Survival analysis
Computing notes and Exercises

Summer School on Modern Methods in Biostatistics and Epidemiology
Cison di Valmarino, Treviso, Italy
20–25 June, 2005

`http://www.bioepi.org/`

Contents

1 Notes on survival analysis using Stata

In order to analyse survival data it is necessary to specify (at a minimum) a variable representing survival time and a variable specifying whether or not the event of interest was observed (called the failure variable). Instead of specifying a variable representing survival time we can specify the entry and exit dates.

In many statistical software programs (such as SAS), these variables must be specified every time a new analysis is performed. In Stata, these variables are specified once using the `stset` command and then used for all subsequent survival analysis (`st`) commands (until the next `stset` command). For example

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

The above code shows how we would `stset` the skin melanoma data in order to analyse cause-specific survival with survival time in completed months (`surv_mm`) as the time variable. The variable `status` takes the values 0=alive, 1=dead due to cancer, and 2=dead due to other causes. We have specified that only `status=1` indicates an event (death due to melanoma) so Stata will consider observations with other values of `status` as being censored. If we wanted to analyse observed survival (where all deaths are considered to be events) we could use the following command

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

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

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

Once the data have been `stset` we can use any of these commands without having to specify the survival time or failure time variables. For example, to plot 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 Exercises using Stata

1. 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.

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

2. Use Stata to confirm the results you obtained in question 1. After starting Stata, you will first have to specify the data set you wish to analyse, that is

   ```
   . use colon_sample
   ```

   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`.

   ```
   . generate csr_fail=0
   . replace csr_fail=1 if status==1
   ```

   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 or the number of censored observations on the Kaplan-Meier plot

   ```
   . sts graph, atrisk
   . sts graph, lost
   ```

   Titles and axis labels can also be specified.

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

3. For the 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
   . keep if stage == 1
   . stset surv_mm, failure(status==1)
   . sts graph, by(year8594)
   ```

   (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)
   ```

   At what point in the follow-up is mortality highest? Does this pattern seem reasonable)? Is this pattern apparent when looking at the plot of the survivor function?

4. In question 3 we studied plots of the survivor function for patients diagnosed with localised skin melanoma by calendar period of diagnosis. 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
   ```

5. Let's now read the melanoma data again, but study all stages.

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

   (a) Plot estimates of the survivor function and hazard function by stage. Does it appear that stage is associated with survival?

   (b) Estimate the mortality rates for each stage using, for example, the `strate` command. What are the units of the estimated rates?

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

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

6. **Diet data: tabulating incidence rates and modelling with Poisson regression** Load the diet data and stset the data using time-on-study as the timescale.

```
. use diet, clear
. stset dox, id(id) fail(chd) origin(doe) scale(365.25)
```

(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
```

(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 $1500, 2500, 3000, 4500$.

(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. These codes are called the levels of the variable.

(e) To find the rate for different levels of eng3 try

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

The option graph will show a graph of rate against levels of exposure.

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

(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
```

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

(j) Use xi: poisson to compare the second and third levels with the first, creating the indicators automatically with i.eng3.

(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). Confirm your answer using stptime.

7. **Localised melanoma: model cause-specific mortality with Poisson regression** Now let's study cause-specific survival (more acurately, cause-specific mortality) of patients diagnosed with localised (`stage==1`) melanoma. The following commands can be used to load and `stset` the data.

```
. use melanoma, clear
. keep if stage == 1
. gen id=_n /* we need to generate an ID variable */
. stset surv_mm, failure(status==1) scale(12) id(id)
```

(a) Later we will use Cox regression to analyse these data. For now we will tabulate mortality rates and model them using Poisson regression. We expect mortality to depend on time since diagnosis so 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
```

(b) 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?

(c) 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?

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

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

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

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

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

What conclusions can you draw from the fitted model? Is there evidence that the effect of follow-up is confounded by age, sex, and calendar period?

(f) Is the effect of sex modified by calendar period? Fit an appropriate interaction term to test this hypothesis.

8. **Localised melanoma: modelling cause-specific mortality using Cox regression** In the previous question 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.

   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

   ```
   . use melanoma
   . keep if stage == 1
   . stset surv_mm, failure(status==1)
   . stcox year8594
   ```

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

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

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

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

   Compare these two test statistics with each other and with the log rank test statistic (which also has a χ^2_1 distribution) calculated in question 4. 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.

   ```
   . xi: stcox sex year8594 i.agegrp
   ```

   i. Interpret the estimated hazard ratio for the parameter labelled `Iagegr_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 _Iagegrp_1 _Iagegrp_2 _Iagegrp_3
   ```

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

   ```
   . xi: 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 8c is similar to the model estimated in question 7e.

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

9. **Examining the proportional hazards hypothesis (localised melanoma)**
Parts (a)–(d) deal with log cumulative hazard plots which were not covered in the lectures. Attempt them if you like or start at part (e).

(a) For the melanoma data, plot the log cumulative hazard function for each calendar period. The following command can be used

```
. use melanoma
. keep if stage == 1
. stset surv_mm, failure(status==1)
. stphplot, by(year8594)
```

Do you think that a proportional hazards assumption is appropriate for these data?

(b) Does the appropriateness of the proportional hazards assumption have any implications for the log rank test?

(c) From the plot, estimate the hazard ratio for patients diagnosed 1985–94 to those diagnosed 1975–84.

(d) Compare the estimated hazard ratio with the one from the fitted Cox model with period as the only explanatory variable. Should the estimates be similar? Are they similar?

(e) Fit the model containing sex, period, and age and test the assumption of proportional hazards.

```
. xi: stcox sex year8594 i.agegrp, schoen(sch*) scaledsch(schs*)
. stphtest, detail
```

Is an assumption of proportional hazards appropriate?

(f) Use graphical methods to explore the assumption of proportional hazards by age. For example,

```
. stphplot, by(agegrp)
. sts graph, hazard by(agegrp)
. stphtest, plot(_Iagegrp_3)
```

What do you conclude?

(g) Use time-varying covariates to 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?

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

10. **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)
    . keep if stage==1
    . xi: stcox sex year8594 i.agegrp
    ```

    (a) Interpret the estimated hazard ratio for the parameter labelled `Iagegr_2`, 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?

11. **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 appear to a difference in survival between males and females?

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

    (d) Decide on a 'most appropriate' model for these data. Be sure to evaluate the proportional hazards assumption.

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

    (a) Fit the following Poisson regression model to the diet data (we fitted this same model in question 6).

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

    Now fit the following Cox model.

    ```
    . stset dox, id(id) fail(chd) origin(doe) scale(365.25)
    . 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?

13. **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, §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, and

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

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

(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(365.25) id(id)
. strate sex, per(1000)
. streg sex, dist(exp)
```

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

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

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

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

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

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

To see the effect on couple 3

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

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

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

```
. streg brv
```

(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.

3 Splitting on two time scales and calculating SMRs using Stata

The standardised mortality ratio (SMR) is the ratio of the number of deaths observed in the study population to the number that would be expected if the study population had the same specific rates as the standard population. It is an indirectly standardised rate. When studying incidence we estimate the standardised incidence ratio (SIR) in the same manner. 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 in the study population and compare this to the expected rate for the standard population (estimated from general population rates and assumed to be fixed rather than random).

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 iodising radiation.

The following example illustrates how the calculations can be performed in Stata. It is not a scientifically realistic example. Our aim is to estimate SMR for the diet data and we are provided with a file containing hypothetical general population mortality rates stratified by age and calendar period.

To study how rates vary jointly on two-time scales we need to split the records twice.

1. Start by breaking the follow–up for the `diet` data into 5 year age bands

   ```
   . use diet, clear
   . stset dox, fail(chd) origin(dob) entry(doe) scale(365.25) id(id)
   . stsplit ageband, at(30(5)70) trim
   ```

2. Now break these new records into 5 year calendar period bands using

   ```
   . stsplit period, after(time=d(1/1/1900)) at(50(5)80) trim
   . replace period=period+1900
   . list id ageband period in 1/20
   ```

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

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

   ```
   . gen _y=_t - _t0 if _st==1
   . table ageband period, c(sum _d)
   . table ageband period, c(sum _y) format(%5.1f)
   ```

4. To make a table of rates per 1000 by `ageband` and `period`, try

   ```
   . gen obsrate=_d/_y*1000
   . table ageband period [iw=_y], c(mean obsrate) format(%5.1f)
   ```

5. To calculate the expected cases for a cohort, using reference 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. Since there are no suitable rates for these data, we have used reference rates which are constant (11 deaths per 1000 person-years) for all combinations of age and calendar period.

   The reference rates are in the file `ref`. Before calculating the expected number of cases it is necessary to add the reference rates to the expanded data with

   ```
   . sort ageband period
   . merge ageband period using ref
   ```

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

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

   ```
   . 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

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

6. 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*refrate/1000
   . list id e _d if in 1/20
   ```

   The SMR is the ratio of the total observed cases to the total number expected, and is most easily obtained with

   ```
   . strate, smr(refrate) per(1000)
   ```

   or with

   ```
   . smrby _d e
   ```

7. To calculate the SMR for the high and low energy groups,try

   ```
   . strate hieng, smr(refrate) per(1000)
   ```

   or

   ```
   . smrby _d e, by(hieng)
   ```

4 The Finnish Cancer Registry

The Finnish Cancer Registry is population-based and covers the whole of Finland (population 5.1 million). The Registry was established in 1952, with 1953 being the first calendar year with complete registration. The Registry obtains information from many different sources: hospitals and other institutions with inpatient beds, physicians working outside hospitals, dentists, and pathological and cytological laboratories. The Finnish Cancer Registry also receives copies of all death certificates where cancer is mentioned. Notification of new cancer cases to the Cancer Registry is mandatory by law. If the reported information is deficient or contradictory, requests are sent to informants in order to ensure accuracy in the following areas: patient details, the primary site of the tumour, and the date of diagnosis.

The diseases registered at the Finnish Cancer Registry include, in addition to all clearly malignant neoplasms, carcinoma in situ lesions (except those of the skin), all neoplasms of the intracranial space and spinal canal irrespective of their malignancy, benign papillomas of the urinary organs, semimalignant tumours of the ovary, basal cell carcinomas of the skin, and cases of polycythaemia vera and myelofibrosis. Various check-ups have shown that the coverage of the Cancer Registry file is almost complete with respect to cancer cases diagnosed in the Finnish population [1, 2]. All independent primary neoplasms in the same person are registered separately. When evaluating whether a new tumour is an independent cancer or a recurrence, attention is focused on, among other aspects, the time interval between the tumours, histology, and knowledge of the general behaviour of each cancer type. In principle, multiple metachronous tumours in the same organ (e.g., in the colon or skin) are registered separately, especially when they have different histologies. However, each case is evaluated individually and a primary site code ‘multiple cancer’ is also available for some organs. The International Classification of Diseases Volume 7 (ICD-7) is used at the Finnish Cancer Registry. Further details of the registry can be found in the annual incidence publications [3].

The Finnish Cancer Registry has kindly provided data on patients diagnosed with skin melanoma and colon carcinoma in Finland during 1975–1994 with follow-up to 31 December 1995. A detailed description and analysis of these data is given in Dickman et al. (1999) [4].

References

[1] Teppo L, Pukkala E, Lehtonen M. Data quality and quality control of a population-based cancer registry. Experience in Finland. *Acta Oncologica* 1994;**33**:365–369.

[2] Hakulinen T. Health care system, cancer registration and follow-up of cancer patients in Finland. In: Berrino F, Sant M, Verdecchia A, Capocaccia R, Hakulinen T, Estève J, eds., *Survival of Cancer Patients in Europe: The EUROCARE Study*, IARC Scientific Publications No. 132. Lyon: International Agency for Research on Cancer, 1995; 53–54.

[3] Finnish Cancer Registry. *Cancer Incidence in Finland 1995*. Cancer Society of Finland Publication No. 58. Helsinki: Cancer Society of Finland, 1997.

[4] Dickman PW, Hakulinen T, Luostarinen T, Pukkala E, Sankila R, Söderman B, Teppo L. Survival of cancer patients in Finland 1955-1994. *Acta Oncologica* 1999;**38 (Suppl. 12)**:1–103.