

  in addition to the outcome and explanatory variables.

- A logistic regression model could be fitted in SAS, for example, using the following commands:

```
proc genmod data=test;
model y/n = x1 x2 / dist=bin link=logit;
run;
```

- The corresponding Stata command is:

```
. glm y x1 x2, family(binomial n) link(logit)
```


Poisson regression is a GLM

$$\begin{aligned} \ln(\text{rate}) &= \mathbf{x}\beta \\ \ln(\text{events/person-time}) &= \mathbf{x}\beta \\ \ln(\text{events}) - \ln(\text{person-time}) &= \mathbf{x}\beta \\ \ln(\text{events}) &= \mathbf{x}\beta + \ln(\text{person-time}) \end{aligned}$$

- $\ln(\text{person-time})$ is known as an offset; it's a constant in the linear predictor.

- Poisson regression can be fitted as a generalised linear model with
  - outcome: number of events
  - link: log
  - error distribution: Poisson
  - offset: logarithm of person-time

Estimation of Poisson model as a GLM

```
. gen risktime=_t-_t0
. glm _d hieng i.ageband  if _st==1, family(poiss) lnoff(risktime) eform
Generalized linear models                   No. of obs      =       755
Optimization     : ML                       Residual df     =       750
                                            Scale parameter =         1
Deviance         =   313.146733             (1/df) Deviance =   .417529
Pearson          =  1938.800828             (1/df) Pearson  =  2.585068
                                            AIC             =  .5498632
Log likelihood   = -202.5733665             BIC             = -4656.892
------------------------------------------------------------------------------
          _d |        IRR   Std. Err.      z    P>|z|     [95% CI]
-------------+----------------------------------------------------------------
       hieng |   .5370429   .1625146    -2.05   0.040    .2967746    .9718319
             |
     ageband |
          30 |   2.07e-06   .0018067    -0.01   0.988           0           .
          50 |   1.225563   .5785056     0.43   0.667    .4858933    3.091224
          60 |   2.108589   .9726144     1.62   0.106    .8538153    5.207387
    risktime | (exposure)
```

Assessing goodness-of-fit of Poisson regression models

- Since Poisson regression is a generalised linear model, methods of assessing goodness-of-fit of GLMs can be applied (many of which you have seen previously with logistic regression).
- For a GLM fitted to non-sparse data, model goodness-of-fit can be assessed using the deviance or the Pearson chi-square statistic.
- Both the deviance and the Pearson chi-square statistic have an approximate χ^2 distribution under the assumption that the model fits, with degrees of freedom equal to the number of observations minus the number of parameters estimated in the model (including the intercept) [22].
- The mean of a χ^2_k distribution is k, so if the model is a reasonable fit the deviance and the Pearson statistic will be close to the residual df.
- The deviance is the difference in twice the log likelihood between the fitted model and what is called the saturated model.

- The saturated model is the model which contains one parameter for every observation, such that the fitted values equal the observed values.
- For data which is cross-classified by k categorical variables, as cancer registry data usually are, the saturated model contains all 2-way, 3-way, up to k-way interactions.
- As such, if a model is fitted containing all main effects, the deviance is essentially a test for interaction (where interaction is equivalent to non-proportional excess hazards).
- The asymptotic χ^2 assumption for the deviance and the Pearson chi-square statistic is only valid for 'non-sparse' data.

- A rule-of-thumb for chi-square based statistics of agreement between observed and fitted values is that both the expected number of successes and the expected number of failures must be 5 or more in at least 80% of the cells and at least 1 in each cell.
- In practice, individual-level data should be grouped.
- The exact distributions of the deviance and the Pearson chi-square statistic are not known, and there is no agreement in the literature regarding which is the best measure of goodness-of-fit.
- However, the two statistics should be similar for a model that provides a good fit to the data, and a large discrepancy between the two statistics is generally indicative of sparse data.
- When data are sparse, we typically see a deviance less than the degrees of freedom and a Pearson chi-square much greater than the degrees of freedom.

- Values of the deviance and Pearson chi-square significantly greater than the associated degrees of freedom can be due to a number of factors, including
  1. an incorrectly specified functional form (an additive rather than a multiplicative model may be appropriate);
  2. overdispersion; or
  3. the absence of important explanatory variables (or interactions) from the model.
- In most cases, lack-of-fit is due to missing explanatory variables (or interactions) from the model.
- Model goodness-of-fit can also be assessed using plots of residuals and influence statistics.

Example: Assessing goodness-of-fit for Poisson regression

```
. use colon if stage==1, clear
. stset surv_mm, failure(status==1) scale(12) id(id) exit(time 120)
. gen risktime=_t-_t0
. glm _d year8594, family(poisson) eform lnoffset(risktime)

Generalized linear models                 No. of obs      =      6274
Optimization     : ML                     Residual df     =      6272
                                          Scale parameter =         1
Deviance         =  9261.056188           (1/df) Deviance =  1.476571
Pearson          =  94685.52343           (1/df) Pearson  =  15.09654
```

- The deviance and Pearson chi-square statistics are not interpretable for individual data. We need to collapse the data into groups with the same combinations of values on all the variables in the model.

Asymptotic properties of the Pearson chi-square statistic (by hand-waving)

- The Pearson chi-square statistic has the form $\sum \frac{(O-E)^2}{E}$ where O and E are the observed and expected number of events for each observation and the sum is over all observations.
- The quantity $r_i = (O_i - E_i)/\sqrt{E_i}$ is the Pearson residual.
- If the r_i follow a normal distribution with mean zero and variance 1 then $\sum_{i=1}^{k} r_i^2$ will be χ^2_k.
- This becomes problematic for individual data, where O is either 0 or 1.
- The distribution of residuals will be bimodal; one group for observations with $O = 0$ one group for observations with $O = 1$.
- That is, residuals will not be standard normal so the sum of the residuals squared will not be χ^2.

- By collapsing, we will have larger values for O and E and it is more likely that the distribution of $r = (O - E)/\sqrt{E}$ is symmetric around zero.
- Here we again refer to the rule-of-thumb for chi-square based statistics of agreement between observed and fitted values; E should be 5 or more in at least 80% of the cells and at least 1 in each cell.
- The same issue exists in logistic regression and the same solution (collapsing) can be used.

Example continued: assessing goodness-of-fit after collapsing

- We have specified that our data are cross-classified by period (2 groups), age (4 groups), and sex (2 groups). So we collapse on the combinations of values from these variables. Then we fit the Poisson model to the collapsed data.

```
. collapse (sum) _d risktime , by(year8594 agegrp sex)

. glm _d year8594, family(poisson) eform lnoffset(risktime)

Generalized linear models                   No. of obs      =        16
Optimization     : ML                       Residual df     =        14
                                            Scale parameter =         1
Deviance         =  248.1137278             (1/df) Deviance =  17.72241
Pearson          =  262.5530674             (1/df) Pearson  =  18.75379
```

- They are not shown above, but parameter estimates are identical for the individual and collapsed data.

- If the model fits, the deviance and Pearson chi-square statistics should follow a χ^2 distribution with 14 degrees of freedom.
- That is, the expected value of these two statistics is 14 (the expected value of the χ^2_k distribution is k).
- The Deviance and the Pearson statistic are far from the number of df's (df=14). There is strong evidence of lack of fit.
- Stata helps us out by presenting the values of the statistic divided by the df.
- If the model fits these should be close to 1 (which is not the case here).

- If we include age and sex in the model

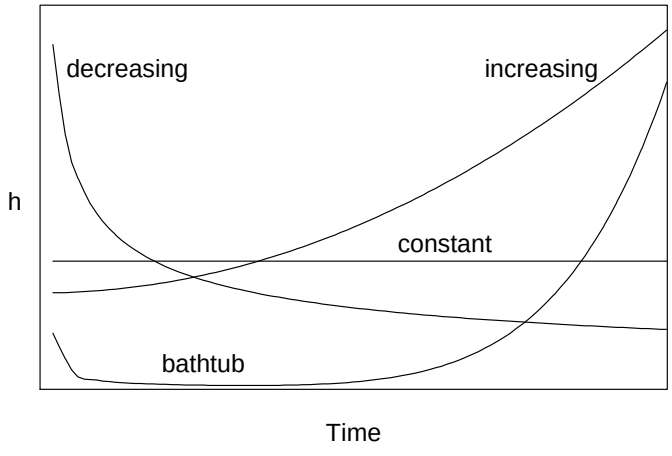
```
. glm _d year8594 i.agegrp sex, family(poisson) eform lnoffset(risktime)

Generalized linear models                   No. of obs      =         16
Optimization     : ML                       Residual df     =         10
                                            Scale parameter =          1
Deviance         =  12.05963472             (1/df) Deviance =   1.205963
Pearson          =   12.0560896             (1/df) Pearson  =   1.205609
```

- There is no longer evidence of lack-of-fit. The scaled deviance is close to 1, i.e. the deviance is almost equal to the number of df's.
- If we fitted all main effects and still saw evidence of lack-of-fit then this might suggest effect modification (i.e., interaction terms are required).

Topics for Day 3

- Rates which vary over time
- Parametric and semi-parametric models
- The Cox model
- Comparison of Cox and Poisson regression
- Choice of time scale in Cox regression
- The proportional hazards assumption
- Assessing the proportional hazards (PH) assumption
- Modelling time-varying exposures

Common forms for the hazard function (time-varying rates)



- A bathtub-shaped hazard is appropriate for mortality in most human populations followed from birth, where the hazard rate decreases to almost zero after an initial period of infant mortality, and then starts to increase again later in life.
- A decreasing hazard function is appropriate for mortality following the diagnosis of most types of cancer, where mortality due to the cancer is highest immediately following diagnosis, and then decreases with time as patients are cured of the cancer.
- An increasing hazard function is appropriate for incidence rates of ageing diseases.

- A constant hazard function is often used for modelling the lifetime of electronic components, but is also appropriate following the diagnosis of some types of cancer, most notably cancers of the breast and prostate, where the level of excess mortality due to the cancer is relatively constant over time and persists even 15-20 years after diagnosis.
- A constant hazard function implies that survival times can be described by an exponential distribution (which has one parameter, the hazard λ). This distribution is 'memoryless' in that the expected survival time for any individual is independent of how long the individual has survived so far.
- The average time to winning a prize for a regular lotto player, for example, can be described by an exponential distribution.
- The survivor function has the same basic shape (a nonincreasing function from 1 to 0) for all types of data and the hazard function is often a more informative means of studying differences between patient groups.

Shape of the hazard in Poisson regression

- The Poisson regression model is

$$\ln(\lambda) = \beta_0 + \beta_1 X$$
$$\lambda = \exp(\beta_0 + \beta_1 X)$$
$$\lambda = \exp(\beta_0)\exp(\beta_1 X)$$

- The baseline hazard is constant in a Poisson regression, $\exp(\beta_0)$.

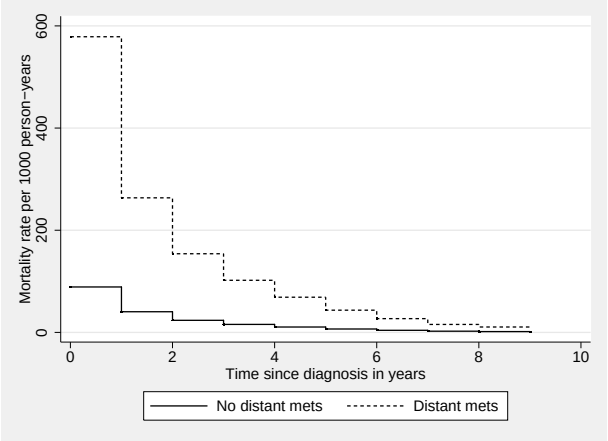
- If we add a categorical variable for time, e.g. time-since-entry in 1-year bands, then the baseline hazard is a step function of time. The hazard is piecewise constant in 1-year bands.

$$\lambda = \exp(\beta_0 + \beta_3 t_{[1,2)} + \beta_4 t_{[2,3)} + \cdots) \exp(\beta_1 X)$$

  where $t_{[1,2)}$ is an indicator for time being in the interval $[1, 2)$, i.e. timeband. Note that $t_{[0,1)}$ if left out from the equation (it is assumed to be the reference time band)

- We can use piece-wise constant hazards to describe most shapes of hazard functions approximately with a step function. (If we split time in finer intervals, then sharper increases/decreases can be captured by the step function.)

Shape of hazard: step function



Parametric models

- If we assume that survival times follow an exponential distribution, then the hazard is constant (overall or piece-wise) and we could model the hazard as a function of one or more covariates using Poisson regression.

- We could then obtain an estimate of the hazard ratio for the treatment group compared to the control group while adjusting for other explanatory variables.

- The disadvantage of this method is that assuming an exponential distribution for survival times implies the assumption of a constant hazard function over time (or within time bands if the data has been splitted), which may not be appropriate.

- The Weibull distribution, which has two parameters, is a more flexible distribution in which the hazard can be either monotonic increasing, decreasing, or constant.

- The Weibull, log-normal and Gompertz distributions have proved to be applicable in several types of medical survival studies.
- If a parametric distribution is appropriate, such models will result in more efficient estimates (narrower confidence limits) of the parameters of interest compared to other models which do not assume a distribution for the survival times.
- Most common statistical procedures are parametric, for example, t-tests, ANOVA, and linear regression all assume normal distributions.
- Inference based on the above procedures is, however, quite robust to violations of the distributional assumptions. For example, application of a standard t-test will generally lead to the correct conclusion even if the two samples are not drawn from populations with normal distributions.
- This is not necessarily the case when assuming a parametric distribution for survival time. The assumption of an inappropriate distribution can result in erroneous conclusions.

- That is, when using parametric survival models, special attention must be paid to testing the appropriateness of the model.
- Still, of all parametric models, Poisson regression is very robust since it allows the hazard to vary freely between timebands.

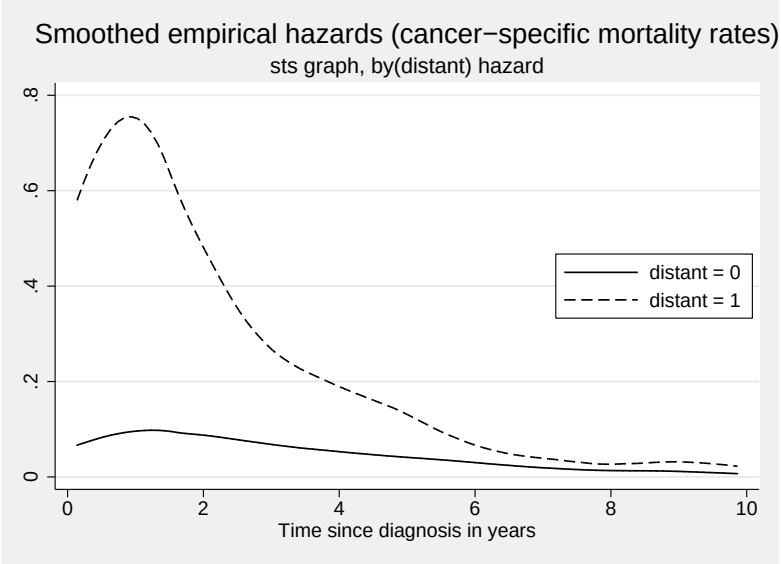
The shape of hazards in a Cox model

- An alternative to parametric models is the Cox model.
- The Cox model does not make any assumption about the shape of the hazard function or the distribution of survival times.
- Instead, the baseline hazard is allowed to vary freely. The baseline hazard is not even estimated (no parameters).
- The Cox model only estimates hazard ratios relative to the baseline hazard.
- Since the baseline hazard is not estimated in the Cox model, it is said to be semi-parametric.
- The hazard ratios are modelled parametrically.

An introduction to the Cox model via an example: Survival of patients diagnosed with colon carcinoma

- Patients diagnosed with colon carcinoma 1975–94. Potential follow-up to end of 1995; censored after 10 years.
- Outcome is death due to colon carcinoma.
- Interest is in the effect of clinical stage at diagnosis (distant metastases vs no distant metastases).
- How might we specify a statistical model for these data?


```
use colon.dta, clear

drop if stage==0 // unknown

stset exit , failure(status==1) enter(dx) origin(dx) ///
scale(365.24) exit(time dx+3650)

gen distant=1 if stage==3
replace distant=0 if stage<3

tab stage distant, miss

sts graph, by(distant) haz noboundary
```

Proportional hazards models

- A proportional hazards model is on the form

$$\lambda(t|X) = \lambda_0(t)\exp(\beta X)$$

- The hazard at time t for an individual with some covariate values, $\lambda(t|X)$, is a multiple of the baseline, $\lambda_0(t)$. The multiple is $\exp(\beta X)$.

- This means that the hazards for different levels of X are proportional: $Y_2 = kY_1$

- It also means that the ratio of hazards is constant and only depends on β and X, regardless of t.

$$\frac{\lambda(t|X)}{\lambda_0(t)} = \exp(\beta X)$$

- This means that a proportional hazards model estimates hazard ratios which are constant over time, and that hazards are assumed to be proportional to each other over time. [KEY message!]

The Cox proportional hazards model

- The Cox model is a proportional hazards model. (And so is the Poisson model, more about that later.)

$$\lambda(t|X) = \lambda_0(t)\exp(\beta X)$$

- However, the Cox model does not estimate the baseline hazard, $\lambda_0(t)$. It only estimates the regression coefficients, β.

- Although the baseline $\lambda_0(t)$ is not estimated, the hazard ratios are adjusted for time t, i.e. time scale.

- The Cox model is said to "automatically adjust for the underlying time scale".

- In a Poisson model, the effect of time (timeband) could be moved from the linear predictor into the baseline, $\lambda_0(t)$. Similarly, for Cox, the baseline hazard includes all the effect of time scale.

- The 'intercept' in the Cox model [10], the hazard (event rate) for individuals with all covariates X at the reference level, is an arbitrary function of time[4], often called the baseline hazard and denoted by $\lambda_0(t)$.

- The Cox model can also be written on the log scale

$$\ln[\lambda(t|X)] = \ln[\lambda_0(t)] + \beta X.$$

  where $X = 1$ for patients with distant metastases at diagnosis and $X = 0$ for patients without distant metastases at diagnosis.

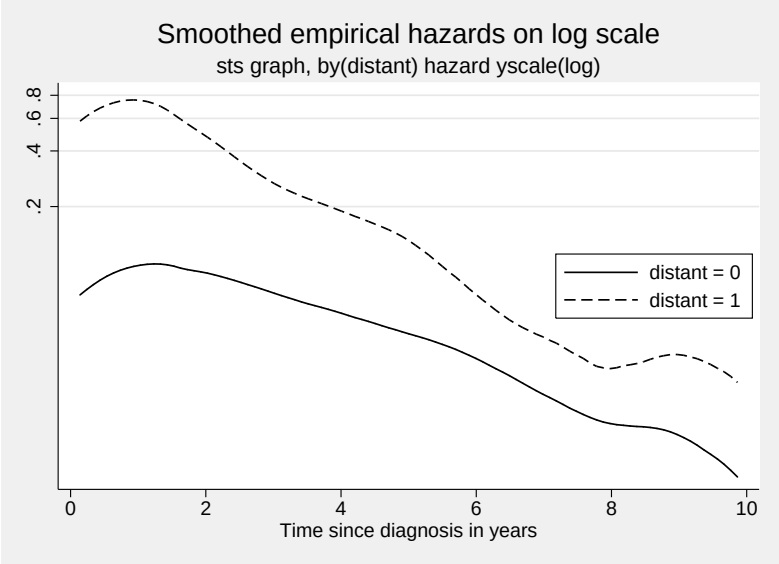
- The difference between two log hazards is a constant β regardless of t

$$\ln[\lambda(t|X)] - \ln[\lambda_0(t)] = \beta X.$$

- The two hazard curves are thus assumed to be parallel, i.e. constant difference across t, on a log scale.

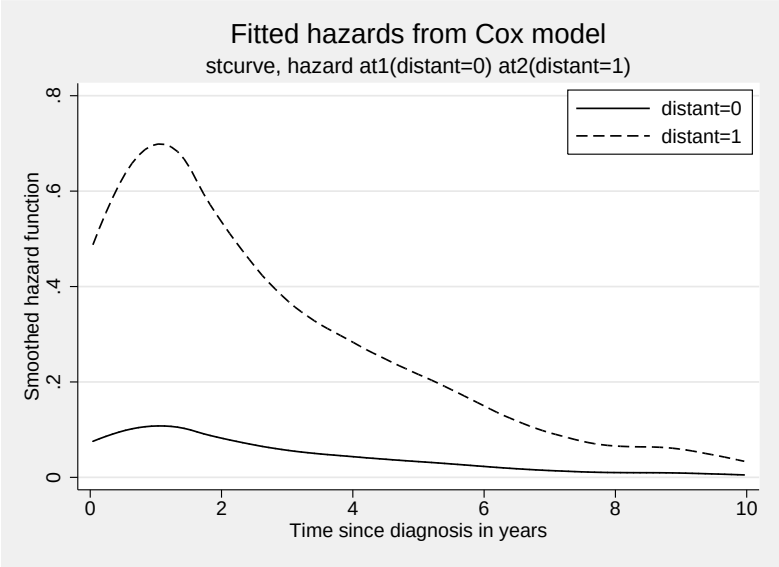
[4] time t is the time scale and can be defined in many ways, e.g., attained age, time-on-study, calendar time, etc.

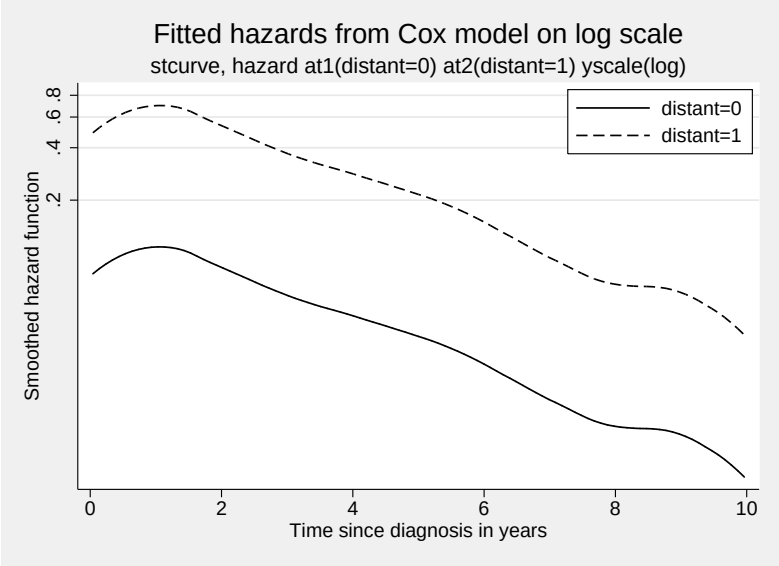
- Hence, if we plot the hazard curves on a log scale, then the curves should be parallel if the assumption of proportional hazards holds.



Fit a Cox model to estimate the hazard ratio

```
. stcox distant
         failure _d:  status == 1
   analysis time _t:  (exit-origin)/365.24
             origin:  time dx
               note:  trim>10 trimmed

Cox regression -- Breslow method for ties

No. of subjects =        13208          Number of obs  =   50666
No. of failures =         7122
Time at risk    =  44013.26215
                                        LR chi2(1)     = 5544.65
Log likelihood  =   -61651.446          Prob > chi2    =  0.0000
------------------------------------------------------------------------------
     _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------+---------------------------------------------------------------------
distant |   6.557777   .1689328    73.00   0.000    6.234895    6.897381
------------------------------------------------------------------------------
```

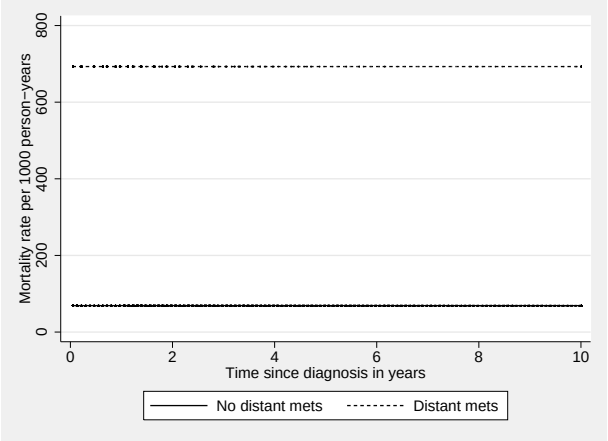


An analogous Poisson regression model?

```
. streg distant, dist(exp)

Exponential regression -- log relative-hazard form

No. of subjects =        13208                     Number of obs   =     13208
No. of failures =         7122
Time at risk    =  44014.07294
                                                   LR chi2(1)      =   8788.80
Log likelihood  =   -19144.094                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     distant |   10.04034   .2473993    93.61   0.000     9.566966    10.53713
       _cons |   .0690183   .0013572  -135.95   0.000     .0664088    .0717304
------------------------------------------------------------------------------
```

- Is this conceptually analogous to the Cox model with one predictor (`distant`)?

Fitted values: Poisson model with one predictor (distant)

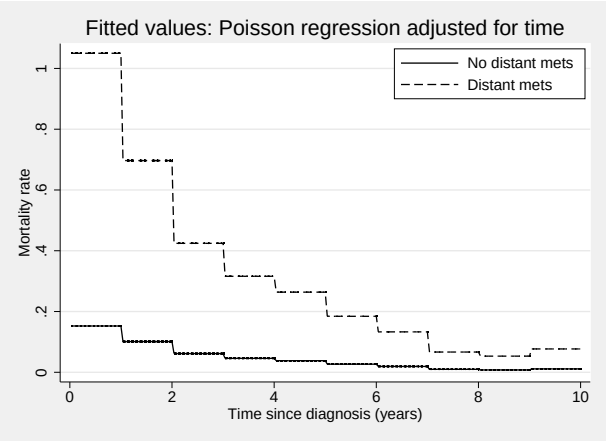


- We haven't controlled for time, whereas the Cox model does.

An analogous Poisson regression model

```
. stsplit fu, at(0(1)10)
(37458 observations (episodes) created)
. streg distant i.fu, dist(exp)
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     distant |   6.890447   .1758378    75.64   0.000     6.554288    7.243847
             |
          fu |
           1 |    .663664   .0204393   -13.31   0.000     .6247888     .704958
           2 |   .4041879   .0178799   -20.48   0.000     .3706202    .4407959
           3 |   .3008835   .0170792   -21.16   0.000     .2692039    .3362912
           4 |   .2511955   .0172521   -20.12   0.000     .2195591    .2873905
           5 |   .1754671   .0157037   -19.45   0.000     .1472368    .2091101
           6 |    .126706   .0145236   -18.02   0.000     .1012112    .1586229
           7 |   .0635093    .011113   -15.75   0.000     .0450705    .0894915
           8 |   .0506029   .0108263   -13.95   0.000     .0332708    .0769638
           9 |   .0732211   .0144196   -13.27   0.000     .0497745    .1077123
             |
       _cons |   .1523782   .0036926   -77.64   0.000     .1453099    .1597902
------------------------------------------------------------------------------
```

Fitted values: Poisson regression model adjusted for time



- Once we adjust for time we get a similar estimate for the effect of distant.

- The shape of the hazard is similar to the predicted hazards from the Cox model.
- Both Cox and Poisson models are proportional hazards models.
- Both will give hazard ratios which are constant over time, $exp(\beta)$.

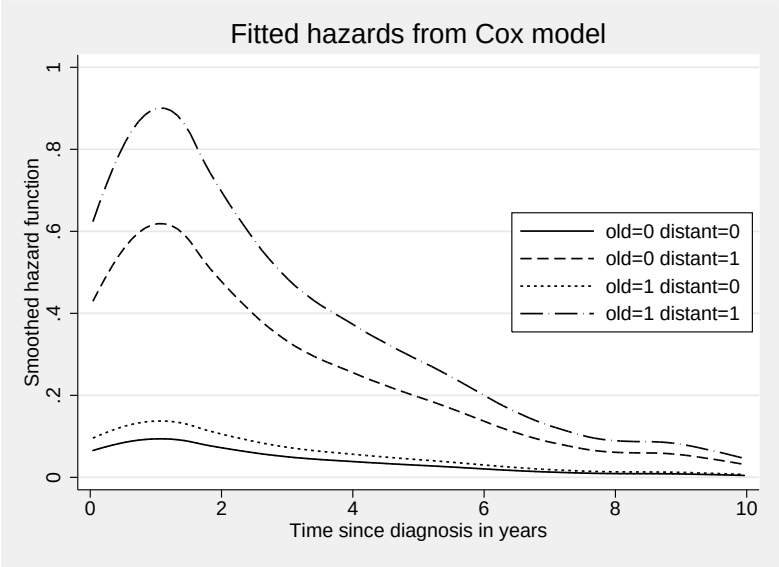
$$\begin{aligned} \text{Cox: } \lambda(t|X) &= \lambda_0(t)\exp(\beta X) \\ \text{Poisson (constant rate): } \lambda(t|X) &= \exp(\beta_0)\exp(\beta X) \\ \text{Poisson (time-varying rate): } \lambda(t|X) &= \exp(\beta_0 + \beta_1 timeband_1 + ...)\exp(\beta X) \end{aligned}$$

Adjusting for confounders in Cox model

- The effect of distant metastases may be confounded by age, since patients with distant metastases tend to be older at diagnosis.
- We can adjust for confounders by including them as covariates in the Cox model.
- This gives a main effects model, where we estimate the effect of distant metastases and the effect of age at diagnosis (young/old). We assume that the effect of distant metastasis is the same for young and old.
- The effect of distant metastases is adjusted for age, while the effect of age is adjusted for distant metastases.

Fit a Cox model adjusted for age at diagnosis

```
. gen old = age>=75
. stcox distant old
         failure _d:  status == 1
   analysis time _t:  (exit-origin)/365.24
             origin:  time dx
               note:  trim>10 trimmed

Cox regression -- Breslow method for ties

No. of subjects =        13208           Number of obs  =    50666
No. of failures =         7122
Time at risk    =  44013.26215
                                         LR chi2(2)     =  5778.91
Log likelihood  =   -61534.317           Prob > chi2    =   0.0000
------------------------------------------------------------------------------
     _t | Haz. Ratio   Std. Err.     z    P>|z|    [95% Conf. Interval]
--------+---------------------------------------------------------------------
distant |    6.65287   .1716121   73.47   0.000   6.324877    6.997871
    old |   1.463653   .0358098   15.57   0.000   1.395124    1.535549
------------------------------------------------------------------------------
```

The Cox proportional hazards model (in detail)

- The most commonly applied model in medical time-to-event studies is the Cox proportional hazards model [10].
- The Cox proportional hazards model does not make any assumption about the shape of the underlying hazards, but makes the assumption that the hazards for patient subgroups are proportional over follow-up time.
- We are usually more interested in studying how the hazard varies as a function of explanatory variables (the relative rates, hazard ratios) rather than the shape of the underlying hazard function (the absolute rate).
- In most statistical models in epidemiology (e.g. linear regression, logistic regression, Poisson regression) the outcome variable (or a transformation of the outcome variable) is equated to the 'linear predictor', $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$.

- $X_1, \ldots, X_k$ are explanatory variables and $\beta_0, \ldots, \beta_k$ are regression coefficients (parameters) to be estimated.
- The Xs can be continuous (age, blood pressure, etc.) or if we have categorical predictor variables we can create a series of indicator variables (Xs with values 1 or 0) to represent each category.
- We are interested in modelling the hazard function, $\lambda(t; \boldsymbol{X})$, for an individual with covariate vector $\boldsymbol{X}$, where $\boldsymbol{X}$ represents $X_1, \ldots, X_k$.
- The hazard function should be non-negative for all $t > 0$; thus, using

$$\lambda(t|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

  may be inappropriate since we cannot guarantee that the linear predictor is always non-negative for all choices of $X_1, \ldots, X_k$ and $\beta_0, \ldots, \beta_k$.

- However, $\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$ is always positive so another option would be

$$\lambda(t|\boldsymbol{X}) = \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$$
$$\ln \lambda(t|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

- In this formulation, both the left and right hand side of the equation can assume any value, positive or negative.
- This formulation is identical to the Poisson regression model. That is,

$$\ln(\frac{\text{no. events}}{\text{person-time}}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

- The one flaw in this potential model is that $\lambda(t|\boldsymbol{X})$ is a function of t, whereas the right hand side will have a constant value once the values of the βs and Xs are known.
- This does not cause any mathematical problems, although experience has shown that a constant hazard rate is unrealistic in most practical situations.

- The remedy is to replace β_0, the 'intercept' in the linear predictor, by an arbitrary function of time — say $\ln \lambda_0(t)$; thus, the resulting model equation is

$$\ln \lambda(t|\boldsymbol{X}) = \ln \lambda_0(t) + \beta_1 X_1 + \cdots + \beta_k X_k.$$

- The arbitrary function, $\lambda_0(t)$, is evidently equal to the hazard rate, $\lambda(t|\boldsymbol{X})$, when the value of $\boldsymbol{X}$ is zero, i.e., when $X_1 = \cdots = X_k = 0$.
- The model is often written as

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t) \exp(\boldsymbol{X}\beta).$$

- It is not important that an individual having all values of the explanatory variables equal to zero be realistic; rather, $\lambda_0(t)$ represents a reference point that depends on time, just as β_0 denotes an arbitrary reference point in other types of regression models.

- This regression model for the hazard rate was first introduced by Cox [10], and is frequently referred to as the Cox regression model, the Cox proportional hazards model, or simply the Cox model.
- Estimates of $\beta_1, \ldots, \beta_k$ are obtained using the method of maximum partial likelihood (slide 418).
- As in all other regression models, if a particular regression coefficient, say β_j, is zero, then the corresponding explanatory variable, X_j, is not associated with the hazard rate of the response of interest; in that case, we may wish to omit X_j from any final model for the observed data.
- As with logistic regression and Poisson regression, the statistical significance of explanatory variables is assessed using Wald tests or, preferably, likelihood ratio tests.
- The Wald test is an approximation to the likelihood ratio test. The likelihood is approximated by a quadratic function, an approximation which is generally quite good when the model fits.

- In most situations, the test statistics will be similar.
- Differences between these two test statistics (likelihood ratio and Wald) indicate possible problems with the fit of the model.
- The assumption of proportional hazards is a strong assumption, and should be tested (see slide 298).
- Because of the inter-relationship between the hazard function, $\lambda(t)$, and the survivor function, $S(t)$, (Equation 7, slide 110) we can show that the PH regression model is equivalent to specifying that

$$S(t|\boldsymbol{X}) = \{S_0(t)\}^{\exp(\beta_1 X_1 + \cdots + \beta_k X_k)} \tag{10}$$

where $S(t|\boldsymbol{X})$ denotes the survivor function for a subject with explanatory variables $\boldsymbol{X}$, and $S_0(t)$ is the corresponding survivor function for an individual with all covariate values equal to zero.

- Most software packages, will provide estimates of $S(t)$ based on the fitted proportional hazards model for any specified values of explanatory variables.
- For example, the Stata `stcurve` can be used after `stcox` to plot the cumulative hazard, survival, and hazard functions at the mean value of the covariates or at values specified by the at() options.

The Estimated Regression Coefficients

- The Cox model can be written as:

$$\lambda(t|X) = \lambda_0(t)\exp(\beta X)$$

$$\frac{\lambda(t|X)}{\lambda_0(t)} = \exp(\beta X)$$

$$\ln(\frac{\lambda(t|X)}{\lambda_0(t)}) = \beta X$$

- The estimated coefficients, β, are log rate ratios. To get the rate ratios we need to exponentiate the coefficients, $\exp(\beta)$.

- The confidence intervals for the β are on the log scale. The CIs are therefore not symmetric around the rate ratios.

Interpreting the Estimated Regression Coefficients

- Recall that the basic proportional hazard (PH) regression model specifies

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\beta_1 X_1 + \cdots + \beta_k X_k)$$

equivalently,

$$\ln\lambda(t|\boldsymbol{X}) = \ln\lambda_0(t) + \beta_1 X_1 + \cdots + \beta_k X_k$$

- Note the similarity to the basic equation for multiple linear regression, i.e.,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

- In ordinary regression we derive estimates of all the regression coefficients, i.e., $\beta_1, \ldots, \beta_k$ and β_0.

- In Cox regression, the baseline hazard component, $\lambda_0(t)$, vanishes from the partial likelihood; we only obtain estimates of the regression coefficients associated with the explanatory variates $X_1, \ldots, X_k$.

- Consider the simplest possible setup, one involving only a single binary variable, X; then the PH regression model is

$$\ln\lambda(t|X) = \ln\lambda_0(t) + \beta X$$

or equivalently,

$$\begin{aligned}\beta X &= \ln\lambda(t|X) - \ln\lambda_0(t) \\ &= \ln\left\{\frac{\lambda(t|X)}{\lambda_0(t)}\right\}\end{aligned} \tag{11}$$

- Since $\lambda_0(t)$ corresponds to the value $X = 0$,

$$\beta = \ln\left\{\frac{\lambda(t|X=1)}{\lambda_0(t)}\right\} \tag{12}$$

- That is, β is the logarithm of the ratio of the hazard rate for subjects belonging to the group denoted by $X = 1$ to the hazard function for subjects belonging to the group indicated by $X = 0$.

- The parameter β is a log relative rate (log hazard ratio) and $\exp(\beta)$ is a relative rate (hazard ratio) of response. PH regression is sometimes called "relative risk regression".

- β is the same for all values of time, i.e. the hazard ratio is constant over t (proportional hazards over time).

- If we conclude that the data provide reasonable evidence to contradict the hypothesis that X is unrelated to response, $\exp(\hat{\beta})$ is a point estimate of the rate at which response occurs in the group denoted by $X = 1$ relative to the rate at which response occurs at the same time in the group denoted by $X = 0$.

- A confidence interval for β, is given by $\hat{\beta} \pm 1.96\text{SE}$.

- Corresponding confidence intervals for the relative rate associated with the same covariate are obtained by transforming the confidence interval for β, i.e.,

$$(\beta_\ell, \beta_u) \Rightarrow \left(e^{\beta_\ell}, e^{\beta_u}\right).$$

- When more than one covariate is involved, the principle is the same; $\exp(\hat{\beta}_j)$ is the estimated relative rate of failure for subjects that differ only with respect to the covariate X_j.

- If X_j is binary, $\exp(\hat{\beta}_j)$ estimates the increased/reduced rate of response for subjects corresponding to $X_j = 1$ versus those denoted by $X_j = 0$.

- When X_j is a numerical (continuous) measurement then $\exp(\hat{\beta}_j)$ represents the estimated change in relative rate associated with a unit change in X_j.

- Since the estimates $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are obtained simultaneously, these estimated relative rates adjust for the effect of all the remaining covariates included in the fitted model.

Example: Localised colon carcinoma 1975–1994

- We will fit a proportional hazards model to study the effect of sex, age (in 4 categories), and calendar period (2 categories) on cause-specific mortality (only deaths due to colon cancer were considered events).

- We'll begin by restricting the data to localised cases only (stage=1).

```
.  use http://www.biostat3.net/download/colon, clear
(Colon carcinoma, all stages, 1975-94, follow-up to 1995)
. keep if stage==1
(9290 observations deleted)
```

- We `stset` the data where only deaths due to colon cancer (status=1) are considered 'failures'.

```
. stset surv_mm, failure(status==1)
     failure event:  status == 1
obs. time interval:  (0, surv_mm]
 exit on or before:  failure

--------------------------------------------------------------------------------
     6274  total observations
        0  exclusions
--------------------------------------------------------------------------------
     6274  observations remaining, representing
     1734  failures in single-record/single-failure data
   427185  total analysis time at risk and under observation
                                             at risk from t =         0
                                  earliest observed entry t =         0
                                       last observed exit t =     251.5
```

- Now we estimate the Cox model.


```
. stcox sex i.agegrp year8594

No. of subjects =         6274                     Number of obs   =      6274
No. of failures =         1734
Time at risk    =       427185
                                                   LR chi2(5)      =    197.23
Log likelihood  =   -14348.889                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9151101   .0451776    -1.80   0.072     .8307126    1.008082
             |
      agegrp |
      45-59  |   .9491689   .1314101    -0.38   0.706      .723597     1.24506
      60-74  |   1.338501   .1682956     2.32   0.020     1.046148    1.712553
        75+  |    2.24848   .2834768     6.43   0.000     1.756199    2.878751
             |
    year8594 |   .7548672   .0372669    -5.70   0.000     .6852479    .8315596
------------------------------------------------------------------------------
```


- The output commences with a description of the outcome and censoring variable and a summary of the number of subjects and number of failures.
- The default method for handling ties (the Breslow method) is used.
- The test statistic `LR chi2(5) = 197.23` is not especially informative. The interpretation is that the 5 parameters in the model (as a group) are statistically significantly associated with the outcome ($P < 0.00005$).
- The variable sex is coded as 1 for males and 2 for females. Since each parameter represents the effect of a one unit increase in the corresponding variable, the estimated hazard ratio for sex represents the ratio of the hazards for females compared to males.
- That is, the estimated hazard ratio is 0.92 indicating that females have an estimated 8% lower colon cancer mortality than males. There is some evidence that the difference is statistically significant ($P = 0.07$).

- The model assumes that the estimated hazard ratio of 0.92 is the same at each and every point during follow-up and for all combinations of the other covariates.
- That is, the hazard ratio is the same for females diagnosed in 1975–1984 aged 0–44 (compared to males diagnosed in 1975–1984 aged 0–44) as it is for females diagnosed in 1985–1994 aged 75+ (compared to males diagnosed in 1985–1994 aged 75+).
- The indicator variable `year8594` has the value 1 for patients diagnosed during 1985–1994 and 0 for patients diagnosed during 1975–1984.
- The estimated hazard ratio is 0.75. We estimate that, after controlling for the time scale, age and sex, patients diagnosed 1985–1994 have a 25% lower mortality than patients diagnosed during 1975–1984. The difference is statistically significant ($P < 0.0005$).
- We chose to group age at diagnosis into four categories; 0–44, 45–59, 60–74, and 75+ years.

- It is estimated that individuals aged 75+ at diagnosis experience 2.25 times higher risk of death due to colon carcinoma than individuals aged 0–44 at diagnosis, a difference which is statistically significant ($P < 0.0005$).
- Similarly, individuals aged 60–74 at diagnosis have an estimated 34% higher risk of death due to colon carcinoma than individuals aged 0–44 at diagnosis, a difference which is statistically significant ($P < 0.02$).

- These significance tests test the pairwise differences and tell us little about the overall association between age and survival – we need to perform a general test.

```
. testparm 1.agegrp 2.agegrp 3.agegrp

 ( 1)  1.agegrp = 0
 ( 2)  2.agegrp = 0
 ( 3)  3.agegrp = 0

           chi2(  3) =  174.13
         Prob > chi2 =    0.0000
```

- This is a Wald test of the null hypothesis that all age parameters are equal to zero, i.e. that age is not associated with the outcome.
- We see that there is strong evidence against the null hypothesis, i.e. we conclude that age is significantly associated with survival time.

- The Wald test is an approximation to the likelihood ratio test, which compares the likelihood between models.
- To perform a likelihood ratio test we fit the reduced model (the model without age) and see that the log likelihood is -14436.387.

```
. stcox sex year8594
No. of subjects =         6274          Number of obs   =      6274
No. of failures =         1734
Time at risk    =    424049.72          LR chi2(2)      =     22.23
Log likelihood  =   -14436.387          Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9978866   .0487896    -0.04   0.965     .9066997    1.098244
    year8594 |     .79287   .0390053    -4.72   0.000     .7199909    .8731261
------------------------------------------------------------------------------
```

- The log likelihood for the model containing age is -14348.889; for the model excluding age it is -14436.387.

- The likelihood ratio test statistic for the association of age with survival is calculated as $2 \times (-14348.889 - (-14436.387)) = 175.0$, which is compared to a χ^2 distribution with 3 degrees of freedom (P=0.0001).
- We see that the Wald test statistic (174.1) is very similar in value to the likelihood ratio test statistic (175.0).
- You can also get Stata to calculate the likelihood ratio test statistic for you (you have to explicitly fit both models and save the estimates for the first).

```
stcox sex i.agegrp year8594
est store A
stcox sex year8594
est store B
lrtest A B
```


- The output of the final command is as follows

```
. lrtest A B
likelihood-ratio test           LR chi2(3)  =    175.00
(Assumption: B nested in A)     Prob > chi2 =    0.0000
```

We might choose to model age as a continuous variable

```
. stcox sex age year8594
No. of subjects =         6274          Number of obs   =      6274
No. of failures =         1734
Time at risk    =    424049.72          LR chi2(3)      =    248.82
Log likelihood  =    -14323.09          Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9029577   .0445694    -2.07   0.039     .8196956    .9946773
         age |   1.033981   .0024182    14.29   0.000     1.029253    1.038732
    year8594 |   .7488609   .0369492    -5.86   0.000     .6798334    .8248973
------------------------------------------------------------------------------
```

- For each (and every) one year increase in age at diagnosis, we estimate that mortality is 3.4% higher.
- For a 10-year increase in age at diagnosis the estimated hazard ratio is $1.033981^{10} = 1.396$.

Comparison of Cox regression to Poisson regression for the analysis of cohort studies

- The methods are very similar; the basic formulation of both models is

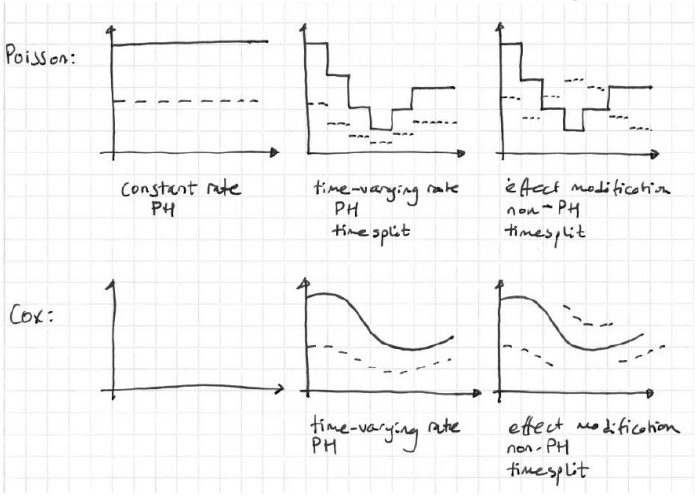
$$\lambda(t|X) = \lambda_0(t)\exp(\beta_1 X_1 + \cdots + \beta_k X_k)$$

- In both cases, the β parameters are interpreted as log rate ratios.
- Both models assume proportional hazards, i.e. constant hazard ratios over time.
- Both models are multiplicative.
- That is, if the RR for males/females is 3 and the RR for smokers to non-smokers is 4, then the RR for male smokers to female non-smokers is 12 (in a model with no interaction terms).

- In Poisson regression, follow-up time is classified into bands and a separate rate parameter is estimated for each band, thereby allowing for the possibility that the rate is changing with time.
- In Poisson regression, the baseline rate $\lambda_0(t)$ has a constant or piece-wise constant shape. It is assumed that the rate is constant within each band, so if the rate is changing rapidly with time we may have to choose very narrow bands.
- In Cox regression, the baseline rate $\lambda_0(t)$ is not estimated but allowed to vary freely.
- In Cox regression, we essentially choose bands of infinitesimal width; each band is so narrow that it includes only a single event.
- Unlike in Poisson regression, we do not estimate the baseline rates within each time band; instead, we estimate the relative rates (rate ratios) for the different levels of the covariates.

- As such, if estimating the effect of time is of interest then Poisson regression is a more natural choice, since Poisson regression will estimate parameters for the time effect.
- Time-by-covariate interactions (i.e., non-proportional hazards) are, in practice, easier to model in the framework of Poisson regression.
- Multiple time scales are typically also easier to include in the framework of Poisson regression, since every time scale will be parameterised separately. (In Cox regression, the main time scale will not be parameterised, while other time scales will.)

Comparison Cox and Poisson regression



Equivalence of Cox and Poisson regression

- The Cox model can be viewed as extending the life-table approach *ad absurdum* by:
  1. splitting time as finely as possible,
  2. modelling one covariate, the time-scale, with one parameter per observed value of time,
  3. profiling these parameters out by maximizing the profile likelihood
- Subsequently recover the effect of the timescale by smoothing an estimate of the parameters that was profiled out!

- If we split time as finely as possible and fit a Poisson regression model we will get the same results as the Cox model.

- Code can be found in `compare_cox_poisson.do` (at `http://www.biostat3.net/download/`).

```
. use http://www.biostat3.net/download/melanoma if stage==1, clear
. stset surv_mm, f(status==1) exit(time 120) id(id)
.
. /* Fit the Cox model */
. stcox sex year8594 i.agegrp
[output ommitted]
```



```
/* split at each failure time */
. stsplit, at(failures) riskset(riskset)
(117 failure times)
(378139 observations (episodes) created)

/* Generate indicator variables for each timeband */
. quietly tab riskset, gen(interval)

. streg interval* sex year8594 i.agegrp, dist(exp)
[output ommitted]
```



```
Hazard ratios and standard errors for various models
-------------------------------------------------------
  Variable |   Cox        Poisson_fine       Poisson
-----------+-------------------------------------------
       sex |   0.588814       0.588814       0.587547
           |   0.038538       0.038538       0.038456
  year8594 |   0.716884       0.716884       0.722411
           |   0.047445       0.047445       0.047813
  1.agegrp |   1.326397       1.326397       1.327795
           |   0.124911       0.124911       0.125042
  2.agegrp |   1.857323       1.857323       1.862376
           |   0.168787       0.168787       0.169244
  3.agegrp |   3.372652       3.372652       3.400287
           |   0.352227       0.352227       0.355140
-------------------------------------------------------
                                          legend: b/se
Poisson_fine: split at each failure time
Poisson: split in annual intervals
```

Choice of time scale in a Cox model

- In Cox regression the hazard ratio (HR) compares the hazard rates of two groups at each t.
- So for any t, the hazards are assumed to be proportional by a multiple, HR.
- Since the comparison is made at each t, the HR is automatically adjusted for t, the underlying time scale.
- For example, if we stset the data so that the underlying time scale is attained age, then all the HRs will be adjusted for age in the Cox model.
- If age is the chosen underlying time scale in the Cox model, the effect of age cannot be estimated directly, since it is incorporated in the shape of the baseline hazard, which is allowed to vary freely.
- The adjustment of the underlying time scale in a Cox model is very efficient, since it adjusts for time in very small intervals.

- Hence, if one of the three possible time scales for your data (time-in-study, attained age, calendar time) is a strong confounder of your exposure — outcome association, then that time scale should preferably be chosen as the underlying time scale.
- For most disease incidences, age is a strong confounder. But since we are generally not interested in estimating the effect of age, the most efficient way to adjust for it is to choose age as the underlying time scale [24, 5, 7].
- Thiebaut and Benichou [24] recommend using age as the timescale and conclude 'we strongly recommend not using time-on-study as the time scale for analysing epidemiologic cohort data [where entry has no clinical or biological relevance]'.

- The choice of time scale in the analysis should be based on:
  1. Your research question: choose the time scale which has the most relevance for your research question, sometimes the exposure is a time scale (e.g. how does the incidence vary over age, how does mortality vary by time-since-diagnosis)
  2. Adjustment for time confounding: choose the time scale which has the strongest confounding effect. In Cox regression, the effect of time is very strongly adjusted for.
- Age: often the strongest confounder in incidence studies
- Calendar time: often a confounder, proxy for other phenomena (including unmeasured confounders)
- Time-since-entry (time on study, follow-up): often relevant in prognosis studies, where entry is at diagnosis, i.e. entry has a meaning
- Other: Time-since-exposure (e.g. time-since-medication or time-since-crime)

Time as confounder, effect modifier or mediator

- Select time scale for the analysis by listing all possible time scales and characterize them as confounders, mediators or effect modifiers. Which time scales are associated with the exposures and which are associated with the outcome rates (plot rates over time scales).
- A time scale is just like any other factor that you need to assess in terms of whether it is an exposure, confounder, mediator or effect modifier.
- A time scale as exposure: You may want to parameterise the time scale (not Cox), since you are interested in the effect of time per se.
- A time scale as confounder/mediator: You may want to choose the time scale as main time scale in analysis. Parameters for time are less important if the effect is well-known. E.g. we know that all-cause mortality increases with age, so we do not need to estimate that effect. We may still want to adjust for age very strongly, e.g. using Cox regression.

- A time scale as effect modifier: You want to include interaction terms with time in the model (non-proportional hazards), easiest if time is parameterised (not Cox).

Multiple time scales

- In some situations, several time scales are confounders for the exposure - outcome association.
- For example, cancer incidence may vary both over age and calendar time.
- In such situations, we must adjust for two time scales. This can be done both in Poisson regression and in Cox regression.
- Data can be split on several time scales.
- In Poisson regression: Data must be split on all time scales that we wish to adjust for.
- In Cox regression: Data does not need to be split on the main time scale, but must be split on all additional time scales we wish to adjust for (we get one time scale adjusted for automatically).

Diet data: Cox regression with age as timescale

- Use attained age as the timescale.

```
. use diet
. stset dox, fail(chd) enter(doe) origin(dob) scale(365.24)
. stcox hieng
```

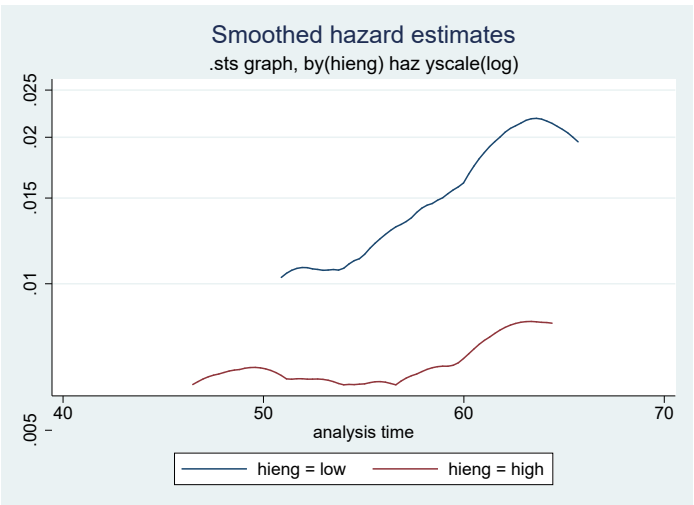
```
        failure _d:  chd
  analysis time _t:  (dox-origin)/365.24
            origin:  time dob
 enter on or after:  time doe
No. of subjects =          337          Number of obs    =       337
No. of failures =           46
Time at risk    =   4603.66872
                                         LR chi2(1)       =      4.20
Log likelihood  =   -234.78217           Prob > chi2      =    0.0405
```



```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |   .5426351   .1643032    -2.02   0.043     .2997606    .9822933
------------------------------------------------------------------------------
```

- This is a test of equality of CHD mortality rates between individuals with a high and low energy intake, adjusted for attained age, and assuming proportional hazards with respect to attained age.
- That is, it is a very similar test to that performed in the framework of Poisson regression on slide 205 (and to the log-rank test).
- The effect estimate and P-value for the test of the effect of `hieng` are very similar in the Cox and Poisson regression models.
- A slight difference is that attained age was categorised in the Poisson regression model and the rate assumed to be constant within each category.

Rates over age by hieng



Summary so far

- We have introduced the Cox model to model survival data.
- The Cox model is an alternative to the Poisson regression model.
- The Cox model does not assume a shape of the baseline hazard, but allows it to vary freely.
- The Cox model assumes proportional hazards (so does the Poisson regression model).
- We need to assess the appropriateness of the proportional hazards assumption.

Assessing the appropriateness of the proportional hazards assumption

- The proportional hazards (PH) assumption is a strong assumption and its appropriateness should always be assessed.
- A PH model assumes that the *ratio* of the hazard functions for any two patient subgroups (i.e. two groups with different values of the explanatory variable $\boldsymbol{X}$) is constant over time.
- Note that it is the hazard ratio which is assumed to be constant. The hazard can vary freely with time.
- When comparing an aggressive therapy vs a conservative therapy, for example, it is not unusual that the patients receiving the aggressive therapy do worse earlier, but then have a lower hazard (i.e. better survival) than those receiving the conservative therapy.

- In this situation, the ratio of the hazard functions will not be constant over time, as is assumed by the PH model.
- Figure 2 (slide 42) shows an example of non-proportional hazards, although this may not be obvious to the untrained eye; it is difficult to assess the PH assumption by looking at the estimates of the survivor function.
- If the hazard functions cross, it is possible that the effect (HR) of treatment will be close to 1 and not statistically significant in a PH model despite the presence of a clinically interesting effect.
- As such, it is important to plot survival and hazard curves before fitting the model and to assess the appropriateness of the proportional hazards assumption after the model has been fitted.
- Note that the hazard functions do not have to cross for the PH assumption to be violated. For example, a hazard ratio of 4 which gradually decreases with time to a value of 1.5 is an example of non-proportional hazards.

- Hess (1995) [18] reviews methods for assessing the appropriateness of the proportional hazards assumption.

- Therneau & Grambsch [23] give a more up-to-date review and include code for implementing the various methods in SAS and R.

Methods to assess the PH assumption

- Following is a list of commonly used methods for assessing the appropriateness of the proportional hazards assumption:

  1. Plotting the cumulative survivor functions and checking that they do not cross. This method is not recommended, since the survivor functions do not have to cross for the hazards to be non-proportional.
  2. Plotting the log cumulative hazard functions over time and checking for parallelism.
  3. Plotting Schoenfeld's residuals against time to identify patterns (for Cox model only).
  4. Including time-by-covariate interaction terms in the model and testing statistical significance. For example, a statistically significant time-by-exposure term would indicate a trend in the hazard ratio with time.

- The first two methods do not allow for the effect of other covariates, whereas the second two methods do. (Not entirely true, for the first two methods, we can do plots for subgroups of patients with given covariate patterns, though

this is less straight-forward.)

- If the PH assumption is violated there are two ways to accommodate non-proportional hazards: (1) including time-by-covariate interaction terms, or (2) fit a stratified Cox model (see Day 4).

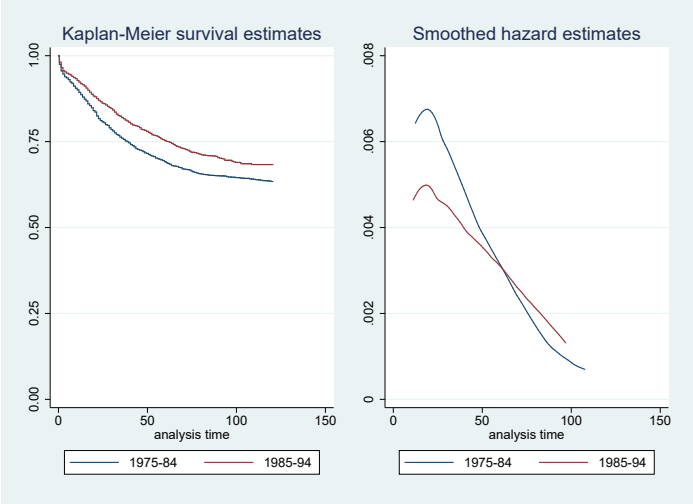
- Including a time-by-covariate interaction in the model has the advantage that we obtain an estimate of the hazard ratio as a function of time, i.e. the hazard ratio will depend on time and differ over time.

1. Plots of the survivor and hazard functions

- The simplest method to assess proportional hazards is simply to plot the survivor function or the hazard function by the exposure groups, and check if the hazards look proportional over time.
- It is often easier to assess PH on the hazard scale.
- In the following graph there is evidence of non-propotional hazard as the hazard curves cross.

```
. sts graph, by(year8594)
. sts graph, by(year8594) haz
```


Plots of S(t) and h(t)



2. Plots of the log cumulative hazard function

- The hazard function and the survivor function are related. One relationship of particular importance is

$$\begin{aligned} S(t) &= \exp\left[-\int_0^t \lambda(s)\,\mathrm{d}s\right] \\ &= \exp(-\Lambda(t)), \end{aligned} \tag{13}$$

  where $\Lambda(t)$ is called the cumulative hazard (or integrated hazard) at time t.

- If we use a proportional hazards model (e.g. Cox or Poisson), then another way to write this equation is

$$S(t|\boldsymbol{X}) = \{S_0(t)\}^{\exp(\beta_1 X_1 + \cdots + \beta_k X_k)}.$$

- I.e. the baseline survivor function is related to the survivor function via the linear predictor.

- Consider the situation where we have only a single binary variable, X, then

$$S(t|X=1) = \{S(t|X=0)\}^r,$$

where $r = \exp(\beta)$ is the hazard ratio.

- Taking natural logarithms of both sides gives

$$\ln S(t|X=1) = r \ln\{S(t|X=0)\}.$$

- Taking natural logarithms of the negatives of both sides gives

$$\ln[-\ln S(t|X=1)] = \ln r + \ln[-\ln\{S(t|X=0)\}].$$

- Consequently, if the proportional hazards model is appropriate, plots of $\ln[-\ln S(t)]$ vs t for each group will be parallel, with the constant difference between them equal to $\ln r$, which is the coefficient β.

- From the equation above, we see that $-\ln S(t)$ is equivalent to the cumulative hazard function, $\Lambda(t)$, and that $\ln[-\ln S(t)] = \ln \Lambda(t)$.

- Consequently, plots of $\ln[-\ln S(t)]$ are often called log cumulative hazard plots. In Stata this can be done by the `stphplot` command.

- Figure 5 was constructed using the following command.

```
stphplot, by(year8594)
```

- The estimated regression coefficient for calendar period is $\ln(0.755) = -0.28$, so we would expect a constant difference of approximately 0.28 between the curves.

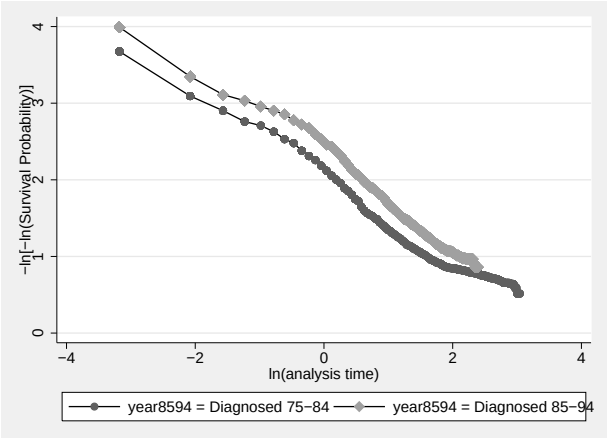


Figure 5: Log cumulative hazard plot by calendar period for the localised colon carcinoma data

- This appears to be the case, except possibly for higher values of $\ln(t)$.
- The proportional hazards assumption for calendar period appears to be appropriate.
- Note that the lines do not have to be straight, it is only necessary for there to be a constant difference between the lines.
- Plotting $\ln(t)$ (as opposed to t) on the x axis results in straighter lines and it is therefore easier to study whether the difference is constant.
- Note that Figure 5 is based on estimates made using the Kaplan-Meier method which, unlike the estimates from the Cox model, are not adjusted for age and sex.
- It is, however, possible to construct adjusted plots in Stata.

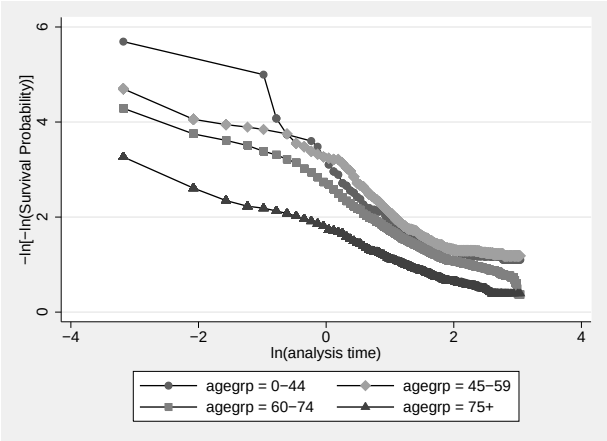


Figure 6: Log cumulative hazard plot by age for the localised colon carcinoma data, showing some evidence of non-proportional hazards

3. Tests of the PH assumption based on Schoenfeld residuals

- If the PH assumption holds then the Schoenfeld residuals (a diagnostic specific to the Cox model) should be independent of time.
- In its simplest form, when there are no ties, the Schoenfeld residual for covariate $x_u, u = 1, ..., p$, and for observation j observed to fail is

$$r_{uj} = x_{uj} - \frac{\sum_{i \in R_j} x_{ui} \exp(\mathbf{x}_i \hat{\beta}_{\mathbf{x}})}{\sum_{i \in R_j} \exp(\mathbf{x}_i \hat{\beta}_{\mathbf{x}})}$$

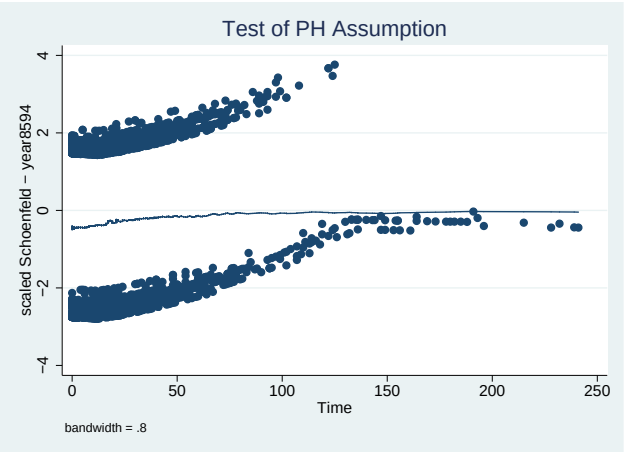
- That is, r_{uj} is the difference between the covariate value for the failed observation and the weighted average of the covariate values over all those subjects at risk of failure when subject j failed.
- A test of the PH assumption can be made by modelling the Schoenfeld residuals as a function of time and testing the hypothesis of a zero slope.

Application to localised colon carcinoma

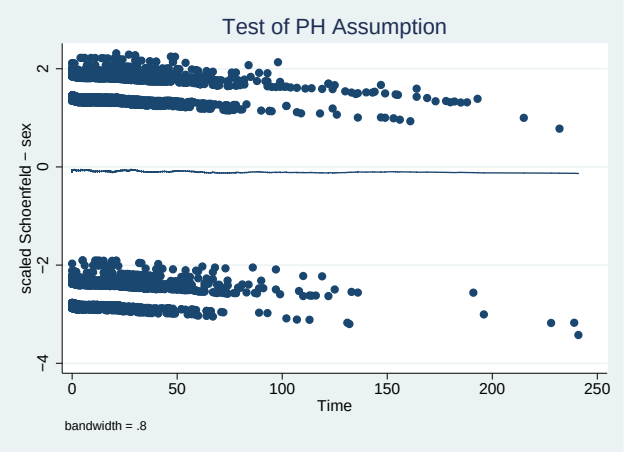
```
. use colon if stage==1, clear
. stset surv_mm, failure(status==1) scale(12)
. quietly stcox sex i.agegrp year8594
. estat phtest, detail

Test of proportional-hazards assumption
Time:  Time
------------------------------------------------------------------
             |       rho            chi2       df       Prob>chi2
-------------+----------------------------------------------------
sex          |      0.00840         0.12        1          0.7262
0b.agegrp    |            .            .        1               .
1.agegrp     |      0.01366         0.32        1          0.5695
2.agegrp     |      0.04178         3.05        1          0.0809
3.agegrp     |     -0.00178         0.01        1          0.9410
year8594     |      0.07231         9.29        1          0.0023
-------------+----------------------------------------------------
global test  |                     27.72        5          0.0000
------------------------------------------------------------------
```


- The tests suggest that there is evidence that the hazards are nonproportional by calendar period (and possibly age).
- Rather than just fitting a straight line to the residuals and testing the hypothesis of zero slope (as is done by `stphtest`) we can study a plot of the residuals along with a smoother to assist us in determining how the mean residual varies as a function of time.
- The smoother illustrates how the log hazard ratio varies as a function of time. We see, for example, that the effect of period is larger during the initial years of follow-up.


```
. estat phtest, plot(year8594)
```



```
. estat phtest, plot(sex)
```



A model including stage

```
. use http://www.biostat3.net/download/colon, clear
. drop if stage == 0 /* remove unknown stage */
. stset surv_mm, failure(status==1)
. stcox sex i.agegrp i.stage year8594
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	.9559269	.0232954	-1.85	0.064	.911342	1.002693
agegrp						
45-59	1.087061	.0693794	1.31	0.191	.9592411	1.231913
60-74	1.308011	.0767528	4.58	0.000	1.165907	1.467436
75+	1.835699	.1089947	10.23	0.000	1.634035	2.062252
stage						
Regional	2.300746	.0945407	20.28	0.000	2.122715	2.493708
Distant	8.072185	.2375035	70.98	0.000	7.619854	8.551367
year8594	.8601408	.0206306	-6.28	0.000	.8206413	.9015415

- Stage is categorised into localised (1), regional (2) and distant (3) tumours.

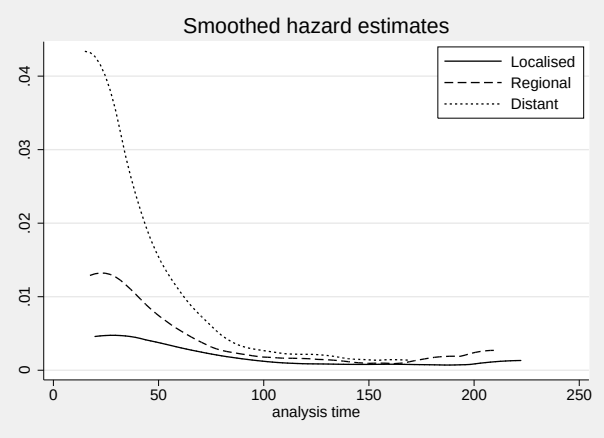

```
. estat phtest, detail
Test of proportional-hazards assumption
Time:  Time
```

	rho	chi2	df	Prob>chi2
sex	-0.00182	0.02	1	0.8773
0b.agegrp	.	.	1	.
1.agegrp	-0.00122	0.01	1	0.9179
2.agegrp	0.02013	2.92	1	0.0876
3.agegrp	-0.00743	0.40	1	0.5296
1b.stage	.	.	1	.
2.stage	-0.04083	11.88	1	0.0006
3.stage	-0.15970	168.33	1	0.0000
year8594	0.02512	4.58	1	0.0323
global test		210.42	7	0.0000

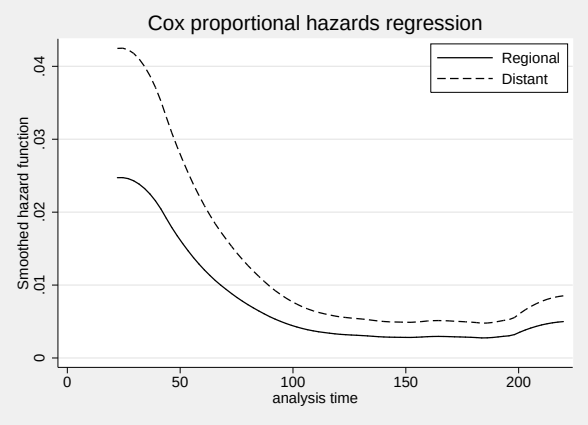
- There is evidence that the hazards are heavily non-proportional by stage.

- A plot of the empirical hazards (slide 319) suggests that individuals diagnosed with distant metastases have proportionally much higher mortality early in the follow-up but once they have survived several years their mortality is not that much higher than the other age groups.
- The plots of the fitted hazards (slide 320) show the effect of the assumption of proportional hazards.
- Exercise: Draw the corresponding hazard ratio across time.

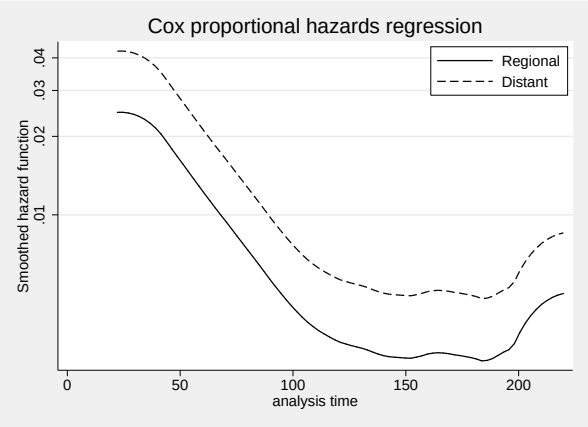

```
/* empirical hazards by stage */
. sts graph, hazard by(stage)
```



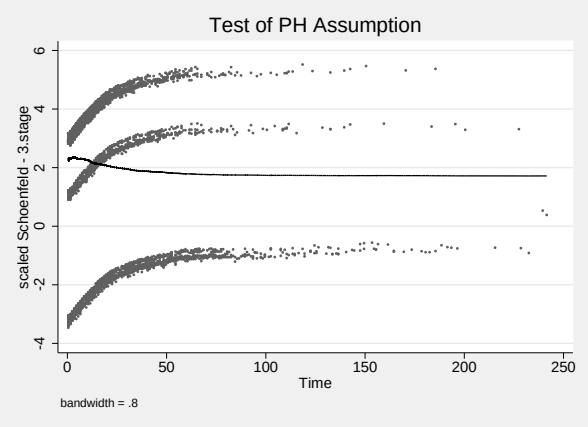

```
/* fitted hazards by stage */
. stcurve, hazard at1(stage=2) at2(stage=3)
```



```
/* fitted hazards by stage (on log scale) */
. stcurve, hazard at1(stage=2) at2(stage=3) yscale(log)
```




```
/* Can also plot a smooth of the scaled Schoenfeld residuals */
. estat phtest, plot(3.stage)
```



4. Modelling interactions with time to test and model non-proportional hazards

- Non proportional hazards is just a special name for 'effect modification by time', i.e. the hazard ratio depends on and differ across time.
- Effect modification is a familiar concept; we can use interaction terms to test for effect modification and to estimate the effect of exposure in each stratum of the modifier.
- To allow for non-proportional hazards we fit time by covariate interaction effects.
- In Poisson regression, we can easily include time by covariate interaction terms in the model after time-splitting.
- The difficulty with the Cox model is that we do not explicitly estimate the effect of time so it is not obvious how to fit a time by covariate interaction.

- We can use one of two approaches in Cox:
  - Split by time, and include time by covariate interaction (using a special parameterisation).
  - Use the options in Stata for modelling 'time-varying covariates' (the `tvc()` option to `stcox`).
- What we are actually interested in is the situation where the *effect* (β) of a covariate varies by time, which is not the same as the *value* of covariate (X) varying with time. We'll discuss the distinction in more detail on slide 344.
- We do not explicitly estimate the effect of the underlying time scale in a Cox model, but we can estimate interactions with the underlying time scale.
- We still allow the baseline hazard to vary freely, but relax the assumption that hazards must be proportional over time, i.e. β can depend on time, $\beta(t)$.
- Note that it is possible to estimate the underlying time-scale (baseline hazard) after fitting a Cox model (type `help stcox postestimation`).

Modelling interactions with time, by splitting time (Cox)

- One way to model interactions with the underlying time scale in a Cox model, is to split time and allow covariates to have different effects over time.
- The Stata `stsplit` divides risktime into several records, one for each timeband we specify.
- We will now model an interaction with time in the colon carcinoma data, to allow for different hazard ratios for calendar period before and after 2 years (24 months) of follow-up since diagnosis.

Colon data: estimating a time by period interaction

- We have seen that mortality depends on calendar period of diagnosis (HR 0.72 for recent/early period).
- Would we expect mortality in the recent period to be 28% lower at all points in the follow-up or is it conceivable that the effect is greater (or even restricted) to the period immediately following diagnosis?
- If the effect is different early in the follow-up, compared to later in the follow-up, then we have a case of non-proportional hazards.
- That is, the effect of calendar period is modified by time since diagnosis.
- Based on clinical knowledge, we choose to estimate the effect separately for the first 24 months of follow-up.

- We start with splitting the data on time, $t < 24$ months, using `stsplit`.

```
. use colon if stage==1, clear
. gen id=_n
. stset surv_mm, failure(status==1) id(id)
. stsplit timeband, at(0,24,1000)
(4611 observations (episodes) created)
```

- We can now fit a model containing the interaction between year of diagnosis (two categories) and time (in two categories).


```
. stcox sex i.agegrp i.year8594##i.timeband
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	.9145475	.0451506	-1.81	0.070	.8302006	1.007464
agegrp						
45-59	.9494187	.1314437	-0.37	0.708	.7237889	1.245385
60-74	1.336923	.1680924	2.31	0.021	1.044923	1.710522
75+	2.250161	.2836501	6.43	0.000	1.757572	2.880806
year8594						
Diagnosed 85-94	.6566005	.0428808	-6.44	0.000	.5777122	.7462612
24.timeband	54.5918	.	.	.	.	.
year8594#timeband						
Diagnosed 85-94 # 24	1.378824	.1362721	3.25	0.001	1.136012	1.673536

- Recall how we interpret interaction effects (in general).
  - `Diagnosed 85-94`; effect of period at the reference of timeband (i.e., the first 24 months).
  - `24.timeband`; effect of time at the reference level of period (the early period).

  - `Diagnosed 85-94 # 24`; additional (multiplicative) effect of period at the second level of timeband (after 24 months).

- Recall how we estimated interaction models in Day 2. The IRRs can be tabulated as

Year	0-24	24+
1975-84	1.00	"54.59" $-$ *meaningless*
1985-94	0.6566	$0.6566 \times 54.59 \times 1.3788$

- `24.timeband` does not have the usual interpretation because we have already adjusted for the effect of time since diagnosis (as the underlying timescale).

- We are effectively trying to adjust for the same confounder in two different ways in the same model. We should ignore this estimate and focus on the other two.

- Since time has no meaning in the Cox model, we choose instead to show the interaction within timebands.

Year	0-24	24+
1975-84	1.00	1.00
1985-94	0.6566	$0.6566 \times 1.3788 = 0.91$

- The estimated hazard ratio for the effect of period of diagnosis is
  - 0.72 when assuming proportional hazards
  - 0.66 for the early period
  - 0.91 for the recent period ($0.656 \times 1.378 = 0.91$)
- We see that there is evidence that the effect of period of diagnosis is more pronounced early in the follow-up (HR=0.66).

- If the interaction effect was zero (HR associated with `Diagnosed 85-94 # 24` equal to one) then there would be no effect modification (proportional hazards).
- We can see that the interaction effect (1.379) is statistically significant (p=0.001) using the Wald test.
- We can reparameterise the model to estimate the effect of period within each timeband, by creating a dummy variable for exposure within each timeband. Variable `year8594_0` will take value 1 for observations where year8594=1 and timeband=0, and value 0 otherwise. Variable `year8594_24` will take value 1 for observations where year8594=1 and timeband=24, and value 0 otherwise.

```
. gen year8594_0 = (year8594==1)*(timeband==0)
. gen year8594_24 = (year8594==1)*(timeband==24)
```

- Then we fit the model again using these indicator variables for the effect of calendar period over time.

- Note that we could also use the Stata `lincom()` command rather than reparameterising the model.

```
. stcox sex i.agegrp year8594_0 year8594_24
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9145475   .0451506    -1.81   0.070     .8302006    1.007464
             |
      agegrp |
      45-59  |   .9494187   .1314437    -0.37   0.708     .7237889    1.245385
      60-74  |   1.336923   .1680924     2.31   0.021     1.044923    1.710522
        75+  |   2.250161   .2836501     6.43   0.000     1.757572    2.880806
             |
  year8594_0 |   .6566005   .0428808    -6.44   0.000     .5777122    .7462612
 year8594_24 |   .9053366   .0673392    -1.34   0.181     .7825236    1.047424
------------------------------------------------------------------------------
```

- The estimated hazard ratio, based on the above model, for patients diagnosed 1985–94 compared to 1975–84 is 0.657 for the period up to 2 years of follow-up and 0.905 for the period after 2 years of follow-up (as we previously saw).

- To test if this interaction is statistically significant we could perform a LR test, comparing the model with the interaction to the model without the interaction.

```
. stcox sex i.agegrp year8594
. est store A
. stcox sex i.agegrp year8594_0 year8594_24
. est store B
. lrtest A B

Likelihood-ratio test                 LR chi2(1)  =  10.54
(Assumption: A nested in B)           Prob > chi2 =   0.0012
```

- Note that the previous z test statistic from the Wald test (slide 328) was 3.25. If we square this we get a test statistic that is χ^2_1. $3.25^2 = 10.56$

- Both of these tests are testing the hypothesis that the interaction effect is zero versus it is non-zero. The reason for the small difference in the test statistic is that one is a likelihood ratio test and one is a Wald test.

A look at the interaction models (for completeness)

- Consider again a proportional hazards model with one single binary variable, X_1, which takes the value 1 if an exposure is present and 0 if it is absent

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\beta_1 X_1).$$

- The hazard ratio for exposed to unexposed is given by $\exp(\beta_1)$.

- We now construct a second variable, $X_2 = X_1 t$ and include this in the model, in addition to X_1. The variable X_2 takes the value t if the exposure is present and 0 if it is absent

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_1 t).$$

- Based on this model, the hazard ratio for exposed to unexposed is given by $\exp(\beta_1 + \beta_2 t)$.

- An estimate for β_2 significantly different from 0 indicates that the hazard ratio is non-constant over time. $\beta_2 > 0$ indicates that the hazard ratio increases with time and $\beta_2 < 0$ indicates it decreases with time.

- This is not a general test of the proportional hazards assumption. It tests against the alternative that the hazard ratio changes monotonically with time.

- Another alternative might be that the hazard ratio is constant for an initial time period, say $t = 2$ years, but takes on a different (constant) value for the remainder of follow-up [17].

- To test against this alternative, we construct a variable X_2 which takes the value 1 if the exposure is present and $t > 2$ years, and 0 otherwise.

- In the resulting model containing the variables X_1 and X_2, the hazard ratio for exposed to unexposed for the period $t \leq 1$ year is given by $\exp(\beta_1)$ and for $t > 2$ years it is given by $\exp(\beta_1 + \beta_2)$.

- An estimate for β_2 significantly different from 0 indicates that the hazard ratio is different between the two time periods.

Modelling interactions, by splitting time (Poisson)

- Similarly to Cox regression, we can test for non-PH also in the Poisson model by including interaction terms for time by exposure interaction.

- In a Poisson model, we already adjust for time (i.e. we have split on time and included timeband in the model as covariate).

- To include time by exposure interactions, we simply include interaction terms for exposure and timeband.

```
. use colon if stage==1, clear
. stset surv_mm, fail(status=1) id(id)
. stsplit timeband, at(0,24,1000)
. streg sex i.agegrp i.year8594##i.timeband, dist(exp)
```

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .8935971   .0441139    -2.28   0.023     .8111866    .9843798
             |
      agegrp |
       45-59 |   .9717692   .1345236    -0.21   0.836     .7408493    1.274666
       60-74 |   1.425765    .179201     2.82   0.005     1.114455    1.824036
         75+ |   2.569885   .3238718     7.49   0.000      2.00743    3.289933
             |
    year8594 |
 Diagnosed 85-94 |   .6514858   .0425449    -6.56   0.000     .5732152    .7404439
  24.timeband |   .2847188   .0190458   -18.78   0.000     .2497333    .3246054
             |
year8594#timeband |
Diagnosed 85-94#24 |   2.045482    .197872     7.40   0.000     1.692208    2.472507
             |
       _cons |   .0064418   .0009341   -34.79   0.000     .0048482    .0085593
------------------------------------------------------------------------------
Note: _cons estimates baseline hazard.
```


- We can present this in a table:

Year	0-24	24+
1975-84	1.00	0.2847
1985-94	0.6514	$0.6514 \times 0.2847 \times 2.0454$

- We can also calculate the hazard ratios associated with period within timebands:

Year	0-24	24+
1975-84	1.00	1.00
1985-94	0.6514	$0.6514 \times 2.0454 = 1.33$

- Testing for interaction is the same in Poisson as for Cox. The p-value for the interaction term is significant ($p < 0.000$).
- The results from the Cox and Poisson models are different. Why?
- One reason for this could be that the Poisson model is not modelling the underlying time scale well enough. Splitting only into two timebands may not capture the underlying shape of the hazard.

```
. use colon if stage==1, clear
. stset surv_mm, fail(status=1) id(id)
. stsplit timeband, at(0,12,24,36,48,60,1000)
. streg sex i.agegrp i.year8594##i.timeband, dist(exp)

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9094825   .0449032    -1.92   0.055      .825598     1.00189
             |
      agegrp |
       45-59 |   .9536832   .1320206    -0.34   0.732       .72706    1.250944
       60-74 |   1.353585   .1701078     2.41   0.016     1.058068    1.731641
         75+ |   2.335439   .2942033     6.73   0.000     1.824482    2.989491
             |
    year8594 |
Diagnosed 85-94 |   .6711657   .0573238    -4.67   0.000     .5677134    .7934696
             |
    timeband |
          12 |   .8503601   .0792878    -1.74   0.082     .7083316    1.020867
          24 |   .5850169   .0636857    -4.92   0.000     .4726128    .7241548
          36 |   .4795016   .0581614    -6.06   0.000     .3780446    .6081869
          48 |   .3715041   .0516819    -7.12   0.000      .282845    .4879537
          60 |   .1447364   .0146098   -19.15   0.000     .1187564    .1764001
             |
```



```
  year8594#timeband |
Diagnosed 85-94#12 |   .9392502    .124205    -0.47   0.636     .7248026    1.217147
Diagnosed 85-94#24 |   1.185639   .1801977     1.12   0.263     .8802047     1.59706
Diagnosed 85-94#36 |   1.194534   .2074875     1.02   0.306     .8498595    1.678998
Diagnosed 85-94#48 |   1.411381   .2817559     1.73   0.084     .9543726    2.087233
Diagnosed 85-94#60 |   2.499684    .399695     5.73   0.000     1.827172    3.419722
             |
       _cons |   .0071894   .0010808   -32.83   0.000     .0053547    .0096527
------------------------------------------------------------------------------
Note: _cons estimates baseline hazard.
```

- If we split time finer, then the POisson model also models the interaction in more categories (and is not comparable to the Cox model with two timebands).

Time-varying covariates and effects

- We have been considering the situation where the *effect* β of a covariate varies with time.
- It is possible that the underlying *values* of covariates X can change during follow-up. For example, blood pressure, occupational exposure to carcinogens, parity, CD4 count, or cumulative exposure to cigarettes.
- Another application is in observational studies where an intervention may occur at any point in the follow-up. At the time of the intervention, the explanatory variable associated with the intervention changes value from 0 (false) to 1 (true).
- We highly recommend the time-splitting approach for modelling such data.
- That is, we split to obtain a separate observation at every value of the time-varying covariate.

- Exercise 125 examines a possible effect of *marital bereavement* (loss of husband or wife) on all–cause mortality in the elderly (see Clayton & Hills, §32.2).

- Bereavement is a time-varying exposure – all subjects enter as not bereaved but may become bereaved at some point during follow–up.

- A distinction is made between internal variables (which relate to an individual and can only be measured while a patient is alive) and external variables (which do not necessarily require survival of the patient for their existence).

- Care should be taken when modelling time-dependent covariates, particularly with internal variables [15, 25].

- A fix exposure can have a constant effect (main effect) or a time-varying effect (interaction). E.g. sex is a fix exposure, but the effect of being woman/man may be different at young and old age.

- A time-varying exposure typically also have a time-varying effect (but in rare cases it can have a constant effect). E.g. smoking is often a time-varying exposure. Usually the risk of a disease depends on the amount of smoking and how it varies over age (time-vary effect), but sometimes having ever smoked (regardless of when and how much) may permanently increase the risk of disease (constant effect).

The tvc (time-varying effects) option in stcox

- The `tvc()` and `texp()` options to `stcox` are used for time-varying exposures but can also be used for estimating time-varying effects of covariates. It does not require time splitting.

- The option will automatically create the dummy variables that we previously coded ourselves after time splitting.

- Let's again fit the model where we allow the effect of period to differ in the first 2 years of follow-up.

```
. stcox sex i.agegrp year8594, tvc(year8594) texp(_t >= 24)
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
         sex |   .9145475   .0451506    -1.81   0.070     .8302006    1.007464
             |
      agegrp |
      45-59  |   .9494187   .1314437    -0.37   0.708     .7237889    1.245385
      60-74  |   1.336923   .1680924     2.31   0.021     1.044923    1.710522
        75+  |   2.250161   .2836501     6.43   0.000     1.757572    2.880806
             |
    year8594 |   .6566005   .0428808    -6.44   0.000     .5777122    .7462612
-------------+----------------------------------------------------------------
tvc          |
    year8594 |   1.378824   .1362721     3.25   0.001     1.136012    1.673536
------------------------------------------------------------------------------
Note: variables in tvc equation interacted with _t>=24
```

- `tvc year8594` (1.3788) is the interaction term.

- The cutoff at 24 months was chosen arbitrarily. For the first 6 months of follow-up the estimated hazard ratio was 0.724, for the first year it was 0.676, and for the first two years it was 0.657.
- Choosing the cutpoint after inspection of the data will invalidate statistical inference (i.e. reported P-values will be too low).
- We have examined only one possible alternative to proportional hazards (a step function with a single step at 24 months).
- In practice, it is possible to fit any model of the form

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_1 f(t)),$$

where $f(t)$ is a function of time.

Summary of Day 3

- We have introduced the Cox proportional hazards regression model and shown how it is very similar to Poisson regression.
- The Cox model assumes proportional hazards (as does Poisson regression), which means that the estimated HRs between groups are constant over time, although we can relax this assumption by modelling interactions.
- The proportional hazards assumption can be tested by fitting time by covariate interactions, which allows effects to vary over time.
- The PH assumption in Cox regression can also be tested using scaled Schoenfeld residuals.

- Poisson regression models assume constant hazards or piecewise constant hazards over time, whereas the Cox model allows the hazard to vary freely over time.
- Can make Poisson regression more 'Cox-like' by making the pieces smaller.
- Hazard ratios from a Cox model are automatically adjusted for confounding by the underlying time scale. One should choose an appropriate timescale.

Exercises for Day 3

120. Localised melanoma: modelling cause-specific mortality using Cox regression. [This is the key exercise]
121. Examining the proportional hazards hypothesis (localised melanoma).
122. Cox regression with observed (all-cause) mortality as the outcome.
123. Cox model for cause-specific mortality for melanoma (all stages).
124. Modelling the diet data using Cox regression.

Topics for Day 4

- Stratified Cox models.
- Flexible parametric survival models.
- More on censoring and truncation, including informative censoring.
- Competing risks analysis (limited coverage).
- Standardised mortality/incidence ratios.
- Biases in survival analysis/cohort studies (not a comprehensive list).

The stratified Cox model

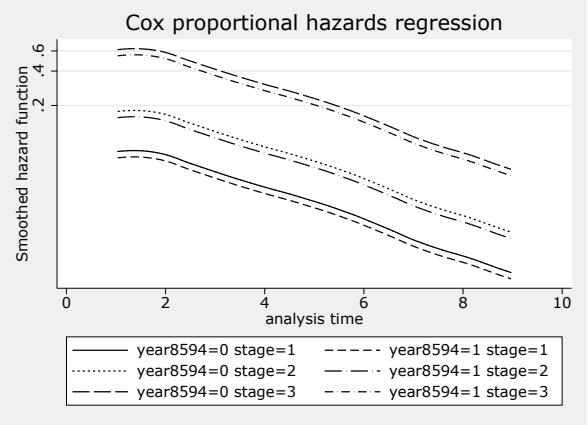
- The Cox model assumes that the baseline hazard (e.g., instantaneous mortality rate in the reference group) is an arbitrary function of time.
- The hazard functions for each of the other groups are assumed to be proportional to the baseline.
- It is possible to relax this assumption to allow separate baseline hazards for different groups, say for each level of stage at diagnosis.
- This is known as a stratified proportional hazards model and is a useful method for modelling data where non-proportional hazards are suspected for a factor that is not of primary interest.
- A model stratified on stage is analogous to including a stage*time interaction in a Poisson regression model.
- Use the `strata()` option in Stata to specify up to 5 strata variables.

- Let's revisit the example of colon cancer.
- We will focus on the HR of cancer-specific death, comparing the two calendar periods.
- Adjusting for stage at diagnosis.
- We previously saw that the effect of stage was non-proportional, which can be taken into account by a stratified Cox model.
- A stratified Cox model should not be confused with stratification in the meaning of performing separate analyses.
- Also in the stratified Cox model, the effect of calendar period is the same across stages. However, the baseline hazard is allowed to differ across stage.

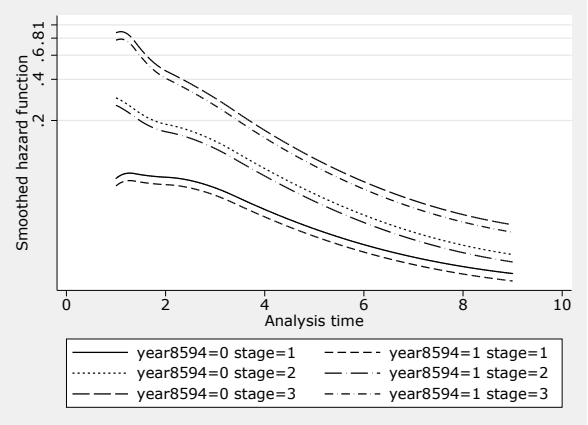
```
. use http://www.biostat3.net/download/colon, clear
. drop if stage == 0
```



```
. stset surv_mm, failure(status==1) scale(12) exit(t 120)
. stcox year8594 i.stage

Cox regression -- Breslow method for ties
No. of subjects =       13,208                 Number of obs    =      13,208
No. of failures =        7,122
Time at risk    =  43950.66667
                                               LR chi2(3)       =     5816.47
Log likelihood  =   -61571.611                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    year8594 |   .8797365   .0210338    -5.36   0.000      .839462    .9219432
             |
       stage |
   Regional  |    2.25628   .0933739    19.66   0.000     2.080496    2.446916
    Distant  |   8.022359   .2377227    70.27   0.000     7.569703    8.502082
------------------------------------------------------------------------------
```



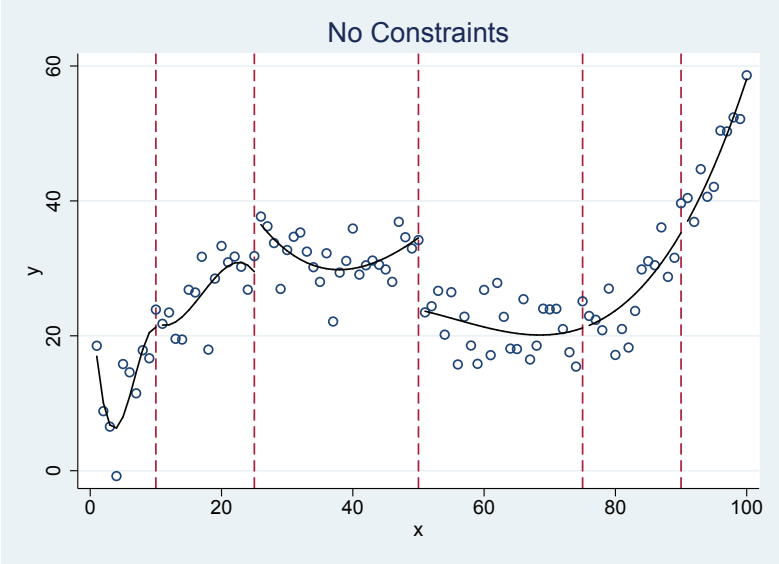

```
. stcox year8594, strata(stage)

Stratified Cox regr. -- Breslow method for ties
No. of subjects =       13,208                  Number of obs    =      13,208
No. of failures =        7,122
Time at risk    =  43950.66667
                                                LR chi2(1)       =       28.11
Log likelihood  =   -55805.194                  Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    year8594 |   .8807531   .0210535    -5.31   0.000     .8404408     .922999
------------------------------------------------------------------------------
                                                        Stratified by stage
```

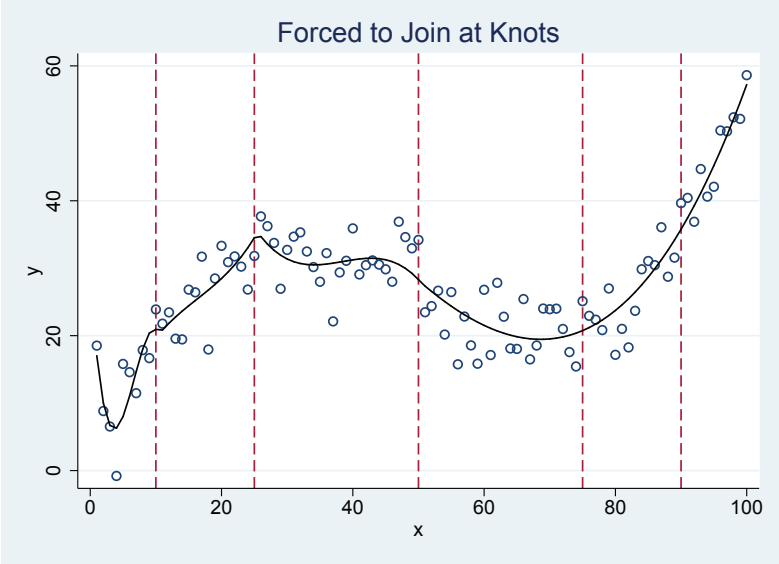



- Stratified Cox models are useful when we want to adjust for a factor and allow for different baseline hazards for different levels of the factor.
- However, stratification should not be done for a covariate of interest.
- Stratified Cox should not be confused with stratification in other models, i.e. fitting separate models
- The HRs obtained from the stratified Cox model are the same across strata.
- Stratified Cox models can be used for multi-center studies, where a treatment effect is assumed the same across centers, but different baseline hazards are assumed for the centers.
- Stratified Cox can also be used with paired/clustered data, such as twin data.
- Stratified Cox can also be used for matched cohorts, by stratifying on the matching strata.

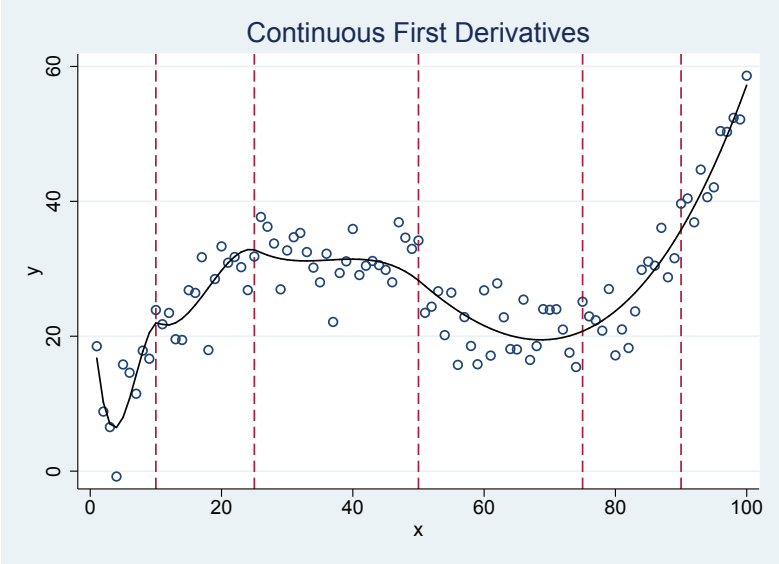
Flexible parametric survival models

- In Cox regression the baseline hazard is not estimated.
- In poisson regression the baseline hazard is estimated as a step function, which is not biologically plausible.
- By fine splitting the steps can be made small and the baseline hazard approximately continuous. The drawback is that it requires the estimation of a lot of parameters.
- One alternative would be to use splines.
- Splines are a way of modeling continuous variables in a flexible way.

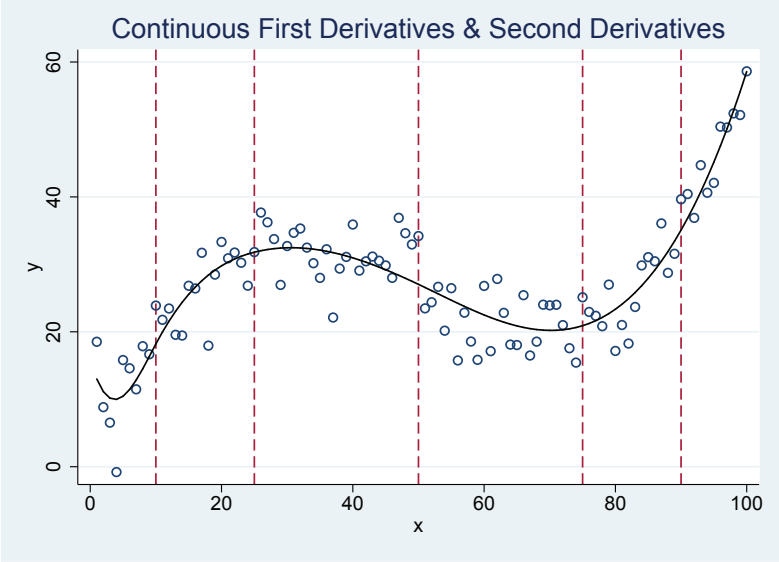



Forced to Join at Knots
y
0
20
40
60
0
20
40
60
80
100
x

363


Continuous First Derivatives
y
0
20
40
60
0
20
40
60
80
100
x

364


Continuous First Derivatives & Second Derivatives
y
0
20
40
60
0
20
40
60
80
100
x

365

- Since the hazard function can have any shape, a lot of knots might be required to model it. Modelling on the hazard scale also requires computationally intensive numerical integration.

- The flexible parametric survival model is instead using the cumulative hazard, which is an increasing function.

- Parameter estimates are still interpreted as hazard ratios (if a PH model).

- Easy to transform to the survival or hazard scale.

- The model can be written as:

$$\ln(H(t;\mathbf{x})) = s(\ln t;\boldsymbol{\gamma_0}, K) + \boldsymbol{\beta}^T\mathbf{x} \tag{14}$$

  where K is the number of knots.

- This is a proportional hazards model, but non-proportional hazards models (time-dependent effects) can be modeled by including interactions between covariates and splines for time.

- Let's again revisit the example of colon cancer.

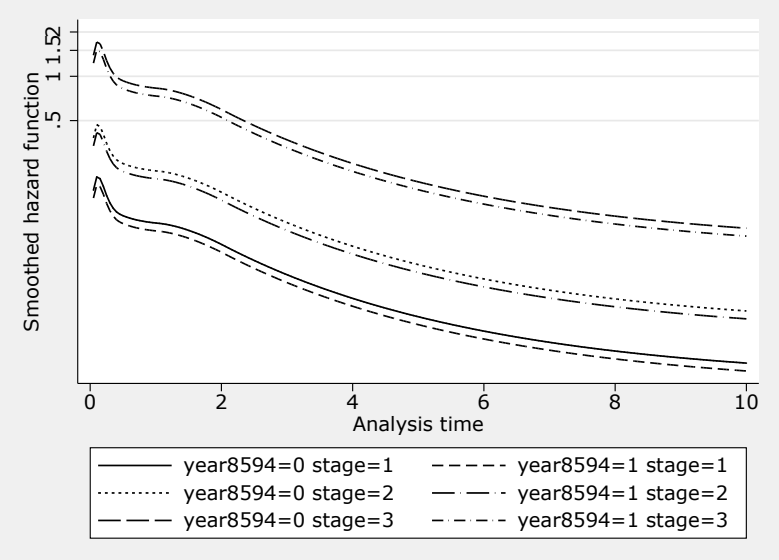
- We will focus on the HR of cancer-specific death, comparing the two calendar periods. Adjusting for stage at diagnosis.

- First a flexible parametric model with proportional hazards.

- Then a flexible parametric model allowing for non-proportional hazards for stage, i.e. including an interaction between time and stage.

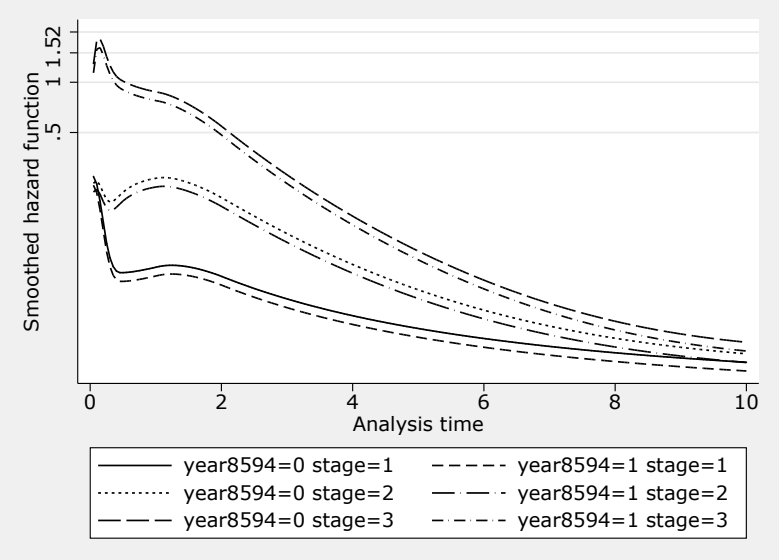
```
. stpm2 year8594  i.stage, scale(hazard) df(5) eform

Log likelihood = -17317.704                     Number of obs     =     13,208

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
    year8594 |   .8838128   .0211523    -5.16   0.000     .8433125    .9262582
             |
       stage |
   Regional  |   2.260989   .0935674    19.71   0.000     2.084841    2.452021
    Distant  |   8.249828   .2439361    71.37   0.000     7.785313     8.74206
             |
       _rcs1 |   3.150462   .0371294    97.37   0.000     3.078524    3.224082
       _rcs2 |   1.302572   .0120809    28.50   0.000     1.279108    1.326467
       _rcs3 |   .9965837   .0059341    -0.57   0.565     .9850207    1.008282
       _rcs4 |   1.048574   .0037925    13.11   0.000     1.041167    1.056034
       _rcs5 |   1.022945   .0028764     8.07   0.000     1.017323    1.028598
       _cons |   .1395332   .0040286   -68.21   0.000     .1318566    .1476567
------------------------------------------------------------------------------
```


- Patients diagnosed in the later calendar period have 12% lower cancer-specific mortality compared to earlier calendar period, after controlling for stage at diagnosis (and the underlying time scale), and this difference is assumed be the same for all stages.
- Patients with regional metastases have more than 2 times the mortality of patients with localised stage, after controlling for calendar period (and the underlying time scale), and the effect is assumed to be the same within both calendar periods.
- Patients with distant metastases have more than 8 times the mortality of patients with localised stage, after controlling for calendar period, and the effect is assumed to be the same within both calendar periods.
- The rest of the parameters are for the splines, and they are not interpreted one by one. However, together they give the function of the baseline.



```
. stpm2 year8594 stage2 stage3, scale(hazard) df(5) tvc(stage2 stage3) dftvc(3) eform

Log likelihood = -17132.061                    Number of obs     =     13,208

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
    year8594 |   .8864017   .0212132    -5.04   0.000     .8457845    .9289694
      stage2 |   1.904373   .1029325    11.92   0.000     1.712947     2.11719
      stage3 |   7.669099   .2661191    58.71   0.000     7.164857    8.208829
        _rcs1 |   2.720272   .0693896    39.23   0.000     2.587615     2.85973
        _rcs2 |   1.072641   .0178232     4.22   0.000     1.038271    1.108149
        _rcs3 |   .9587595   .0098322    -4.11   0.000     .9396812    .9782252
        _rcs4 |   1.046235   .0055504     8.52   0.000     1.035413    1.057171
        _rcs5 |   1.024975   .0028437     8.89   0.000     1.019417    1.030564
_rcs_stage21 |   1.369303   .0736305     5.85   0.000     1.232334    1.521495
_rcs_stage22 |   1.130488   .0418797     3.31   0.001     1.051315    1.215625
_rcs_stage23 |   1.142751   .0245353     6.21   0.000      1.09566    1.191865
_rcs_stage31 |   1.126492    .032725     4.10   0.000     1.064144    1.192493
_rcs_stage32 |   1.337784   .0264739    14.71   0.000      1.28689    1.390691
_rcs_stage33 |   1.034256   .0130236     2.67   0.007     1.009043      1.0601
       _cons |   .1445927   .0048611   -57.52   0.000     .1353723     .154441
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
```


- Patients diagnosed in the later calendar period have 11% lower cancer-specific mortality compared to earlier calendar period, after controlling for stage at diagnosis with non-proportional hazards (and the underlying time scale), and this difference is assumed be the same for all stages.
- Since stage is allowed to have non-proportional hazards, i.e. an interaction between stage and the time-scale, the HR changes over time, and is not one number found in the output.
- However, the HR for stage can be plotted as a function of time (see later example).

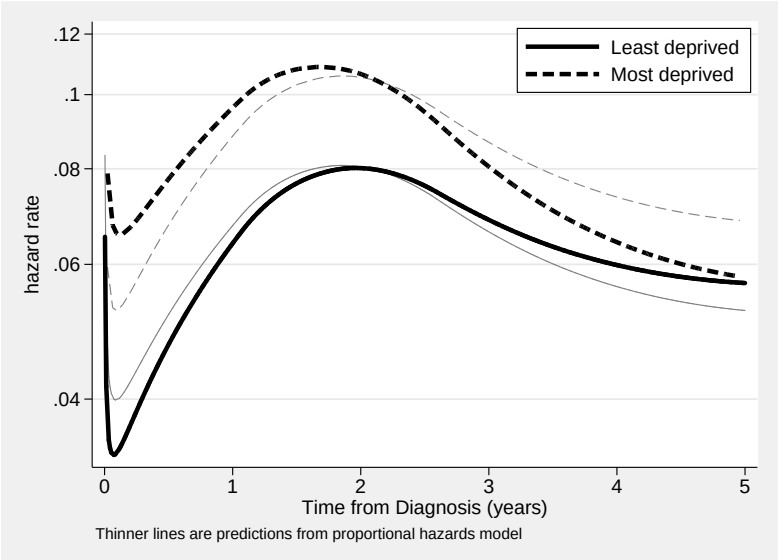


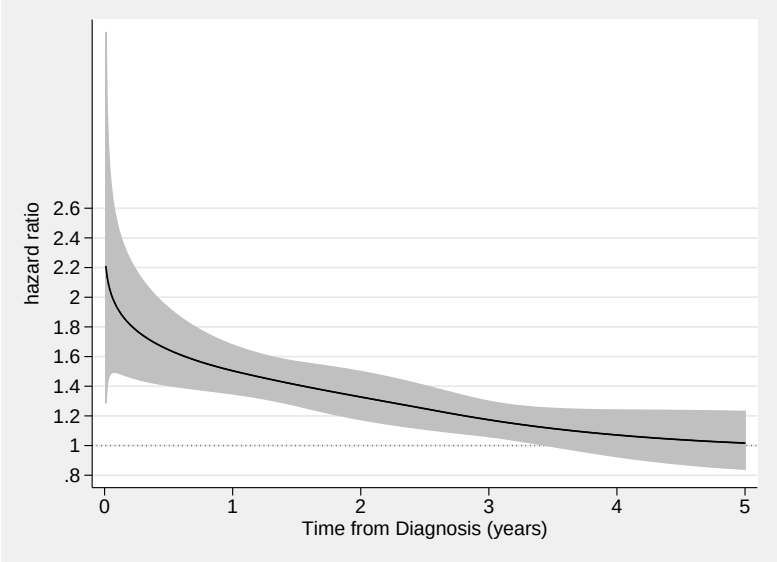
Hazard ratios and standard errors for various models

Variable	Cox	Poisson_fine	Poisson	FPM
sex	0.588814	0.588814	0.587547	0.587710
	0.038538	0.038538	0.038456	0.038467
year8594	0.716884	0.716884	0.722411	0.721818
	0.047445	0.047445	0.047813	0.047780
1.agegrp	1.326397	1.326397	1.327795	1.327769
	0.124911	0.124911	0.125042	0.125040
2.agegrp	1.857323	1.857323	1.862376	1.862001
	0.168787	0.168787	0.169244	0.169211
3.agegrp	3.372652	3.372652	3.400287	3.400808
	0.352227	0.352227	0.355140	0.355178

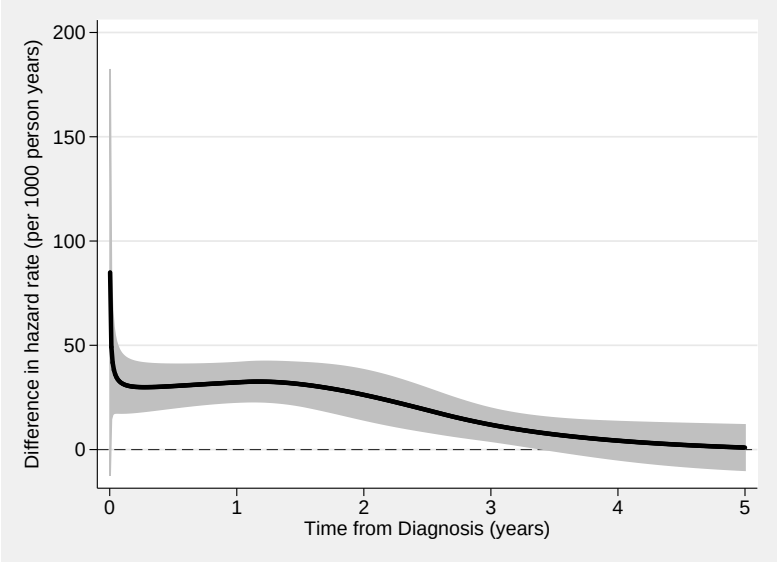
- The HRs from Cox, Poisson and FPM are generally very similar

- Since the baseline hazard is estimated as a continuous function in the flexible parametric survival model it is easy to present results using graphs, and to present results on the hazard scale, as hazard ratios, or the survival scale.
- This is illustrated in the following graphs.
- A flexible parametric survival model fitted to data on breast cancer patients in England, with breast cancer death as the outcome.
- The variable of interest is deprivation status, and results are shown for the lowest and highest group.

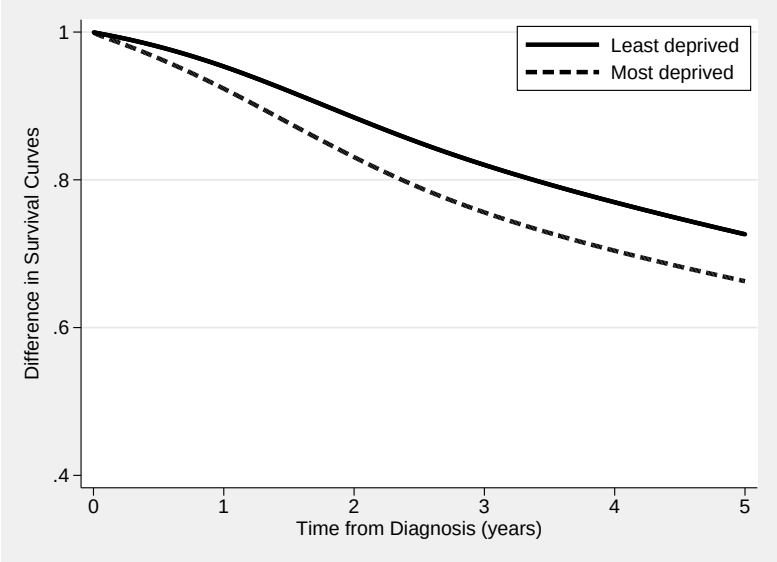



hazard ratio
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
.8
0
1
2
3
4
5
Time from Diagnosis (years)

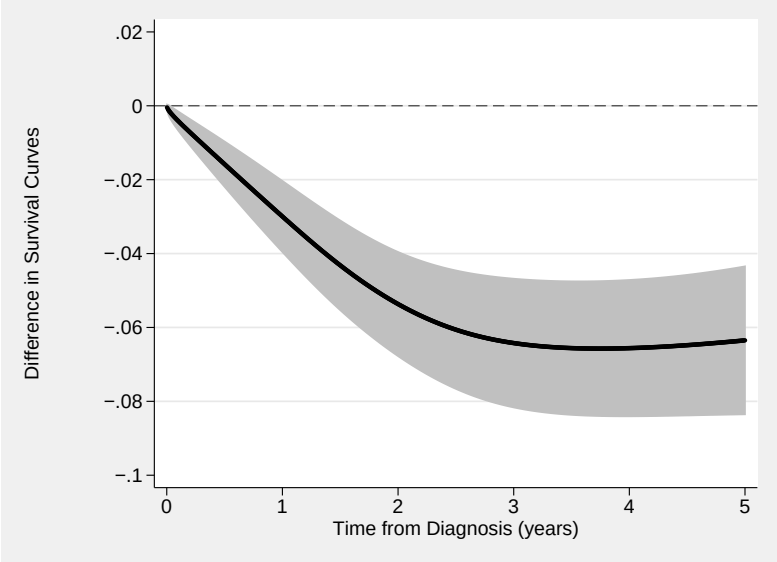
378


Difference in hazard rate (per 1000 person years)
200
150
100
50
0
0
1
2
3
4
5
Time from Diagnosis (years)

379


Difference in Survival Curves
1
.8
.6
.4
Least deprived
Most deprived
0
1
2
3
4
5
Time from Diagnosis (years)

380



Summary, flexible parametric model

- Hazard ratios are very similar to hazard ratios from a Cox model and Poisson model.
- Since the baseline hazard is modelled it is easy to include non-PH, interaction.
- The time-scale is included as a continuous variable, more plausible than step function.
- Easy to present results using graphs.
- The parametric approach enables predictions and extrapolations.

Censoring and truncation

- With right censoring, the most common form of censoring in medical studies, we know that the event has not occurred during follow-up, but we are unable to follow-up the patient further. We know only that the true survival time of the patient is greater than a given value.
- Less common is left-censoring, where we know the event has occurred prior to the time of observation but we don't know exactly when.
- Interval censoring occurs when we know that the event has occurred between two time points but don't know the exact date (e.g. HIV infection between two test dates, or cancer between two screens).
- Standard methods for survival analysis assume that all censored data are right censored and we have only used right censored data in the course.
- Special methods are required for analysing left censored and interval censored data, which is covered in this course.

- Censoring, in general, refers to the situation where we can identify the individuals in our study but we do not have precise information on the event time for all individuals (we know only that it is in some interval).
- A second feature of survival studies, often confused with censoring, is *truncation*.
- Truncation refers to the situation where certain subjects are not observed such that the investigator is not aware of their existence.
- Left truncated data occurs when we only observe the individual if they are event free after a certain follow-up time. For example, late entry to the study or using age as the primary time scale.
- Left truncated data is common. As long as the time-scale is adressed properly, so all subjects are not assumed to be followed from time 0, left truncated data can be analysed with Cox and poisson regression.

- Right truncated data occurs when only individuals who experience the event of interest are included in the study.
- Special methods of analysis are required for analysing right truncated data, such as use of a conditional likelihood or a method which uses a selective risk set (see Klein & Moeschberger (1997) [20]).

Estimating AIDS incubation time: An example of right truncated data

- Knowledge of the time between HIV infection and development of AIDS (called the incubation period) is important in AIDS research.
- The first reliable estimates of incubation time were obtained in the early 1980's by studying individuals who developed AIDS from blood transfusions (before prospective donors were screened for HIV).
- Only individuals who experienced the event could be studied. That is, the data were right truncated.
- Not all blood recipients were exposed to HIV, and not everyone who was exposed had developed AIDS at the time of the analysis.
- Nevertheless, by studying those individuals who developed AIDS as a result of HIV exposure at transfusion, using appropriate statistical methods, it was possible to estimate incubation time.

Informative right-censoring

- To make it possible for statistical analysis we make the crucial assumption that, conditional on the values of any explanatory variables, censoring is unrelated to prognosis (the probable course and outcome of the disease).
- The statistical methods used for survival analysis assume that the prognosis for an individual censored at time t will be no different from those individuals who were alive at time t and were under follow-up past time t.
- One way to think of this is that, conditional on the values of any explanatory variables, the individuals censored at time t should be a random sample of the individuals at risk at time t.
- This is known as noninformative censoring. Under this assumption, there is no need to distinguish between the different reasons for right-censoring.
- When withdrawal from follow-up is associated with prognosis, this is known

as informative censoring and standard methods of analysis will result in biased estimates.

- Common methods for controlling for informative censoring are to stratify or condition on those explanatory factors on which censoring depends.
- Censoring due to termination of the study, or accidental death, are usually uninformative, but careful consideration must be given to other forms of censoring.
- Determining whether or not censoring is informative is not a statistical issue — it must be made based on subject matter knowledge.

Example of informative censoring — dairy cows

- Consider a cohort study of dairy cows where the aim is to study whether the type of feed and type of housing (indoor/outdoor) are associated with incidence of a disease.
- If a cow is slaughtered without the disease being diagnosed, is it appropriate to consider the survival time as censored in the analysis?
- No, not if the disease, or a precursor to the disease, affects milk production and cows with low milk yield are sacrificed.
- If this were the case then the sacrificed cows would be more likely to be diagnosed with the disease than cows who were not sacrificed.

Example of informative censoring — colon cancer in IBD patients

- In a historical cohort study, 19,500 individuals with inflammatory bowel disease (IBD) were identified in the Swedish hospital inpatient separations register and IBD registers maintained in Uppsala and Stockholm.
- We were interested in risk factors for cancer of the colon; the cohort was followed up using the Swedish cancer register.
- Some patients had their colon surgically removed (colectomy) without being diagnosed with colon cancer, so were not at risk for colon cancer.
- These were the patients with the most extensive type of IBD, and it is known that risk of colon cancer is proportional to the extent of the IBD.
- Therefore, censoring due to colectomy is informative.

Competing risks

- Competing risks are outcomes that prevent a person from experiencing another outcome, for example dying from CVD precludes someone of dying from cancer.
- But it doesn't have to be death, for example being discharged from hospital when event of interest is infection during hospital stay.
- When estimating cause-specific survival, survival times from individuals who experience a competing event are usually censored.
- Since censoring assumes that people are still at risk of experiencing the event (but we will not be able to observe when), censoring when people die of another cause will assume that they could still die from the cause of interest. This is in practice impossible.
- Problem therefore arise when looking at and interpreting the survival function (or $1 - S(t)$).

Competing risks, an example

	week	n	infection	discharge
1.	0-1	306	11	15
2.	1-2	280	12	19
3.	2-3	249	16	16
4.	3-4	217	21	16
5.	4-5	180	19	19
6.	5-6	142	21	22
7.	6-7	99	12	14
8.	>7	73	0	73

- What proportion of patients have an infection within the first 5 weeks of hospitalisation? How is this estimated? Censor when discharged?
- Depends on in what context you want to interpret the estimate.

- Without censoring (11+12+16+21+19)/306=0.26
- Life table estimate (censoring for discharge) gives 0.63
- Which estimate would you use if you had to make a budget for hospital costs due to infection? To compare to another hospital to draw conclusions about differences in care?

Competing risks, another example

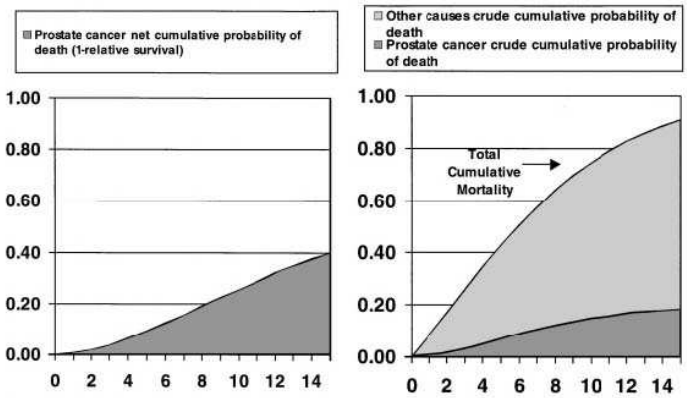


Figure 7: From Cronin and Feuer (2000) [11]

Competing risks

- When there are competing events (risks), we can still estimate and interpret the hazard (and HR), since the hazard is, for each time point, based on those still alive.
- If there are competing risks the hazard rate and HR has to be interpreted as the hazard rate and HR when the competing risks exists.
- If the two competing risks are independent, within variables adjusted for, (non-informative censoring) the survival function can still be interpreted as 'net survival'. This can be thought of as the survival in the absence of competing events.
- Net survival is the proportion of people who would survive up to a certain point in time in the hypothetical scenario where the event of interest is the only possible event (if we could eliminate competing events).

- Can be the measure you are interested in when for example studying temporal trends in cancer patient survival or making comparisons between groups.
- For some research questions it is instead the 'crude survival' which is of interest. This can be thought of as the survival, in the presence of competing events.
- There is a lot of literature on comepting risks, unfortunately most is difficult to read, and sometimes even misleading.
- Special methods are available to estimate measures in the presence of competing risks, but is not covered in this course.

Standardised mortality/incidence ratios

- Sometimes you only have an exposed population, and no unexposed population to compare with.
- We can then not follow-up both the exposed and the unexposed population and compare the two estimated rates.
- We instead only estimate the rate (or number of events) in the exposed population and compare this to the expected rate (expected number of events) for the standard population.
- For example, we might study disease incidence or mortality among individuals with a certain occupation (farmers, painters, airline cabin crew) or cancer incidence in a cohort exposed to ionising radiation.
- The standardized mortality ratio (SMR) is the ratio of the observed number of deaths in the study population to the number that would be expected if the study population experienced the same mortality as the standard population.

- It is an indirectly standardized rate.
- When studying disease incidence the corresponding quantity is called a standardized incidence ratio (SIR).
- Example, estimating relative risk of cancer among organ transplant recipients compared to the general population [21].

Estimation in Stata

- To estimate an SIR/SMR, we typically assume the expected rates depend on age, sex, and calendar period.
- We use `stsplit` to split follow-up time in the same way the file of background rates is classsified. That is, if the population rates are for 5-year age groups we split using the same categories.
- Multiply the expected rate by person-time, Y, to get the expected number of events in each category.
- Collapse to get the total number of observed and expected events in each category; SMR=O/E.
- Modelling is done in the same way as for rates except we have $\ln(E)$ as the offset instead of $\ln(Y)$.

Biases in survival analysis, conditioning on the future

- It is commonly seen in observational studies that exposure is not known at start of follow-up, e.g.
- In these cases be careful how you define (the time-varying) exposure to not condition on the future.
- Those with longer follow-up have a higher chance if being registered as exposed, which could introduce a bias.
- Even when you know that some subjects must have been exposed during the whole period, the reason you know it is because you have information from further follow-up.
- NEVER condition on the future.

Biases in survival analysis, when comparing treatments

- A common question is whether a combination treatment (e.g. surgery followed by radiation therapy) is preferable to the single treatment (e.g. surgery alone).
- Survival time is usually measured from date of diagnosis, date of first hospital admission, or date of first treatment.
- In order to receive the combination treatment, one must survive a sufficient period after surgery in order to receive the radiation therapy.
- Those who die during, or immediately after, surgery are included in the 'surgery only' group.
- A naive analysis would show that the group receiving combination therapy experience superior survival.

Biases in survival analysis, lead time bias

- The survival time is measured from the start and end of follow-up.
- As such, any factor which affects either the start date, or the end date will also affect the survival estimation.
- It is possible, for example, to increase patient survival time by bringing forward the date of diagnosis without altering the date of death.
- The implementation of a mass-screening program leads to cancers being detected earlier than they would have been without screening.
- This difference in the time of diagnosis is called the lead time (Figure 8) and can bias the comparison of survival between patient groups, the so-called lead time bias [12].
- The implementation of a mass-screening program is not the only way to introduce lead time bias.

- Increased contact with the health care system for any reason may lead to early clinical diagnosis of a disease, so comparisons of groups that have different health care seeking behaviour could suffer from lead time bias.

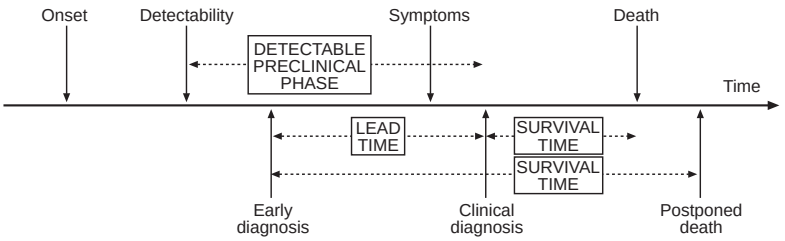


Figure 8: Natural history of chronic illnesses.

Summary of Day 4

- Stratified Cox models can be used to adjust for a covariate with non-proportional hazards.
- Right-censored, and left truncated data can be analysed with the tools given in this course, but other methods are needed for left- or interval censoring and right truncated data.
- The assumption of non- informative censoring.
- Competing risks.
- Never condition on the future.
- Flexible parametric models are an alternative to Poisson and Cox regression when interest lies in estimating the baseline.

Key concepts of the course

- Special methods (i.e., survival analysis) are required when the outcome of interest has a time dimension.
- Epidemiological cohort studies can (and should) be analysed in the framework of survival analysis.
- Survival data often include observations with censored survival times, most commonly right-censoring.
- 'Time' may be a confounder, a mediator or an effect modifier.
- The outcome can be presented as a survival proportion or an event rate. The two measures are mathematically related.
- When comparing groups, HRs are often presented and estimated within a modelling framework.

- Cox regression and Poisson regression are very similar.
- The methods presented assume non-informative censoring.
- Most methods assume proportional hazards, byt this assumption can often be relaxed.
- Reinforcing key concepts in statistical modelling of epidemiological data
  - Studying confounding and effect modification in a modelling framework
  - Reparameterising a statistical model to estimate interaction effects[5]

[5] In this course, we tend to use "effect modification" and "interaction" synonymously.

Exercises for Day 4

125. Estimating the effect of a time-varying exposure – the bereavement data
181. Estimating SMRs
130. Understanding splines
131. Model cause-specific survival using flexible parametric models
132. Flexible parametric models with time-dependent effects
140. Probability of death in a competing risks framework

Appendix 1 Day 4: An introduction to likelihood inference and the Cox likelihood

- The aim of statistical inference is to estimate population parameters of interest from observed data. For example,
- Estimate the hazard ratio for exposed/unexposed in a study of cancer patient survival. The parameter of interest is the log hazard ratio β in the population from which the sample is drawn. This parameter is estimated using the sample of patients observed.
- Similar with logistic regression (the parameter of interest is the log odds ratio β).
- Estimate the recombination fraction, θ, in parametric linkage analysis from the observed pedigrees (marker genotypes and phenotypes).
- A simple example: Imagine we are interested in estimating the proportion, p, that a toss of a coin will result in heads.

- We toss the coin 10 times and observe 4 heads.
- We wish to estimate the parameter of interest, p, from the observed data (the 10 tosses of the coin). Issues of interest are
  - What is the most likely value for p?
  - What is a range of likely values for p?
  - Is $p = 0.5$ a plausible value?
- The likelihood approach is to calculate the probability of observing the observed data, given the probability model, for all possible values of the parameter(s) of interest and choosing the values of the parameter(s) that make the data most likely.
- That is, for what value of p is the probability of tossing 4/10 heads most likely?
- We will calculate the probability of observing 4 heads in 10 tosses for a range of possible values of p.

- If the true value is $p = 0$, what is the probability of observing 4 heads in 10 tosses?
- That one was easy (the probability is zero), but what if $p = 0.1$?
- If $p = 0.1$ then the number of observed heads can theoretically be any integer between 0 and 10 and the probability of each is described by the binomial distribution.
- Recall that if X is a random variable described by a binomial distribution with parameters n and p then the probability distribution of X is given by

$$\Pr(X = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}, \text{ for } r = 0, 1, 2, 3, \ldots n.$$

- $\Pr(X = r)$ is the probability of obtaining r 'successes' (e.g., toss heads) in a sample of size n where the true proportion is p.

- For $p = 0.1$ and $n = 10$ the probability of observing each of the possible outcomes is as follows.

r	Prob(r heads)
0	0.35
1	0.39
2	0.19
3	0.06
4	0.01
5	0.00
6	0.00
7	0.00
8	0.00
9	0.00
10	0.00
Σ	1.00

Binomial distribution with $n = 10$ for various values of p

	Assumed value of p										
r	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
0	1.00	0.35	0.11	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.39	0.27	0.12	0.04	0.01	0.00	0.00	0.00	0.00	0.00
2	0.00	0.19	0.30	0.23	0.12	0.04	0.01	0.00	0.00	0.00	0.00
3	0.00	0.06	0.20	0.27	0.21	0.12	0.04	0.01	0.00	0.00	0.00
4	0.00	0.01	0.09	0.20	0.25	0.21	0.11	0.04	0.01	0.00	0.00
5	0.00	0.00	0.03	0.10	0.20	0.25	0.20	0.10	0.03	0.00	0.00
6	0.00	0.00	0.01	0.04	0.11	0.21	0.25	0.20	0.09	0.01	0.00
7	0.00	0.00	0.00	0.01	0.04	0.12	0.21	0.27	0.20	0.06	0.00
8	0.00	0.00	0.00	0.00	0.01	0.04	0.12	0.23	0.30	0.19	0.00
9	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.12	0.27	0.39	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.11	0.35	1.00
Σ	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

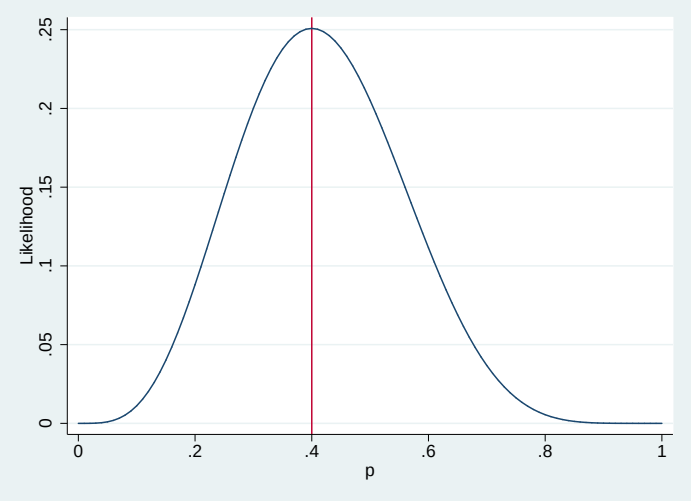
'Likelihood' for a range of values of p

p	Prob($r = 4$)
0.00	0.00
0.10	0.01
0.20	0.09
0.30	0.20
0.40	0.25
0.50	0.21
0.60	0.11
0.70	0.04
0.80	0.01
0.90	0.00
1.00	0.00

- This is the likelihood function[6]. The value of p for which the likelihood is greatest is $p = 0.4$. This is called the maximum likelihood estimate.

[6] In practice, we would exclude the constant $\frac{10!}{4!(10-4)!}$ from the likelihood function.

Plot of the binomial likelihood



What are other likely values for p

- We can see that $p = 0.5$ is also quite likely. The probability of the data is 0.21 when $p = 0.5$ compared to a probability of 0.25 when $p = 0.4$ (the MLE).
- We can test whether $p = 0.5$ is a likely value by studying the ratio of the likelihoods.

$$L(0.5)/L(0.4) = 0.21/0.25 = 0.8176$$

- A result in mathematical statistics tells us that, if the true value of p was 0.5, then minus twice the log likelihood ratio will have a chi square distribution with 1 degree of freedom.

$$-2\ln[L(0.5)/L(0.4)] = -2[l(0.5) - l(0.4)] = 0.40279$$

where l is the log likelihood (the natural logarithm of the likelihood).


```
. di chi2tail(1,0.403)
.52554398
```

- We see that, if the true value of p was 0.5, then we would observe a test statistic at least as large as that we observed 53% of the time. That is, we cannot reject the hypothesis that the true value of p is 0.5.

Mathematically

- We wish to find the value of p that maximises the likelihood function

$$L(p) = \frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}, \text{ for } r = 0, 1, 2, 3, \ldots n.$$

- It is generally easier to maximise the log likelihood (the maximum will occur at the same value). Ignoring the constant,

$$l(p) = \ln[L(p)] = r\ln(p) + (n-r)\ln(1-p).$$

- The derivative of $l(p)$ wrt p is $l'(p) = r/p - (n-r)/(1-p)$.
- The maximum value of $l(p)$ will occur when $l'(p) = 0$ which $\hat{p} = r/n$.

Likelihood calculations for the Cox model

- Estimation is based on the concept of *risk sets*. Understanding this is central to understanding risk set sampling (e.g., nested case-control and case-cohort studies).
- The risk set at each failure time is the collection of subjects who were at risk of failing at that time.
- In theory, only one individual can fail at each failure time and we can calculate the conditional probability of failure for the subject who actually failed.
- The partial likelihood function is the product of these conditional probabilities.
- Imagine 5 individuals at risk at time t of which one fails.
- These individuals have hazards $\lambda_1, \lambda_2, \ldots, \lambda_5$ which may be different since the individuals have different covariate values.

- Conditional on one of the five failing, the probability it is number 2 is

$$\frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5}$$

- Since $\lambda(t) = \lambda_0(t)\exp(x\beta)$ we can write this as

$$\frac{\lambda_0(t)\exp(x_2\beta)}{\lambda_0(t)\exp(x_1\beta) + \lambda_0(t)\exp(x_2\beta) + \ldots + \lambda_0(t)\exp(x_5\beta)}$$

- The baseline hazard, $\lambda_0(t)$, cancels and we have

$$\frac{\exp(x_2\beta)}{\sum_{i\in R}\exp(x_i\beta)}$$

  where R represents the risk set.

- The likelihood function is the product of these conditional probabilities.

- If we have k distinct failure times then

$$L(\beta) = \prod_{j=1}^{k} \left[\frac{\exp(x_j\beta)}{\sum_{i \in R_j} \exp(x_i\beta)} \right] \tag{15}$$

- Note that these calculations do not depend on the underlying failure times; only the ordering of failure times is important.
- Although this is not a likelihood in the strict sense, it is a partial likelihood, it can for all intents and purposes be treated as a likelihood.
- In practice we often observe multiple failures at the same time (ties) and need to use an approximation to equation 15.
- Conceptually similar to a matched (on time) case-control study. Cox partial likelihood is similar to the likelihood for conditional logistic regression (used for analysing matched case-control studies).

References

[1] Paul D. Allison. *Survival Analysis Using the SAS System: A Practical Guide*. Cary, NC: SAS Institute Inc., 1996.

[2] D. G. Altman. *Practical Statistics for Medical Research*. London: Chapman and Hall, 1991.

[3] P. E. Böhmer. Theorie der unabhängigen Wahrscheinlichkeiten. *Rapports, Mémoires et Procès-verbaux de Septième Congrès International d'Actuaires, Amsterdam*, 2:327–343, 1912.

[4] N. E. Breslow and N. E. Day. *Statistical Methods in Cancer Research: Volume II - The Design and Analysis of Cohort Studies*. IARC Scientific Publications No. 82. Lyon: IARC, 1987.

[5] N. E. Breslow, J. H. Lubin, and B. Langholz. Multiplicative models and cohort analysis. *Journal of the American Statistical Association*, 78, 1983.

[6] Alan Cantor. *SAS Survival Analysis Techniques for Medical Research*. BBU Press, second edition, 2003.

[7] Yin Bun Cheung, Fei Gao, and Kei Siong Khoo. Age at diagnosis and the choice of survival analysis methods in cancer epidemiology. *J Clin Epidemiol*, 56(1):38–43, Jan 2003.

[8] D. Clayton and M. Hills. *Statistical Models in Epidemiology*. Oxford: Oxford University Press, 1993.

[9] M. A. Cleves, W. W. Gould, R. G. Gutierrez, and Y. Marchenko. *An Introduction to Survival Analysis Using Stata*. Stata Press, third edition, 2010.

[10] D. R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

[11] KA Cronin and EJ Feuer. Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. *Stat Med*, 19(13):1729–40, Jul 2000.

[12] N. E. Day. The assessment of lead time and length bias in the evaluation of screening programmes. *Maturitas*, 7:51–58, 1985.

[13] F. Ederer, L. M. Axtell, and S. J. Cutler. The relative survival rate: A statistical methodology. *National Cancer Institute Monograph*, 6:101–121, 1961.

[14] Regina C. Elandt-Johnson. Definition of rates: Some remarks on their use and misuse. *American Journal of Epidemiology*, 102:267–271, 1975.

[15] L. D. Fisher and D. Y. Lin. Time-dependent covariates in the cox proportional-hazards regression model. *Annu Rev Public Health*, 20:145–57, 1999.

[16] M. Greenwood. *The Errors of Sampling of the Survivorship Table*, volume 33 of *Reports on Public Health and Medical Subjects*. London: Her Majesty's Stationery Office, 1926.

[17] Miguel A. Hernán. The hazards of hazard ratios. *Epidemiology*, 21(1):13–15, Jan 2010.

[18] Kenneth R. Hess. Graphical methods for assessing vioations of the proportional hazards assumption in Cox regression. *Statistics in Medicine*, 14:1707–1723, 1995.

[19] E. L. Kaplan and P Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[20] John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, 1997.

[21] Britta Krynitz, Gustaf Edgren, Bernt Lindelöf, Eva Baecklund, Christina Brattström, Henryk Wilczek, and Karin E. Smedby. Risk of skin cancer and other malignancies in kidney, liver, heart and lung transplant recipients 1970 to 2008–a swedish population-based study. *Int J Cancer*, 132(6):1429–1438, Mar 2013.

[22] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. London: Chapman and Hall, 2 edition, 1989.

[23] T. M. Therneau and P. M. Grambsch. *Modelling Survival Data: Extending the Cox Model*. Springer: New York, 2000.

[24] Anne C M Thiébaut and Jacques Bénichou. Choice of time-scale in cox's model analysis of epidemiologic cohort data: a simulation study. *Stat Med*, 23(24):3803–3820, Dec 2004.

[25] R. A. Wolfe and R. L. Strawderman. Logical and statistical fallacies in the use of cox regression models. *Am J Kidney Dis*, 27:124–9, 1996.