Biostatistics III: Survival analysis for epidemiologists
Computing notes and exercises

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

Karolinska Institutet
6—15 March, 2017

http://biostat3.net

Contents

1 Notes on survival analysis using Stata **3**

2 Downloading user-written Stata commands and data files **4**

2.1 Downloading the course files 4

2.2 Installing Stata user-written commands 4

3 Exercises **6**

1. Hand calculation: Life table and Kaplan-Meier estimates of survival 6

2. Comparing cause-specific and all-cause survival 10

3. Comparing estimates of cause-specific survival between periods; log rank test 12

4. Comparing actuarial and Kaplan-Meier approaches with discrete-time data 14

6. Reviewing the Poisson regression example from the lecture notes (diet data) 15

7. Model cause-specific mortality using Poisson regression 17

8. Poisson regression with the diet data; choice of timescale 20

9. Model cause-specific mortality using Cox regression 22

10. Examining the proportional hazards hypothesis (localised melanoma) 24

11. Cox regression for all-cause mortality 26

12. Examining the effect of sex on melanoma survival 27

13. Modelling the diet data using Cox regression 28
22. Time-varying exposures – the bereavement data 29
23. Calculating SMRs/SIRs 31
25. Generating and analysing a nested case-control study 34
28. Model cause-specific mortality using flexible parametric models 35
39. Probability of death in a competing risks framework (cause-specific survival) . . . 38

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 directory, change the Stata working directory to the new directory (e.g., `cd c:\survival\`), and then download the files. You can create a new directory in Windows Explorer or you can do it from within Stata as follows.

```
mkdir c:\survival
cd c:\survival
```

Use the `pwd` command to confirm you are in the working directory you wish to use for the course. The course files are available on the web as a ZIP archive:

```
http://biostat3.net/download/biostat3.zip
```

Save and extract this folder in your created working directory. You can also read the data files directly from the web from within Stata. For example,

```
. use http://www.biostat3.net/download/colon, clear
(Colon carcinoma, all stages, 1975-94, follow-up to 1995)
```

2.2 Installing Stata user-written commands

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

1 (a) **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.

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

ACTUARIAL APPROACH

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

time	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				

1 (b) **Using Stata to validate the hand calculations done in part 1 (a)** We will now use Stata to reproduce the same analyses done by hand calculation in part 1 (a) 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)
```

2. **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?

3. **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.

4. **Localised melanoma: Comparing actuarial and Kaplan-Meier approaches with discrete time data**

   The aim of this exercise is to examine the effect of heavily grouped data (i.e., data with lots of ties) on estimates of survival made using the Kaplan-Meier method and the actuarial method.

   For the patients diagnosed with localised skin melanoma, use Stata to estimate the 10-year cause-specific survival proportion. Use both the Kaplan-Meier method and the actuarial method. Do this both with survival time recorded in completed years and survival time recorded in completed months. That is, you should obtain 4 separate estimates of the 10-year cause-specific survival proportion to complete the cells of the following table. The purpose of this exercise is to illustrate small differences between the two methods when there are large numbers of ties.

   In order to reproduce the results in the printed solutions you'll need to restrict to localised stage (`stage==1`) and estimate cause-specific survival (`status==1` indicates an event). Look at the Stata code in the previous questions if you are unsure.

	Actuarial	Kaplan-Meier
Years		
Months		

   (a) Of the two estimates (Kaplan-Meier and actuarial) made using time recorded in years, which do you think is the most appropriate and why?
   [HINT: Consider how each of the methods handle ties.]

   (b) Which of the two estimates (Kaplan-Meier or actuarial) changes most when using survival time in months rather than years? Why?

6. **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)
```

(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) 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
```

(d) 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.

(e) 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.

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

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

(g) 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
```

(h) 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 part 6e.

(i) Use `poisson` to compare the first and third levels with the second.

(j) 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
```

(k) 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.]

7. 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 9 we model cause-specific mortality using Cox regression and in exercise 28 we use flexible parametric models. The aim is to illustrate that these three methods are very similar.

The aim of these exercises is to explore the similarities and differences to these three approaches to modelling. We will be comparing the results (and their interpretation) as we proceed through the exercises so you may wish to save your commands in a do file to facilitate comparison.

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

```
. use melanoma if stage==1, clear
. 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 7a and 7b was confounding by time since diagnosis. The comparison in part 7a was adjusted for time since diagnosis (since we compare the differences between the curves at each point in time) whereas the comparison in part 7b 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 if stage==1, clear
. 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 of the latter) is mortality the lowest? Is this consistent with what you found in part 7b?

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 year8594, dist(exp)
```

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 year8594 if _st==1, exp(risktime) irr
. glm _d 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?

(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 c).

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

Is the estimated effect of calendar period of diagnosis consistent with what you expected? Add an interaction between follow-up and calendar period of diagnosis and interpret the results.

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

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

i. Interpret the estimated hazard ratio for the parameter labelled `agegrp 2`, 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.

(k) Based on the interaction model you fitted in exercise 7j, 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 7j.

ii. Using the estimates from exercise 7j 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 year8594 sex_early sex_latter, dist(exp)
```

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

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

8. **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 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?

   ```
   . gen bmi=weight/(height/100*height/100)
   . poisson chd hieng 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 hieng i.ageband, e(risktime) irr
. poisson _d 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 hieng i.fuband, e(risktime) irr
. poisson _d 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?

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

In exercise 7 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.

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 7.

```
. use melanoma
. keep if stage == 1
. stset surv_mm, failure(status==1) exit(time 120)
. stcox 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 3c (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 sex year8594 i.agegrp
```

i. Interpret the estimated hazard ratio for the parameter labelled `agegrp 2`, 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
```

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

```
. stcox sex year8594 i.agegrp
. est store A
. stcox sex 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 9c is similar to the model estimated in question 7i.

i. Both models adjust for `sex, year8594,` and `i.agegrp` but the Poisson regression model in question 7i 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.

Following is some code for estimating and comparing the Cox and Poisson regression models (you might wish to copy this code from the PDF version of this document and paste it into the Stata do file editor or use `compare_cox_poisson.do`.

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

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

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

(f) ADVANCED: By splitting at each failure time we can estimate a Poisson regression model that is identical to the Cox model. Code is available in the file `compare_cox_poisson.do` on the course website. This model takes several minutes to estimate and you may need to reset the values of `memory` and `matsize`.

(g) 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* year8594 sex i.agegrp, dist(exp)
predict xb, xb
twoway line xb fu if year8594==0 & sex==1 & agegrp==1, sort
```

10. **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. These plots were not given extensive coverage in the lectures, so attempt this if you like or continue to part (d). 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 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 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 sex 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 sex 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 sex year8594 i.fuband i.fuband#i.agegrp
```

	0–2 years	2+ years
Agegrp1	1.00	1.00
Agegrp2		
Agegrp3		
Agegrp4		

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

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

11. **Cox regression with observed (all-cause) mortality as the outcome**

    Now fit a model to the localised melanoma data where the outcome is observed survival (i.e. all deaths are considered to be events).

```
. stset surv_mm, failure(status==1,2) exit(time 120)
. keep if stage==1
. stcox sex year8594 i.agegrp
```

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

    (b) On comparing the estimates between the observed and cause-specific survival models it appears that only the parameters for age have changed substantially. Can you explain why the estimates for the effect of age would be expected to change more than the estimates of the effect of sex and period?

12. **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 by other factors (e.g. age, stage, subsite, period)? After controlling for potential confounders, does there still appear to 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. Be sure to evaluate the proportional hazards assumption.

13. **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 6).

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

Now fit the following Cox model.

```
. stset dox, id(id) fail(chd) entry(doe) origin(doe) scale(365.24)
. stcox 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?

22. **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.

23. Calculating SMRs/SIRs

The standardized mortality ratio (SMR) is the ratio of the observed number of deaths in the study population to the number that would be expected if the study population experienced the same mortality as the standard population. It is an indirectly standardized rate. When studying disease incidence the corresponding quantity is called a standardized incidence ratio (SIR). These measures are typically used when the entire study population is considered 'exposed'. Rather than following-up both the exposed study population and an unexposed control population and comparing the two estimated rates we instead only estimate the rate (or number of events) in the study population and compare this to the expected rate (expected number of events) for the standard population. For example, we might study disease incidence or mortality among individuals with a certain occupation (farmers, painters, airline cabin crew) or cancer incidence in a cohort exposed to ionising radiation.

In the analysis of cancer patient survival we typically estimate *excess mortality* (observed - expected deaths). The SMR (observed/expected deaths) is a measure of *relative mortality*. The estimation of observed and expected numbers of deaths are performed in an identical manner for each measure but with the SMR we assume that the effect of exposure is multiplicative to the baseline rate whereas with excess mortality we assume it is additive. Which measure, relative mortality or excess mortality, do you think is more homogeneous across age?

The following example illustrates the approach to estimating SMRs/SIRs using Stata. Specifically, we will estimate SMRs for the melanoma data using the general population mortality rates stratified by age and calendar period (derived from `popmort.dta`) to estimate the expected number of deaths. The expected mortality rates depend on current age and current year so the approach is as follows

- Split follow-up into 1-year age bands
- Split the resulting data into 1-year calendar period bands
- For each age-period band, merge with `popmort.dta` to obtain the expected mortality rates
- Sum the observed and expected numbers of deaths and calculate the SMR (observed/expected) and a 95% CI

(a) Start by stsetting the data with age as the timescale and splitting the follow-up into 1 year age bands

```
use melanoma, clear
stset exit, fail(status == 1 2) origin(bdate) entry(dx) scale(365.24) id(id)
stsplit _age, at(0(1)110) trim
```

(b) Now split these new records into 1 year calendar period bands using

```
stsplit _year, after(time=d(1/1/1900)) at(70(1)100) trim
replace _year=1900+_year
list id _age _year in 1/15
```

Note that we have used the second syntax for `stsplit` and set the origin for calendar period as 1/1/1900 for convenience in setting the breaks.

(c) Each subject's follow–up is now divided into small pieces corresponding to the age-bands and calendar periods the subject passes through. We can make tables of deaths and person-years by age and calendar period with

```
gen _y = _t - _t0 if _st==1
table _age _year, c(sum _d)
table _age _year, c(sum _y) format(%5.3f)
```

As the data have been split in 1-year intervals on both time scales the table created above is not so informative. Grouped variables will provide a better overview.

```
egen ageband_10=cut(_age), at (0(10)110)
egen period_5=cut(_year), at(1970(5)2000)

table ageband_10 period_5, c(sum _d)
table ageband_10 period_5, c(sum _y) format(%4.1f)
```

(d) To make a table of rates by age and calendar period, try

```
gen obsrate=_d/_y
table ageband_10 period_5 [iw=_y] , c(mean obsrate) format(%5.3f)
```

(e) To calculate the expected cases for a cohort, using reference mortality rates classified by age and calendar period, it is first necessary to break the follow–up into parts which correspond to these age bands and calendar periods, as above.

Before calculating the expected number of cases it is necessary to add the reference rates to the expanded data with

```
sort _year sex _age
merge _year sex _age using popmort
```

This is a matched merge on age band and calendar period and will add the appropriate survival probability to each record. The system variable `_merge` takes the following values:

1– record in the master file but no match in `popmort`
2– record in `popmort` but no match in the master file
3– record in the master file with a match in `popmort`

```
tab _merge
```

should show mostly 3's with some 2's but no 1's. You can now drop the records with no match in the master file and the system variable

```
drop if _merge==2
drop _merge
```

(f) The mortality rates for the standard population are derived by transforming the survival probabilities

```
gen mortrate=(-ln(prob))
```

and to calculate the expected number of cases, multiply the follow-up time for each record by the reference rate for that record

```
gen e=_y*mortrate
list id e _d in 1/15
```

(g) The SMR is the ratio of the total observed cases to the total number expected. The total numbers are obtained through

```
egen obs=total(_d)
egen exp=total(e)
```

from which the manually calculated SMR with corresponding 95% confidence interval are obtained (`preserve` and `restore` are used to speed up the processing)

```
preserve
keep in 1
gen SMR = obs/exp
gen LL = ( 0.5*invchi2(2*obs, 0.025)) / exp
gen UL = ( 0.5*invchi2(2*(obs+1), 0.975)) / exp

display "SMR(95%CI)=" round(SMR,.001) " ///
        (" round(LL,.001) ":" round(UL,.001) ")"
restore
```

An easier approach is to let the `strate` command perform these calculations for us (after first splitting and merging in the standard rates).

```
strate, smr(mortrate)
```

(h) To calculate the SMR for the different stages, try

```
strate stage, smr(mortrate)
```

Summary

The following commands can be used to calculate an SMR for the melanoma patients:

```
use melanoma, clear
stset exit, fail(status == 1 2) origin(bdate) entry(dx) scale(365.25) id(id)
stsplit _age, at(0(1)110) trim
stsplit _year, after(time=d(1/1/1900)) at(70(1)100) trim
replace _year=1900+_year
sort _year sex _age
merge _year sex _age using popmort
drop if _merge==2
gen mortrate=-ln(prob)
strate, smr(mortrate)
```

25. **Localised melanoma: Generating and analysing a nested case-control study**

The data on patients diagnosd with localised melanoma were obtained from the Finnish cancer registry. We might be interested in studying additional prognostic factors not available in the cancer registry data. For example, we might wish to extract additional information from the medical records (e.g., treatment) or genotype archived tumour samples. A nested case control design will restrict the number of individuals for whom we need to collect additional information. We will first fit a model to the full cohort and then generate a nested case-control study and fit the same model.

We will first analyse the full cohort using a Cox model. We will only study the first 10 years subsequent to diagnosis.

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

/* Cox model to full cohort */
stcox sex year8594 i.agegrp
```

(a) How many individuals are in our study? That is, if we were collecting additional information, on how many individuals would we need to collect it?

(b) How many experience the event (i.e., death due to cancer).

(c) Now generate a nested case-control study with 1 control per case and fit a model with the same explanatory variables.

```
sttocc, n(1)
clogit _case sex year8594 i.agegrp, group(_set) or
```

(d) How many unique individuals are in our study? That is, if we were collecting additional information, on how many individuals would we need to collect it?

(e) Compare the estimated parameters and standard errors between the full cohort analysis and the nested case-control study. What is the relative efficiency (ratio of variances) of the nested case-control compared to the full cohort design?

(f) If you are interested in exploring this further you can repeat the exercise. That is, generate another nested case-control study and analyse it. If you did this a large number of times, and plotted a histogram of the parameter estimates, you will find that the parameter estimates from the nested case-control study have a symmetric distribution with mean equal to the the full-cohort estimate. In the solutions you can find Stata code for looping.

28. Modelling cause-specific mortality using flexible parametric models

Stata addon required! This exercise requires the Stata user-written command `stpm2`. See Section 2.2 (page 4) 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 the Melanoma data and refit the Cox model to use as a comparison.

```
. use melanoma, clear
. keep if stage == 1
. stset surv_mm, failure(status==1) exit(time 120) scale(12)
. stcox year8594,
```

(a) Fit a 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.

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

```
. predict s1, survival
. predict h1, hazard per(1000)

. twoway  (line s1 _t if year8594 == 0, sort) ///
          (line s1 _t if year8594 == 1, sort) ///
          , legend(order(1 "1975-1984" 2 "1985-1994") ring(0) pos(1) col(1)) ///
          xtitle("Time since diagnosis (years)")

. twoway  (line h1 _t if year8594 == 0, sort) ///
          (line h1 _t if year8594 == 1, sort) ///
          , legend(order(1 "1975-1984" 2 "1985-1994") ring(0) pos(1) col(1)) ///
          xtitle("Time since diagnosis (years)")
```

(c) 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?

(d) 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.

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

Compare the hazard ratios, AIC and BIC from the different models. On the following page you can find a brief description of the AIC and BIC.

```
. 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 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.

Which measure – AIC or BIC – is best?

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. We suggest using the AIC rather than BIC.

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

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

```
. stpm2 i.sex year8594 i.agegrp, df(4) scale(hazard) eform
. estimates store ph
```

Compare the estimates to those obtained in question 9(c). Perform a Wald test (use the `test` command) for the overall effect of age and compare it to that obtained for the Cox model.

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

(h) We will now extend the model to allow the effect of age group to be time-dependent using 2 degrees of freedom for each age category. Note that you can't use Stata's factor variables in the `tvc()` option, so create your own dummy variables.

```
. tab agegrp, gen(agegrp)
. stpm2 i.sex 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
```

(i) Predict and plot the baseline hazard function for this model. For which group of patients (i.e., which covariate values) does the baseline hazard apply?

```
. predict h0, hazard zeros ci
. line h0 _t, sort
```

(j) 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
```

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 largest age group is so high early on in the time-scale (hint: look at the baseline hazard)

(k) Obtain and plot with 95% confidence intervals the difference in the hazard rates between the oldest and youngest agegroups 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")
```

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

(l) Obtain and plot with 95% confidence intervals the difference in the survival functions between the oldest and youngest agegroups for females diagnosed in 1985-1994.

```
. predict sdiff4, sdiff1(agegrp4 1 sex 2 year8594 1) ///
                sdiff2(agegrp4 0 sex 2 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")
```

(m) 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.

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

Stata addon required! This exercise requires the Stata user-written commands `stpm2`, `stcompet`, `stcompadj` and `stpm2cif`. 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. We will be estimating similar quantities to q34 and q38 but we are now working in a cause-specific survival framework. 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 `stcompadj` command estimates the CIF using a competing risks analogue of the Cox model. Finally, the `stpm2cif` command estimates the CIF through postestimation after fitting a flexible parametric model.

(a) Load the melanoma data. 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 (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. This then assumes that the patients dying from other causes would have been at no systematically higher or lower risk of dying from cancer (independent censoring). Patients that die due to competing causes can no longer die from cancer which essentially violates this independence assumption. Therefore, the resulting estimates would not be interpretable as the probability of death from cancer.

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 along with the complements of the Kaplan-Meier estimates from part (a). What do you notice?

```
. stset surv_mm, failure(status==1) scale(12)
. stcompet CIF=ci, compet1(2)
. gen CIFcancer=CIF if status==1
. gen CIFother=CIF if status==2
```

(c) We now want to consider a competing risks analysis that will take into account covariate effects. Use the `stcompadj` command to estimate the cumulative incidence function for cancer and other causes taking into account the effect of sex. First we need to generate a binary variable for females. We shall assume that the effect of females is the same for both cancer and other causes. You will need to run the command twice in order to evaluate the cumulative incidence functions for each sex. Plot the cumulative incidence functions for cancer and other causes for both males and females. Are there any differences between males and females?

```
. gen female=sex==2
. stset surv_mm, failure(status==1) scale(12)
. stcompadj female=0, compet(2) gen(CIFcancermale CIFothermale)
. stcompadj female=1, compet(2) gen(CIFcancerfemale CIFotherfemale)
```

(d) We will now consider the same model but using the flexible parametric approach. In order to do this we will first need to expand the data set so that each patient has two rows of data - one for each cause of death.

i. Expand the data and have a look at the new data set.

```
. expand 2
. bysort id: gen cause=_n
. gen cancer=(cause==1)
. gen other=(cause==2)
```

ii. Fit a flexible parametric model for cancer and other causes simultaneously. Include sex as a covariate assuming that the effect of sex is the same for both cancer and other causes. Interpret the effect of sex.

```
. stset surv_mm, failure(event) scale(12)
. stpm2 cancer other female, scale(hazard) ///
  rcsbaseoff dftvc(3) nocons tvc(cancer other) eform nolog
```

By including the two cause indicators (`cancer` and `other`) as both main effects and time-dependent effects (using `tvc` option) we have fitted a stratified model with two separate baselines, one for each cause. For this reason we have used the `rcsbaseoff` option together with the `nocons` option which excludes the baseline hazard from the model.

iii. Use the `stpm2cif` postestimation command to obtain the cumulative incidence functions for cancer and other causes for each sex. You will need to run this command twice - once for each sex. Notice that a new time variable is generated called `_newt`. You will need to use this instead of `_t` in your plots. Do the results look the same as the ones from the Cox model approach?

```
. stpm2cif cancermale othermale, cause1(cancer 1) ///
  cause2(other 1)
. stpm2cif cancerfemale otherfemale, cause1(cancer 1 female 1) ///
  cause2(other 1 female 1)
```

iv. Think about an alternative way to present the results. Try stacking the cumulative incidence functions for cancer and other causes.

(e) So far the model only included one covariate (male/female). We now want to adjust the model for age. However, we don't believe that the effect of age is the same for both cancer and other causes of death.

i. To allow the effect of age to vary for the two causes create interaction terms between age group and the causes of death.

```
. gen age0ca=(agegrp==0 & cancer==1)
. gen age1ca=(agegrp==1 & cancer==1)
. gen age2ca=(agegrp==2 & cancer==1)
. gen age3ca=(agegrp==3 & cancer==1)
. gen age0oth=(agegrp==0 & other==1)
. gen age1oth=(agegrp==1 & other==1)
. gen age2oth=(agegrp==2 & other==1)
. gen age3oth=(agegrp==3 & other==1)
```

ii. Fit a flexible parametric model including sex and the interaction terms between age group and cause.

```
. stpm2 cancer other female age1ca age2ca age3ca ///
        age1oth age2oth age3oth , scale(hazard) ///
        rcsbaseoff dftvc(3) nocons tvc(cancer other) eform nolog
```

iii. Now incorporate sex as a time-dependent effect in the model.

```
. stpm2 cancer other female age1ca age2ca age3ca ///
        age1oth age2oth age3oth , scale(hazard) ///
        rcsbaseoff dftvc(3) nocons tvc(cancer other female) eform nolog
```

iv. Estimate the cumulative incidence functions for each sex and for the age groups 0-44 and 75+. Plot the cumulative incidence functions against time.

```
. stpm2cif cancermaleage0 othermaleage0, ///
  cause1(cancer 1) cause2(other 1)

. stpm2cif cancermaleage3 othersex1age3, ///
  cause1(cancer 1 age3ca 1) cause2(other 1 age3oth 1)

. stpm2cif cancerfemaleage0 othefemaleage0, ///
  cause1(cancer 1 female 1) cause2(other 1 female 1)

. stpm2cif cancerfemaleage3 otherfemaleage3, ///
  cause1(cancer 1 age3ca 1 female 1) cause2(other 1 age3oth 1 female 1)
```