

Statistical methods for population-based cancer survival analysis

Computing notes and exercises

Paul W. Dickman¹, Paul C. Lambert^{1,2}, Sandra Eloranta¹,
Therese Andersson¹, Mark J Rutherford², Anna Johansson¹,
Caroline E. Weibull¹, Sally Hinchliffe², Hannah Bower¹, Michael Crowther²

(1) Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden

(2) Department of Health Sciences
University of Leicester
Leicester, UK.

paul.dickman@ki.se
paul.lambert@leicester.ac.uk
sandra.eloranta@ki.se
therese.m-l.andersson@ki.se
mjr40@le.ac.uk
anna.johansson@ki.se
caroline.weibull@ki.se
srh20@leicester.ac.uk
hannah.bower@ki.se
mjc76@leicester.ac.uk

June 2018

Contents

1	Notes on survival analysis using Stata	4
2	Downloading user-written Stata commands and data files	5
2.1	The quick and easy way	5
2.2	Downloading the course files	5
2.3	Installing Stata user-written commands for relative survival	6
3	For SAS users	7
4	For R users	7
5	Exercises	9
100.	Hand calculation: Life table and Kaplan-Meier estimates of survival	9
101.	Using Stata to validate the hand calculations done in question 100	11
102.	Comparing actuarial and Kaplan-Meier approaches with discrete-time data	13
103.	Comparing cause-specific and all-cause survival	14
104.	Comparing estimates of cause-specific survival between periods; log rank test	16
110.	Reviewing the Poisson regression example from the lecture notes (diet data)	18
111.	Model cause-specific mortality using Poisson regression	20
112.	Poisson regression with the diet data; choice of timescale	23
120.	Model cause-specific mortality using Cox regression	25
121.	Examining the proportional hazards hypothesis (localised melanoma)	27
122.	Cox regression for all-cause mortality	29
123.	Examining the effect of sex on melanoma survival	30
124.	Modelling the diet data using Cox regression	31
125.	Time-varying exposures – the bereavement data	32
130.	Understanding splines	34
131.	Model cause-specific mortality using flexible parametric models	39
132.	Flexible parametric models with time-dependent effects	43
133.	Modelling on other scales using <code>stpm2</code>	47
140.	Probability of death in a competing risks framework (cause-specific survival)	49
150.	Adjusted/standardized survival curves	56
180.	Outcome-selective sampling designs	59
181.	Calculating SMRs/SIRs	64
182.	Calculating SMRs/SIRs using <code>strs</code>	67
200.	Hand calculation of expected survival	69

201. Life table estimates of relative survival using <code>strs</code>	70
202. Life table estimates of cause-specific survival using <code>strs</code>	72
203. Period estimation of relative survival	73
204. Evaluating period predictions	74
210. Model excess mortality using Poisson regression	75
211. Model excess mortality using Poisson regression with a smooth baseline	76
230. Model excess mortality using flexible parametric models	79
231. Non-linear effects in relative survival I – Proportional hazards	82
232. Non-linear effects in relative survival II – Time-dependent effects	84
240. Age-standardised estimates of relative survival (internal standard)	86
241. Age-standardised estimates of relative survival	88
242. Age standardization using flexible parametric models	89
243. Age-standardised estimates of relative survival (external standard)	91
250. Probability of death in a competing risks framework (life table relative survival) .	92
251. Probability of death in a competing risks framework (relative survival model) . .	93
260. Fitting cure models	94
261. Fitting cure models using flexible parametric models	96
280. Creating a <code>popmort</code> file from the Human Mortality Database	98
281. Constructing a <code>popmort</code> file by modelling cohort data	100
282. Excess and ‘avoidable’ deaths from life tables	102
283. Simulating relative survival	104
284. Estimating loss in expectation of life	107
285. Missing covariate data (using official Stata commands)	110
286. Understanding frailty	114

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.

<code>stset</code>	Declare data to be survival-time data
<code>stsplit</code>	Split time-span records
<code>sts</code>	Generate, graph, list, and test the survivor and cumulative hazard functions
<code>strate</code>	Calculate person-time at risk and failure rates
<code>stcox</code>	Estimate Cox proportional hazards model
<code>streg</code>	Estimate parametric survival models
<code>strs</code>	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. Standard Stata does not contain commands for relative survival, so we must extend Stata with user-written commands. Note that there are two separate steps; downloading the course files and installing the user-written commands. We have written an automated script that does both these steps (see section 2.1) or if you prefer more control, there are instructions in sections 2.2 and 2.3.

2.1 The quick and easy way

Enter the following at the Stata command line to download the course files (e.g., data files and solution do files) and install all Stata user-written commands:

```
do http://www.pauldickman.com/survival/install_packages.do
```

NOTE: You do not need to change the working directory before running this command. This will create a directory, `c:\survival`, and install the course files into that directory. If the directory `c:\survival` already exists, or you prefer to install the files into another directory then you will need to download `install_packages.do`, edit the directory reference, and then run the file from within Stata.

2.2 Downloading the course files

You do not have to do this if you already used the ‘quick and easy way’ described in section 2.1; the course files will already be downloaded for you.

The course files (e.g., data files and solution do files) are distributed as a Stata package so should be downloaded from within Stata. 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 and then issue the following command from the Stata command line to install the course files.

```
net install http://www.pauldickman.com/survival/course_files, all replace
```

`net install` downloads the files and copies them to appropriate directories according to the way Stata is setup. Ancillary files (e.g., PDF, XLS, DTA) are copied to the current working directory; ADO and HLP files are installed into the appropriate directory according to the way Stata is configured.

2.3 Installing Stata user-written commands for relative survival

You do not have to do this if you already used the ‘quick and easy way’ described in section 2.1. Standard Stata does not contain any commands for estimating and modelling relative survival so we must extend Stata using commands written by users. Download and installation 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.

2.3.1 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 1.6.3 14Jan2016
```

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.

2.3.2 strs - estimating and modelling relative survival

The `strs` command, written by Paul Dickman and Enzo Coviello can be downloaded by typing the following:

```
. net install http://www.pauldickman.com/rsmodel/stata_colon/strs, all replace
```

Note that some of the data files are contained in both the `strs` and the `course_files` packages, hence the need for the `replace` option. See http://pauldickman.com/rsmodel/stata_colon/ for further details about the command or read the Stata help file after installation. The command is described in a Stata Journal article [1].

2.3.3 stpm2 - flexible parametric models

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

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

The command is described in a Stata Journal article [2]. `rcsgen` is a command for generating basis vectors for restricted cubic splines and is required by `stpm2`. Flexible parametric cure models (fitted using an option to `stpm2`) are described in another Stata Journal article [3].

2.3.4 strsmix and strsnmix - cure models

To install `strsmix` and `strsnmix` (commands for fitting cure models) first type `findit lambert cure` then click on the Stata Journal link followed by *click to install*. These commands are described in a Stata Journal article [4].

2.3.5 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
```

The `stpm2cif` command is described in a Stata Journal article [5].

3 For SAS users

Paul Dickman has written SAS code for estimating and modelling relative survival, see http://pauldickman.com/rsmodel/sas_colon/. The code was written in 2004 and has not been updated to incorporate recent methods. It implements life table estimation of relative survival using the Ederer II method (cohort or period approach) and modelling excess mortality using Poisson regression.

Ron Dewar has written SAS macros that implement everything that can be done using Paul Dickman's code, along with many newly developed methods (most notably estimation using the Pohar Perme method and modelling using flexible parametric models). The macros are not publicly available, but you can request them by writing to Ron (epiman46@gmail.com). The macro for flexible parametric models requires a licence for SAS/STAT and SAS/IML, whereas the other macros require only SAS/STAT. All of the macros can be run using SAS University Edition, which is free for academic and non-commercial use.

Hermann Brenner and colleagues have also published SAS code, see http://www.imbe.med.uni-erlangen.de/cms/software_period.html although we do not have any experience in using it.

4 For R users

Maja Pohar has written an R package, `relnsurv`, for relative survival that is easily found on CRAN [6].

At Karolinska Institutet we run a postgraduate course called 'Biostatistics III: Survival analysis for epidemiologists'. The exercises are based on the same data sets and many exercises are

similar (if not identical) to this course. Information and R code can be found at the following link:

<http://biostat3.net/download/R/index.html>

'Biostatistics III' is a general course on survival analysis for epidemiologists so there are no exercises on net survival. There is, however, an exercise on flexible parametric models in R which can be extended to relative survival.

5 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.

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				

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)
```

102. **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?

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 estimate 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 clinical/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)
```

- (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 110e.

- (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.]

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 three methods are very similar. Cox regression cannot be used to model excess mortality, so we use Poisson regression (exercise 210) or flexible parametric models (exercise 230).

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 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 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 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 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 111c.

```
. 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 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 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)
```

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 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 `stsplot` to generate agebands.

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

. stsplot 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`?

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.

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)
. 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_1^2 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 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 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) 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 year8594 sex i.agegrp
est store Cox

/* split on time since diagnosis (1-year intervals) */
stsplot 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)
```

- (g) **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 `q120.do` on the course website. This model might take several minutes to estimate and you may need to reset the values of `memory` and `matsize`.
- (h) **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) */
stsplot 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
```

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. 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 `stsplot`:

```
. stsplot 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.

122. 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?

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

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 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?

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 (see Clayton & Hills [7], §32.2). 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 bereavement 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 2.3 (page 6) 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)
. stsplitt 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 stabilise the estimated function. A good derivation of how the linear restrictions are imposed can be found in Appendix B of Royston and Parmar (2002) [8].

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 according to the distribution of event times, i.e. in this case at the 0th (minimum), 25th, 50th, 75th and 100th (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*
. rcsген 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 2.3 (page 6) 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 $\log(\text{time})$. 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 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 i.sex 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 we 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` option 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.

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

. 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 add-on 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 `tvf()` 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 its 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 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") ///
  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 your 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))

```

133. **Modelling cause-specific mortality on other scales (proportional odds and Aranda-Ordaz link function) using stpm2**

This question uses the Melanoma data. Load and `stset` the data.

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

- (a) Fit a proportional hazards model to the melanoma data with age group, sex and calendar year as covariates. Predict the survival and hazard functions for the youngest and oldest age groups for those diagnosed in 1975-1984. Store the model estimates.

```
stpm2 female i.agegrp year8594, scale(hazard) df(4) eform
forvalues i = 0/3 {
    predict s_age'i'_ph, surv at(agegrp 'i') zeros
    predict h_age'i'_ph, hazard at(agegrp 'i') zeros
}
estimates store ph
```

- (b) Now fit a proportional odds model and predict the survival and hazard functions. You just need to change the `scale(hazard)` option to `scale(odds)`

```
stpm2 female i.agegrp year8594, scale(odds) df(4) eform
forvalues i = 0/3 {
    predict s_age'i'_po, surv at(agegrp 'i') zeros
    predict h_age'i'_po, hazard at(agegrp 'i') zeros
}
estimates store po
```

Interpret the effect of the covariate `female`

- (c) Compare the predict survival and hazard function between the proportional odds and proportional hazards models. Explain why they are not the same.

```
twoway (line s_age0_ph _t, sort) ///
       (line s_age0_po _t, sort) ///
       (line s_age3_ph _t, sort) ///
       (line s_age3_po _t, sort) ///
       , name(survcomp, replace)

twoway (line h_age0_ph _t, sort) ///
       (line h_age0_po _t, sort) ///
       (line h_age3_ph _t, sort) ///
       (line h_age3_po _t, sort) ///
       , name(hazcomp, replace)
```

- (d) According to the BIC and AIC, which is the best fitting model?

```
count if _d == 1
estimates stats ph po, n('r(N)')
```

- (e) For the proportional odds model the hazards will not be proportional. Predict and plot the hazard ratio for females in the youngest age group diagnosed in 1975-1984.

```
predict hr_female_age0_7584, hrnum(female 1) hrdenom(female 0) ci
twoway (rarea hr_female_age0_7584_lci hr_female_age0_7584_uci _t, sort pstyle(ci)) ///
      (line hr_female_age0_7584 _t, sort) ///
      ,legend(off) ///
      xtitle("Years since diagnosis") ///
      ytitle("Hazard Ratio") ///
      title("HR for sex (age<45, diagnosed 1975-1984)") ///
      name(HR1, replace)
```

- (f) The hazard ratio for females will be different at different levels of other covariates. Show this by now calculating the hazard ratio for females in the oldest age group diagnosed in 1975-1984.

```
predict hr_female_age3_7584, hrnum(female 1 agegrp 3) hrdenom(female 0 agegrp 3) ci
twoway (line hr_female_age0_7584 _t, sort) ///
      (line hr_female_age3_7584 _t, sort) ///
      ,name(HR2, replace)
```

- (g) Now fit a model using the Aranda-Ordaz link function using the `scale(theta)` option. Compare the AIC/BIC with the proportional hazard and proportional odds model.

```
stpm2 female i.agegrp year8594, scale(theta) df(4)
estimates store ao
count if _d == 1
estimates stats ph po ao, n('r(N)')
```

- (h) The proportional odds model provides a better fit. Calculate the estimated value of θ with 95% confidence intervals. Explain why this is the case.

```
lincom [ln_theta] [_cons], eform
```


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

Stata add-on required! This exercise requires the Stata user-written commands `stpm2`, `stcompet`, `stpepemori`, `stcompadj`, and `stpm2cif`. See Section 2.3 (page 6) 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
```

(e) We will now fit a Fine and Gray model with death due to cancer as the main outcome of interest, and death from other causes as the competing event.

i. Fit a model that includes sex as a covariate.

```
stset surv_mm, failure(status==1) scale(12) exit(time 120.5)
stcrreg i.sex, compete(status == 2)
```

How would you interpret the estimated effect of sex (SHR)?

ii. Based on the results from the fitted model, plot the cancer-specific CIF for males and females, respectively, and test if there is evidence of a difference by sex.

```
predict cif_males, basecif
gen cif_females = 1 - (1-cif_males)^exp(_b[2.sex])
```

```
graph twoway line cif_males cif_females _t, ///
sort connect(step step) yscale(range(0 0.6)) ylabel(0(0.2)0.6) ///
legend(order(1 "Males" 2 "Females")) ///
yttitle(Cause-specific cumulative incidence) ///
xtitle(Time since diagnosis (years))
```

```
stpepemori sex, compet(2)
```

iii. The cause-specific CIFs can also be retrieved via the `stcurve` function in Stata. Using, `stcurve`, re-plot the cancer-specific CIFs by sex and verify that the estimates are the same by plotting them side by side the estimates that you calculated from first principles.

```
stcurve, cif at1(sex = 1) at2(sex=2)
```

(f) Now, fit another Fine and Gray model but this time with death due to other causes as the main event of interest, and death due to cancer as the competing event. Interpret the parameter estimate and state your conclusion about the effect of sex based on the model output.

(g) We will now fit a competing risks model 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)
. gen event=(cause==status)
. list id status cause sex event in 1/8, sepby(id)
```

- 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. We usually would like to allow the effect of covariates to be different for the different causes. Now fit a model where the effect of sex is allowed to be different for cancer and other causes.

```
. gen fem_can = female*cancer
. gen fem_other = female*other
. stpm2 cancer other fem_can fem_other, scale(hazard) ///
    rcsbaseoff dftvc(3) nocons tvc(cancer other) eform nolog
```

Test (using a Wald test) where there is evidence that the effect of sex differs between cancer and other causes.

```
. test fem_can = fem_other
```

- iv. Explain how the parameter estimates for the effect of sex on cancer mortality and other cause mortality are conceptually different from the ones you estimated in the Fine and Gray models.
- (h) 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 empirical estimates (overlay these to make the comparison easier).

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

See the do file for the relevant Stata code. What are potential reasons for any disagreement between the empirical and model estimates?

- (i) Think about an alternative way to present the results. Try stacking the cumulative incidence functions for cancer and other causes (see do file for code to plot this).
- (j) 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.

```
. forvalues i = 0/3 {
    gen age'i'can=(agegrp=='i' & cancer==1)
    gen age'i'oth=(agegrp=='i' & other==1)
}
```

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

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

Interpret the hazard ratios

- (k) Predict the cause-specific CIFs for males in the youngest and oldest age groups.

```
. stpm2cif cancermale_age0 othermale_age0, cause1(cancer 1) ///
    cause2(other 1)
. stpm2cif cancermale_age3 othermale_age3, cause1(cancer 1 age3can 1) ///
    cause2(other 1 age3oth 1)
```

- (l) Now incorporate sex and age as time-dependent effects for cancer in the model (use 3 df). Estimate the CIFs and compare to the model that assumes proportional hazards.

```
. stpm2 cancer other fem_can fem_oth ///
    age1can age2can age3can age1oth age2oth age3oth , scale(hazard) ///
    rcsbaseoff dftvc(cancer:4 other:4 3) nocons ///
    tvc(cancer other fem_can age1can age2can age3can) eform nolog

. stpm2cif cancermale_age0_tvc othermale_age0_tvc, cause1(cancer 1) ///
    cause2(other 1)
. stpm2cif cancermale_age3_tvc othermale_age3_tvc, cause1(cancer 1 age3can 1) ///
    cause2(other 1 age3oth 1)
```

- (m) **OPTIONAL EXERCISE** By default `stpm2` defines knot positions for all events and does not distinguish between events types. We will now refit the model, but use the default knot positions obtained when fitting each cause separately.

- i. Produce a histogram of the event times separately by event status. Is the distribution of events similar for each cause?

```
. hist _t if _d==1, by(status)
```

- ii. Fit separate models for each cause and store the knot locations. As we are fitting time-dependent effects for sex and age for cancer, we will include these in the model.

```
. stpm2 fem_can age1can age2can age3can if cancer == 1, ///
df(4) scale(hazard) dftvc(3) ///
tvc(fem_can age1can age2can age3can) eform nolog
. global knots_cancer 'e(bhknots)'
. global knots_cancer_tvc 'e(tvcknots_age1can)'
```

```
. stpm2 fem_oth age1oth age2oth age3oth if other == 1, ///
df(4) scale(hazard) eform nolog
. global knots_other 'e(bhknots)'
```

- iii. Refit the model using these knot locations. Has it made any difference to the estimated CIFs?

```
. stpm2 cancer other fem_can fem_oth ///
    age1can age2can age3can age1oth age2oth age3oth , scale(hazard) ///
    rcsbaseoff nocons ///
    tvc(cancer other fem_can age1can age2can age3can) eform nolog ///
    knotstvc(cancer $knots_cancer other $knots_other) ///
    fem_can $knots_cancer_tvc ///
```

```

    age1can $knots_cancer_tvc ///
    age2can $knots_cancer_tvc ///
    age3can $knots_cancer_tvc)
. stpm2cif cancermale_age0_tvc2 othermale_age0_tvc2, cause1(cancer 1) ///
  cause2(other 1)
. stpm2cif cancermale_age3_tvc2 othermale_age3_tvc2, cause1(cancer 1 age3can 1) ///
  cause2(other 1 age3oth 1)

```

- (n) We will now estimate cause-specific CIFs in a Cox regression framework using the `stcompadj` command. First we need to reload the colon cancer data as the `stcompadj` uses as input the data in its original format. The data will be expanded for us in the background (and we will see later how it is possible to save a copy of the expanded data that can be used for testing or viewing the results from the Cox regression etc.)
- i. Read in the data and `stset` it with cancer as the main outcome of interest.

```

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

```

```

stset surv_mm, failure(status==1) scale(12) exit(time 120.5)

```

- ii. Estimate the cause-specific CIFs by sex from a Cox model that assumes that the effect of sex on mortality from cancer and other causes, respectively is the same. Plot the CIFs for males and females in separate graphs.

```

stcompadj sex=1 , compet(2) gen(Main_males Compet_males)
stcompadj sex=2 , compet(2) gen(Main_females Compet_females)

```

```

graph twoway line Main_males Compet_males _t, ///
sort connect(step step) yscale(range(0 1)) ylabel(0(0.2)1) ///
ytitle(Probability of Death) xtitle(Time Since Diagnosis (Years)) ///
title(Males) ///
legend(order(1 "Cancer" 2 "Other")) ///
name(Cox_males, replace)

```

```

graph twoway line Main_females Compet_females _t, ///
sort connect(step step) yscale(range(0 1)) ylabel(0(0.2)1) ///
ytitle(Probability of Death) xtitle(Time Since Diagnosis (Years)) ///
title(Females) ///
legend(order(1 "Cancer" 2 "Other")) ///
name(Cox_females, replace)

```

```

graph combine Cox_males Cox_females, ycommon

```

- (o) We've seen in the earlier exercises that the effect of sex is not the same for both outcomes under investigation. Estimate another set of CIFs from a model where the assumption of a shared effect of sex between the outcomes is relaxed. Check how much the new estimates differ from those that you estimated in the previous exercise (Overlay the new estimates to make comparisons easier).

```

stcompadj sex=1 , compet(2) maineffect(sex) competeffect(sex) ///
gen(Main_males_2 Compet_males_2)
stcompadj sex=2 , compet(2) maineffect(sex) competeffect(sex)///
gen(Main_females_2 Compet_females_2)

```

```

graph twoway (line Main_males Compet_males _t, lpattern(dash dash) ///
  lcolor(navy red) sort connect(step step)) ///
(line Main_males_2 Compet_males_2 _t, lpattern(solid solid) ///
  lcolor(navy red) sort connect(step step)), ///
  yscale(range(0 0.6)) ylabel(0(0.2)0.6) ///
ytitle(Probability of Death) xtitle(Time Since Diagnosis (Years)) ///
title(Males) ///
legend(order(3 "Cancer" 4 "Other")) ///
name(Cox_males_2, replace)

graph twoway (line Main_females Compet_females _t, lpattern(dash dash) ///
  lcolor(navy red) sort connect(step step)) ///
(line Main_females_2 Compet_females_2 _t, lpattern(solid solid) ///
  lcolor(navy red) sort connect(step step)), ///
  yscale(range(0 0.6)) ylabel(0(0.2)0.6) ///
ytitle(Probability of Death) xtitle(Time Since Diagnosis (Years)) ///
title(Females) ///
legend(order(3 "Cancer" 4 "Other")) ///
name(Cox_females_2, replace)

graph combine Cox_males_2 Cox_females_2, ycommon

```

- (p) To test if the effect of sex differs between cancer mortality and other cause mortality we can fit a Cox regression model that includes an interaction between sex and the variable that indicates which outcome is being modelled (`stcompadj` creates a variable called `stratum` in the expanded data set for this purpose). The expanded data set created by `stcompadj` can be saved and read into Stata for us to use for further analyses. Run the code below to generate and store an expanded data set, use it to fit a Cox regression with the relevant interaction. Is there statistical evidence that the effect of sex is different for the two outcomes?

```

preserve
stcompadj sex=1 , compet(2) savexp(silong,replace)
use silong,clear
xi:stcox i.sex*i.stratum, strata(stratum) nohr nolog
restore

```

- (q) Lastly, we will now refit a Cox model where the effect of sex is no longer assumed to be shared between the outcomes. This can be done by reading in the expanded data set and fitting a Cox model directly to it. Based on the model output, what is the estimated hazard ratio for sex on cancer-specific and other-cause mortality, respectively?

```

preserve
stcompadj sex=1 , compet(2) maineffect(sex) competeffect(sex) savexp(silong,replace)
use silong,clear
xi: stcox Main_sex Compet_sex stratum, nolog
restore

```

150. Adjusted/standardized survival curves

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

This question uses the Rotterdam breast cancer data, comprising information on 2,982 patients with primary breast cancer. The outcome will be relapse-free survival, which is defined as the time from primary surgery to disease recurrence or death from breast cancer. Time to death from other causes were treated as censored.

- (a) Load and `stset` the data. Restrict the follow-up time to 10 years.

```
. use rott2
. stset rf, f(rfi==1) scale(12) exit(time 120)
```

Plot the Kaplan-Meier estimate of the survival function by hormonal treatment group (no hormonal therapy vs hormonal therapy).

```
. sts graph, by(hormon)
. sts gen S_km = s, by(hormon)
```

Will the unadjusted hazard ratio for hormonal therapy be less than or greater than 1?

- (b) Now fit a proportional hazards flexible parametric model using `stpm2`. Use 3 df for the baseline.

```
. stpm2 hormon, scale(hazard) df(3) eform
```

- (c) Compare the model-based and Kaplan-Meier survival curves. Comment on the extent of agreement between the two (did you expect agreement? is there agreement).

```
. twoway (line S_km _t if hormon == 0, sort lcolor(black) lpattern(dash) ///
connect(stepstair)) ///
        (line S_km _t if hormon == 1, sort lcolor(red) lpattern(dash) ///
connect(stepstair)) ///
        (line s _t if hormon==0,sort lcolor(black) lwidth(thick)) ///
        (line s _t if hormon==1, sort lcolor(red) lwidth(thick)) ///
        , xtitle("Years from surgery") ///
        ytitle("S(t)") ///
        legend(order(3 "No hormonal therapy" 4 "hormonal therapy") ring(0) ///
        pos(1) cols(1)) caption("Dashed lines show KM estimates")
```

- (d) In a previous analysis of this data, it was proposed to incorporate the effect of the number of positive lymph nodes using the following transformation[9]. Add `enodes` in the model.

```
. stpm2 hormon enodes, scale(hazard) df(3) eform
```

What has happened to the hazard ratio for `hormon`?

- (e) Now add further covariates to the model. Include the effect of age (as a restricted cubic spline with 3 df), and tumour size.

```
. rcsgen age, df(3) gen(agercs) orthog
. stpm2 hormon i.size enodes agercs*, scale(hazard) df(3) eform
```


- (f) Obtain the predicted survival function at 1 year and 5 years. Produce a histogram for each measure.

```
. gen t1 = 1
. gen t5 = 5
. predict s1, surv timevar(t1)
. predict s5, surv timevar(t5)
. hist s1, name(hist_1yr, replace) xlabel(0(0.1)1)
. hist s5, name(hist_5yr, replace)xlabel(0(0.1)1)
```

- (g) Predict a prognostic index. This is the predicted values of the linear predictor without the spline terms. This can be used to classify into risk groups. We will plot from the 10th to the 90th centile of the prognostic index to show the range in predicted survival probability in the study population.

First predict the prognostic index and then refit the model with this as the only covariate.

```
. predict xb, xbnobaseline
. stpm2 xb, scale(h) df(3)
```

Compare the likelihood to the previous model. Why are they the same?

Now obtain predictions from the 10th to the 90th centile and plot the resulting functions.

```
. forvalues i = 10(10)90 {
    centile xb, centile('i')
    predict s_xb'i', surv at(xb 'r(c_1)')
}
. twoway (line s_xb??_t, sort lcolor(black ..)) ///
, legend(off) ///
ylabel(0(0.2)1, angle(h)) ///
xlabel("Years from surgery") ///
ylabel("S(t)") ///
text(0.8 8 "10th centile") ///
text(0.1 8 "90th centile")
```

- (h) We will now move on to obtaining adjusted survival curves. Refit the original model and obtain the average survival curve for the study population as a whole. Note that in large datasets the prediction can take a long time so use the `timevar` option to ensure that predictions are made at selected values of time.

```
. stpm2 hormon i.size enodes ageracs*, scale(hazard) df(3) eform
. range timevar 0 10 100
. predict s_mean, meansurv timevar(timevar) ci
. twoway (rarea s_mean_lci s_mean_uci timevar, sort pstyle(ci)) ///
(line s_mean timevar, sort) ///
, xlabel("Years from surgery") ///
ylabel("S(t)") ///
legend(off)
```

- (i) Obtain the adjusted survival curves by hormonal therapy status standardising over the covariate pattern of the whole study population. Use the `meansurv` option combined with the `at()` option.

```
. predict s_h0, meansurv at(hormon 0) timevar(timevar) ci
. predict s_h1, meansurv at(hormon 1) timevar(timevar) ci
. twoway (line s_h0 timevar, sort) ///
  (line s_h1 timevar, sort) ///
  , xtitle("Years from surgery") ///
  ytitle("S(t)") ///
  ylabel(0(.2)1,angle(h)) ///
  legend(order(1 "No hormonal therapy" 2 "hormonal therapy") ring(0) ///
  pos(1) cols(1)) name(adj1, replace)
```

- (j) Obtain the adjusted survival curves by hormonal therapy status standardising over the covariate pattern of those not on hormonal therapy.

```
. predict s_h0b if hormon==0, meansurv at(hormon 0) timevar(timevar) ci
. predict s_h1b if hormon==0, meansurv at(hormon 1) timevar(timevar) ci
. twoway (line s_h0b timevar, sort) ///
  (line s_h1b timevar, sort) ///
  , xtitle("Years from surgery") ///
  ytitle("S(t)") ///
  ylabel(0(.2)1,angle(h)) ///
  legend(order(1 "No hormonal therapy" 2 "hormonal therapy") ring(0) ///
  pos(1) cols(1)) name(adj2, replace)
```

- (k) Obtain the adjusted survival curves by hormonal therapy status standardising over the covariate pattern of those on hormonal therapy.

```
. predict s_h0c if hormon==1, meansurv at(hormon 0) timevar(timevar) ci
. predict s_h1c if hormon==1, meansurv at(hormon 1) timevar(timevar) ci
. twoway (line s_h0c timevar, sort) ///
  (line s_h1c timevar, sort) ///
  , xtitle("Years from surgery") ///
  ytitle("S(t)") ///
  ylabel(0(.2)1,angle(h)) ///
  legend(order(1 "No hormonal therapy" 2 "hormonal therapy") ring(0) ///
  pos(1) cols(1)) name(adj3, replace)
```

Explain why the curve in parts (j) and (k) look so different?

- (l) Now calculate and plot the difference in adjusted survival curves. To do this you can use the `predictnl` command. This is a very powerful postestimation Stata command that performs the delta method in order to obtain standard errors and 95% confidence intervals of complex functions of model parameters.

```
. predictnl sdiff = predict(meansurv at(hormon 0) timevar(timevar)) - ///
  predict(meansurv at(hormon 1) timevar(timevar)) ///
  , ci(sdiff_lci sdiff_uci)

twoway (rarea sdiff_lci sdiff_uci timevar, sort pstyle(ci)) ///
  (line sdiff timevar, sort) ///
  , xtitle("Years from surgery") ///
  ytitle("Difference in S(t)") ///
  legend(off)
```

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
. stsplitt 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
. stsplitt 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 several 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%.

181. 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 `stset`ting 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)
. stsplrit _age, at(0(1)110) trim
```

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

```
. stsplrit _year, after(time=d(1/1/1975)) at(0(1)22) trim
. replace _year=1975+_year
. list id _age _year in 1/15
```

Note that we have used the second syntax for `stsplrit` and set the origin for calendar period as 1/1/1975 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 m:1 _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 [10] 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)
. stsplrit _age, at(0(1)110) trim
. stsplrit _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)
```

182. Using `strs` for calculating SMRs

Stata addon required! This exercise requires the Stata user-written command `strs`. See Section 2.3 (page 6) for details and installation instructions.

The Stata command for estimating relative survival (`strs`) can also be used for calculating SMR. The primary purpose of this command is to estimate excess mortality (observed minus expected deaths). We can therefore use `strs` to estimate the numbers of observed and expected deaths for us and then calculate the ratio of these two numbers to get an SMR/SIR. That is, `strs` does the splitting and merging for us. We start by `stset`ting the data with time since diagnosis as the time scale.

```
. use melanoma, clear
. stset exit, fail(status == 1 2) origin(dx) entry(dx) scale(365.24) id(id)
```

Note that `strs` expects there to be a variable in the dataset containing age at diagnosis and calendar year at diagnosis (the default names for these variables are `age` and `yydx`) and uses these variables to keep track of current age and current year.

```
. strs using popmort, br(0(1)21) mergeby(_year sex _age) notables save(replace)
```

`strs` saves two datasets: one containing individual subject-band data (`individ.dta`) and one with grouped life-table data (`grouped.dta`). Either can be used for calculating the SMR and confidence interval but using the grouped data is faster.

We suppressed the printing of the lifetable in the previous step (using the `notables` option) but the same information is stored in `grouped.dta`.

```
. use grouped.dta, clear
. list start n d w p cp d_star, sum(d d_star)
```

The observed number of events is in the variable `d` and the expected number of events is in the variable `d_star`. Now we can sum these variables and take the ratio to obtain the SMR.

```
. collapse (sum) obs=d exp=d_star
. gen LL=( 0.5*invchi2(2*obs, 0.025)) / exp
. gen UL=( 0.5*invchi2(2*(obs+1), 0.975)) / exp
. gen smr=obs/exp
. list obs exp smr LL UL
```

The results obtained using `strs` are not identical to those obtained in question 181. This is because in question 181 we split on both current age and current year whereas `strs` only splits on age and assumes that current year increments by one each and every time current age increments by one. Using the `calyear` option with `strs` will cause it to also split on calendar year, thereby giving the exact same results as in question 181. Splitting on calendar year increases the computational time and is therefore not done by default.

The following code illustrates the differences.

- Splitting on both current age and current year.


```
. use melanoma, clear
. gen bdate = dx-(age*365.25)
. stset exit, fail(status == 1 2) origin(bdate) entry(dx) scale(365.24) id(id)
. stsplrit _age, at(0(1)110) trim

. stsplrit _year, after(time=d(1/1/1975)) at(0(1)22) trim
. replace _year=1975+_year

. sort _year sex _age
. merge m:1 _year sex _age using popmort
. drop if _merge==2
. gen mortrate=-ln(prob)
. strate, smr(mortrate)
```
- Splitting only on current age and approximating current year.


```
. use melanoma, clear
. gen bdate = dx-(age*365.25)
. stset exit, fail(status == 1 2) origin(bdate) entry(dx) scale(365.24) id(id)
. stsplrit _age, at(0(1)110) trim

. gen _year=year(dx)+(_age-age)

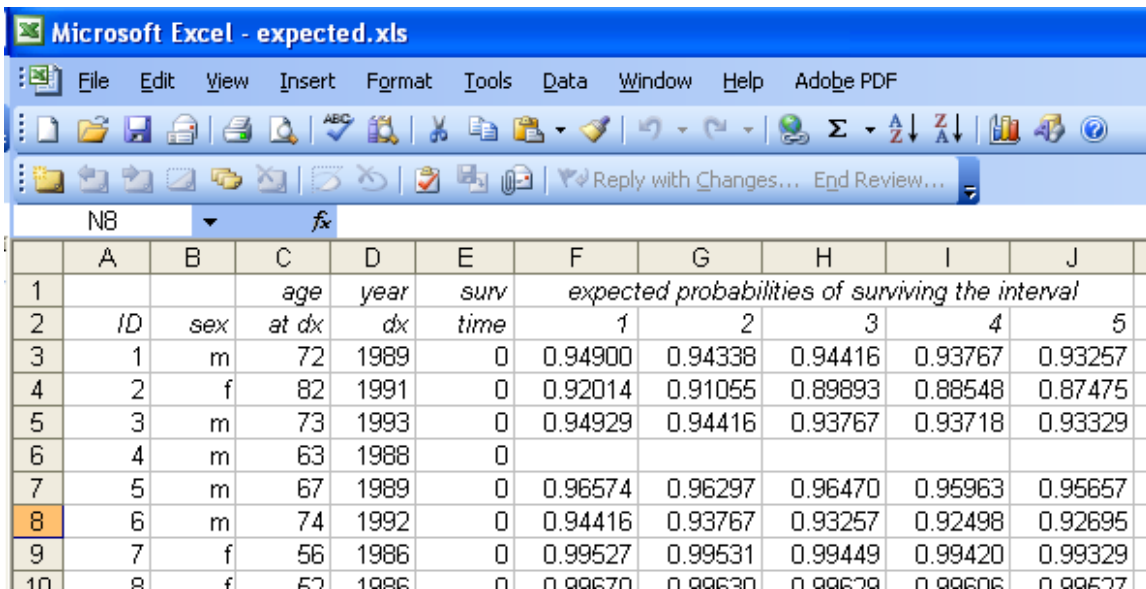
. sort _year sex _age
. merge m:1 _year sex _age using popmort
. drop if _merge==2
. gen mortrate=-ln(prob)
. strate, smr(mortrate)
```

If for example, a man is diagnosed with cancer in June 1990 aged 81 then the approximate approach assumes that for the entire first year of follow-up he will experience the 1990 rates and for the entire second year of follow-up he will experience the 1991 rates. The exact approach (splitting on both age and period) results in this man, during the second year of follow-up experiencing 1991 rates for 6 months and 1992 rates for 6 months. If population mortality is decreasing over time then the approximation will result in an overestimate of the expected number of deaths. This bias should not, however, be of practical significance.

	D	E	SMR	Lower	Upper
Exact approach	3047	1261.02	2.416	2.332	2.504
Approximation	3047	1268.08	2.403	2.319	2.490

200. Calculating expected survival by hand (with help from MS Excel)

The Microsoft Excel file `expected.xls` contains annual probabilities of survival for members of the general population matched to the sample of 35 patients diagnosed with colon carcinoma.



The screenshot shows a Microsoft Excel spreadsheet titled "Microsoft Excel - expected.xls". The spreadsheet has columns labeled A through J and rows numbered 1 through 10. The data is as follows:

	A	B	C	D	E	F	G	H	I	J
1			<i>age</i>	<i>year</i>	<i>surv</i>	<i>expected probabilities of surviving the interval</i>				
2	<i>ID</i>	<i>sex</i>	<i>at dx</i>	<i>dx</i>	<i>time</i>	1	2	3	4	5
3	1	m	72	1989	0	0.94900	0.94338	0.94416	0.93767	0.93257
4	2	f	82	1991	0	0.92014	0.91055	0.89893	0.88548	0.87475
5	3	m	73	1993	0	0.94929	0.94416	0.93767	0.93718	0.93329
6	4	m	63	1988	0					
7	5	m	67	1989	0	0.96574	0.96297	0.96470	0.95963	0.95657
8	6	m	74	1992	0	0.94416	0.93767	0.93257	0.92498	0.92695
9	7	f	56	1986	0	0.99527	0.99531	0.99449	0.99420	0.99329
10	8	f	57	1986	0	0.99570	0.99530	0.99579	0.99595	0.99577

These annual probabilities of survival were obtained by matching with the data in `popmort.dta` (a Stata data file).

- Five expected probabilities are listed for patient number 1. Locate the corresponding probabilities in `popmort.dta`.
- The expected probabilities are missing for patient number 4. Complete the missing cells using the data in `popmort.dta`.
- Calculate (using Excel) the five-year expected survival probabilities for the 35 patients using both the Ederer I and Ederer II methods.
- Confirm your results using `strs`.

```
. use colon_sample, clear
. gen id = _n
. stset surv_mm, failure(status==1 2) scale(12) id(id)
. strs using popmort, breaks(0(1)5) mergeby(_year sex _age) ///
    ederer1 list(n d w p cp_e1 cp_e2)
```

201. **Localised melanoma: life table estimates of relative survival by calendar period of diagnosis using `strs`**

Stata addon required! This exercise requires the Stata user-written command `strs`. See Section 2.3 (page 6) for details and installation instructions.

Use `strs` to calculate life table estimates of relative survival (Ederer II) for patients diagnosed with localised skin melanoma for each of the two calendar periods 1975–1984 and 1985–1994. Use annual intervals up to 10 years. Following is the required Stata code. It is strongly suggested that you run these commands from a do file.

```
. use melanoma if stage==1, clear
. stset surv_mm, fail(status==1 2) id(id) scale(12)
. strs using popmort, br(0(1)10) mergeby(_year sex _age) by(year8594) save(replace)
```

- (a) Confirm that you understand what each of columns represents and check, for example, that the relative survival is equal to the ratio of observed to expected survival. Confirm that cumulative relative survival is equal to the product of interval-specific survival (for the first couple of intervals).
 - i. During which point in the follow-up is excess mortality highest?
 - ii. Does the peak in excess mortality occur where you would expect from a biological/clinical perspective?
 - iii. Is there evidence that the patients reach a point of statistical cure?
- (b) Now estimate the life table using intervals of length 6 months rather than 1 year. Does the estimate of 10-year relative survival change a great deal as a result?
- (c) Now estimate the life table using intervals of length 3 months for the first year followed by intervals of 1 year. Does the estimate of 10-year relative survival change a great deal as a result?
- (d) Now estimate 20-year relative survival (using annual intervals). Why is it not possible to estimate 20-year survival for both periods?
- (e) Plot the estimates of cumulative relative survival for each calendar period against follow-up time (estimate survival using annual intervals up to at least 20 years).

```
. use grouped, clear
. twoway (connected cr end if year8594==0) (connected cr end if year8594==1)
```

Following is some alternative code for producing a graph (for you to copy and paste from the PDF file into the do file editor).

```
. twoway (scatter cr end if year8594==0, msymbol(0)) ///
        (scatter cr end if year8594==1, msymbol(0)) ///
        (rcap lo_cr hi_cr end if year8594==0, lcolor(black)) ///
        (rcap lo_cr hi_cr end if year8594==1, lcolor(black)) , ///
        yti("Relative Survival") yscale(range(0.4 1)) ///
        ylabel(0.4(0.2)1, format(%3.1f)) ///
        xti("Years from diagnosis") xla(0(2)20) ///
        legend(order(1 "1975-84" 2 "1985-94") ring(0) pos(7) col(1))
```

- (f) Plot the estimates of interval-specific relative survival for each calendar period against follow-up time.

```
. twoway (rcap lo_r hi_r end) (scatter r end), ///
  by(year8594, legend(off)) yti("Interval-specific RSR") ///
  xti("Years from diagnosis") xla(0(2)20) yla(0.8(0.1)1)
```

Revisit questions (i)–(iii) in part (a). Do you gain additional insight if you redraw the graph with narrower (e.g., 6 month) intervals? Note that to change the interval widths you will need to go back to the patient data and rerun `strs`.

- (g) Estimate and compare cumulative relative survival (stratified by calendar period) using three alternative methods for estimating expected survival (Hakulinen, Ederer I, and Ederer II).

```
. use melanoma if stage==1, clear
. stset exit, origin(dx) fail(status==1 2) id(id) scale(365.24)
. gen long potfu = date("31/12/1995", "DMY") /* "dmy" if using Stata 9 */
. strs using popmort if stage==1, br(0(1)20) mergeby(_year sex _age) ///
by(year8594) list(start n d w cr_e1 cr_e2 cr_hak) ederer1 potfu(potfu)
```

Are there large differences in the estimates of cumulative relative survival depending on the method used to estimate expected survival (Hakulinen, Ederer I, and Ederer II)? Study the differences at 1, 5, 10, 15, and 20 years subsequent to diagnosis.

- (h) Use the `pohar` option to get estimates of net survival using the Pohar Perme *et al.* [11] approach and compare the estimates to relative survival (Ederer II).

```
. use melanoma if stage==1, clear
. stset exit, origin(dx) fail(status==1 2) id(id) scale(365.24)
. strs using popmort, br(0(=1/12')20) mergeby(_year sex _age) ///
  by(year8594) pohar list(start n d w cr_e2 cns_pp) save(replace)
. use grouped, clear
. list start end cr_e2 cns_pp if mod(end,1)==0 & year8594, noobs
```

- (i) Use `strs` to generate a single lifetable for all patients diagnosed with localised skin melanoma (annual intervals up to 10 years). Use the `save` option, that will cause the results to be saved to `grouped.dta` (one observation for each life table interval) and `individ.dta` (one observation for each person for each life table interval).

- i. Use the data saved to `grouped.dta` to estimate observed, expected, and excess mortality rates for each life table interval. Before doing the calculations, consider how you expect expected mortality to vary as a function of time since diagnosis (i.e., do you expect it to increase, decrease, or remain constant). Was the pattern what you expected? We will investigate this further in part (iii) by looking at the individual data.
- ii. Plot the excess mortality rates as a function of time since diagnosis. Is the pattern similar to the pattern for cause-specific mortality (e.g., the graph you plotted in exercise 111f)?
- iii. Open the individual data and calculate the average age of the patients during each interval.

```
. use individ, clear
. collapse (mean) age _age, by(end)
. list
```

`_age` is the average age of the patients still at risk at the start of each interval and `age` is the average age of these same patients were at diagnosis. Here we see how the age distribution of the patients varies with follow-up and that the elderly are more likely to die earlier.

202. **Localised melanoma: life table estimates of cause-specific survival by calendar period of diagnosis using `ltable` and `strs`**

The purpose of this exercise is to illustrate how the `strs` command can be used for estimating cause-specific survival (results should be identical to those obtained from the `ltable` command) and to compare estimates of cause-specific to relative survival.

- (a) Use `strs` to estimate cause-specific survival (using annual life table intervals) for patients diagnosed with localised skin melanoma (a single life table for all patients).
HINT: To estimate cause-specific survival using `strs` you will need to define the failure event as death due to melanoma when you `stset` the data. The figures in the `CP` column will then be the cumulative cause specific survival. The estimates of relative survival (the `R` and `CR` columns) will be meaningless.
- (b) Repeat the previous exercise but instead use the `ltable` command. Confirm that both commands give identical estimates.
HINT: Use the online help if you are unsure of the command syntax. You will need to create an indicator variable for cause-specific death. Note that `ltable` is not an `st` command (so does not require or respect `stset`).
- (c) Compare the estimated relative survival ratios to the estimated cause-specific survival rates. Would you expect substantial differences? What could explain the differences?

203. Localised melanoma: period estimation of relative survival

Stata addon required! This exercise requires the Stata user-written command `strs`. See Section 2.3 (page 6) for details and installation instructions.

The commands in `q203.do` produce period estimates (using the window 01jan1994–31dec1995) of relative survival for each sex. Check that you understand each of the commands in the do file and compare the period estimates of survival with cohort estimates of relative survival.

204. **Localised melanoma: period estimation of relative survival using data up to the end of 1983**

Imagine it is early 1984. You have data for patients diagnosed and followed up until the end of 1983 and wish to predict the 5-year relative survival of newly diagnosed patients. We will calculate period estimates and cohort estimates and compare these to the actual survival of patients diagnosed in 1984.

- (a) Using the knowledge that patient survival is improving with time:
 - i. which do you expect will be higher; the period estimate or the cohort estimate of relative survival?
 - ii. which do you expect will be a better predictor of the survival of patients diagnosed in 1983; the period estimate or the cohort estimate?
- (b) Calculate the 5-year relative survival (Ederer II, annual life table intervals) using the traditional cohort method based on patients diagnosed 1975-1983 followed up until the end of 1983. Throughout this exercise you should specify the dates of diagnosis and exit when using `stset`.
- (c) Calculate the 5-year relative survival (Ederer II, annual life table intervals) using the traditional cohort method based on patients diagnosed 1977-1983 followed up until the end of 1983. That is, repeat the last part but without including patients diagnosed 1975 & 1976. Before performing the calculations, use your knowledge of how survival is changing over time to predict whether the estimated survival will be lower or higher than the estimate you obtained in the previous part.
- (d) Calculate the 5-year relative survival (Ederer II, annual life table intervals) using period analysis with a 'window' of 01jan1983-31dec1983.
- (e) Calculate the 5-year relative survival (Ederer II, annual life table intervals) using period analysis with a 'window' of 01jan1982-31dec1983. How would you expect the estimates to differ from the previous part?
- (f) Calculate the actual 5-year relative survival for patients diagnosed 1983.
- (g) Calculate the actual 5-year relative survival for patients diagnosed 1984.
- (h) Compare the estimates of 5-year relative survival (and associated standard errors) calculated in the previous parts. Did the relative magnitudes of the estimates conform to your expectations?
- (i) If relative survival does not vary according to calendar period of diagnosis, which of the following is most likely to be true?
 - i. the period estimate of relative survival will be lower than the cohort estimate.
 - ii. the period estimate of relative survival will be equal to the cohort estimate.
 - iii. the period estimate of relative survival will be higher than the cohort estimate.
- (j) **ADVANCED:** For each and every calendar year between 1981 and 1990, calculate the following and plot them on a graph (with calendar year on the X axis)
 - i. the most up-to-date cohort estimate of 5-year relative survival for patients diagnosed up until the end of the year.
 - ii. the most up-to-date period estimate of 5-year relative survival for patients diagnosed up until the end of the year.
 - iii. the actual 5-year relative survival for patients diagnosed during the subsequent year.

What do you conclude?

210. Model excess mortality using Poisson regression

We will now model relative survival (excess mortality) in patients diagnosed with localised melanoma as a function of follow-up time, sex, age at diagnosis, and period for the first 5 years of follow-up. The first step is to estimate relative survival for each combination of sex, age at diagnosis, and period which can be done by running the commands in `q210.do`. Be sure to study this file in the Stata do-file editor (or some other text editor) to ensure you understand the commands it contains.

We can model excess mortality using Poisson regression [12] using the following commands.

```
. use grouped if end < 6, clear
. glm d i.end i.sex i.year8594 i.agegrp, fam(pois) link(rs d_star) lnoff(y) eform
```

The `eform` option requests that the estimates be presented as the exponential of the estimated parameters (i.e. relative excess risks), rather than the estimated parameters.

- During which year of follow-up was excess mortality highest? Is this what you would expect?
- Compare the estimates (and their interpretation) to those obtained from the Cox model for cause-specific mortality (and also to the Poisson regression model for cause-specific mortality if you did it).
- Fit the model without the `eform` option and ensure you understand why some values in the output change and some do not. Note that you can obtain these estimates simply by typing

```
. glm
```

which results in Stata displaying the estimates from the last model estimated using the `glm` command.

- Test the assumption of proportional excess hazards across age groups by fitting an appropriate interaction term in the model.
- The model in the previous part was estimated based on collapsed data (`d`, `d_star` and `y` were summed across covariate patterns). Now fit the model to individual subject-band data by using the data set `individ` rather than `grouped`. Do the estimates change considerably? Would you expect them to?
- Now model relative survival using the full likelihood approach (Estève et al. [13]) and compare the estimates with the Poisson regression model estimated using individual data (the estimates should be identical).

```
. use individ if end < 6, clear
. ml model lf esteve (d=i.end sex year8594 i.agegrp)
. ml maximize, eform("RER")
```

- Now model relative survival using the Hakulinen-Tenkanen approach [14] and compare the estimates with previous models.

```
. use grouped if end < 6, clear
. glm ns i.end i.sex i.year8594 i.agegrp, fam(bin n_prime) link(ht p_star) eform
```

- Time permitting, fit similar models to the data for all stages (include stage in the model). Test the assumption of proportional excess hazards for stage.

211. Model excess mortality using Poisson regression with a smooth baseline

Stata addons required! This exercise requires the Stata user-written command `rcsgen`. Type `ssc install rcsgen` to install.

We will now extend the Poisson modelling approach by defining narrow intervals and fitting a smooth function (e.g., splines or fractional polynomials) of time for the excess hazard rate rather than a piecewise estimate [15, 16].

It is possible to use individual level data in these situations, but having many rows of data per subject can lead to very large file sizes and computation can be very slow (less so now than when this exercise was first written). If you are using categorical covariates then you can collapse the data over these covariates and follow-up interval. For large datasets this is particularly useful.

- (a) Load the melanoma data and use `strs` to split the time scale using 1 month intervals. Use the option `by(agegrp sex year8594)` as these are the covariates we are interested in modelling.

```
. use melanoma
. stset surv_mm, failure(status=1 2) scale(12) exit(time 120.5) id(id)
. strs using popmort, br(0(0.08333)6) mergeby(_year sex _age) ///
    by(agegrp sex year8594) notables ///
    savind(vnarrowint_ind, replace) ///
    savgroup(vnarrowint_grp, replace)
```

Open the `vnarrowint_ind` and `vnarrowint_grp` data file and compare the number of observations in each dataset.

For computational speed we will use the grouped data.

- (b) Create new variables for the average follow-up time for each interval, dummy variables for the covariates of interest and the restricted cubic spline variables (for the moment we will have knots at 0.05, 0.25, 0.75, 1.5, 2.5, and 4 years.

```
. use vnarrowint_grp, clear
. tab agegrp, gen(agegrp)
. gen female = sex == 2

. gen midtime = (start+end)/2
. rcsgen midtime, knots(0.05 0.25 0.75 1.5 2.5 4) gen(rcs)
```

Fit a proportional excess hazards model using restricted cubic splines for the baseline hazard.

```
. glm d rcs1-rcs5 agegrp2 agegrp3 agegrp4 female year8594 ///
    , family(poisson) link(rs d_star) lnoffset(y)
. estimates store M_sp_peh
```

Use the `glm`, `eform` option to obtain excess hazard ratios. Compare these to the estimated excess hazard ratios obtained in exercise 210 where follow-up time was modelled using a step function.

- (c) Calculate predicted values for the excess hazard rate and plot for the oldest and youngest groups (for males in the 1985-1994 period of diagnosis).

```
. predict lh, xb nooffset
. gen h=exp(lh)
. twoway (line h midtime if agegrp1 == 1 & female == 0 & year8594 == 1, sort) ///
        (line h midtime if agegrp4 == 1 & female == 0 & year8594 == 1, sort)
```

Use the `yscale(log)` to plot the excess hazard on the log scale. Why are these lines parallel?

- (d) Create dummy variables for the interaction between the age groups and the spline variables and then include these in the model to allow for non-proportionality of the excess hazards (for age group).

```
. forvalues i = 1/4 {
    forvalues j = 1/5 {
        gen age'i'rcs'j' = agegrp'i' * rcs'j'
    }
}

. glm d rcs1-rcs5 agegrp2 agegrp3 agegrp4 female year8594 ///
    age2rcs1-age2rcs5 age3rcs1-age3rcs5 age4rcs1-age4rcs5 ///
    ,family(poisson) link(rs d_star) lnoffset(y)
. lrtest M_sp_peh
```

Is there evidence of non-proportionality of the excess hazard rates?

- (e) Obtain estimates of the time-dependent excess hazard ratios for age group. Use the `partpred` command to obtain 'partial' predictions. Plot the excess hazard ratios against follow-up time.

```
. forvalue i = 2/4 {
    partpred agehr'i', for(agegrp'i' age'i'rcs1-age'i'rcs5) eform ///
    ci(agehr'i'_lci agehr'i'_uci)
}

. twoway (line agehr2 midtime if agegrp2==1, sort) ///
        (line agehr3 midtime if agegrp3==1, sort) ///
        (line agehr4 midtime if agegrp4==1, sort) ///
        ,legend(order(1 "45-59" 2 "60-74" 3 "75+") ring(0) pos(1) cols(1)) ///
        yline(1) ytitle(Excess Hazard Ratio) xtitle(Years from Diagnosis)
```

- (f) Now plot the time-dependent excess hazard ratio with 95% confidence intervals for the oldest age group (compared to the youngest).

```
. twoway (rarea agehr4_lci agehr4_uci midtime if agegrp4==1, sort pstyle(ci)) ///
        (line agehr4 midtime if agegrp4==1, sort) ///
        (function y = 2.9, range(0 6) lcolor(black) lpattern(dash)), ///
        legend(off) yline(1) ytitle(Excess Hazard Ratio)
```

(g) THIS IS AN ADVANCED QUESTION!

Use the code in the do file 211.do (part (j)) to obtain predicted relative survival curves with 95% confidence intervals. Compare the predicted relative survival estimates for the youngest and oldest age groups (for males in the the 1985-1994 period) to those obtained using life tables.

This code is fairly complex; relative survival is a function of the model parameters, $RSR = \exp(-\text{cumulative hazard})$, and we use the the delta method (see <http://www.stata.com/support/faqs/stat/deltam.html>) to obtain standard errors. The code could be improved by rewriting in Stata's matrix language, Mata.

If you complete this exercise you will gain an insight into why we prefer `stpm2`. There is a lot of code in this question, much of it advanced. If you have any questions about the code – please ask (Paul Lambert wrote the code for this part and is the best person to ask).

(h) An alternative to using splines is fractional polynomials. The following code fits a PEH model allowing an 3 degree fractional polynomial (FP3) model for the baseline excess hazard. Compare the excess hazard ratios with those obtain in part (b). Compare the fitted hazard rates with those obtained in part (c).

```
/* Note that this will take a few minutes to fit */
. mfp glm d midtime (agegrp2 agegrp3 agegrp4 female year8594) ///
  , family(poisson) link(rs d_star) lnoffset(y) eform ///
  df(4, midtime:6) alpha(-1) xorder(n)

. predict lh_fp, xb nooffset
. gen h_fp = exp(lh_fp)
```

(i) Fractional polynomials models can also be extended to time dependent effects. The following code will fit FP functions for the time-dependent effects. At least a linear effect will be incorporated for the time-dependent effect, but backward selection can be incorporated if desired using the `select` option.

```
. forvalues i = 1/4 {
  gen age'i'midtime = midtime*agegrp'i'
}

. mfp, df(4, midtime:6) alpha(-1) ///
xorder(n): glm d midtime ///
(agegrp2 agegrp3 agegrp4 female year8594) ///
, family(poisson) link(rs d_star) lnoffset(y) eform
```

(j) Using individual data takes a longer time, but the model estimates should be very similar. If your computer is fast enough, run the code in the solution Do file to fit a proportional excess hazards model. Compare the excess hazard ratios to those obtained in part (b).

230. Modelling excess mortality using flexible parametric models

Stata add-on required! This exercise requires the Stata user-written command `stpm2`. See Section 2.3 (page 6) for details and installation instructions.

We will now fit some models with the linear predictor on the log cumulative **excess** hazard scale, i.e. flexible parametric relative survival models (an extension of Royston-Parmar survival models).

- (a) Load the Melanoma data. We need to merge in the expected hazard rate at the time of death. This can be done making use of `stset`.

```
. use melanoma, clear
. stset surv_mm, failure(status=1 2) scale(12) exit(time 120.5) id(id)
. gen _age = min(int(age + _t),99)
. gen _year = int(yydx + _t)
. sort _year sex _age
. merge m:1 _year sex _age using popmort, keep(match master)
. tab agegrp, gen(agegrp)
. gen female = sex == 2
```

Use `stpm2` to fit a flexible parametric relative survival model (an extension of Royston-Parmar models) incorporating the expected mortality rate. We will start by ignoring the effects of covariates and use 3 degrees of freedom (2 internal knots). Use `predict` to obtain the predicted hazard and survival functions.

```
. stpm2, df(3) scale(hazard) bhazard(rate)
. predict h1, hazard per (1000) ci
. predict s1, survival ci
```

Use the data editor to look at the data. Notice that six new variables have been created. These are the spline variables, `_rcs1–_rcs3` and their derivatives, `_d_rcs1–_d_rcs3`.

- (b) Plot the predicted hazard and survival functions.

```
. twoway (rarea h1_lci h1_uci _t, sort pstyle(ci)) ///
        (line h1 _t, sort), name(h1)
. twoway (rarea s1_lci s1_uci _t, sort pstyle(ci)) ///
        (line s1 _t, sort), name(s1)
```

- (c) Now compare the estimated hazard and survival functions using 2 4 and 6 degrees of freedom.

```
. foreach df in 2 4 6 {
    stpm2, bhazard(rate) df(`df') scale(hazard)
    predict h_df`df', hazard ci
    replace h_df`df' = h_df`df' * 1000
    predict s_df`df', survival ci
    estimates store df`df'
}
. twoway (line h_df2 h_df4 h_df6 _t, sort lcolor(red blue black))
. twoway (line s_df2 s_df4 s_df6 _t, sort lcolor(red blue black))
```

Compare the models using the AIC and BIC. See exercise 131 for notes on AIC and BIC.

```
. estimates stats df2 df4 df6
```

Which is the best fitting model?

- (d) Fit a proportional excess hazards model (using 3 degrees of freedom for the baseline) including sex, age group and period of diagnosis.

```
. stpm2 agegrp2 agegrp3 agegrp4 female year8594, ///
      bhazard(rate) df(3) scale(hazard) eform
. predict h2, hazard per(1000) ci
. predict s2, survival ci
```

Compare the estimates hazard ratios with those from other proportional excess hazard models (piecewise, spline and fractional polynomial models).

- (e) Plot the predicted excess hazard rate for the oldest and youngest groups (for males in the 1985-1994 period of diagnosis).

```
. twoway (line h2 _t if agegrp1 == 1 & female == 0 & year8594 == 1, sort) ///
        (line h2 _t if agegrp4 == 1 & female == 0 & year8594 == 1, sort)
```

- (f) Extend the model by allowing time-dependent effects for age group.

```
. stpm2 agegrp2 agegrp3 agegrp4 female year8594, bhazard(brate) df(3) ///
      scale(hazard) tvc(agegrp2 agegrp3 agegrp4) dftvc(2)
. predict h3, hazard per(1000) ci
. predict s3, survival ci
```

Plot the predicted excess hazard rate for the oldest and youngest groups (for males in the 1985-1994 period of diagnosis). Note how these differ from those obtained from the proportional excess hazards model.

```
. twoway (line h3 _t if agegrp1 == 1 & female == 0 & year8594 == 1, sort) ///
        (line h3 _t if agegrp4 == 1 & female == 0 & year8594 == 1, sort)
```

- (g) Use the `hrnumerator()` and `hrdenominator()` options to obtain the time-dependent predicted hazard ratios for age group. Plot these for age groups 2, 3 and 4.

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

. twoway (line hr2 _t if agegrp2 == 1 & female == 0 & year8594 == 1, sort) ///
        (line hr3 _t if agegrp3 == 1 & female == 0 & year8594 == 1, sort) ///
        (line hr4 _t if agegrp4 == 1 & female == 0 & year8594 == 1, sort)
```

Plot the hazard ratio for age group 4 with 95% confidence intervals.

```
. twoway (rarea hr4_lci hr4_uci _t, sort pstyle(ci)) ///
        (line hr4 _t, sort), yline(1) ///
        xtitle("Years from Diagnosis") ///
        ytitle("Excess Mortality Rate Ratio")
```


- (h) Now obtain and plot the difference in estimated relative survival curves for the oldest versus the youngest age group (for males in the 1985-1994 period of diagnosis).

```
. predict sdiff4, sdiff1(agegrp4 1 female 0 year8594 1) ///
      sdiff2(agegrp4 0 female 0 year8594 1) ci
. twoway (rarea sdiff4_lci sdiff4_uci _t, sort pstyle(ci)) ///
      (line sdiff4 _t, sort), yline(0) legend(off) ///
      xtitle("Years from Diagnosis") ///
      ytitle("Difference in Relative Survival")
```

- (i) Now obtain and plot the difference in estimated excess mortality rates for the oldest versus the youngest age group (for males in the 1985-1994 period of diagnosis).

```
. predict hdiff4, hdiff1(agegrp4 1 female 0 year8594 1) ///
      hdiff2(agegrp4 0 female 0 year8594 1) ci
. replace hdiff4 = hdiff4*1000
. replace hdiff4_lci = hdiff4_lci*1000
. replace hdiff4_uci = hdiff4_uci*1000

. twoway (rarea hdiff4_lci hdiff4_uci _t, sort pstyle(ci)) ///
      (line hdiff4 _t, sort), yline(0) legend(off) ///
      xtitle("Years from Diagnosis") ///
      ytitle("Difference in Excess Mortality Rate")
```

231. **Modelling non-linear effects in relative survival I**
Proportional hazards models

- (a) Load the colon cancer data and merge in the background mortality rates as in question 230. Only keep those aged 90 and under at diagnosis.

```
. keep if age<=90
```

- (b) Fit a flexible parametric relative survival model with no covariates.

```
. stpm2 , scale(hazard) df(5) bhazard(rate)
```

It is possible to obtain ‘Martingale-like’ residuals from parametric models which can be used in a similar way to Cox models to assess the functional form of continuous covariates. Obtain the martingale-like residuals for this model and add a non-parametric smoother when plotting the residuals vs age.

```
. predict mg1, martingale
. lowess mg1 age, name(mg1, replace)
```

What does this tell you about the functional form needed to model age?

- (c) Now add `age` to the model assuming a linear effect. Interpret the coefficient for `age`.
 (d) We can decide on a reference age and the plot the excess mortality rate ratio as a function of age. Use the `partpred` command to get the predictions and confidence intervals with age 50 as the reference.

```
. partpred hr_age_lin, for(age) ref(age 50) ci(hr_age_lin_lci hr_age_lin_uci) eform
```

Plots the results and interpret.

- (e) Calculate a new set of Martingale-like residuals and plot vs age with a lowess smoother. What does this tell you about the assumption of linearity in your model?
 (f) Now use restricted cubic splines `splines` to model the effect of age. We will initially use 4 df. We will store the locations of the knots and the projection matrix for the orthogonalization for future use.

```
. rcsagen age, gen(rcsage) df(4) orthog
. matrix Rage = r(R)
. global knotsage 'r(knots)'
. stpm2 rcsage1-rcsage4, scale(hazard) df(5) bhazard(rate)
```

Calculate a third set of Martingale-like residuals and plot vs age. Have the splines captured the non-linearity in the data?

- (g) The parameters in the model are difficult to interpret individually, but we can still get useful predictions. Use the following code to obtain the predicted excess mortality rates and predicted relative survival functions for subjects aged 40, 60 and 80 at diagnosis.

```
. range temptime 0 5 200
. foreach age in 40 60 80 {
  rcsagen , scalar('age') rmatrix(Rage) gen(c) knots(\$knotsage)
. predict h'age', hazard ///
  at(rcsage1 '=c1' rcsage2 '=c2' rcsage3 '=c3' rcsage4 '=c4') ///
  timevar(temptime) per(1000)
. predict s'age', survival ///
  at(rcsage1 '=c1' rcsage2 '=c2' rcsage3 '=c3' rcsage4 '=c4') ///
  timevar(temptime)
}
```

Note the use of the `rcsgen` command. We need to know the value of the spline variables for the ages of interest so we make sure we use the same knot locations and projection matrix.

Plot the predicted excess hazard and relative survival functions. Which age group is the most different?

- (h) Obtain predicted 1-year relative survival with 95% confidence intervals as a function of age using the following code.

```
. gen t1 = 1
. predict s1, survival timevar(t1) ci
```

Plot this function (vs age) and interpret.

- (i) Repeat the above, but now obtain predicted 5-year relative survival as a function of age.
- (j) Calculate the 5-year relative survival conditional on survival to 1 year. (HINT Recall that survival to time t conditional on survival to time t_0 is $S(t)/S(t_0)$).

Plot the conditional relative survival versus age. Explain the similarity between 2-year, 5-year and the conditional relative survival.

Note that if you want to obtain confidence intervals for conditional relative survival you can use Stata's `predictnl` command that uses the delta-method with numerical derivatives.

```
. predictnl condsurv2 = predict(survival timevar(t5)) / ///
                        predict(survival timevar(t1)) ///
                        , ci(condsurv2_lci condsurv2_uci)
```

- (k) Obtain the excess mortality rate ratio as a function of age at diagnosis using the following code.

```
. rcsgen , scalar(50) rmatrix(Rage) gen(c) knots(\$knotsage)
. partpred hr_age_rcs, for(rcsage1-rcsage4) ///
  ref(rcsage1 '=c1' rcsage2 '=c2' rcsage3 '=c3' rcsage4 '=c4') ///
  eform ci(hr_age_rcs_lci hr_age_rcs_uci)
```

Plot this function (vs age) and compare to the graph you produced in part (d). Interpret the graph and explain why it was important to model the effect of age at diagnosis as a non-linear function.

- (l) Optional Extra Question. Investigate the effect of using a different number of knots to model the effect of age at diagnosis using restricted cubic splines. E.g. Try 3, 4 and 5 df using the `rcsgen` command. You can plot the different functions and also compare the fit of the models using AIC or BIC.

232. Modelling non-linear effects in relative survival II – Time-dependent effects

- (a) Load the colon cancer data and merge in the background mortality rates as in question 230. Remember to drop those aged over 90 years.
- (b) Fit a proportional excess hazards flexible parametric relative survival model using splines for the effect of age (see part (f) of question 231).

```
. stpm2 rcsage1-rcsage4, scale(hazard) df(5) bhazard(rate)
```

Use `estimates store peh` to save this model so we can later perform a likelihood ratio test.

- (c) Now fit a model with time-dependent effects for the (non-linear) effect of age. Use 2 df for the time-dependent effects. Perform a likelihood ratio test comparing this model with proportional excess hazards model.

```
. stpm2 rcsage1-rcsage4, scale(hazard) df(5) bhazard(rate) ///
      tvc(rcsage1-rcsage4) dftvc(2)
. estimates store timedep
. lrtest peh timedep
```

Is there evidence to reject the proportional excess hazards assumption?

- (d) Predict the excess mortality rates and predicted relative survival functions for subjects aged 40, 60 and 80 at diagnosis. Note that you can use exactly the same code as part (g) of question 231 since the `predict` command of `stpm2` will recognize which covariates have been specified as having time-dependent effects. Compare these graphs with those obtained in question 231. Are the functions different?
- (e) Obtain the 1 year relative survival as a function of age and plot. You can use exactly the same code as part (h) of question 231. Compare the estimates from the proportional excess hazards model and the model allowing time-dependent effects.
- (f) Obtain the 5 year relative survival as a function of age and plot. You can use exactly the same code as part (i) of question 231. Compare the estimates from the proportional excess hazards model and the model allowing time-dependent effects.
- (g) Obtain the 5 year relative survival conditional on survival to 1 year as a function of age and plot. You can use exactly the same code as part (j) of question 231. Compare the estimates from the proportional excess hazards model and the model allowing time-dependent effects. Why is the shape different to that obtained in question 231?
- (h) We can still summarize our data as excess mortality rate ratios, but now not only is the effect of age non-linear, the excess mortality ratios vary as a function of follow-up time. Using age 50 as the reference age calculate the time-dependent excess mortality rate ratios for subjects aged 40, 60, 70 and 80. The following code will help.

```
. rcsgen , scalar(50) rmatrix(Rage) gen(ref) knots(\$knotsage)
. foreach age in 40 60 70 80 {
  rcsgen , scalar('age') rmatrix(Rage) gen(c'age'_) knots(\$knotsage)
  predict hr'age', ///
    hrnum(rcsage1 '=c'age'_1' rcsage2 '=c'age'_2' ///
          rcsage3 '=c'age'_3' rcsage4 '=c'age'_4') ///
    hrdenom(rcsage1 '=ref1' rcsage2 '=ref2' ///
            rcsage3 '=ref3' rcsage4 '=ref4') ///
    timevar(temptime) ci
}
```

Interpret the excess mortality rate ratios.

- (i) By adapting the code in part (h) obtain differences the excess mortality rate for the same ages with age 50 as the reference age. Plot the results and think about how the relative effects seen in part(h) relate the absolute effects seen here.

Adapt the code once more to obtain differences in relative survival as a function of follow-up time for the same ages with age 50 as the reference age.

- (j) Perform a sensitivity analysis for the number of degrees of freedom you use to model the time-dependence, i.e. using the `dftvc` option. Try between 1 and 3 df and predict the time-dependent excess mortality rate ratio comparing a subject aged 70 with a subject aged 50. Is the time-dependent effects sensitive to the number of df? Unless you are an experienced Stata programmer you may want to look at the solution Do file to help you.

240. **Localised melanoma: age-standardised estimates of relative survival (for a single cohort using an internal standard)**

Stata addon required! This exercise requires the Stata user-written command `strs`. See Section 2.3 (page 6) for details and installation instructions.

- (a) By calculating a single life table with 1 year intervals, estimate the 10-year relative survival ratio (Ederer II method) for all patients diagnosed with localised melanoma 1975-1994.
- (b) Now calculate an age-standardised estimate of the 10-year RSR (Ederer II) using traditional direct standardisation using an internal standard **without** using the standardisation options in `strs`. That is, calculate a life table for each age group to obtain age-specific estimates and make a weighted average of these estimates (use the proportion in each age group as weights) to obtain the age-standardised estimate. You might wish to use MS Excel to help you perform the calculations where the worksheet might contain the following cells.

agegrp	cr_i	w_i	$cr_i \times w_i$
0-44			
45-59			
60-74			
75+			
Sum			

The age standardised estimate is then calculated as $\sum_i(cr_i \times w_i) / \sum_i w_i$.

Does the age-standardised estimate differ from the crude estimate? Did you expect it to differ?

- (c) Now recalculate the age-standardised 10-year RSR (traditional direct standardisation) using the standardisation options in `strs`. Verify that you get the same answer. HINT: Specifying the `standstrata()` option results in `strs` first producing stratified life tables for each level of the variables specified in `standstrata()` and then producing standardised estimates using the weights contained in the variable specified in the `iweights()` option. We therefore need to first generate a variable containing the weights. The weights must be specified as proportions so we will use the number first at risk in each age stratum divided by the total number first at risk as weights.

```
. local totalobs = _N
. bysort agegrp: gen standwei = _N/'totalobs'
. strs using popmort [iw=standwei], br(0(1)15)
      mergeby(_year sex _age) standstrata(agegrp)
```

- (d) Now calculate the age-standardised 10-year RSR (Ederer II) using the Brenner 'alternative method' [17] using the standardisation options in `strs`. The approach is identical to the previous part except we specify the `brenner` option.

Does the age-standardised estimate (alternative method) differ from the crude estimate? Did you expect it to differ?

- (e) In question 201g we estimated and compared cumulative relative survival using three alternative methods for estimating expected survival (Hakulinen, Ederer I, and Ederer II). Repeat this exercise but standardise by age (using traditional direct standardisation with an internal standard).

How does standardisation affect the magnitude of the differences between the 3 methods? Is this what you expected?

- (f) Now calculate the Pohar Perme estimate of relative survival. As this estimate does not need to be estimated separately in each age group, do not include the `by(agegrp)` option. You should make sure to split the time-scale into monthly intervals.

Compare the Pohar Perme estimates with the other estimates.

Plot the Pohar Perme estimates together with the Ederer II estimates.

241. **Localised melanoma: age-standardised comparisons of relative survival between two periods of diagnosis**

Stata addon required! This exercise requires the Stata user-written command `strs`. See Section 2.3 (page 6) for details and installation instructions.

The aim of this exercise is to compare 10-year relative survival (Ederer II) between the two calendar periods of diagnosis (1975–1984 & 1985–1994) whilst standardising for age. We will use both the traditional and the alternative methods for age standardisation. The age distribution for the first calendar period will be used as the standard.

- (a) **Crude estimates.** Use `strs` to estimate the 10-year relative survival ratio (Ederer II method, 1-year intervals) for each of the two calendar periods of diagnosis (1975–1984 & 1985–1994).
- (b) **Traditional method by hand.** Without using the standardisation options of `strs`, estimate age-standardised 10-year relative survival for patients diagnosed 1975–1984 (i.e., the early period) using the internal age distribution as the standard. That is, repeat exercise 240b with the only difference being that we are restricting to the first period of diagnosis.

Now estimate the age-standardised 10-year relative survival for patients diagnosed 1985–1994 (i.e., the latter period) using the age distribution of the early period as the standard.

- i. Did age-standardisation have a large impact (compared to the crude analysis)?
 - ii. Based on the stratum-specific survival estimates and stratum-specific weights would you expect age-standardisation to have a large effect on the comparisons?
 - iii. Under what conditions would you expect the comparison of age-standardised estimates to differ from the comparison of crude estimates?
- (c) **Traditional method using `iweights` option in `strs`.** Repeat the previous part using the `iweights` option in `strs` and confirm you get the same results.
- (d) **Alternative method using `iweights` and `brenner` options in `strs`.** Now repeat the exercise using the ‘alternative’ method of standardisation rather than the traditional method. How does the choice of method for standardisation affect the inference.
- (e) **Pohar Perme estimate.** The Pohar Perme estimate of relative survival is an internally standardized estimate. Explain the dangers of comparing the crude estimates between the two calendar periods.

Obtain the age standardized estimates of the Pohar Perme estimate of relative survival in the two calendar periods using 1975–1984 as the reference age.

How do these compare to the Ederer II estimates?

242. Age standardization using flexible parametric models

This question uses `stpm2` to obtain model based age standardized estimates of relative survival. The results should be compared to the life table based estimates from questions 240 and 241.

This question assumes familiarity with using flexible parametric models to model excess mortality and does not provide complete details of the required Stata code; we therefore suggest you look at question 230 before attempting this question if you are unfamiliar with flexible parametric models.

- (a) Load the melanoma data, keep those with localized melanoma, merge in the expected mortality at death/censoring and fit a flexible parametric relative survival model with no covariates using 5 df for the baseline.

Create a temporary time variable.

```
. range temptime 0 10 100
```

Predict the all age combined relative survival curve (use the option `timevar(temptime)`).

What is the estimated relative survival at 10 years post diagnosis? How does this compare to question 240(a).

- (b) Fit a proportional excess hazards model for age group. Use the `predict` command to estimate a relative survival curve for each age group. Plot these on a graph together with the all age estimate to compare them.
- (c) Use `tab agegrp` to calculate the the proportion of subjects in each age group. Calculate a weighted average of the four age specific relative survival curves to obtain the age standard relative survival curve. Note this is traditional direct standardization using an internal standard. However, now the relative survival curves are model-based rather than calculated separately in life tables.

Add the age standardized curve to the figure produced in part (b).

- (d) Compare the 10 year age standardized estimate of relative survival to that obtained in question 240.
- (e) Calculate age-standardized relative survival using the `meansurv` option.

```
. predict rs_stand2, meansurv timevar(temptime)
```

Check that the estimate is the same as the calculated in part (c)

- (f) You can calculate confidence intervals for the standardised estimate using the `ci` option. This uses the delta method (with numerical derivatives) and will be slow if you do not use the `timevar` option.

```
. predict rs_stand3, meansurv timevar(temptime) ci
```

Compare the confidence interval of the age standardized relative survival to that obtained in question 240.

We now compare two periods of diagnosis in a similar way to question 241. If the age distribution has changed then we need to standardize so that both time periods are forced to have the same age distribution. As for question 241, the age distribution for the first calendar period will be used as the standard

- (g) Fit a proportional excess hazards model for age group and calendar period. Predict and compare the 10 year relative survival in the eight groups.

- (h) Is there evidence that the age distribution has changed between the two calendar periods? If so, how has it changed?
- (i) Use the `meansurv` option to obtain the age standardized estimate of relative survival for each calendar period with the first calendar period as the reference.

```
. predict rs_7584 if year8594 == 0, meansurv at(year8594 0) timevar(temptime)
. predict rs_8594 if year8594 == 1, meansurv at(year8594 1) timevar(temptime)
```

Plot these on a graph and compare the 10 year age standardized relative survival to the estimates in question 241.

- (j) Now obtain the age-standardized estimate for 1985–1994 with this same calendar period as the reference. Add this to the graph in (i). Explain the differences between the curves.

243. **Localised melanoma: age-standardised estimates of relative survival (for a single cohort using an external standard)**

Stata addon required! This exercise requires the Stata user-written command `strs`. See Section 2.3 (page 6) for details and installation instructions.

This question is similar to question 240, but here we will standardise to an external population. For external standardisation it is common practice to standardise to a international standard population. There are different standard populations depending on cancer site (see [18].) The International Cancer Survival Standard (ICSS) for wide age-groups are given in the table below. Weights are also available for 5-year age groups.

Age Group	Weights		
	ICSS 1	ICSS 2	ICSS 3
0-44	0.07	0.28	0.60
45-54	0.12	0.17	0.10
55-64	0.23	0.21	0.10
65-74	0.29	0.20	0.10
75+	0.29	0.14	0.10

- (a) Calculate the age-standardised 5-year RSR (traditional direct standardisation - Ed-erer II method) using the standardisation options in `strs` for all patients diagnosed with localised melanoma 1975-1994. Use the age groups defined in the table above. HINT: to create the age-groupings and the internal weights to internally standardise, you can use the code below.

```
.recode age (min/44=1) (45/54=2) (55/64=3) (65/74=4) (75/max=5), gen(agegrpICSS)
. label define agegrpICSS 1 "0-44" 2 "45-54" 3 "55-64" 4 "65-74" 5 "75+"
. label values agegrpICSS agegrpICSS
. label variable agegrpICSS "Age groups for ICSS"
. local totalobs = _N
. bysort agegrpICSS: gen standwei = _N/'totalobs'
. label variable standwei "Internal age group weights"
```

- (b) For melanoma, we use the International Cancer Survival Standard (ICSS) 2 weights. Calculate the externally age-standardised 5-year RSR using the standardisation options in `strs` by using the ICSS 2 weights given in the table above.
- (c) Compare the estimates using the two different weights. Are they similar? Did you expect them to be?

HINT: look at the values for the two weights for the 5 specified age groups.

- (d) Repeat part (b) using the ICSS 1 weights instead.

What do you expect to happen to the standardised estimate when standardising to an older age distribution?

250. **Probability of death in a competing risks framework (life table relative survival)**

`strs` implements the approach proposed by Cronin and Feuer (2000) [19] for estimating the crude probability of death based on life table estimates of relative survival. We explore the life table approach in this question. Lambert *et al.* (2010) [20] subsequently showed how the estimates can be obtained after fitting a relative survival model, namely a flexible parametric models for relative survival, which use restricted cubic splines for the baseline cumulative excess hazard and for any time-dependent effects. The approach using flexible parametric models for relative survival is covered in question 251. Although the two approaches estimate the same quantity, the life table approach provides estimates for grouped data so we get an estimated probability for an age group rather than an estimate for a specific age as can be obtained in the model-based approach.

- (a) Load the Melanoma data, drop subjects diagnosed 1975–1984 and then use `strs` to obtain life-tables stratified by age group and sex. Use the `cuminc` option to obtain the crude probabilities of death due to cancer and due to other causes.
- (b) How is the probability of death due to all causes, `F`, calculated?
- (c) Why is the crude probability of death due to cancer, `ci_dc` similar to the all-cause probability of death for subjects aged 0-44?
- (d) For both males and females aged 60-74 what is the probability of death due to all-causes at 5 years post diagnosis? What two variables can be added together to give the probability of death due to all-causes?
- (e) What proportion of the all-cause deaths at 5 years post diagnosis are due to cancer and due to other causes for males? Compare these figures for the different age groups.
- (f) The age groups are fairly wide, explain how you would expect the crude probability of death due to cancer to differ between a 60 and 74 year old, even if the relative survival was identical.
- (g) Plot the net probability of death, the crude probability of death due to cancer and the overall probability of death for males by age group. Try to understand the relationship between these various measures.

251. Probability of death in a competing risks framework (relative survival model)

In exercise 250 we explored how one could estimate crude probabilities of death based on life table estimates of relative survival making use of the `strs` implementation of the approach proposed by Cronin and Feuer (2000) [19]. Lambert *et al.* (2010) [20] subsequently showed how the estimates can be obtained after fitting a relative survival model, namely a flexible parametric models for relative survival, which use restricted cubic splines for the baseline cumulative excess hazard and for any time-dependent effects. Although the two approaches estimate the same quantity, the life table approach provides estimates for grouped data so we get an estimated probability for an age group rather than an estimate for a specific age as can be obtained in the model-based approach.

- (a) Load the Melanoma data and merge in the background mortality rates as in question 230. Fit a flexible parametric relative survival model including age group with time-dependent effects.

```
. tab agegrp, gen(agegrp)
. stpm2 agegrp2-agegrp4, scale(hazard) bhazard(rate) df(5) ///
    tvc(agegrp2-agegrp4) dftvc(3)
```

Calculate the estimated net mortality (1 - relative survival) and plot the four curves on a single graph. Interpret the plot.

- (b) Use the `stpm2cm` command to estimate the crude probability of death. Note that `stpm2cm` will predict for individual covariate patterns and for ages at diagnosis. Perform the predictions for males aged 40, 55, 70 and 80 diagnosed in 1985. The prediction for a 40 year old (the first age group) can be obtained using,

```
. stpm2cm using popmort, at(agegrp2 0 agegrp3 0 agegrp4 0) ///
    mergeby(_year sex _age) ///
    diagage(40) diagyear(1985) ///
    sex(1) stub(cm1) nobs(1000) ///
    tgen(cm1_t)
```

Plot the estimated crude probability of death due cancer for each of the selected ages on the same graph. Contrast these with the estimated net probability of death from part (a).

- (c) Generate a similar plot but for the crude probability of death due to other causes.
- (d) A useful way of presenting crude probabilities is through stacked graphs. Generate the stacked graphs for each of the selected ages. Use the solution Do file for help.
- (e) Advanced: Now fit a model using splines for the effect age with the spline terms allowed to be time-dependent. Calculate the crude probabilities of death and compare these to the model where age is categorized.

260. Fitting cure models

Stata addon required! This exercise requires the Stata user-written command `strsmix`. See Section 2.3 (page 6) for details and installation instructions.

We will now apply cure fraction models [21, 22] to the colon cancer data. In this exercise we fit mixture cure models and in exercise 261 we fit flexible parametric cure models. The cure fraction models treat time as continuous and thus there is no need to split the time scale. However, the expected hazard (mortality) rate at the time of death (or censoring) is required. Use the following commands to merge in the expected mortality rate.

```
. use colon
. stset surv_mm, failure(status=1 2) scale(12) exit(time 120)
. gen _age = min(int(age + _t),99)
. gen _year = int(yydx + _t)
. sort _year sex _age
. merge m:1 _year sex _age using popmort, keep(match master)
```

The `scale(12)` option converts survival time to years. The `exit(time 120.5)` option creates a maximum follow-up time of 10 years (120 months).

- (a) Explain the purpose of the two `gen` statements in the above stata code.
- (b) Fit a mixture cure fraction model to those diagnosed between 1975-1984 using the following command.

```
. strsmix if year8594==0, dist(weibull) link(identity) bhazard(rate)
```

- i. What is the estimate of the cure fraction?

Use the following commands to obtain prediction of the relative survival curve and the survival distribution of the ‘uncured’ and then plot these estimates against time (`_t`)

```
. predict rs7584, survival
. predict rs7584u, survival uncured
```

- ii. Does the relative survival curve appear to reach a plateau at the cure fraction? Would you expect it to?
 - iii. Approximately what proportion of the ‘uncured’ group have died after 2 years?
 - iv. Approximately what is the median survival time of the ‘uncured’?
- (c) Repeat the above for those diagnosed between 1985–1994. Contrast the estimates for the two time periods.
 - (d) Now we will compare the two time periods more formally by including (`year8594`) as a covariate. First just allow the cure fraction to vary between time periods.

```
. strsmix year8594, dist(weibull) link(identity) bhazard(rate)
```

- i. What is the estimated difference in the cure fraction between the two time periods? Contrast this to the estimates obtained in b(i) and (c).

- ii. This model is making a fairly strong assumption regarding the survival distribution of the ‘uncured’ for the two periods. What is this assumption?

Now allow the two Weibull parameters (λ and γ) to vary between the two time periods.

```
. strsmix year8594, dist(weibull) link(identity) bhazard(rate)
      k1(year8594) k2(year8594)
```

- iii. What is the estimated difference in the cure fraction between the two time periods? Contrast this with d(i).
 - iv. Test the assumption that the survival distribution of the ‘uncured’ is the same for the two time periods.
- (e) Now fit a model including age group and time period of diagnosis using a logit link (use option `link(logit)`).
- i. Interpret the parameter estimates (you may want to display the exponentiated coefficients by using `strsmix, eform`).
 - ii. Obtain predictions of the median survival of the ‘uncured’.
Hint, use `predict med, centile` to obtain predicted values of the median.

261. **Fitting cure models using flexible parametric survival models**

Stata addon required! This exercise requires the Stata user-written command `stpm2`. See Section 2.3 (page 6) for details and installation instructions.

We will now apply flexible parametric cure models to the same data as in exercise 260, where we fitted mixture cure models. Read in the data, `stset` and merge on expected mortality rates in the same way as in exercise 260.

- (a) Compare the cure proportion in the two time periods by including the variable `year8594` as a covariate in the `stpm2` command. Assume proportional hazards.

```
. stpm2 year8594, df(6) bhazard(rate) scale(hazard) cure
```

- i. How do you interpret the coefficient for the effect of the time period?
- ii. Use the coefficients in the output to calculate the estimated cure proportions for the two time periods.
- iii. Predict the cure proportions using the `predict` command to check your calculations.

```
. predict cure1, cure
. list cure1 if year8594==0, constant
. list cure1 if year8594==1, constant
```

- iv. What is the estimated difference in the cure proportion between the two time periods? Compare this to the estimates obtained in exercise 260. Are the results similar? Would you expect them to be similar?
- v. Predict the median survival time of uncured. Is the median survival time the same in the two groups? Should it be?

```
. predict med1, centile(50) uncured
. list med1 if year8594==0, constant
. list med1 if year8594==1, constant
```

- (b) Now allow time-dependent effect.

```
. stpm2 year8594, df(6) tvc(year8594) dftvc(4) bhazard(rate) scale(hazard) cure
```

- i. How do you interpret the coefficient for the effect of the time period?
- ii. Use the coefficients in the output to calculate the estimated cure proportions for the two time periods.
- iii. Predict the cure proportions using the `predict` command to check your calculations.

```
. predict cure2, cure
. list cure2 if year8594==0, constant
. list cure2 if year8594==1, constant
```

- iv. Are the cure proportions similar to what was estimated in (a)?

- v. Predict the median survival time of uncured. Is the median survival time the same in the two groups? Should it be? Is the difference between the periods smaller or larger than in (a)? Why?
- ```
. predict med2, centile(50) uncured
. list med2 if year8594==0, constant
. list med2 if year8594==1, constant
```
- (c) Plot the estimated overall relative survival and the relative survival among uncured for the two periods. Do the survival curves reach a plateau? Should they?

280. **Constructing a popmort file for Stata using data from the Human Mortality Databases**

The Human Mortality Database (HMD) was created to provide detailed mortality and population data to researchers, students, journalists, policy analysts, and others interested in the history of human longevity. The project began as an outgrowth of earlier projects in the Department of Demography at the University of California, Berkeley, USA, and at the Max Planck Institute for Demographic Research in Rostock, Germany. The HMD currently maintains data from 37 different countries. In this exercise we will provide a step-by-step introduction to how country-specific popmort files can be downloaded from the HMD web site and converted into Stata format.

*The data in the HMD project are provided free of charge to all individuals who request access to the database. However, before gaining full access to the database, you must become a registered user, which requires accepting our user agreement and answering just a few questions.*

- Go to the HMD web-page <http://www.mortality.org> and complete the registration form for new users. After having completed the registration you will receive an e-mail with your user name and password.
- In this exercise we will create a popmort file for Sweden. In order to be directed to the Swedish raw data, go to the main home page and click on **Sweden**.
- Several types of data sets, with different levels of resolution, are generally available for each country. The country-specific data sets are documented by HMD and the meta data is available from the web-page. To create a popmort file of the same structure as those used throughout the course click on the cell that contains 1 x 1 period estimates of the death rates. *You will be asked to enter your username and password to open the text file that contains these rates.*

|                           | Available dates | Age interval |     |      |
|---------------------------|-----------------|--------------|-----|------|
|                           |                 | 1x1          | 1x5 | 1x10 |
| <b>Period data</b>        |                 |              |     |      |
| Births                    | 1749 - 2011     | 1-year       |     |      |
| Deaths                    | 1751 - 2011     | 1x1          | 1x5 | 1x10 |
| Deaths by Lexis triangles | 1751 - 2011     | Lexis        |     |      |
| Population size           | 1751 - 2012     | 1-year       |     |      |
| Exposure-to-risk          | 1751 - 2011     | 1x1          | 1x5 | 1x10 |
| Death rates               | 1751 - 2011     | 1x1          | 1x5 | 1x10 |
| Life tables               | 1751 - 2011     |              |     |      |
| Females                   |                 | 1x1          | 1x5 | 1x10 |

- (d) Look through the data file and make sure that you understand the structure of this file. The first few lines are reproduced below. Download the text file to your computer and save it at some convenient location. Note that you will, in a later step, need to type the file reference so saving it on the desktop may not be the best option.

Sweden, Death rates (period 1x1)      Last modified: 23-Aug-2012, (May07)

| Year | Age | Female   | Male     | Total    |
|------|-----|----------|----------|----------|
| 1751 | 0   | 0.212235 | 0.241105 | 0.226774 |
| 1751 | 1   | 0.049412 | 0.052949 | 0.051169 |
| 1751 | 2   | 0.032247 | 0.034587 | 0.033409 |
| 1751 | 3   | 0.026006 | 0.027883 | 0.026936 |
| 1751 | 4   | 0.023696 | 0.025692 | 0.024681 |
| 1751 | 5   | 0.018761 | 0.020801 | 0.019766 |
| 1751 | 6   | 0.012956 | 0.014635 | 0.013788 |

- (e) Before the data can be read into Stata you need to make sure that Stata will be able to recognize all numeric variables. In the Swedish data the death rates for people older than 109 years are denoted as 110+. Use the find and replace function (of a text editor or the Stata do file editor) to replace 110+ with 110. Also, remove the first line of text and the second blank line in the text file (but keep the variable names).
- (f) Read the data into Stata using the `infile` command (see below). Note that we restrict the file to years 1950 onwards and ages below 100.

```
. infile _year _age female male total using "Name of file.txt" ///
 if (_year > 1949 & _age < 100), clear
```

- (g) The rates for males and females are stored in separate columns. In the final popmort file, we would like to have the rates for males and females stacked on top of each other and a variable called `sex` that enables us to separate the rates. This can be achieved by running the following code:

```
. drop total
. rename male rate1
. rename female rate2
. reshape long rate, i(_year _age)
. rename _j sex
```

- (h) Next, we will use the mathematical relationship between mortality rates and survival probabilities to create a new variable, `prob`, that contains the survival probabilities for each covariate pattern (in addition to the mortality rates that were available in the raw data file).

```
. gen prob=exp(-rate)
```

- (i) The popmort file is now ready to be used in survival analysis. The following command can be used to add some final touches to the file. Do not forget to save the cleaned popmort file.

```
. label data "Swedish death rates from http://www.mortality.org/"
. label variable rate "Death rate"
. label variable prob "Survival probability"
. label variable _year "Year of death"
. label variable _age "Age"
. label variable sex "Sex"
. sort _year sex _age
. save popmort2011, replace
```

Voila!

## 281. Constructing a popmort file by modelling cohort data

This code illustrates how one can create a ‘popmort’ file using data from a cohort. If you are interested in a ‘standard’ popmort file then the first place to look is the Human Mortality Database or your national statistics office. The approach described here is for when those approaches don’t suffice. For example, if you wish to create a popmort file stratified by region or social class and you have access to a cohort of individuals followed up for death.

In this illustration we use data on cancer patients (because that is what we have) and use non-cancer mortality as the outcome. In a real life application, however, you would ideally have data on individuals from the general population and use all-cause mortality as the outcome.

```
. use colon, clear

/* We need a variable for date of birth */
. gen dob=dx-age*365.25
. format dob %d

/* stset using attained age as the timescale */
. stset exit, fail(status==2) enter(dx) origin(dob) scale(365.25) id(idnr)

/* graph the age-specific hazards for each sex */
. sts graph if _t < 100, haz by(sex) name(observed, replace) kernel(epan2) ///
 xscale(range(40 100)) xlabel(40(5)100) ///
 yscale(range(0 1.5)) ylabel(0(0.5)1.5)

/* Tabulate age-specific mortality rates */
. preserve
. stsplot attage, at(0(1)110)
. strate sex attage
. restore
```

Model the age-specific hazards for each sex. First using a proportional hazards model.

```
. stpm2 sex, df(5) scale(hazard)

/* Predict the hazards and plot by sex*/
. predict h, haz

. twoway (line h _t if sex==1 & _t>40,sort) ///
 (line h _t if sex==2 & _t>40, sort), ///
 name(fitted, replace) title("Fitted values from fpm (prop hazards)") ///
 xscale(range(40 100)) xlabel(40(5)100) ///
 yscale(range(0 1.5)) ylabel(0(0.5)1.5)

/* Plot the empirical and fitted hazards (FPM PH model) together */
. sts graph if _t < 100, haz by(sex) kernel(epan2) name(overlay1, replace) ///
 xscale(range(40 100)) xlabel(40(5)100) ///
 title("Empirical and fitted hazards (fpm PH)") ///
 addplot(line h _t if sex==1 & _t>40,sort ||) ///
 line h _t if sex==2 & _t>40, sort)
```

```

/* Same plot but on the log scale */
. sts graph if _t < 100, haz by(sex) kernel(epan2) name(overlay2, replace) ///
 xscale(range(40 100)) xlabel(40(5)100) ///
 title("Empirical and fitted hazards (fpm PH)") ///
 yscale(log) ylabel(0 0.01 0.1 0.5 1.5) ///
 addplot(line h _t if sex==1 & _t>40,sort || ///
 line h _t if sex==2 & _t>40, sort)

```

Now allow non-proportional hazards.

```

. stpm2 sex, df(5) scale(hazard) tvc(sex) dftvc(3)
. predict h2, haz

. sts graph if _t < 100, haz by(sex) kernel(epan2) name(overlay1_tvc, replace) ///
 xscale(range(40 100)) xlabel(40(5)100) ///
 title("Empirical and fitted hazards (fpm tvc)") ///
 addplot(line h2 _t if sex==1 & _t>40,sort || ///
 line h2 _t if sex==2 & _t>40, sort)

. sts graph if _t < 100, haz by(sex) kernel(epan2) name(overlay2_tvc, replace) ///
 xscale(range(40 100)) xlabel(40(5)100) ///
 title("Empirical and fitted hazards (fpm tvc)") ///
 yscale(log) ylabel(0 0.01 0.1 0.5 1.5) ///
 addplot(line h2 _t if sex==1 & _t>40,sort || ///
 line h2 _t if sex==2 & _t>40, sort)

```

Create a table of predicted rates and probabilities. These figures could be used as the popmort file for relative survival analysis. For a real-life example we would also stratify by calendar year and race (if applicable). Could also model rates as a function of social class or ethnicity. To estimate rates for a small population one could include data for a larger population to get the shape correct.

```

. range attage 40 100 61
. drop if attage==.

/* Use non PH model to predict hazards for these ages (both sexes)*/
. predict rate_m, hazard at(sex 1) timevar(attage)
. predict rate_f, hazard at(sex 2) timevar(attage)

/* Generate survival probabilities*/
. gen prob_m=exp(-rate_m)
. gen prob_f=exp(-rate_f)

. format rate_m rate_f prob_m prob_f %8.5f
. keep attage prob_m prob_f rate_m rate_f
. list

```

## 282. Calculating excess and ‘avoidable’ deaths from life tables

- (a) Load the Melanoma data, drop subjects diagnosed 1975-1984 and then use `strs` to obtain life-tables stratified by age group and sex. Load the grouped data and keep the following variables.

```
. keep start end n cp cp_e2 cr_e2 sex agegrp
```

- (b) What is the difference in five-year relative survival between males and females in each age group?
- (c) We will now investigate excess deaths and ‘avoidable’ deaths. The question of interest is how many fewer deaths we would expect to see if males could achieve the same relative survival as females. To do this we will reshape the data from long form to wide form to make calculations easier.

```
. bysort sex (agegrp start): gen j = _n
. gen sexlab =cond(sex==1, "_m", "_f")
. drop sex
. reshape wide start end n cp cp_e2 cr_e2 agegrp, i(j) j(sexlab) string
. rename agegrp_m agegrp
. rename start_m start
. rename end_m end
. drop agegrp_f start_f end_f
```

Look at the data in the data browser to make sure you understand what the `reshape` command has done.

- (d) In order to calculate the predicted number of deaths we need to define how many subjects were at risk at the the start of follow-up. For simplicity, we will use the average number of cases per year over the 10 year diagnosis period. This can be calculated as follows.

```
. bys agegrp: gen Nrisk_m = n_m[1]/10
```

Calculate the overall (all-cause) probability of death,  $1 - S^*(t)R(t)$ , for males.

```
. gen p_dead_m = 1 - cp_e2_m * cr_e2_m
```

For males, calculate the expected number of all-cause deaths, `Nd_m`, the expected number of deaths if the study population were free of cancer, `NExp_d_m` and the excess deaths associated with a diagnosis of cancer, `ED_m`.

```
. gen Nd_m = Nrisk_m*p_dead_m
. gen NExp_d_m = Nrisk_m*(1-cp_e2_m)
. gen ED_m = Nd_m - NExp_d_m
```

- i. How many all cause deaths would we expect to see in each age group at 5 years post diagnosis?
  - ii. How many more deaths are there than would be expected in a similar cancer free group in the population?
  - iii. How many excess deaths by 5 years are associated with a diagnosis of melanoma over all age groups?
- (e) Repeat the above calculations for females. How do the excess deaths for females compare to the males?

- (f) We will now apply the relative survival estimates for females to the males' expected survival in order to calculate the 'avoidable' deaths.

```
. gen Nd_m_f = Nrisk_m*(1 - cp_e2_m * cr_e2_f)
. gen AD_m = Nd_m - Nd_m_f
```

How many deaths would be avoided if males could achieve the same relative survival as females for Melanoma?.

- (g) List the avoidable deaths for the oldest age group over all follow-up times. Why are the number of avoidable deaths decreasing as follow-up time increases?

## 283. Simulating relative survival

In this exercise, the aim is to learn how to simulate relative survival data. We will go through this step-by-step and show the process in a fairly simple setting. To carry out the exercises you will need to install some user-written commands from within Stata. The `survsim` command is used to simulate survival data from a range of distributions in Stata.

```
. ssc install survsim
```

The process of generating relative survival data through simulation involves simulating two times of death from differing mortality rates. Firstly, we generate the cause-specific time of death; when patients would die in the hypothetical scenario where they can only die of the disease of interest. Secondly, we generate the time of death that patients would experience if they were only at risk of other causes of death (based on general population mortality rates). We then take the minimum of these two times to be our all-cause time of death.

- (a) Firstly, use the code below to set the observations to 1000 and create an age and year of diagnosis variable. Here, we use a normal distribution for age with a mean of 60 and a standard deviation of 13. This means that age-at-diagnosis will be a random draw from a normal distribution for these 1000 patients. We set year of diagnosis to 1990 for all patients and make them all males for simplicity. We also create an age-group variable using the standard 5 age-groups.

```
. clear
. set obs 1000
. gen ageddiag=floor(rnormal(60,13))
. gen yydx=1990
. gen sex=1
. egen agegrp=cut(ageddiag), at(0 45 55 65 75 110) icodes
. tab agegrp, gen(agegrp)
```

Plot a histogram for age to see the distribution for your sample.

- (b) We now need to use `survsim` to generate cause-specific survival times for our patients. These times will relate to times of death due to the disease of interest only meaning that we will calculate the time that the patient would have died of their disease for all patients. We will assume that the survival times follow a Weibull distribution with  $\lambda = 0.2$  and  $\gamma = 0.5$ . Use the code below to generate the survival times. We will also assume a proportional effect for categorised age. The hazard ratios for the 5 groups will be 0.8, 0.9, 1, 1.2 and 1.4 respectively. Remember that the baseline will be defined by the  $\lambda$  and  $\gamma$  above.

```
. survsim timere1, lambdas(0.2) gammas(0.5) ///
covariates(agegrp1 '=ln(0.8)' agegrp2 '=ln(0.9)' agegrp3 '=ln(1)' ///
agegrp4 '=ln(1.2)' agegrp5 '=ln(1.4)')
```

Use `twoway` function to plot the survival function for the 5 age-groups.

HINT: `twoway` function `y=exp(-0.2*x^0.5)`, `range(0 5)` will plot the Weibull survival curve for the baseline age-group.

- (c) Now we have the cause-specific times of death, we also need the time of death due to other causes for each patient. Using both these times, we can create an all-cause death time variable for our `stset`. We will use the population mortality file to draw survival times for our patients (assuming an exponential distribution); we will need to do this a number of times because our patients will age over follow-up and therefore their mortality rates will change as they age. Therefore, for each year of follow-up;



we will calculate an expected survival time and evaluate if the patient survives the year. We will let them age by a year and recalculate the times based on the new rates. Here, we do this 5 times so that we have a minimum of 5 years of follow-up.

```
. gen _age=.
. gen _year=.
. forvalues i=0/4 {
capture drop _merge prob
replace _age=agediag+'i'
 replace _year=yydx+'i'
quietly merge m:1 _age _year sex using popmort, keep(matched master)
gen timeback'i'=(-log(runiform()))/(-ln(prob))
replace timeback'i'=1 if timeback'i'>=1
}

. gen timeback=.
. forvalues i=0/4 {
quietly replace timeback=timeback'i'+i' if timeback'i'<1 & timeback==.
}

. drop timeback? _age _year _merge rate
```

`timeback` now contains the background survival time based on the population mortality rates. `timeback` is created from the five numbered `timeback` variables by adding a year for each interval a patient survives, or stopping the sum once the patient dies in any given interval.

- (d) We now take the minimum of the cause-specific survival times and the background survival times to obtain our all-cause survival time estimates. We censor patients at 5 years should they survive for 5 years or more for both time variables; creating a death indicator in the process.

```
. gen time=min(timeback,timerel)

. generate died = time <= 5
. replace time = 5 if died == 0

. stset time, failure(died = 1)
```

We can now `stset` our data as though we have the all-cause data that we would have in real-life. However, we now know the true values for the age-group specific survival curves from our Weibull parameters. We also know the true distribution of age is normal and so can calculate the proportion in each age-group using  $z$ -values. We can also create an indicator variable indicating whether each patient died of cancer or other causes; we have perfect cause of death information because we generated the data. We could have just fitted a cause-specific analysis to the initial survival times from `survsim`. However, it is also useful to be able to simulate data suitable for a relative survival analysis; using `stset` on the all-cause data, so that we can evaluate the properties of our relative survival estimation methods.

- (e) Now that we have generated the survival data, we can treat this as any normal dataset and perform the analysis that we wish to evaluate through simulation. As usual, we now must merge in the background mortality data before performing an excess mortality model for age using `stpm2`.

```
. gen _age = min(int(agediag + _t),99)
. gen _year = int(yydx + _t)

. quietly merge m:1 _age _year sex using popmort, keep(matched master)

. stpm2 ib2.agegrp, bhaz(rate) df(5) scale(hazard) ///
 eform
```

Compare the estimated hazard ratios to those you used in the generation of the survival data using `survsim`. They may well be quite different - this is a single run of the simulation with quite a small sample size.

- (f) We now have all of the components to create a Stata program to perform a simulation. We can run the simulation 1000 times, for instance, to evaluate the unbiasedness of our parameter estimates for the flexible parametric excess mortality model. See the end of the solution do file for how this can be done.

```
. simulate agegrp0=r(agegrp0) agegrp1=r(agegrp1) ///
 agegrp2=r(agegrp2) agegrp3=r(agegrp3) ///
 agegrp4=r(agegrp4), reps(1000): relsurvsim, hazratio(0.8 0.9 1 1.2 1.4)
```

## 284. Estimating loss in expectation of life

In this exercise the aim is to estimate the loss in expectation of life for the melanoma cohort as a function of age, year and sex. This can be used to estimate the total number of life years lost for a given cohort of cancer patients. We will also use loss in expectation of life as a way of quantifying the sex difference in melanoma survival, as an alternative to using avoidable deaths (exercise 282).

Loss in expectation of life, together with life expectancy in absence of cancer and life expectancy in presence of cancer can be estimated after fitting a flexible parametric model by using the `lifelost` option of the `predict` postestimation command after using `stpm2` to fit a model. All options used together with `lifelost` are described below:

---

|                                                       |                                                                                                                                                                                                                                                                                                      |
|-------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>mergeby(string)</code>                          | specifies the variables by which the file of general population survival probabilities is sorted.                                                                                                                                                                                                    |
| <code>diagage(name)</code>                            | specifies the variable containing age at diagnosis. Default is <code>diagage</code> .                                                                                                                                                                                                                |
| <code>diagyear(name)</code>                           | specifies the variable containing calendar year of diagnosis. Default is <code>diagyear</code> .                                                                                                                                                                                                     |
| <code>maxage(int 99)</code>                           | specifies the maximum age for which general population survival probabilities are provided in the using file. Probabilities for individuals older than this value are assumed to be the same as for the maximum age. Default is 99.                                                                  |
| <code>attage(name)</code>                             | specifies the variable containing attained age in the popmort file. This variable cannot exist in the patient data file. Default is <code>_age</code> .                                                                                                                                              |
| <code>attyear(name)</code>                            | specifies the variable containing attained calendar year in the popmort file. This variable cannot exist in the patient data file. Default is <code>_year</code> .                                                                                                                                   |
| <code>survprob(name)</code>                           | specifies the variable containing survival probabilities in the popmort file. This variable cannot exist in the patient data file. Default is <code>prob</code> .                                                                                                                                    |
| <code>using(string)</code><br><code>by(string)</code> | specifies the popmort file to be used for expected survival probabilities. specifies stratification variables. Survival probabilities are averaged for each combination of these variables and assumed the same within each combination. Can only be used together with the <code>grp</code> option. |
| <code>maxyear(int 2050)</code>                        | specifies the maximum age for which general population survival probabilities are provided in the using file. Probabilities for years beyond this value are assumed to be the same as for the maximum year. Default is 2050.                                                                         |
| <code>nodes(int 50)</code>                            | specifies the number of nodes to be used for the numerical integration. Default is 50.                                                                                                                                                                                                               |
| <code>tinf(int 50)</code>                             | specifies the end year used for the numerical integration. Both observed and expected survival is assumed to be 0 after this point. Default is 50.                                                                                                                                                   |
| <code>tcond(real 0)</code>                            | specifies the starting year used for the numerical integration. This is used to retrieve conditional estimates. Default is 0.                                                                                                                                                                        |
| <code>grp</code>                                      | specifies that average survival probabilities should be used, as opposed to individual probabilities. If this is used together with the <code>by</code> option, the average is calculated within each combination of the specified by variables.                                                     |
| <code>stub(string)</code>                             | stubname for estimated life expectancy in absence and presence of cancer.                                                                                                                                                                                                                            |

---

- (a) Load the melanoma data and `stset` the data for relative survival.

```
. use melanoma, clear
. gen patid = _n
. stset surv_mm, failure(status=1 2) scale(12) exit(time 120.5) id(patid)
```

- (b) Fit a flexible parametric model including year, age and sex. Include age and year as continuous variables using splines. Allow all covariates to have a time-dependent effect. Remember to merge on the expected mortality at the exit times.

```
. rcsngen age, df(4) gen(sag) orthog
. rcsngen yydx, df(4) gen(syr) orthog
. gen fem= sex==2

. gen _age = min(int(age + _t),99)
. gen _year = int(yydx + _t)
. sort _year sex _age
. merge m:1 _year sex _age using popmort, keep(match master) keepusing(rate)
. drop _age _year _merge

. stpm2 sag1-sag4 syr1-syr4 fem, scale(hazard) df(5) ///
 bhazard(rate) tvc(sag1-sag4 syr1-syr4 fem) dftvc(3)
```

- (c) We will now estimate the loss in expectation of life. To save time we don't estimate confidence intervals, although they can be obtained by removing the comments around the `ci` option. (NOTE! Don't attempt to run this with the `ci` option during the lab session. This would take more than an hour, and the only way to stop Stata is to force the program to shut down completely.)

```
. predict ll, lifelost mergeby(_year sex _age) diagage(age) ///
 diagyear(yydx) nodes(40) tinf(80) using(popmort) ///
 stub(surv) maxyear(2000) /*ci*/
```

- (d) Create a graph that shows how the loss in expectation of life varies over age, for males diagnosed in 1994.

```
. twoway (line ll age if sex==1 & yydx==1994, sort) , legend(off) ///
 scheme(sj) name(q41_d, replace) ytitle("Years", size(*0.8)) ///
 xtitle("Age at diagnosis", size(*0.8)) xlabel(, labsize(*0.7)) ///
 ylabel(0 5 10 15 20 25 30 35 40 45, labsize(*0.7) angle(0)) ///
 yscale(range(0 45))
```

- (e) List the life expectancy and the loss in expectation of life for someone aged 50, 60, 70 and 80 at diagnosis, both males and females. Also calculate the total number of life years lost among patients diagnosed in 1994.

```
. foreach age in 50 60 70 80
 foreach sex in 1 2
 list age sex yydx survexp survobs ll if age=='age' & ///
 sex=='sex' & yydx==1994, constant

. qui summ ll if yydx==1994
. display r(sum)
```

- (f) Now estimate the loss in expectation of life if male patients had the same mortality due to melanoma as female patients, but the expected survival of males.

```
. replace fem=1

. predict ll_alt, lifelost mergeby(_year sex _age) diagage(age) ///
 diagyear(yydx) nodes(40) tinf(80) using(popmort) ///
 stub(surv_alt) maxyear(2000) /*ci*/
```

- (g) How many life years could potentially be saved if males diagnosed in 1994 had the same survival from melanoma as female patients diagnosed in 1994?

```
. gen lldiff= ll-ll_alt
. summ lldiff if yydx==1994
. display r(sum)

. foreach age in 50 60 70 80
 list ll ll_alt lldiff age if sex==1 & age=='age' & yydx==1994, constant
```

## 285. Multiple imputation for missing covariate data

This exercise gives an introduction to approaches for dealing with missing covariate data in population-based cancer studies. Falcaro *et al.* (2015) [23] published a nice overview of methods for modelling net survival when covariate data are missing.

In this exercise we use the official Stata commands for missing data that were introduced with Stata release 12 (July 2011). Users of earlier Stata versions can perform the same analyses with the user-written commands `ice` and `mim`.

Mechanisms for missingness in survival analysis can be classified as follows:

- Missing completely at random (MCAR)
- Covariate-dependent missing at random (CD-MAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

Much of the literature on missing data uses three classifications (MCAR, MAR, and MNAR). Falcaro *et al.* (2015) [23] distinguish between MAR (where the probability of missingness may depend on known covariates or the outcome) and CD-MAR (where the probability of missingness may depend on known covariates, but **not** the outcome).

The analytic approach depends on the mechanism. Multiple imputation requires the assumption that the data are MAR (MCAR and CD-MAR are special cases of MAR). If the data are MNAR, then the problem is more difficult. We cannot use the observed data to identify the mechanism, we need substantive scientific knowledge of the processes that gave rise to the data. The validity of inference depends on the (untestable) assumptions inherent in the mechanism.

- (a) Load the colon data and merge in the background mortality information as in question 230. Tabulate `stage`. What proportion of observations have stage missing (i.e., unknown stage)?
- (b) Investigate the distribution of unknown stage across age group and gender. Are older patients more likely to have an unknown recorded stage?
- (c) Study the estimated relative survival as a function of age group and stage (including unknown).

```
stpm2 ib1.stage##i.agegrp , df(5) bhaz(rate) scale(hazard) eform nolog
predict survival, surv
```

```
line survival _t if stage==0, sort lpattern(dash) || ///
line survival _t if stage==1, sort || ///
line survival _t if stage==2, sort || ///
line survival _t if stage==3, sort ///
 legend(order(1 "Unknown" 2 "Localised" 3 "Regional" 4 "Distant")) ///
 by(agegrp)
```

Do you see anything notable about the survival of the patients with unknown stage?

- (d) Which mechanism do you feel is applicable for stage in these data? The following questions are relevant to arrive at an answer.
- i. Will the probability of missingness depend on survival time?
  - ii. Upon what known factors (if any) do you think missingness might depend?
  - iii. Do you think the probability of missingness might depend on factors that are not available in our data? Which factors?

Note that this question (and the next) cannot be answered based on the observed data. It requires knowledge of the process by which stage is assessed and recorded. Answer the question based on your knowledge of such processes in your own jurisdiction. Talk to a classmate or teacher if you don't have knowledge in this area.

- (e) Multiple imputation requires an assumption of MAR (i.e., it is not valid for MNAR). Do you think a MAR assumption is appropriate for the available data?

Note that we will proceed with multiple imputation, an approach that requires an assumption that stage is MAR. Stata will provide estimates irrespective of what we think about the MAR assumption.

- (f) Using the code below, fit a flexible parametric model for excess mortality with explanatory variables stage (localised as the reference category) and age group. For pedagogic simplicity, we will not include a stage by age interaction, despite the fact that the previous graph suggests it might be needed. This approach to analysing incomplete data is known as the 'missing indicator' approach (since individuals with missing stage are included in a separate category).

```
. stpm2 ib1.stage i.agegrp, df(5) bhazard(rate) scale(hazard) eform
```

- (g) Now change the coding of the variable `stage` so that the unknown category is coded as missing (`replace stage=. if stage==0`). Refit the model from the previous part, using the exact same code, and compare the number of observations and the estimated excess hazard ratios. This approach is known as the 'complete records' or 'complete case' approach for analysing incomplete data.

Both the missing indicator approach and the complete records approach are known to be biased unless the data are missing completely at random (a scenario which rarely occurs in practice). We will now explore an approach, multiple imputation, that is preferred over the naïve approaches.

- (h) Before asking Stata to impute missing values we need to set the data for multiple imputation (`mi set`) and declare those variables that contain missing values (stage in this example). Stata recommends that we also register the ‘regular’ variables. `_rcs*` and `_d_rcs*` are the spline variables (and their derivatives) created later by `stpm2`.

```
mi set flong
mi register imputed stage
mi register regular subsite agegrp sex
mi register passive _rcs* _d_rcs*
```

Before performing the multiple imputation, give your best guess of what you think the distribution of stage should be for each of the following three observations.

| id   | agegrp | sex    | subsite              | stage | _t        | _d |
|------|--------|--------|----------------------|-------|-----------|----|
| 2287 | 45-59  | Female | Coecum and ascending | .     | .04166667 | 1  |
| 3362 | 75+    | Female | Coecum and ascending | .     | 6.2083333 | 1  |
| 3501 | 75+    | Female | Coecum and ascending | .     | 10        | 0  |

That is, if you think each of the three stages are equally likely then specify 33.33% for each. If you think the only possibility for stage is distant then specify 100% for distant and 0% for the other two alternatives.

There is no correct answer to this exercise; the goal of the exercise is to provide you with insight into what multiple imputation does. In particular, rather than impute a single value of stage for each observation we impute the distribution of stage. This exercise, we hope, will also help you understand the role of the outcome in the imputation model. If you find yourself thinking along the lines ‘I cannot provide a reasonable guess of the missing value of stage without knowing treatment’ then you have identified a violation of the MAR assumption.

| id   | agegrp | _t   | _d | localised | regional | distant |
|------|--------|------|----|-----------|----------|---------|
| 2287 | 45-59  | 0.04 | 1  |           |          |         |
| 3362 | 75+    | 6.21 | 1  |           |          |         |
| 3501 | 75+    | 10.0 | 0  |           |          |         |

We now multiply impute missing values using chained equations. Theory dictates that the imputation model should contain the outcome. Recent research suggests this be specified using the event indicator and estimated cumulative hazard. We therefore generate the Nelson-Aalen estimate of the cumulative hazard and store it in the variable `H`. We set the seed so as to obtain reproducible results (i.e., so you get the same answers as in the solutions) but that step is not necessary in practice. Carpenter and Kenward [24, p. 55] suggest 30 imputations but we will use only 10 in order to save time.

```
sts gen H=na
set seed 29390
mi impute chained (mlogit) stage = i.subsite sex i.agegrp H _d, add(10)
```

What type of model has been fitted for the imputation model? What covariates are included in the imputation model?



- (i) Examine the imputed values for the three observations we previously tried to predict the distribution of stage. The variable `_mi_m` is zero for the observations in the original data and a sequential integer for the imputed observations.

```
. list id _mi_m agegrp sex stage _t _d if id==2287
. list id _mi_m agegrp sex stage _t _d if id==3362
. list id _mi_m agegrp sex stage _t _d if id==3501
```

How closely did your predictions match the distribution of imputed values?

- (j) We now refit the flexible parametric model, this time to the imputed data. The `mi estimate` command effectively fits the model to each of the 10 imputed data sets and then combines the resulting estimates. Since `stpm2` is not an official Stata command we need to specify the `cmdok` option to specify that the command is OK for use with imputed data. We also save the resulting estimates in order to make predictions in a later step.

```
mi estimate, dots cmdok saving(mi_stpm2,replace): ///
 stpm2 ib1.stage i.agegrp, df(5) bhaz(rate) scale(hazard) nolog
```

Now predict the survival function based on the fitted model.

```
mi predictnl survimp2 = predict(survival at(agegrp 2) timevar(_t)) ///
 using mi_stpm2
```

We need to include the `timevar()` option for technical reasons. The spline variables (`_rcs*`) are needed for the predictions, but these are not stored when `stpm2` is run with `mi`. By using the `timevar()` option with `predictnl` we force the spline variables to be recalculated.

- (k) Using the fact that the original data is still in the new dataset (with `_mi_m==0`), we can refit the model using the complete records approach and obtain predictions of survival. We will graphically compare the predicted relative survival curves from the imputation model and the complete records model for the 60 – 74 age-group.

```
stpm2 ib1.stage i.agegrp if _mi_m==0, df(5) scale(h) bhaz(rate)
predict surv, survival at(agegrp 2)

line surv survimp2 _t if stage==1 & _mi_m==0, sort || ///
line surv survimp2 _t if stage==2 & _mi_m==0, sort || ///
line surv survimp2 _t if stage==3 & _mi_m==0, sort ///
title("Predicted survival for agegrp==2 (60-74)") ///
legend(order(1 "Localised (Complete)" 2 "Localised (Imputed)" ///
 3 "Regional (Complete)" 4 "Regional (Imputed)" ///
 5 "Distant (Complete)" 6 "Distant (Imputed)")) ///
 name(imputed, replace)
```

- (l) If you want to try these methods further with this dataset then you can artificially create missing data. For example, to randomly make `agegrp` missing for 25% of the study population use,

```
. replace agegrp = . if runiform()<0.25
```

286. **Understanding frailty**

**Stata addon required!** This exercise requires the Stata user-written commands `survsim`, `stpm2`, and `moremata`; all of which can be installed using `ssc install -packagename-`. See Section 2.3 (page 6) for details and installation instructions. The `survsim` package for simulating survival data is described in exercise 283.

This question is a placeholder for some code written by Paul Lambert illustrating the concept of frailty. The code simulates survival data with frailty and illustrates how the estimated survival function is biased if one does not account for the frailty. The code illustrates some of the concepts discussed in the paper by Aalen and colleagues [25], which we recommend you read if you are interested in this topic.

In short, if you are interested in learning about frailty we recommend you read the paper [25] and look at the code.

## References

- [1] Dickman PW, Coviello E. Estimating and modelling relative survival. *The Stata Journal* 2015;**15**:186–215.
- [2] Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *The Stata Journal* 2009;**9**:265–290.
- [3] Andersson TML, Lambert PC. Fitting and modeling cure in population-based cancer studies within the framework of flexible parametric survival models. *The Stata Journal* 2012;**12**:623–628.
- [4] Lambert PC. Modeling of the cure fraction in survival studies. *The Stata Journal* 2007;**7**:351–375.
- [5] Hinchliffe SR, Lambert PC. Extending the flexible parametric survival model for competing risks. *The Stata Journal* 2013;**13**:344–355.
- [6] Pohar M, Stare J. Relative survival analysis in r. *Comput Methods Programs Biomed* 2006;**81**:272–278.
- [7] Clayton D, Hills M. *Statistical Models in Epidemiology*. Oxford: Oxford University Press, 1993.
- [8] Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 2002;**21**:2175–2197.
- [9] Sauerbrei W, Royston P, Bojar H, Schmoor C, Schumacher M. Modelling the effects of standard prognostic factors in node-positive breast cancer. german breast cancer study group (gbsg). *British Journal of Cancer* 1999;**79**:1752–1760.
- [10] Ulm K. A simple method to calculate the confidence interval of a standardized mortality ratio (SMR). *American Journal of Epidemiology* 1990;**131**:373–5.
- [11] Pohar Perme M, Stare J, Estève J. On estimation in relative survival. *Biometrics* 2012;**68**:113–120.
- [12] Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Stat Med* 2004;**23**:51–64.
- [13] Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine* 1990;**9**:529–538.
- [14] Hakulinen T, Tenkanen L, Abeywickrama K, Paivarinta L. Testing equality of relative survival patterns based on aggregated data. *Biometrics* 1987;**43**:313–325.
- [15] Lambert PC, Smith LK, Jones DR, Botha JL. Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine* 2005;**24**:3871–3885.
- [16] Remontet L, Bossard N, Belot A, Estève J, French network of cancer registries FRANCIM. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Stat Med* 2007;**26**:2214–2228.

- [17] Brenner H, Hakulinen T. Are patients diagnosed with breast cancer before age 50 years ever cured? *Journal of Clinical Oncology* 2004;**22**:432–438.
- [18] Corazziari I, Quinn M, Capocaccia R. Standard cancer patient population for age standardising survival ratios. *Eur J Cancer* 2004;**40**:2307–2316.
- [19] Cronin KA, Feuer EJ. Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. *Statistics in Medicine* 2000; **19**:1729–1740.
- [20] Lambert PC, Dickman PW, Nelson CP, Royston P. Estimating the crude probability of death due to cancer and other causes using relative survival models. *Stat Med* 2010;**29**:885 – 895.
- [21] Lambert PC, Dickman PW, Åsterlund P, Andersson TML, Sankila R, Glimelius B. Temporal trends in the proportion cured for cancer of the colon and rectum: a population-based study using data from the Finnish cancer registry. *International Journal of Cancer* 2007; **121**:2052–2059.
- [22] Lambert PC, Thompson JR, Weston CL, Dickman PW. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics* 2007;**8**:576–594.
- [23] Falcaro M, Nur U, Rachet B, Carpenter JR. Estimating excess hazard ratios and net survival when covariate data are missing: strategies for multiple imputation. *Epidemiology* 2015;**26**:421–428.
- [24] Carpenter JR, Kenward MG. *Multiple imputation and its application*. Chichester: John Wiley & Sons, 2013.
- [25] Aalen OO, Valberg M, Grotmol T, Tretli S. Understanding variation in disease risk: the elusive concept of frailty. *Int J Epidemiol* 2014;.