Biostatistics III: Survival analysis for epidemiologists
Computing notes and exercises

Paul W. Dickman, Sandra Eloranta, Therese Andersson, Caroline Weibull, Anna Johansson,
Hannah Bower and Mark Clements
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden

`http://biostat3.net`

Contents

1 Notes on survival analysis using Stata

A general introduction to Stata (`stataintro.pdf`) can be downloaded from:
`http://biostat3.net/download/`.

If you are not familiar with Stata you should start by downloading and reading this introduction. The same document includes an extensive description of the `stset` command that is central to survival analysis.

In order to analyse survival data it is necessary to specify (at a minimum) a variable representing the time at risk (e.g., 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 time at risk we may instead specify the entry and exit dates.

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. The variable `status` takes the values 0=alive, 1=dead due to cancer, and 2=dead due to other causes. We have specified that only `status=1` indicates an event (death due to melanoma) so Stata will consider observations with other values of `status` as being censored. 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
sts         Generate, graph, list, and test the survivor and cumulative
                hazard functions
strate      Calculate person-time at risk and failure rates
stcox       Estimate Cox proportional hazards model
streg       Estimate parametric survival models
strs        Life table estimation of relative survival
```

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 the estimated 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
```

2 Downloading user-written Stata commands and data files

Stata will be used throughout the course. This section describes how to download and install the files required for the computing exercises (e.g., data files) as well as how to install user-written commands for extending Stata. If you are working in a computer lab during a course it's possible these files may have already been installed for you.

2.1 Downloading the course files

It is suggested that you create a new folder where you can put all course files, e.g. `c:\survival\`. The course files are available on the web as a ZIP archive.

Save and extract this folder in the course folder you created. Within Stata, you can change the Stata working directory to the new directory (e.g., `cd c:\survival\`). You can at any point use the `pwd` command to confirm you are in the working directory you wish to use for the course.

2.2 Installing Stata user-written commands

For some exercises you will need to install user-written commands. For each exercise where this is needed we have added a note stating that a Stata addon is required. You can get back to this section when you get to the first exercise requiring a Stata addon. If you have reached the first exercise where you need to download user-written commands you can read further, otherwise skip the rest of this section on user-written commands for now. Download and installation of user-written commands is done within Stata. It is recommended that you change the Stata working directory to the course directory (e.g., `cd c:\survival\`) before issuing these commands.

2.2.1 How can I check if these commands are already installed?

You can use the `which` command to check if (and where) a Stata command is installed.

```
. which stpm2
c:\ado\plus\s\stpm2.ado
*! version 1.5.0 29Jul2014
```

Use the `adoupdate` command to update previously installed user-written commands (note that this is distinct from the `update` command that updates official Stata commands). Simply type `adoupdate, update` to update all user-written commands.

2.2.2 stpm2 - flexible parametric models

The `stpm2` command, written by Paul Lambert and Patrick Royston, fits flexible parametric survival models (so called Royston-Parmar models). It is installed from within Stata using the following commands:

```
ssc install stpm2
ssc install rcsgen
```

`rcsgen` is a command for generating basis vectors for restricted cubic splines and is required by `stpm2`.

2.2.3 Estimating probability of death in a competing risks framework

The `stcompet` command estimates the cumulative incidence function (CIF) non-parametrically. The `stcompadj` command estimates the CIF using a competing risks analogue of the Cox model. The `stpm2cm` command estimates the crude probabilities of death (i.e. CIF) after fitting a relative survival model using `stpm2`. The `stpm2cif` command estimates the CIF through postestimation after fitting a cause-specific competing risks model using `stpm2`.

```
ssc install stcompet
ssc install stcompadj
ssc install stpm2cm
ssc install stpm2cif
```

3 Exercises

100. **Hand calculation: Life table and Kaplan-Meier estimates of survival**
Using hand calculation (i.e., using a spreadsheet program or pen, paper, and a calculator) estimate the cause-specific survivor function for the sample of 35 patients diagnosed with colon carcinoma (see the table below) using both the Kaplan-Meier method (up to at least 30 months) and the actuarial method (at least the first 5 annual intervals).

In the lectures we estimated the observed survivor function (i.e. all deaths were considered to be events) using the Kaplan-Meier and actuarial methods; your task is to estimate the cause-specific survivor function (only deaths due to colon carcinoma are considered events) using the same data. The next page includes some hints to help you get started.

		Age	Clinical	dx date	Surv. time		
ID	Sex	at dx	stage	mmyy	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

ACTUARIAL APPROACH

We suggest you start with the actuarial approach. Your task is to construct a life table with the following structure.

year of fo-up	l	d	w	l'	p	$S(t)$
[0-1)	35					
[1-2)						
[2-3)						
[3-4)						
[4-5)						
[5-6)						

We have already entered l_1 (number of people alive at the start of interval 1). The next step is to add the number who experienced the event (d) and the number censored (w) during the first year. From l, d, and w you will then be able to calculate l' (effective number at risk), followed by p (conditional probability of surviving the interval) and finally $S(t)$, the cumulative probability of surviving from time zero until the end of the interval.

KAPLAN-MEIER APPROACH

To estimate survival using the Kaplan-Meier approach you will find it easiest to add a line to the table at each and every time there is an event or censoring. We should use time in months. The first time at which there is an event or censoring is time equal to 2 months. The trick is what to do when there are both events and censorings at the same time.

time	# at risk	d	w	p	$S(t)$
2	35				

101. **Using Stata to validate the hand calculations done in question 100**

We will now use Stata to reproduce the same analyses done by hand calculation in question 100 although you can do this part without having done the hand calculations, since this question also serves as an introduction to survival analysis using Stata. Our aim is to estimate the cause-specific survivor function for the sample of 35 patients diagnosed with colon carcinoma using both the Kaplan-Meier method and the actuarial method. In the lectures we estimated the all-cause survivor function (i.e. all deaths were considered to be events) using the Kaplan-Meier and actuarial methods whereas we will now estimate the cause-specific survivor function (only deaths due to colon carcinoma are considered events).

After starting Stata, you will first have to specify the data set you wish to analyse, that is

```
. use colon_sample, clear
```

Stata will search for this file in the current working directory. The `pwd` command will return the name of the current working directory. If you need to change to another directory you can use, for example, `cd c:\survival\`. The `describe` command will return a summary of the data set structure (e.g., variable names) whereas the `list` command will display the values of variables.

In order to use the Stata `ltable` command (life table estimates of the survivor function) we must construct a new variable indicating whether the observation period ended with an event (the new variable is assigned code 1) or censoring (the new variable is assigned code 0). We will call this new variable `csr_fail` (cause-specific failure). The `ltable` command is not a standard Stata survival analysis (`st`) command and does not require that the data be `stset`.

```
. recode status (1=1) (nonmissing=0), gen(csr_fail)
```

There are many ways to create the new variable, the above approach is preferred because missing values of status will remain missing. Even though we don't have any missing values, it is good programming practice to always write code that will handle missing values appropriately.

The following command will give the actuarial estimates

```
. ltable surv_yy csr_fail
```

Alternatively, we could use

```
. ltable surv_mm csr_fail, interval(12)
```

Before most Stata survival analysis commands can be used (`ltable` is an exception) we must first `stset` the data using the `stset` command (see Section 1).

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

A listing of the Kaplan-Meier estimates is then obtained as follows

```
. sts list
```

To graph the Kaplan-Meier estimates

```
. sts graph
```

Note that we only have to `stset` the data once. You can also tell Stata to show the number at risk either on the curve or in a table.

```
. sts graph, atrisk
. sts graph, risktable
```

Titles and axis labels can also be specified.

```
. sts graph, risktable ///
       title(Kaplan-Meier estimates of cause-specific survival) ///
       xtitle(Time since diagnosis in months)
```

103. Melanoma: Comparing survival proportions and mortality rates by stage for cause-specific and all-cause survival

The purpose of this exercise is to study survival of the patients using two alternative measures - survival proportions and mortality rates. A second purpose is to study the difference between cause-specific and all-cause survival.

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

(a) Plot estimates of the survivor function and hazard function by stage.

```
. sts graph, by(stage)
. sts graph, hazard by(stage)
```

By default, the `sts graph` command plots Kaplan-Meier estimates of survival. If we add the `hazard` option it shows estimates of the hazard function. Does it appear that stage is associated with patient survival?

Stata tip: You may have found that each time you produce a graph Stata overwrites the previous graph in the graph window. You can instruct Stata to open each graph in a separate window by naming the graphs. This will give you the possibility to compare graphs side by side.

```
. sts graph, by(stage) name(survival)
. sts graph, by(stage) name(hazard) hazard
```

You can use `set autotabgraphs` to control whether multiple graphs are created as tabs within one window or as separate windows. Issue the following command to make Stata present graphs as tabs within a single window (and store the setting permanently).

```
set autotabgraphs on, permanently
```

(b) Estimate the mortality rates for each stage using, for example, the `strate` command.

```
. strate stage
```

What are the units of the estimated rates?

[The `strate` command, as the name suggests, is used to estimates rates. Look at the help pages if you are not familiar with the command.]

(c) If you haven't already done so, estimate the mortality rates for each stage per 1000 person-years of follow-up.

[HINT: consider the `scale()` option to `stset` and the `per()` option to `strate`.]

(d) Study whether survival is different for males and females (both by plotting the survivor function and by tabulating mortality rates).

```
. sts graph, by(sex)
. sts graph, hazard by(sex)
```

Is there a difference in survival between males and females? If yes, is the difference present throughout the follow up?

(e) The plots you made above were based on cause-specific survival (i.e., only deaths due to cancer are counted as events, deaths due to other causes are censored). In the next part of this question we will estimate all-cause survival (i.e., any death is counted as an event). First, however, study the coding of vital status and tabulate vital status by age group.

How many patients die of each cause? Does the distribution of cause of death depend on age?

```
. codebook status
. tab status agegrp
```

(f) To get all-cause survival, specify all deaths (both cancer and other) as events in the stset command.

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

Now plot the survivor proportion for all-cause survival by stage. We name the graph to be able to separate them in the graph window. Is the survivor proportion different compared to the cause-specific survival you estimated above? Why?

```
. sts graph, by(stage) name(anydeath, replace)
```

(g) It is more common to die from a cause other than cancer in older ages. How does this impact the survivor proportion for different stages? Compare cause-specific and all-cause survival by plotting the survivor proportion by stage for the oldest age group (75+ years) for both cause-specific and all-cause survival. We suggest you copy the code from the PDF file into the Stata do editor and run the code from there.

```
. stset surv_mm, failure(status==1)
. sts graph if agegrp==3, by(stage) ///
      name(cancerdeath_75, replace) subtitle("Cancer")
. stset surv_mm, failure(status==1,2)
. sts graph if agegrp==3, by(stage) ///
      name(anydeath_75, replace) subtitle("All cause")
. graph combine cancerdeath_75 anydeath_75
```

(h) Now estimate both cancer-specific and all-cause survival for each age group.

```
. use melanoma, clear
. stset surv_mm, failure(status==1,2)
. sts graph, by(agegrp) name(anydeathbyage, replace) subtitle("All cause")

. stset surv_mm, failure(status==1)
. sts graph, by(agegrp) name(cancerdeathbyage, replace) subtitle("Cancer")

. graph combine anydeathbyage cancerdeathbyage
```

Are there bigger differences between the age groups for cause-specific or for all-cause survival?

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

We will now analyse the full data set of patients diagnosed with localised skin melanoma.

Use Stata to estimate the cause-specific survivor function, using the Kaplan-Meier method with survival time in months, separately for each of the two calendar periods 1975–1984 and 1985–1994. The following commands can be used

```
. use melanoma if stage == 1, clear
. stset surv_mm, failure(status==1)
. sts graph, by(year8594)
```

The variable `year8594` takes the value 1 for patients diagnosed 1985–1994 and 0 for those diagnosed 1975–1984.

(a) Without making reference to any formal statistical tests, does it appear that patient survival is superior during the most recent period?

(b) The following commands can be used to plot the hazard function (instantaneous mortality rate):

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

i. At what point in the follow-up is mortality highest?

ii. Does this pattern seem reasonable from a clinicial/biological perspective? [HINT: Consider the disease with which these patients were classified as being diagnosed along with the expected fatality of the disease as a function of time since diagnosis.]

(c) Use the log rank test to determine whether there is a statistically significant difference in patient survival between the two periods. The following command can be used:

```
. sts test year8594
```

What do you conclude?

An alternative test is the generalised Wilcoxon, which can be obtained as follows

```
. sts test year8594, wilcoxon
```

Haven't heard of the log rank (or Wilcoxon) test? It's possible you may reach this exercise before we cover the details of these tests during lectures. You should nevertheless do the exercise and try and interpret the results. Both of these tests (the log rank and the generalised Wilcoxon) are used to test for differences between the survivor functions. The null hypothesis is that the survivor functions are equivalent for the two calendar periods (i.e., patient survival does not depend on calendar period of diagnosis).

(d) Estimate cause-specific mortality rates for each age group, and graph Kaplan-Meier estimates of the cause-specific survivor function for each age group. Are there differences between the age groups? Is the interpretation consistent between the mortality rates and the survival proportions?

```
. strate agegrp, per(1000)
. sts graph, by(agegrp)
```

What are the units of the estimated hazard rates? HINT: look at how you defined time when you `stset` the data.

(e) Repeat some of the previous analyses after using the `scale()` option to stset to rescale time from months to years. This is equivalent to dividing the time variable by 12 so all analyses will be the same except the units of time will be different (e.g., the graphs will have different labels).

```
. stset surv_mm, failure(status==1) scale(12)
. sts graph, by(agegrp)
. strate agegrp, per(1000)
```

(f) Study whether there is evidence of a difference in patient survival between males and females. Estimate both the hazard and survival function and use the log rank test to test for a difference.

110. Diet data: tabulating incidence rates and modelling with Poisson regression

Load the `diet` data and `stset` the data using time-on-study as the timescale.

```
. use diet, clear
. stset dox, id(id) fail(chd) origin(doe) scale(365.24) enter(doe)
```

(a) Use the `strate` command to tabulate CHD incidence rates per 1000 person-years for each category of `hieng`. Calculate (by hand) the ratio of the two incidence rates.

(b) Use the command `poisson` to find the incidence rate ratio for the high energy group compared to the low energy group and compare the estimate to the one you obtained in the previous question:

```
. poisson chd hieng, e(y) irr
```

NOTE: Rates are calculated as events/person-time so when modelling rates we need to let Stata know both of these quantities. `chd` is the event indicator and `y` is the person time at risk for each individual. The `irr` option results in the estimates being presented as estimated incidence rate ratios rather than parameter estimates (log incidence rate ratios).

(c) Write out the model formulation of the poisson model in 110b .

(d) Grouping the values of total energy into just two groups does not tell us much about how the CHD rate changes with total energy. It is a useful exploratory device, but to look more closely we need to group the total energy into perhaps 3 or 4 groups. In this example we shall use the cut points 1500, 2500, 3000, 4500. To check if these cutpoints seem reasonable, type:

```
. histogram energy, normal
. sum energy, detail
```

(e) Use the commands

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

to create a new variable `eng3` coded 1500 for values of `energy` in the range 1500–2499, 2500 for values in the range 2500–2999, and 3000 for values in the range 3000–4500.

(f) To estimate and plot the rates for different levels of `eng3` try

```
. strate eng3, per(1000) graph
```

Calculate (by hand) the ratio of rates in the second and third levels to the first level.

(g) Create your own indicator variables for the three levels of `eng3` with

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

(h) Check the indicator variables with

```
. list energy eng3 X1 X2 X3 if eng3==1500
. list energy eng3 X1 X2 X3 if eng3==2500
. list energy eng3 X1 X2 X3 if eng3==3000
```

(i) Use `poisson` to compare the second and third levels with the first, as follows:

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

Compare your estimates with those you obtained in 110f.

(j) Write out the model formulation of the poisson model in 110i .

(k) Use `poisson` to compare the first and third levels with the second. Write out the model formulation for this poisson model.

(l) Repeat the analysis comparing the second and third levels with the first but this time have Stata create the indicators automatically via the `i.` syntax. That is

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

(m) Without using `st` commands, calculate the total number of events during follow-up, person-time at risk, and the crude incidence rate (per 1000 person-years), for example with Stata commands for descriptive statistics (e.g., summarize). Confirm your answer using `strate` or `stptime`.

[HINT: Remember that the total number of person-years is the number of persons at risk multiplied with the mean follow up time among those persons.]

111. Localised melanoma: model cause-specific mortality with Poisson regression

In this exercise we model, using Poisson regression, cause-specific mortality of patients diagnosed with localised (`stage==1`) melanoma.

In exercise 120 we model cause-specific mortality using Cox regression and in exercise 131 we use flexible parametric models. The aim is to illustrate that these methods are very similar.

The aim of these exercises is to explore the similarities and differences between these approaches to modelling.

The following commands can be used to load and `stset` the data.

```
. use melanoma, clear
. keep if stage==1
. stset surv_mm, failure(status==1) scale(12) id(id)
```

(a) Plot Kaplan-Meier estimates of cause-specific survival as a function of calendar period of diagnosis.

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

i. During which calendar period (the early or the latter) is survival best?

ii. Now plot the estimated hazard function (cause-specific mortality rate) as a function of calendar period of diagnosis.

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

During which calendar period (the early or the latter) is mortality the lowest?

iii. Is the interpretation (with respect to how prognosis depends on period) based on the hazard consistent with the interpretation of the survival plot?

(b) Use the `strate` command to estimate the cause-specific mortality rate for each calendar period.

```
. strate year8594, per(1000)
```

During which calendar period (the early or the latter) is mortality the lowest? Is this consistent with what you found earlier? If not, why the inconsistency?

(c) The reason for the inconsistency between parts 111a and 111b was confounding by time since diagnosis. The comparison in part 111a was adjusted for time since diagnosis (since we compare the differences between the curves at each point in time) whereas the comparison in part 111b was not. Understanding this concept is central to the remainder of the exercise so please ask for help if you don't follow.

Two approaches for controlling for confounding are 'restriction' and 'statistical adjustment'. We will first use restriction to control for confounding. That is we will `stset` the data again but use the `exit(time 120)` option to restrict the potential follow-up time to a maximum of 120 months. Individuals who survive more than 120 months are censored at 120 months.

```
. use melanoma, clear
. keep if stage==1
. stset surv_mm, failure(status==1) scale(12) id(id) exit(time 120)
```

i. Use the `strate` command to estimate the cause-specific mortality rate for each calendar period.

```
. strate year8594, per(1000)
```

During which calendar period (the early or the latter) is mortality the lowest? Is this consistent with what you found in part 111b?

ii. Calculate by hand the ratio (85–94/75–84) of the two mortality rates (i.e., a mortality rate ratio) and interpret the estimate (i.e., during which period is mortality higher/lower and by how much).

iii. Now use Poisson regression to estimate the same mortality rate ratio.

```
. streg i.year8594, dist(exp)
```

iv. Write out the model formulation from 111(c)iii.

NOTE: `streg` is one of several Stata commands for performing Poisson regression. The model could also be fitted using the `poisson` or `glm` commands.

```
. gen risktime=_t-_t0
. poisson _d i.year8594 if _st==1, exp(risktime) irr
. glm _d i.year8594  if _st==1, family(poisson) eform lnoffset(risktime)
```

However, if you have stset/stsplit the data it is recommended that you use `streg` since `streg` understands and respects the internal `st` variables (`_st`, `_t`, `_t0`, and `_d`). In particular, 'trimmed' person-time will be ignored by `streg` but not by the `poisson` command.

Strictly speaking, `streg` fits parametric survival models. A parametric survival model assuming survival times are exponentially distributed (`dist(exp)`) implies a constant hazard and a Poisson process for the number of events (i.e., Poisson regression).

(d) In order to adjust for time since diagnosis (i.e., adjust for the fact that we expect mortality to depend on time since diagnosis) we need to split the data by this timescale. We will restrict our analysis to mortality up to 10 years following diagnosis.

```
. stsplit fu, at(0(1)10) trim
```

NOTE: The `trim` option instructs Stata to ignore time-at-risk outside the interval [0,10] (i.e., after 10 years subsequent to diagnosis). Since we have already made this restriction using `stset` there should not be any time 'trimmed'.

(e) Now tabulate (and produce a graph of) the rates by follow-up time.

```
. strate fu, per(1000) graph
```

Mortality appears to be quite low during the first year of follow-up. Does this seem reasonable considering the disease with which these patients have been diagnosed?

(f) Compare the plot of the estimated rates to a plot of the hazard rate as a function of continuous time.

```
. sts graph, hazard
```

Is the interpretation similar? Do you think it is sufficient to classify follow-up time into annual intervals or might it be preferable to use, for example, narrower intervals?

(g) Use Poisson regression to estimate incidence rate ratios as a function of follow-up time.

```
. streg i.fu, dist(exp)
```

Does the pattern of estimated incident rate ratios mirror the pattern you observed in the plots?

i. Write out the model formulation.

(h) Now estimate the effect of calendar period of diagnosis while adjusting for time since diagnosis. Before fitting this model, predict what you expect the estimated effect to be (i.e., will it be higher, lower, or similar to the value of 0.8831852 we obtained in part 111c.

```
. streg i.fu i.year8594, dist(exp)
```

Is the estimated effect of calendar period of diagnosis consistent with what you expected?

(i) Now control for age, sex, and calendar period.

```
. streg i.fu i.agegrp i.year8594 i.sex, dist(exp)
```

i. Interpret the estimated hazard ratio for the parameter labelled `agegrp 60-74`, including a comment on statistical significance.

ii. Is the effect of calendar period strongly confounded by age and sex? That is, does the inclusion of sex and age in the model change the estimate for the effect of calendar period?

iii. Perform a Wald test of the overall effect of age and interpret the results.

```
. test 1.agegrp 2.agegrp 3.agegrp
```

(j) Is the effect of sex modified by calendar period (whilst adjusting for age and follow-up)? Fit an appropriate interaction term to test this hypothesis and write out the model formulation for the model that you fit.

(k) Based on the interaction model you fitted in exercise 111j, estimate the hazard ratio for the effect of sex (with 95% confidence interval) for each calendar period.

ADVANCED: Do this with each of the following methods and confirm that the results are the same:

i. Using hand-calculation on the estimates from exercise 111j.

ii. Using the estimates from exercise 111j and the `lincom` command.

```
. lincom 2.sex + 1.year8594#2.sex, eform
```

iii. Creating appropriate dummy variables that represent the effects of sex for each calendar period.

```
. gen sex_early=(sex==2)*(year8594==0)
. gen sex_latter=(sex==2)*(year8594==1)
. streg i.fu i.agegrp i.year8594 sex_early sex_latter, dist(exp)
```

iv. Using Stata 11 syntax to repeat the previous model.

```
. streg i.fu i.agegrp i.year8594 i.year8594#i.sex, dist(exp)
```

(l) Now fit a separate model for each calendar period in order to estimate the hazard ratio for the effect of sex (with 95% confidence interval) for each calendar period. Why do the estimates differ from those you obtained in the previous part?

```
. streg i.fu i.agegrp i.sex if year8594==0, dist(exp)
. streg i.fu i.agegrp i.sex if year8594==1, dist(exp)
```

Can you fit a single model that reproduces the estimates you obtained from the stratified models? Try:

```
. streg i.fu##i.year8594 i.agegrp##i.year8594 i.year8594##i.sex, dist(exp)
```

112. **Diet data: Using Poisson regression to study the effect of energy intake adjusting for confounders on two different timescales**

Use Poisson regression to study the association between energy intake (`hieng`) and CHD adjusted for potential confounders (job, BMI). We know that people who expend a lot of energy (i.e., are physically active) require a higher energy intake. We do not have data on physical activity but we are hoping that occupation (job) will serve as a surrogate measure of work-time physical activity (conductors on London double-decker busses expend energy walking up and down the stairs all day).

Fit models both without adjusting for 'time' and by adjusting for attained age (you will need to split the data) and time-since-entry and compare the results.

(a) Rates can be modelled on different timescales, e.g., attained age, time-since-entry, calendar time. Plot the CHD incidence rates both by attained age and by time-since-entry. Is there a difference? Do the same for CHD hazard by different energy intakes (hieng).

```
. use diet, clear

.* Timescale: Attained age
. stset dox, id(id) fail(chd) origin(dob) enter(doe) scale(365.24)
. sts graph, hazard
. sts graph, by(hieng) hazard

.* Timescale: Time-since-entry
. stset dox, id(id) fail(chd) origin(doe) enter(doe) scale(365.24)
. sts graph, hazard
. sts graph, by(hieng) hazard
```

(b) Model the rate using Poisson regression, without adjusting for any timescale. What is the effect of hieng on CHD? What assumption does this model make on the shape of the underlying incidence rate over time?

```
. poisson chd i.hieng, e(y) irr
```

(c) Adjust for BMI and job. Is there evidence that the effect of energy intake on CHD is confounded by BMI and job? Write out the model formulation.

```
. gen bmi=weight/(height/100*height/100)
. poisson chd i.hieng i.job bmi, e(y) irr
```

(d) Firstly, let's adjust for the timescale attained age. To do this in Poisson regression you must split the data on timescale age. First use `stset` (with origin date of birth) and then use `stsplit` to generate agebands.

```
. stset dox, id(id) fail(chd) origin(dob) enter(doe) scale(365.24)
. stsplit ageband, at(30,50,60,72) trim
. list id _t0 _t ageband y in 1/10
```

As the poisson command is not an `st` command, you must keep track of the risktime yourself. Why is the `y` variable not correct anymore? Generate a new variable, `risktime`, which contains the risktime for each split record.

```
. gen risktime=_t-_t0
. list id _t0 _t ageband y risktime in 1/10
```

You must also keep track of the event variable, as chd will not be valid after the split.

```
. tab ageband chd, missing
. tab ageband _d, missing
```

Now fit the model for CHD, both without and with the adjustment for job and bmi. Is the effect of hieng on CHD confounded by age, BMI or job?

```
. poisson _d i.hieng i.ageband, e(risktime) irr
. poisson _d i.hieng i.job bmi i.ageband, e(risktime) irr
```

What assumption is being made about the shape of the baseline hazard (HINT: the baseline hazard takes the shape of the timescale)?

(e) Secondly, do the same analysis, but now adjust for the timescale time-since-entry. (You must read the data in again, as you now want to split on another timescale. This is strictly not necessary, but to avoid mistakes it is generally a good idea to start over again.)

```
. use diet, clear
. gen bmi=weight/(height/100*height/100)
```

Specify time-since-entry as the timescale by specifying date of entry as the time origin.

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

. stsplit fuband, at(0,5,10,15,22) trim
. list id _t0 _t fuband y in 1/10

. gen risktime=_t-_t0
. list id _t0 _t fuband y risktime in 1/10

. tab fuband chd, missing
. tab fuband _d, missing

. poisson _d i.hieng i.fuband, e(risktime) irr
. poisson _d i.hieng i.job bmi i.fuband, e(risktime) irr
```

Compare the results with the analysis adjusted for attained age. Are there any differences? Why (or why not)? Go back to the graphs at the beginning of the exercise and look for explanations.

(f) Repeat the exercise using streg. What is the advantage/disadvantage of using streg?

120. Localised melanoma: modelling cause-specific mortality using Cox regression

In exercise 111 we modelled the cause-specific mortality of patients diagnosed with localised melanoma using Poisson regression. We will now model cause-specific mortality using Cox regression and compare the results to those we obtained using the Poisson regression model.

We will start with modelling the effect of calendar period. To fit a Cox proportional hazards model (for cause-specific survival) with calendar period as the only explanatory variable, the following commands can be used. Note that we are censoring all survival times at 120 months (10 years) in order to facilitate comparisons with the Poisson regression model in exercise 111.

```
. use melanoma
. keep if stage == 1
. stset surv_mm, failure(status==1) exit(time 120) id(id)
. stcox i.year8594
```

(a) Interpret the estimated hazard ratio, including a comment on statistical significance.

(b) (This part is more theoretical and is not required in order to understand the remaining parts.)

Stata reports a Wald test of the null hypothesis that survival is independent of calendar period. The test statistic (and associated P-value) is reported in the table of parameter estimates (labelled `z`). Under the null hypothesis, the test statistic has a standard normal (Z) distribution, so the square of the test statistic will have a chi square distribution with one degree of freedom.

Stata also reports a likelihood ratio test statistic of the null hypothesis that none of the parameters in the model are associated with survival (labelled `LR chi2(1)`). In general, this test statistic will have a chi-square distribution with degrees of freedom equal to the number of parameters in the model. For the current model, with only one parameter, the test statistic has a chi square distribution with one degree of freedom.

Compare these two test statistics with each other and with the log rank test statistic (which also has a χ^2_1 distribution) calculated in question 104c (you should, however, recalculate the log rank test since we have restricted follow-up to the first 10 years in this exercise). Would you expect these test statistics to be similar? Consider the null and alternative hypotheses of each test and the assumptions involved with each test.

(c) Now include sex and age (in categories) in the model.

```
. stcox i.sex i.year8594 i.agegrp
```

i. Interpret the estimated hazard ratio for the parameter labelled `agegrp 2`, including a comment on statistical significance.

ii. Perform a Wald test of the overall effect of age and interpret the results.

```
. test 1.agegrp 2.agegrp 3.agegrp
```

(d) Perform a likelihood ratio test of the overall effect of age and interpret the results. The following commands can be used

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

Compare your findings to those obtained using the Wald test. Are the findings similar? Would you expect them to be similar?

(e) The model estimated in question 120c is similar to the model estimated in question 111i.

   i. Both models adjust for `sex, year8594,` and `i.agegrp` but the Poisson regression model in question 111i appears to adjust for an additional variable (`i.fu`). Is the Poisson regression model adjusting for an additional factor? Explain.

   ii. Would you expect the parameter estimate for sex, period, and age to be similar for the two models? Are they similar?

   iii. Do both models assume proportional hazards? Explain.

(f) Write out the model formulation for the following models or predicted rates

   i. the Poisson model used in 111i.

   ii. the Cox models used in 120a and 120c. Make sure you understand what each parameter means. What is the intercept in the Poisson regression model? What is the intercept in the Cox models in 120a and 120c?

   iii. using the model from 120c, write down the mathematical expression (linear predictor) for the rate of males diagnosed 1985-1994 in agegroup 2. Do the same for the rate of females diagnosed 1985-1994 in agegroup 2. Using these two rates, write down the mathematical expression for the hazard ratio of females to males diagnosed 1985-1994 in agegroup 2. Comment on the proportional hazard assumption and how that relates to the expression and/or parameter(s) you obtained for the hazard ratio.

(g) Following is some code for estimating and comparing the Cox and Poisson regression models.

```
use melanoma if stage==1, clear
stset surv_mm, failure(status==1) id(id) exit(time 120)
stcox i.year8594 i.sex i.agegrp
est store Cox

/* split on time since diagnosis (1-year intervals) */
stsplit fu, at(0(12)120) trim

streg i.fu i.year8594 i.sex i.agegrp, dist(exp)
est store Poisson
est table Cox Poisson, eform equations(1)
```

(h) ADVANCED: By splitting at each failure time we can estimate a Poisson regression model that is identical to the Cox model. This model might take several minutes to estimate and you may need to reset the values of `memory` and `matsize`.

```
use melanoma, clear
keep if stage == 1
stset surv_mm, failure(status==1) exit(time 120) id(id) noshow
stsplit, at(failures) riskset(riskset)
quietly tab riskset, gen(interval)

streg interval* i.sex i.year8594 i.agegrp, dist(exp)
est store Poisson_fine

/* Compare the estimates and SEs */
est table Cox Poisson_fine Poisson, eform equations(1) ///
keep(2.sex 1.year8594 1.agegrp 2.agegrp 3.agegrp) ///
se b(%9.6f) se(%9.6f) modelwidth(12) ///
title("Hazard ratios and standard errors for various models")
```

(i) ADVANCED: Split the data finely (e.g., 3-month intervals) and model the effect of time using a restricted cubic spline.

```
use melanoma if stage==1, clear
stset surv_mm, failure(status==1) id(id) exit(time 120)
/* split on time since diagnosis (1-month intervals) */
stsplit fu, at(0(1)120) trim
/* Create basis for restricted cubic spline */
mkspline fu_rcs=fu, cubic
streg fu_rcs* i.year8594 i.sex i.agegrp, dist(exp)
predict xb, xb
twoway line xb fu if year8594==0 & sex==1 & agegrp==1, sort
```

121. **Examining the proportional hazards hypothesis (localised melanoma)**

(a) For the localised melanoma data with 10 years follow-up, plot the instantaneous cause-specific hazard for each calendar period. The following commands can be used

```
. use melanoma if stage == 1, clear
. stset surv_mm, failure(status==1) id(id) exit(time 120) scale(12)
. sts graph, hazard by(year8594)
```

Make a rough estimate of the hazard ratio for patients diagnosed 1985–94 to those diagnosed 1975–84. In part (d) you will fit a Cox model and check your estimate.

(b) Now plot the instantaneous cause-specific hazard for each calendar period using a log scale for the y axis (use the option `yscale(log)`). What would you expect to see if a proportional hazards assumption were appropriate? Do you see it?

(c) Another graphical way of checking the proportional hazards assumption is to plot the log cumulative cause specific hazard function for each calendar period. The command for plotting this function is

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

What would you expect to see if a proportional hazards assumption were appropriate? Do you see it?

(d) Compare your estimated hazard ratio from part (a) with the one from a fitted Cox model with calendar period as the only explanatory variable. Are they similar?

(e) Now fit a more complex model and use graphical methods to explore the assumption of proportional hazards by calendar period. For example,

```
. stcox i.sex i.year8594 i.agegrp
. estat phtest, plot(1.year8594)
```

What do you conclude?

(f) Do part (a)–(e) but now for the variable `agegrp`. What are your conclusions regarding the assumption of proportional hazards?

(g) Now formally test the assumption of proportional hazards using

```
. stcox i.sex i.year8594 i.agegrp
. estat phtest, detail
```

Are your conclusions from the test coherent with your conclusions from the graphical assessments?

(h) Estimate separate age effects for the first two years of follow-up (and separate estimates for the remainder of the follow-up) while controlling for sex and period. Do the estimates for the effect of age differ between the two periods of follow-up?

There are two ways to fit time-varying effects: 1) the tvc option in stcox or 2) by splitting on time using stsplit.
Using tvc:

```
. tab(agegrp), gen(agegrp)
. stcox i.sex i.year8594 agegrp2 agegrp3 agegrp4, ///
          tvc(agegrp2 agegrp3 agegrp4) texp(_t>=2)
```

Using stsplit:

```
. stsplit fuband, at(0,2)
. list id _t0 _t fu in 1/10

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

These are simply two alternative syntaxes for fitting the same model with the same parameterizations. They give the so-called default parameterizations for interaction effects. We see effects of age (i.e., the hazard ratios) for the period 0–2 years subsequent to diagnosis along with the interaction effects. An advantage of the default parameterisation is that one can easily test the statistical significance of the interaction effects. Before going further, test whether the age*follow-up interaction is statistically significant (using a Wald and/or LR test).

(i) Often we wish to see the effects of exposure (age) for each level of the modifier (time since diagnosis). That is, we would like to complete the table below with relevant hazard ratios. To get the effects of age for the period 2+ years after diagnosis, using the default parametrization, we must multiply the hazard ratios for 0–2 years by the appropriate interaction effect. Now let's reparameterise the model to directly estimate the effects of age for each level of time since diagnosis. This is easily done in Stata (version 11 or later) using single #'s

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

	0–2 years	2+ years
Agegrp0	1.00	1.00
Agegrp1		
Agegrp2		
Agegrp3		

Fill in the table above. Does the effect of age appear different before and after 2 years?

(j) Write down the model formulation for the Cox model used in (h).

i. Use that model formula to express the rates (in terms of the linear predictor) in each of the eight cells of the table in (i). Express the rates for males diagnosed 1975-84.

ii. Use the rates from i. to obtain an expression for the hazard ratio of agegroup3 to agegroup0 in early followup 0-2years, for males diagnosed 1975-84.

iii. Use the rates from i. to obtain an expression for the hazard ratio of agegroup3 to agegroup0 in late followup 2+years, for males diagnosed 1975-84.

(k) ADVANCED: Fit an analogous Poisson regression model. Are the parameter estimates similar? HINT: You will need to split the data by time since diagnosis.

123. **Cox model for cause-specific mortality for melanoma (all stages)**

Use Cox regression to model the cause-specific survival of patients with skin melanoma (including all stages).

(a) First fit the model with sex as the only explanatory variable. Does there appear to be a difference in survival between males and females?

(b) Is the effect of sex confounded or mediated by other factors (e.g. age, stage, subsite, period)? After controlling for potential confounders or mediators, does there still appear to be a difference in survival between males and females?

(c) Consider the hypothesis that there exists a class of melanomas where female sex hormones play a large role in the etiology. These hormone related cancers are diagnosed primarily in women and are, on average, less aggressive (i.e., prognosis is good). If such a hypothesis were true we might expect the effect of sex to be modified by age at diagnosis (e.g., pre versus post menopausal). Test whether this is the case.

(d) Decide on a 'most appropriate' model for these data, given the research question described above. Be sure to evaluate the proportional hazards assumption.

(e) Write down the model formulation for the Cox models used in (a), (b) and (c).

   i. Use the formula for model (c), to express the linear predictor for the rate among females in the highest agegroup while all other variables are at the reference level.

   ii. Write down the expression for the hazard ratio comparing the females in the highest agegroup to the males in the highest agegroup, while all other variables are at the reference level.

124. **Modelling the diet data using Cox regression**

(a) Fit the following Poisson regression model to the diet data (we fitted this same model in question 110).

```
. use diet, clear
. poisson chd i.hieng, e(y) irr
```

Now fit the following Cox model.

```
. stset dox, id(id) fail(chd) enter(doe) origin(doe) scale(365.24)
. stcox i.hieng
```

i. On what scale are we measuring 'time'? That is, what is the timescale?

ii. Is it correct to say that both of these models estimate the effect of high energy on CHD *without controlling for any potential confounders*? If not, how are these models conceptually different?

iii. Would you expect the parameter estimates for these two models to be very different? Is there a large difference?

(b) `stset` the data with attained age as the timescale and refit the Cox model. Is the estimate of the effect of high energy different? Would we expect it to be different?

(c) Write down themodel formulation for the Poisson and Cox models used in (a).

i. Comment on the difference between the Poisson and Cox model, and explain why (or why not) they are different conceptually.

ii. Write down the model formulation for the Cox model used in (b). Comment on how this Cox model differs from the Cox model in (a), in particular the interpretation of the parameters.

125. Estimating the effect of a time-varying exposure – the bereavement data

These data were used to study a possible effect of *marital bereavement* (loss of husband or wife) on all–cause mortality in the elderly. The dataset was extracted from a larger follow-up study of an elderly population and concerns subjects whose husbands or wives were alive at entry to the study. Thus all subjects enter as not bereaved but may become bereaved at some point during follow–up. The variable `dosp` records the date of death of each subject's spouse and takes the value 1/1/2000 where this has not yet happened.

(a) Load the data with

```
. use brv, clear
. desc
```

To see how the coding works for couples try

```
. list id sex doe dosp dox fail if couple==3
```

for a couple, both of whom die during follow–up. Draw a picture showing the follow–up for both subjects, and mark the dates of entry exit and death of spouse on it. Try

```
. list id sex doe dosp dox fail if couple==4
```

for a couple, one of whom dies during follow–up,

```
. list id sex doe dosp dox fail if couple==19
```

for a couple, neither of whom die during follow–up, and

```
. list id sex doe dosp dox fail if couple==7
```

for a couple where only data on one individual is available.

(b) Set the `st` variables, calculate the mortality rate per 1000 years for men and for women, and find the rate ratio comparing women (coded 2) with men (coded 1), using

```
. stset dox, fail(fail) origin(dob) entry(doe) scale(365.24) id(id)
. strate sex, per(1000)
. streg sex, dist(exp)
```

i. What dimension of time did we use as the timescale when we `stset` the data? Do you think this is a sensible choice?

ii. Which gender has the highest mortality? Is this expected?

iii. Could age be a potential confounder? Does age at entry differ between males and females? Later we will estimate the rate ratio while controlling for age.

(c) **Breaking records into pre and post bereavement.** In these data a subject changes exposure status from not bereaved to bereaved when his or her spouse dies. The first stage of the analysis therefore is to partition each follow–up into a record describing the period of follow-up pre–bereavement and (for subjects who were bereaved during the study) the period post–bereavement.

This can be done using `stsplit`:

```
. stsplit brv, after(time=dosp) at(0)
. recode brv -1=0 0=1
```

This syntax of `stsplit` splits the records at the death of spouse (or 1/1/2000 if the spouse is still alive). The variable `brv` takes the values -1 for the pre bereavement part and 0 for the post bereavment part and the `recode` command changes these to 0 and 1 respectively.
To see the effect on couple 3

```
. list id sex doe dosp dox brv _t0 _t _d fail if couple==3
```

We see that, of this couple, only the woman was bereaved during follow-up (it is impossible for both of a couple to contribute person-time to the bereaved category). This woman was classified as 'not bereaved' during age 83.87 and 84.41 and 'bereaved' during ages 84.41 and 84.82. Study the data for the other couples mentioned above.

(d) Now find the (crude) effect of bereavement

```
. streg brv, dist(exp)
```

(e) Since there is a strong possibility that the effect of bereavement is not the same for men as for women, use `streg` to estimate the effect of bereavement separately for men and women. Do this both by fitting separate models for males and females (e.g. `streg brv if sex==1`) as well as by using a single model with an interaction term (you may need to create dummy variables). Confirm that the estimates are identical for these two approaches.

(f) **Controlling for age.** There is strong confounding by age. Use `stsplit` to expand the data by 5 year age–bands, and check that the rate is increasing with age. Use `streg` to find the effect of bereavement controlled for age. If you wish to study the distribution of age then it is useful to know that age at entry and exit are stored in the variables `_t0` and `_t` respectively.

(g) Now estimate the effect of bereavement (controlled for age) separately for each sex.

(h) We have assumed that any effect of bereavement is both *immediate* and *permanent*. This is not realistic and we might wish to improve the analysis by further subdividing the post–bereavement follow–up. How might you do this? (you are not expected to actually do it)

(i) **Analysis using Cox regression.** We can also model these data using Cox regression. Provided we have stset the data with attained age as the time scale and split the data (using `stsplit`) to obtain separate observations for the bereaved and non-bereaved person-time the following command will estimate the effect of bereavement adjusted for attained age.

```
. stcox brv
```

That is, we do not have to split the data by attained age (although we can fit the model to data split by attained age and the results will be the same).

(j) Use the Cox model to estimate the effect of bereavement separately for males and females and compare the estimates to those obtained using Poisson regression.

130. **Melanoma: Understanding splines**

Stata addon required! This exercise requires the Stata user-written command `rcsgen`. See Section 2.2 for details and installation instructions.

This question is for those who want to understand the calculations made when using splines and how the constraints enable a smooth function to be fitted. This will be demonstrated by fitting a Poisson model with no covariates (other than follow-up time). First load the melanoma data with follow-up to 10 years and split the time scale with intervals of one month.

```
. use melanoma
. gen female = sex == 2
. stset surv_mm, failure(status=1,2) scale(12) exit(time 120) id(id)
. stsplit fu, every(‘=1/12’)
. gen risktime = _t - _t0
. collapse (sum) d = _d risktime (min) start=_t0 (max) end=_t, ///
    by(fu female year8594 agegrp)
```

(a) Fit a Poisson model for all cause survival with one parameter for each interval. Predict the hazard function and plot this against follow-up time.

```
. egen interval = group(start)
. gen midtime = (start + end)/2
. glm d ibn.interval, family(poisson) link(log) lnoffset(risktime) nocons

// predict the baseline (one parameter for each interval)
. predict haz_grp, nooffset
. replace haz_grp = haz_grp*1000
. twoway (scatter haz_grp midtime)  ///
        , xtitle("Years from diagnosis") ///
        ytitle("Baseline hazard (1000 pys)") ///
        ylabel(5 10 20 50 100 150, angle(h)) ///
       name(piecewise, replace)
```

How many parameters have been used to estimate the baseline hazard?

(b) We will now use piecewise linear splines. To simplify things we will only have one knot at 1.5 years. We will first fit the equivalent of two separate linear functions, one before the knot and one after the knot. The model is as follows,

$$\ln h(t) = \beta_0 + \beta_1 t + \beta_2 (t > 1.5) + \beta_3 (t - 1.5)_+$$

Note the use of the '+' notation, where $u_+ = u$ if $u > 0$ and 0 otherwise. Write down the functional form of the linear functions before and after the knot at 1.5 years.

Now fit this model and compare the fitted values to the piecewise estimate in part 130a.

```
. gen lin_s1 = midtime
. gen lin_int2 = (midtime>1.5)
. gen lin_s2 = (midtime - 1.5)*(midtime>1.5)
```

```
// Fit two separate linear regression lines (4 parameters)
. glm d lin_s1 lin_int2 lin_s2 , family(poisson) link(log) lnoffset(risktime)

. predict haz_lin1, nooffset
. replace haz_lin1 = haz_lin1*1000
. twoway  (scatter haz_grp midtime)  ///
          (line haz_lin1 midtime if midtime<=1, lcolor(red)) ///
          (line haz_lin1 midtime if midtime>1, lcolor(red)) ///
          , xtitle("Years from diagnosis") ///
          ytitle("Baseline hazard (1000 pys)") ///
          xline(1.5, lcolor(black) lpattern(dash)) ///
          ylabel(5 10 20 50 100 150, angle(h)) ///
          legend(off) ///
          name(linear1, replace)
```

Calculate the intercept and gradient before the knot and after the knot at 1.5 years.

(c) Now we will force the function to be continuous at the knot. This can be done by dropping the second intercept term. Thus the model is,

$$\ln h(t) = \beta_0 + \beta_1 t + \beta_2 (t - 1.5)_+$$

Fit this model and plot the estimated hazard against the piecewise estimates.

```
// Force the functions to join at the knot (3 parameters)
. glm d lin_s1 lin_s2 , family(poisson) link(log) lnoffset(risktime)

. predict haz_lin2, nooffset
. replace haz_lin2 = haz_lin2*1000
. twoway (scatter haz_grp midtime)  ///
         (line haz_lin2 midtime, lcolor(red)) ///
         , xtitle("Years from diagnosis") ///
         ytitle("Baseline hazard (1000 pys)") ///
         xline(1.5, lcolor(black) lpattern(dash)) ///
         ylabel(5 10 20 50 100 150, angle(h)) ///
         legend(off) ///
         name(linear2, replace)
```

Calculate the gradient before the knot and after the knot at 1.5 years.

(d) We will now perform a similar exercise for cubic splines. Again we will have a single knot, this time at 2 years. We will fit the equivalent two separate cubic functions, one before the knot and one after the knot, and then start to introduce the constraints that ensure the function is smooth. There will be eight parameters in the model (3 polynomial terms and an intercept for each of the two intervals).

$$\ln h(t) = \sum_{k=0}^{3} \beta_k t^k + \sum_{k=0}^{3} \beta_{k+4} (t - 2)_+^k$$

Fit this model, predict the hazard and plot, comparing the fit to the piecewise estimates.

```
. gen cubic_s1 = midtime
. gen cubic_s2 = midtime^2
. gen cubic_s3 = midtime^3
. gen cubic_int = midtime>2
. gen cubic_lin = (midtime - 2)*(midtime>2)
. gen cubic_quad = ((midtime - 2)^2)*(midtime>2)
. gen cubic_s4 = ((midtime - 2)^3)*(midtime>2)

. glm d cubic* , family(poisson) link(log) lnoffset(risktime)
. predict haz_cubic1, nooffset
. replace haz_cubic1 = haz_cubic1*1000
. twoway (scatter haz_grp midtime)  ///
        (line haz_cubic1 midtime if midtime<=2, lcolor(red)) ///
        (line haz_cubic1 midtime if midtime>2, lcolor(red)) ///
        , xtitle("Years from diagnosis") ///
        ytitle("Baseline hazard (1000 pys)") ///
        xline(2, lcolor(black) lpattern(dash)) ///
        ylabel(5 10 20 50 100 150, angle(h)) ///
        legend(off) ///
        name(cubic1, replace)
```

(e) We will now constrain the function to be continuous at the knot. This can be done by dropping the second intercept term. The model becomes,

$$\ln h(t) = \sum_{k=0}^{3} \beta_k t^k + \sum_{k=1}^{3} \beta_{k+3}(t-2)_+^k$$

(The subscript for the second sum, starts at $k = 1$ rather than $k = 0$.

Fit this model by excluding the variable `cubic_int` from the model. Plot the predicted hazard to ensure that it is continuous at the knot. Explain why the function does not look smooth.

(f) Now we will force the first derivative to be continuous. This can be done by dropping the second linear term from the model.

$$\ln h(t) = \sum_{k=0}^{3} \beta_k t^k + \sum_{k=2}^{3} \beta_{k+2}(t-2)_+^k$$

Fit this model by excluding the variable `cubic_lin` from the model. Plot the predicted hazard to ensure that it is continuous at the knot. Does the function look smoother?

(g) Finally we will force the second derivative to also be continuous. This can be done by dropping the second quadratic term from the model.

$$\ln h(t) = \sum_{k=0}^{3} \beta_k t^k + \beta_4(t-2)_+^3$$

Fit this model by excluding the variable `cubic_quad` from the model. Plot the predicted hazard. Compare the fits of the models with the different constraints.

(h) For the models we fit we usually use restricted cubic splines. These are constrained to be linear before the first knot and after the final knot. We nearly always put the boundary knots at the minimum and maximum event times and so the restriction of linearity is not actually within the range of our data, but the linear restrictions help to stabalise the estimated function. A good derivation of how the linear restrictions are imposed can be found in Appendix B of Royston and Parmar (2002).

Generate the restricted cubic spline basis functions with 4 degrees of freedom (5 knots). As we have collapsed data we need to use the `fw(d)` (frequency weights) option as we place the knots evenly acording to the distribution of events times, i.e. in this case at the 0^{th} (minimum), 25^{th}, 50^{th}, 75^{th} and 100^{th} (maximum) centiles of the distribution of event times.

```
. rcsgen midtime, gen(rcs) df(4) fw(d)
. global knots ‘r(knots)’
```

We have stored the location of the knots in a global macro, so we can add them to later plots.

(i) The first spline variable, `rcs1`, is just copy of our x variable, `midtime`. Fit a model where you assume that the log hazard function is a linear function of log time and plot the fitted function.

```
. glm d rcs1, family(poisson) link(log) lnoffset(risktime)
. estimates store rcs1
. predict haz_rcs1, nooffset
. replace haz_rcs1 = haz_rcs1*1000
. twoway (scatter haz_grp midtime)  ///
         (line haz_rcs1 midtime, lcolor(red)) ///
         , xtitle("Years from diagnosis") ///
         ytitle("Baseline hazard (1000 pys)") ///
         ylabel(5 10 20 50 100 150, angle(h)) ///
         legend(off) ///
         name(rcs1, replace)
```

Does this model look a good fit?

(j) Now add the remaining spline variables, `rcs2-rcs4`, to the model and perform a likelihood ratio test to see if there is evidence of non linearity.

```
. glm d rcs*, family(poisson) link(log) lnoffset(risktime)
. estimates store rcs2
. lrtest rcs1 rcs2
```

Plot the fitted function against the piecewise estimates and show the location of the knots as reference lines.

```
. predict haz_rcs2, nooffset
. replace haz_rcs2 = haz_rcs2*1000
. twoway (scatter haz_grp midtime)  ///
         (line haz_rcs2 midtime, lcolor(red)) ///
         , xtitle("Years from diagnosis") ///
         ytitle("Baseline hazard (1000 pys)") ///
         xline(\$knots, lcolor(black) lpattern(dash)) ///
         ylabel(5 10 20 50 100 150, angle(h)) ///
         legend(off) ///
         name(rcs2, replace)
```

(k) We will now show the restriction of linearity beyond the boundary knots by moving them within the range of the data. Recalculate the restricted cubic splines with knots at 1, 2 and 3 years. Plot the estimated hazard function.

```
. drop rcs*
. rcsgen midtime, gen(rcs) knots(1 2 3) fw(d)
. global knots ‘r(knots)’
. glm d rcs*, family(poisson) link(log) lnoffset(risktime)
. predict haz_rcs3, nooffset
. replace haz_rcs3 = haz_rcs3*1000
. twoway (scatter haz_grp midtime)  ///
        (line haz_rcs3 midtime, lcolor(red)) ///
        , xtitle("Years from diagnosis") ///
        ytitle("Baseline hazard (1000 pys)") ///
        xline(\$knots , lcolor(black) lpattern(dash)) ///
        ylabel(5 10 20 50 100 150, angle(h)) ///
        legend(off) ///
        name(rcs3, replace)
```

131. Modelling cause-specific mortality using flexible parametric models

> **Stata addon required!** This exercise requires the Stata user-written command `stpm2`. See Section 2.2 for details and installation instructions.

We will now fit some models with the linear predictor on the log cumulative hazard scale using flexible parametric survival models (Royston-Parmar models).

Load and `stset` the Melanoma data.

```
. use melanoma, clear
. keep if stage == 1
. stset surv_mm, failure(status==1) exit(time 120.5) scale(12)
```

(a) Plot a Kaplan-Meier curve for the study population as a whole.

```
sts graph
```

(b) Fit a Weibull model using `stpm2`. In a Weibull model the log cumulative hazard function is a linear function of log(time). This can be fitted with the `scale(hazard)` and `df(1)` options.

```
stpm2, scale(hazard) df(1)
predict s1, surv
predict h1, hazard
```

The two prediction commands will estimate the survival and hazard functions respectively.

Overlay the estimated survival function from a Weibull model with the Kaplan-Meier curve.

```
sts graph, addplot(line s1 _t, sort) name(km1, replace)
```

Does the Weibull model fit well? What assumptions are made about the shape of the hazard function with a Weibull model? Is it possible to assess this from looking at the survival curve?

(c) In Stata we can obtain an estimate of the hazard function using weighted kernel-density estimation using `sts graph, hazard`. I usually use the `kernel(epan2)` option as this generally works better than the default.

Plot this hazard estimate, overlaying the estimated hazard from the Weibull model.

```
sts graph, hazard kernel(epan2) addplot(line h1 _t, sort) name(hazard1, replace)
```

You should now understand why the Weibull model does not fit very well.

(d) We will now relax the assumption of linearity of the log cumulative hazard with respect to log time by incorporating restricted cubic splines into the model. We will initially use 4 df (5 knots) using the default knot locations.

```
stpm2, scale(hazard) df(4)
predict s4, surv
predict h4, hazard
```

Now add the predictions of the survival and hazard functions to the non-parametric survival and hazard functions.

```
sts graph, addplot(line s4 _t, sort) name(km4, replace) ///
    xline(‘e(bhknots)’ ‘e(boundary_knots)’)
sts graph, hazard kernel(epan2) addplot(line h4 _t, sort) name(hazard4, replace) ///
    line(‘e(bhknots)’ ‘e(boundary_knots)’)
```

Are the fitted curves better than the Weibull model?

(e) Fit a Cox model with the diagnosis period (`year8594` as the only covariate.

```
. stcox year8594
```

(f) Fit the equivalent flexible parametric survival model on the log cumulative hazard scale with 4 degrees of freedom for the baseline.

```
. stpm2 year8594, scale(hazard) df(4) eform
```

Compare the estimated hazard ratio, 95% confidence interval and statistical significance to the Cox model.

(g) Obtain predicted values of the survival and hazard functions and plot these functions by calendar period of diagnosis

```
. predict s1ph, survival
. predict h1ph, hazard per(1000)
. twoway (line s1ph _t if year8594 == 0, sort) ///
    (line s1ph _t if year8594 == 1, sort) ///
    , legend(order(1 "1975-1984" 2 "1985-1994") ring(0) pos(1) col(1)) ///
    xtitle("Time since diagnosis (years)") ///
    ytitle("Survival")

. twoway (line h1ph _t if year8594 == 0, sort) ///
    (line h1ph _t if year8594 == 1, sort) ///
    , legend(order(1 "1975-1984" 2 "1985-1994") ring(0) pos(1) col(1)) ///
    xtitle("Time since diagnosis (years)") ///
    ytitle("Cause specific mortality rate (per 1000 py’s)")
```

(h) Add the option `yscale(log)` to the hazard plot to display the hazard function on the log scale. Why is the difference between the two lines constant over the time scale?

(i) Note that there are 4 `_rcs` terms because of the `df(4)` option. We can investigate more or less degrees of freedom for the baseline. It is easiest to do this in a loop. We will also store the model estimates using `estimates store` and predict the baseline survival and hazard functions.

```
forvalues i = 1/6 {
    stpm2 year8594, scale(hazard) df(‘i’) eform
    estimates store df‘i’
    predict h_df‘i’, hazard per(1000) zeros
    predict s_df‘i’, survival zeros
}
```

Compare the hazard ratios, AIC and BIC from the different models.

```
. estimates table df*, eq(1) keep(year8594) se stats(AIC BIC)
```

According to the AIC and BIC how many degrees of freedom should be used for the baseline? Does it matter for the interpretation of the estimated hazard ratio?

About AIC and BIC AIC (Akaike information criterion) and BIC (Bayesian information criterion) are two popular measures for comparing the relative goodness-of-fit of statistical models. The AIC and BIC are defined as:

$AIC = -2\ln(\text{likelihood}) + 2k$

$BIC = -2\ln(\text{likelihood}) + \ln(N)k$

where k = number of parameters estimated and N = number of observations.

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC/BIC value. Hence, the measures not only reward goodness of fit, but also include a penalty that is an increasing function of the number of estimated parameters. AIC uses a fixed constant, 2, in the penalty term whereas the penalty in BIC is a function of the number of observations. It is not always obvious how 'number of observations' should be defined for time-to-event data, particularly for grouped or split data. Volinsky and Raftery (2000) suggest using the number of events for N in the BIC penalty term for survival models. The `estimates stats` command contains an option `n(#)` for specifying N. In this exercise, the estimates table command will give the BIC based on number of events as preferred.

In many circumstances both the AIC and BIC will suggest the same model. For population-based survival data, the number of observations is large so BIC will penalize models with additional parameters more strongly than AIC.

(j) Compare the estimated baseline survival and hazard functions for the models with varying degrees of freedom.

```
. line s_df* _t, sort  ///
    legend(ring(0) cols(1) pos(1)) ///
    xtitle("Time since diagnosis (years)") ///
    ytitle("Survival") ///
    name(compsurv, replace)

. line h_df* _t, sort ///
    xtitle("Time since diagnosis (years)") ///
    ytitle("Cause specific mortality rate (per 1000 py's)") ///
    name(comphazard, replace)
```

Comment on the agreement between the different choices of degrees of freedom.

(k) The previous question compares using a different number of knots placed at their default locations. However, it is also reasonable to ask whether placing the knots in different places leads to different model fits. Run the code in the solution Do file to fit 10 different models with 5df (6 knots) with the 4 internal knots placed at random centiles of the distribution of event times. Don't worry about understanding the code (unless you want to).

Comment on the agreement in the hazard ratios, and the baseline survival and hazard functions between the different models.

(l) Now include sex and age (in categories) in the model. First fit a Cox model and compare the parameter estimates

```
. stcox female year8594 i.agegrp
. estimate store cox
. stpm2 female year8594 i.agegrp, df(4) scale(hazard) eform
. estimates store stpm2_ph
```

(m) Explain why the estimates from the Cox model and the flexible parametric model are so similar.

(n) As models become more complex, we may want predictions for specific combinations of covariates. This is particularly the case when be have continuous covariates. `stpm2`'s `predict` command has useful `at()` and `zeros` options. The `at()` option requests predictions at specified values of covariates and the `zeros` options requests that any variables not listed in the `at()` option are set to zero.

Another useful option is the `timevar({\it varname})` option. This requests that the predictions are at values of time specified in *varname* rather than the default `\_t`. This is useful for very large data sets or when wanting to predict survival for various combinations of covariates at one specific time, e.g., 5 years.

i. Define a new variable, `temptime` with 200 observations taking values of time between 0 and 10 and predict the baseline survival function.

```
. estimates restore ph
. range temptime 0 10 200
. predict S0, survival zeros timevar(temptime)
. line S0 temptime, sort
```

What combination of covariates does the baseline survival represent?

ii. Using the `at()` and `zeros` options predict the hazard function for females aged 75+ diagnosed in 1985-1994 for values of `temptime` with a 95% confidence interval.

```
. predict S_F_8594_age75, survival ///
    at(female 1 year8594 1 agegrp 3) timevar(temptime) ci

. twoway (rarea S_F_8594_age75_lci S_F_8594_age75_uci temptime, pstyle(ci)) ///
    (line S_F_8594_age75 temptime) ///
    , legend(off) ///
    xtitle("Time since diagnosis (years)") ///
    ytitle("S(t)") ///
    title("Female age 75+ diagnosed 1985-1994")
```

132. Modelling time-dependent effects using flexible parametric models

Stata addon required! This exercise requires the Stata user-written command `stpm2`.

We will now fit a model where the effect of age group is time-dependent. Reload and `stset` the data.

```
. use melanoma, clear
. keep if stage == 1
. gen female = sex == 2
. stset surv_mm, failure(status==1) exit(time 60.5) scale(12)
```

We have restricted follow-up to five years.

(a) First we will fit a Cox model and assess the proportional hazards assumption using Schoenfeld residuals. One can obtain a plot of the scaled Schoenfeld residuals, with a smoother, for a chosen predictor using the following command.

```
. estat phtest, plot(3.agegrp)
```

We will use a loop to produce plots for each agegrp and add a horizontal line at the value of the estimated log hazard ratio.

```
. stcox female year8594 i.agegrp,
. forvalue i = 1/3 {
    local beta = _b[‘i’.agegrp]
    estat phtest, plot(‘i’.agegrp) name(sch_age‘i’, replace) ///
        yline(0 ‘beta’) msize(small) msymbol(Oh) bw(0.4)
}
. estat phtest, detail
```

Is the effect of age proportional?

(b) Now fit a flexible parametric proportional hazards model with 4 df for the baseline. Note that when we go on to use the `tvc()` option you can’t use Stata’s factor variables, so we create dummy variables for age group.

```
. tab agegrp, gen(agegrp)
. stpm2 female year8594 agegrp2-agegrp4, df(4) scale(hazard) eform
. estimates store ph
```

Predict and plot the hazard function for each age group for males diagnosed in 1975-1984. Note the use of the `at()` and `zeros` options.

```
. predict h_age1, hazard zeros per(1000)
. predict h_age2, hazard at(agegrp2 1) zeros per(1000)
. predict h_age3, hazard at(agegrp3 1) zeros per(1000)
. predict h_age4, hazard at(agegrp4 1) zeros per(1000)

. twoway (line h_age1 _t, sort) ///
     (line h_age2 _t, sort) ///
     (line h_age3 _t, sort) ///
     (line h_age4 _t, sort) ///
     ,xtitle("Time since diagnosis (years)") ///
     ytitle("Cause specific mortality rate (per 1000 py’s)") ///
     legend(order(1 "<45" 2 "45-59" 3 "60-74" 4 "75+") ring(0) pos(1) cols(1))
```

(c) Now fit a model with time-dependent effects for age group. We will do this using 2 degrees of freedom for each age category.

```
. stpm2 female year8594 agegrp2-agegrp4, df(4) scale(hazard) ///
       tvc(agegrp2 agegrp3 agegrp4) dftvc(2)
. estimates store nonph
```

Perform a likelihood ratio test comparing the proportional hazards model with the non-proportional hazards (for age) model. Is there evidence of a non-proportional effect?

```
. lrtest ph nonph
```

(d) Now predict the hazard function for each age group. Note that the prediction command is identical to that used in the proportional hazards model as `stpm2` knows which variables have time-dependent effects and takes this into account when predicting.

```
. predict h_age1_tvc, hazard zeros per(1000)
. predict h_age2_tvc, hazard at(agegrp2 1) zeros per(1000)
. predict h_age3_tvc, hazard at(agegrp3 1) zeros per(1000)
. predict h_age4_tvc, hazard at(agegrp4 1) zeros per(1000)

. twoway (line h_age1 h_age1_tvc _t, sort lcolor(red red) lpattern(solid dash)) ///
    (line h_age2 h_age2_tvc _t, sort lcolor(blue blue) lpattern(solid dash)) ///
    (line h_age3 h_age3_tvc _t, sort lcolor(magenta magenta) lpattern(solid dash)) ///
    (line h_age4 h_age4_tvc _t, sort lcolor(green green) lpattern(solid dash)) ///
    ,xtitle("Time since diagnosis (years)") ///
    ytitle("Cause specific mortality rate (per 1000 py's)") ///
    legend(order(1 "<45" 2 "45-59" 3 "60-74" 4 "75+") ring(0) pos(1) cols(1)) ///
    name(hazard_tvc, replace)
```

(e) Obtain a prediction of the hazard ratio as a function of time for each age group.

```
. predict hr2, hrnumerator(agegrp2 1) ci
. predict hr3, hrnumerator(agegrp3 1) ci
. predict hr4, hrnumerator(agegrp4 1) ci
```

Note that by default the `hrdenominator` option sets all covariates to zero. As we only have one covariate with a time-dependent effect we can leave this unspecified.

Plot these hazard ratios versus follow-up time on the same graph. What happens to the hazard ratios as follow-up time increases? Also plot the hazard ratio for the oldest group with a 95% confidence interval. Explain why the hazard ratio for the oldest age group is so high early on in the time-scale (hint: look at the baseline hazard)

```
. twoway (line hr2 hr3 hr4 _t, sort), ///
     yscale(log)  ylabel(1 2 10 20 50) ///
     legend(order(1 "Age 45-59" 2 "Age 60-74" 3 "Age 75+") ring(0) pos(1) cols(1)) ///
     xtitle("Time since diagnosis (years)") ///
     ytitle("Hazard ratio") ///
     name(hr, replace)

. twoway (rarea hr4_lci hr4_uci _t, sort pstyle(ci)) ///
    (line hr4 _t, sort) ///
    ,legend(off) yscale(log) ylabel(1 2 10 20 50) ///
    xtitle("Time since diagnosis (years)") ///
    ytitle("Hazard ratio") ///
    name("hr_age4", replace)
```

(f) Obtain and plot with 95% confidence intervals the difference in the hazard rates between the oldest and youngest age groups for males in 1975-1984.

```
. predict hdiff4, hdiff1(agegrp4 1) ci per(1000)
. twoway (rarea hdiff4_lci hdiff4_uci _t, sort) ///
    (line hdiff4 _t, sort) ///
    ,legend(off)  ///
    xtitle("Time since diagnosis (years)") ///
    ytitle("Difference in hazard rate") ///
    name(hdiff, replace)
```

Explain why the hazard difference is small early on in the time-scale, when the hazard ratio is at is greatest.

(g) Predict and plot the survival function for the youngest and oldest age groups for females diagnosed in 1985-1994.

```
. predict s1, surv at(female 1 year8594 1) zeros
. predict s2, surv at(agegrp4 1 female 1 year8594 1) zeros
```

Obtain and plot with 95% confidence intervals the difference in the survival functions between the oldest and youngest age groups for females diagnosed in 1985-1994.

```
. predict sdiff4, sdiff1(agegrp4 1 female 1 year8594 1) ///
               sdiff2(agegrp4 0 female 1 year8594 1) ci
. twoway (rarea sdiff4_lci sdiff4_uci _t, sort) ///
    (line sdiff4 _t, sort) ///
    ,legend(off)  ///
    xtitle("Time since diagnosis (years)") ///
    ytitle("Difference in survival functions") ///
    name(sdiff, replace)
```

(h) Fit models with 1, 2 and 3 df for the time-dependent effect of age. Use the AIC and BIC to compare models. Compare the estimated time-dependent hazard ratio for the oldest age group compared to the youngest (also compare the 95% confidence intervals). You may want to exclude the first month from you plot as the hazard ratio is very high close to zero.

```
forvalues i = 1/3 {
  stpm2 i.sex year8594 agegrp2-agegrp4, df(4) scale(hazard) ///
    tvc(agegrp2 agegrp3 agegrp4) dftvc(‘i’)
  estimates store dftvc‘i’
  predict hr4_df‘i’, hrnumerator(agegrp4 1) ci
}
count if _d==1
estimates stats dftvc*, n(‘r(N)’)

twoway (line hr4_df1 hr4_df1_lci hr4_df1_uci _t, sort lcolor(red..) ///
        lpattern(solid dash dash) lwidth(medthick thin thin)) ///
    (line hr4_df2 hr4_df2_lci hr4_df2_uci _t, sort lcolor(midblue..) ///
      lpattern(solid dash dash) lwidth(medthick thin thin)) ///
    (line hr4_df3 hr4_df3_lci hr4_df3_uci _t, sort lcolor(midgreen..) ///
        lpattern(solid dash dash) lwidth(medthick thin thin)) ///
    if _t>0.1, ///
    yscale(log) ///
    ylabel(1 2 4 8 20 50, angle(h)) ///
    legend(order(1 "1 df" 4 "2 df" 7 "3 df") ring(0) pos(1) cols(1)) ///
    xtitle("Time since diagnosis (years)") ///
    ytitle("Hazard Ratio") ///
    yscale(log) ///
    name(tvc_df_comp, replace)
```

(i) We will now also let the effect of sex be time-dependent to illustrate when there are two time-dependent effects the hazard ratios are not the exactly the same. Add `female` to the `tvc` option.

```
. stpm2 female  agegrp2-agegrp4, df(4) scale(hazard) ///
    tvc(agegrp2 agegrp3 agegrp4 female) dftvc(3)
```

Use the `hrnumerator` and `hrdenominator` options of the `predict` command to obtain a prediction of the hazard ratio for `female` for the youngest and the oldest age groups. Add the `ci` option to obtain confidence intervals. Compare the resulting curves and their 95% confidence intervals to show that the curves are similar, but not identical.

```
. predict hr_f_age1, hrnum(female 1) ci
. predict hr_f_age4, hrnum(female 1 agegrp4 1) hrdenom(agegrp4 1) ci

. twoway (line hr_f_age1* hr_f_age4* _t if _t>0.1, sort yscale(log))
```

(j) **additional topic - not covered in lectures** If we were modelling on the log hazard scale, then including exactly the same covariates and time-dependent effect, the hazard ratio would be equivalent. Such a model can be fitted using the `strcs` command. This command is much slower than `stpm2` as it requires numerical integration (using Gauss-Legendre quadrature) to estimate the parameters.

```
. strcs female  agegrp2-agegrp4, df(4) ///
    tvc(agegrp2 agegrp3 agegrp4 female) dftvc(3) nodes(50)
. predict hr_f_age1b, hrnum(female 1) ci
. predict hr_f_age4b, hrnum(female 1 agegrp4 1) hrdenom(agegrp4 1) ci

. twoway (line hr_f_age1b* hr_f_age4b* _t if _t>0.1, sort yscale(log))
```

140. Probability of death in a competing risks framework (cause-specific survival)

Stata addon required! This exercise requires the Stata user-written command `stcompet`. See Section 2.2 (page 4) for details and installation instructions.

This question gives an introduction to some of the methods available for competing risks analyses. To carry out the exercises you will need to install some user-written commands from within Stata. The `stcompet` command estimates the cumulative incidence function (CIF) non-parametrically. The `stpm2cif` command estimates the CIF through post-estimation after fitting a flexible parametric model.

(a) Load the colon data dropping those with missing stage.

```
use colon, clear
drop if stage ==0
gen female = sex==2
```

If you summarize status you will notice that there are deaths from both cancer and other causes. Plot the complement of the Kaplan-Meier estimate for males (i.e., 1 minus Kaplan-Meier survival estimate) for both cancer and other causes. Describe what you see.

A common confusion when competing risks are present is to think that the probability of death from cancer can be obtained by taking the complement of the Kaplan-Meier estimate (1-KM). By doing this we treat deaths from other causes as censored. If we can assume independence, that is the patients dying from other causes would have been at no systematically higher or lower risk of dying from cancer, then we estimate the marginal probability of death, i.e., the probability of death in the hypothetical world where it is not possible to die of other causes.

The appropriate estimate for the "real world" probability of death from cancer when competing risks are present is the cumulative incidence function. That is the proportion of patients that have died from cancer at a certain time in the follow-up period taking into account competing causes of death.

(b) Use the `stcompet` command to estimate the cumulative incidence function for both cancer and other causes. Plot the cumulative incidence functions for males along with the complements of the Kaplan-Meier estimates from part (a). What do you notice?

```
stset surv_mm, failure(status==1) scale(12) exit(time 120.5)
stcompet CIF_sex=ci, compet1(2) by(sex)
gen CIF_sex_cancer=CIF_sex if status==1
gen CIF_sex_other=CIF_sex if status==2
```

(c) Obtain estimates of the CIF for cancer and other causes by age group. Plot and interpret the curves.

```
stset surv_mm, failure(status==1) scale(12) exit(time 120.5)
stcompet CIF_age=ci, compet1(2) by(agegrp)

twoway (line CIF_age _t if agegrp == 0 & status == 1, sort connect(stepstair)) ///
    (line CIF_age _t if agegrp == 1 & status == 1, sort connect(stepstair)) ///
    (line CIF_age _t if agegrp == 2 & status == 1, sort connect(stepstair)) ///
    (line CIF_age _t if agegrp == 3 & status == 1, sort connect(stepstair)) ///
    , legend(order(1 "<45" 2 "45-59" 3 "60-74" 4 "75+") ring(0) pos(5) cols(1)) ///
    xtitle("Years since diagnosis") ///
    ytitle("CIF") ///
    title("Cancer") ///
    name(CIF_age1,replace)

twoway (line CIF_age _t if agegrp == 0 & status == 2, sort connect(stepstair)) ///
    (line CIF_age _t if agegrp == 1 & status == 2, sort connect(stepstair)) ///
    (line CIF_age _t if agegrp == 2 & status == 2, sort connect(stepstair)) ///
    (line CIF_age _t if agegrp == 3 & status == 2, sort connect(stepstair)) ///
    , legend(order(1 "<45" 2 "45-59" 3 "60-74" 4 "75+") ring(0) pos(11) cols(1)) ///
    xtitle("Years since diagnosis") ///
    ytitle("CIF") ///
    title("Other causes") ///
    name(CIF_age2,replace)

graph combine CIF_age1 CIF_age2, nocopies ycommon
```

(d) Now obtain the CIF for cancer and other causes by stage group. Plot the results. Explain why those diagnosed with regional and distant stage are less likely to die from other causes when compared to those with localized disease.

```
. stcompet CIF_stage=ci, compet1(2) by(stage)

. twoway (line CIF_stage _t if stage == 1 & status == 1, sort connect(stepstair)) ///
    (line CIF_stage _t if stage == 2 & status == 1, sort connect(stepstair)) ///
    (line CIF_stage _t if stage == 3 & status == 1, sort connect(stepstair)) ///
    , legend(order(1 "local" 2 "regional" 3 "distant") ring(0) pos(5) cols(1)) ///
    xtitle("Years since diagnosis") ///
    ytitle("CIF") ///
    title("Cancer") ///
    name(CIF_stage1,replace)

. twoway (line CIF_stage _t if stage == 1 & status == 2, sort connect(stepstair)) ///
    (line CIF_stage _t if stage == 2 & status == 2, sort connect(stepstair)) ///
    (line CIF_stage _t if stage == 3 & status == 2, sort connect(stepstair)) ///
    , legend(order(1 "local" 2 "regional" 3 "distant") ring(0) pos(1) cols(1)) ///
    xtitle("Years since diagnosis") ///
    ytitle("CIF") ///
    title("Other causes") ///
    name(CIF_stage2,replace)

. graph combine CIF_stage1 CIF_stage2, nocopies ycommon
```

180. Outcome-selective sampling designs (nested case-control and case-cohort)

In this exercise we compare a full cohort analysis of the melanoma data to analyses using nested case-control (NCC) and case-cohort designs. For the purpose of the exercise, we will assume that the main exposure of interest is sex and we will adjust for age at diagnosis (`agegrp`), year of diagnosis (`year8594`) and `stage`.

We would not use outcome-selective sampling in practice for this research question, but do it here for pedagogic purposes. In practice, we might use such designs if we were interested in collecting additional information on an expensive or time-consuming exposure or confounder/effect modifier (e.g., biomarker information or collecting information from medical records), which may not be feasible on the full cohort of 7,775 patients.

(a) We start by performing a full cohort analysis on all 7,775 patients. We `stset` using death due to melanoma as the outcome and time-since-diagnosis as the timescale.

```
. use melanoma, clear
. stset exit, fail(status==1) enter(dx) origin(dx) scale(365.24) id(id)
```

How many deaths are there among the patients? These deaths will be the ‘cases’ in the NCC and the case-cohort designs.

(b) Is there evidence of a difference in mortality between women and men? Estimate Kaplan-Meier curves and fit a Cox regression model with sex as the main exposure. Also adjust the model for age, year and stage. Is there evidence of confounding by age, year and stage?

```
. * Kaplan-Meier curves
. sts graph, by(sex)

. * Cox regression
. stcox i.sex
. stcox i.sex i.agegrp i.year8594 i.stage
```

(c) Now fit the same model, but as a flexible parametric model and as a Poisson regression model. To adjust for time-since-cancer as the underlying timescale in the Poisson regression model, we must time-split the data on follow-up (`fuband`) and add time-since-cancer in the model as a covariate (`fuband`). This step is not needed for the flexible parametric model, which automatically uses the timescale specified in `stset`.

```
* Flexible parametric model
. stpm2 i.sex i.agegrp i.year8594 i.stage, df(5) scale(hazard) eform

* Poisson regression
. stsplit fuband, at(0(5)20)
. streg i.sex i.agegrp i.year8594 i.stage i.fuband, dist(exp)
```

Compare the estimates of the sex effect for Cox regression, FPM and Poisson regression. Are they different? Would you expect them to differ? Why (or why not)?

You may wish to make use of Stata's estimate machinery for storing, manipulating, and displaying estimation results.

```
use melanoma, clear
stset exit, fail(status==1) enter(dx) origin(dx) scale(365.24) id(id)

stcox i.sex i.agegrp i.year8594 i.stage, nolog
estimates store cox

stpm2 i.sex i.agegrp i.year8594 i.stage, df(5) scale(hazard) eform nolog
estimates store fpm

stsplit fuband, at(0(5)20)
streg i.sex i.agegrp i.year8594 i.stage i.fuband, dist(exp) nolog
estimates store poiss

estimates table cox fpm poiss, eq(1) b(%5.3f) eform
```

(d) We will now generate and analyse a nested case-control study with 1 control per case. First reload and stset the data using melanoma-specific death as the outcome and time-since-diagnosis as the timescale.

```
. use melanoma, clear
. stset exit, fail(status==1) enter(dx) origin(dx) scale(365.24) id(id)
```

How many events (deaths due to melanoma) were observed during follow-up. If we generated a nested case-control study with 1 control per case, how many unique individuals do you expect would be in the NCC?

Let's now generate the NCC using the `sttocc` command. We will match on age group and select 1 control per case.

```
. set seed 339487731
. sttocc, match(agegrp) n(1)
```

The sampling will take a few minutes. For each riskset a dot will appear in the results window. Since there are 1913 deaths, there will be 1913 dots. We have chosen to specify a seed for the Stata random number functions in order to make the sampling reproducible (i.e., force everyone to get the same results).

(e) The data set in memory will now be the nested case-control dataset. Use the `describe` command to examine its contents and list the first 10 observations.

The variable `_case` include the case-control status (1=case, 0=control), `_set` is a unique identifier for the matched set (i.e., the case and its matched control will have the same value on the `_set` variable).

i. What information is represented by the variable `_time`?

ii. Confirm that there are an equal number of cases and controls and that the age distribution of the cases and controls is the same (due to matching on age).

iii. How many unique individuals are there in the nested case-control study?

(f) Nested case-control studies are analysed using conditional logistic regression. We must condition on the matching strata (by including the _set variable in the `group()` option).

```
. clogit _case i.sex i.year8594 i.stage, group(_set) or
```

i. What underlying quantity is being estimated by the estimates in the column labelled ‘Odds ratio’?

ii. Is the estimated effect measure for sex similar to the hazard ratio for the full cohort? Would you expect it to be?

iii. Are the confidence intervals (standard errors) similar?

(g) We will now generate and analyse a case-cohort study. Reload and stset the data using cause-specific death as the outcome and time-since-diagnosis as the timescale

```
. use melanoma, clear
. stset exit, fail(status==1) enter(dx) origin(dx) scale(365.24) id(id)
```

Now we will generate a case-cohort design with a sampling fraction of around 25%. We will generate a new outcome variable based on the event indicator variable (_d) from the `stset`.

```
. gen case=_d
```

First, we must sample the subcohort. To make sure the sampling is reproducible, we use a seed. Then we assign a random number between 0 and 1 to all observations in the dataset (using the `runiform` command). We select a subcohort by creating an indicator variable subcoh which takes the value 1 for 25% observations in the subcohort and value 0 for the remaining 75% observations outside the subcohort.

```
. set seed 339487731  // makes sampling reproducible
. gen u = runiform()   // assign random number to all obs
. gen subcoh = 1 if (u <= 0.25) // generate dummy subcohort
. replace subcoh = 0 if (u > 0.25)
```

Check that the sampling worked by tabulating the number of cases and non-cases inside and outside the subcohort? Complete the following table.

```
. tab case subcoh
```

	Outside subcohort	Inside subcohort	Total
Non-cases			
Cases			
Total			7,775

(h) Calculate the exact sampling fraction of the subcohort. Also calculate the exact sampling fraction of non-cases, i.e. the proportion of non-cases in the subcohort compared to non-cases in the full cohort.

(i) The analysis of case-cohort samples is identical to that of cohort designs with the addition of (1) weights and (2) robust standard errors. Generate the weights by creating a wt variable, which takes the value 1 for cases and 'the inverse of the sampling fraction for non-cases' for non-cases.

```
. gen wt = 1 if case==1
. replace wt = 1 / (1470/5862) if case==0 & subcoh==1
. tab wt, missing
```

Are the weights as you expected? Which observations have missing values for `wt`?

(j) To include weights in the analysis in Stata, simply include them in the `stset` command using the `pweight` option `[pw=wt]`. The weights will now be included in all `st` commands. Any observation with missing values for `wt` will not be included in the analysis.

```
. stset exit [pw=wt], fail(status==1) enter(dx) origin(dx) ///
                      scale(365.24) id(id)
```

(k) Estimate the effect of sex on mortality by fitting the models using the weighted data, i.e. Cox regression, FPM and Poisson regression. Compare the estimates of the sex effect. What do you conclude?

```
. * Cox model for case-cohort - Borgan II weights
. stcox i.sex i.agegrp i.year8594 i.stage, vce(robust)

. * FPM model for case-cohort - Borgan II weights
. stpm2 i.sex i.agegrp i.year8594 i.stage, scale(h) df(5) eform ///
                                          vce(robust) nolog

. * Poisson regression - Borgan II weights
. stsplit fuband, at(0(5)20)
. streg i.sex i.agegrp i.year8594 i.stage i.fuband, dist(exp) vce(robust)
```

(l) Investigate the extent of sampling variation by generating multiple nested case-control studies. The following code generates, analyses, and reports a table of results for 5 repetitions. Note that this code will take sevarl minutes to run.

```
set more off
use melanoma, clear
stset exit, fail(status==1) enter(dx) origin(dx) scale(365.24) id(id)

forvalues i=1/5 {
preserve
display as text _newline "Now processing iteration " `i'  _newline
sttocc, match(agegrp) n(1)
clogit _case i.sex i.year8594 i.agegrp i.stage, group(_set) or nolog
estimates store ncc`i'
restore
}

est table ncc1 ncc2 ncc3 ncc4 ncc5, eform equations(1) ///
b(%9.6f) se modelwidth(10)
```

(m) Now see what happens if you increase the number of controls to, for example, 3 and then 5. What would you expect with 5 controls per case?

(n) Investigate generate and analyse multiple case-cohort studies. The following code generates, analyses, and reports a table of results for 5 repetitions with a subcohort of 25%.

```
set more off
use melanoma, clear

gen case=(status==1)

forvalues i=1/5 {
preserve
display as text _newline "Now processing iteration " ‘i’  _newline
gen subcoh = (runiform() <= 0.25)
gen wt = 1 if case==1
replace wt = 1 / 0.25 if case==0 & subcoh==1
stset exit [pw=wt], fail(status==1) enter(dx) origin(dx) scale(365.24) id(id)
stcox i.sex i.agegrp i.year8594 i.stage, vce(robust)
estimates store cc‘i’
restore
}

est table full_cox cc1 cc2 cc3 cc4 cc5, eform equations(1) ///
b(%9.6f) se modelwidth(10)
```

(o) Now investigate the effect of changing the size of the subcohort to, for example, 10% and 50%.