Biostatistics III: Survival analysis for epidemiologists
Solutions to exercises

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

Karolinska Institutet
6–15 March, 2017
http://biostat3.net

Exercise solutions

1.(a) **Life table and Kaplan-Meier estimates of survival**

The results are contained in the Excel file `\solutions\exercise1.xls` and are also shown in the Stata output below.

(b) Following are the life table estimates. Note that in the lectures, when we estimated all-cause survival, there were 8 deaths in the first interval. One of these died of a cause other than cancer so in the cause-specific survival analysis we see that there are 7 'deaths' and 1 censoring (Stata uses the term 'lost' for lost to follow-up) in the first interval.

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

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
0	12	35	7	1	0.7971	0.0685	0.6210	0.8977
12	24	27	1	3	0.7658	0.0726	0.5856	0.8755
24	36	23	5	4	0.5835	0.0901	0.3887	0.7356
36	48	14	2	1	0.4971	0.0953	0.3023	0.6647
48	60	11	0	1	0.4971	0.0953	0.3023	0.6647
72	84	10	0	3	0.4971	0.0953	0.3023	0.6647
84	96	7	0	1	0.4971	0.0953	0.3023	0.6647
96	108	6	1	4	0.3728	0.1292	0.1403	0.6091
108	120	1	0	1	0.3728	0.1292	0.1403	0.6091

```
. stset surv_mm, failure(status==1)
[output omitted]
```

Following is a table of Kaplan-Meier estimates. Although it's not clear from the table, the person censored (lost) at time 2 was at risk when the other person dies at time 2. On the following page is a graph of the survival function.

```
. sts list

         failure _d:  status == 1
   analysis time _t:  surv_mm
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
2	35	1	1	0.9714	0.0282	0.8140	0.9959
3	33	1	0	0.9420	0.0398	0.7873	0.9852
5	32	1	0	0.9126	0.0482	0.7528	0.9709
7	31	1	0	0.8831	0.0549	0.7178	0.9545
8	30	1	0	0.8537	0.0605	0.6835	0.9364
9	29	1	0	0.8242	0.0652	0.6499	0.9170
11	28	1	0	0.7948	0.0692	0.6171	0.8965
13	27	0	1	0.7948	0.0692	0.6171	0.8965
14	26	0	1	0.7948	0.0692	0.6171	0.8965
19	25	0	1	0.7948	0.0692	0.6171	0.8965
22	24	1	0	0.7617	0.0738	0.5788	0.8733
25	23	0	1	0.7617	0.0738	0.5788	0.8733
27	22	1	1	0.7271	0.0781	0.5394	0.8482
28	20	1	0	0.6907	0.0823	0.4989	0.8213
32	19	2	1	0.6180	0.0882	0.4229	0.7641
33	16	1	0	0.5794	0.0908	0.3837	0.7327
35	15	0	1	0.5794	0.0908	0.3837	0.7327
37	14	0	1	0.5794	0.0908	0.3837	0.7327
43	13	1	0	0.5348	0.0941	0.3376	0.6972
46	12	1	0	0.4902	0.0962	0.2944	0.6600
54	11	0	1	0.4902	0.0962	0.2944	0.6600
77	10	0	1	0.4902	0.0962	0.2944	0.6600
78	9	0	1	0.4902	0.0962	0.2944	0.6600
83	8	0	1	0.4902	0.0962	0.2944	0.6600
85	7	0	1	0.4902	0.0962	0.2944	0.6600
97	6	0	1	0.4902	0.0962	0.2944	0.6600
100	5	0	1	0.4902	0.0962	0.2944	0.6600
102	4	1	0	0.3677	0.1284	0.1377	0.6035
103	3	0	1	0.3677	0.1284	0.1377	0.6035
105	2	0	1	0.3677	0.1284	0.1377	0.6035
108	1	0	1	0.3677	0.1284	0.1377	0.6035

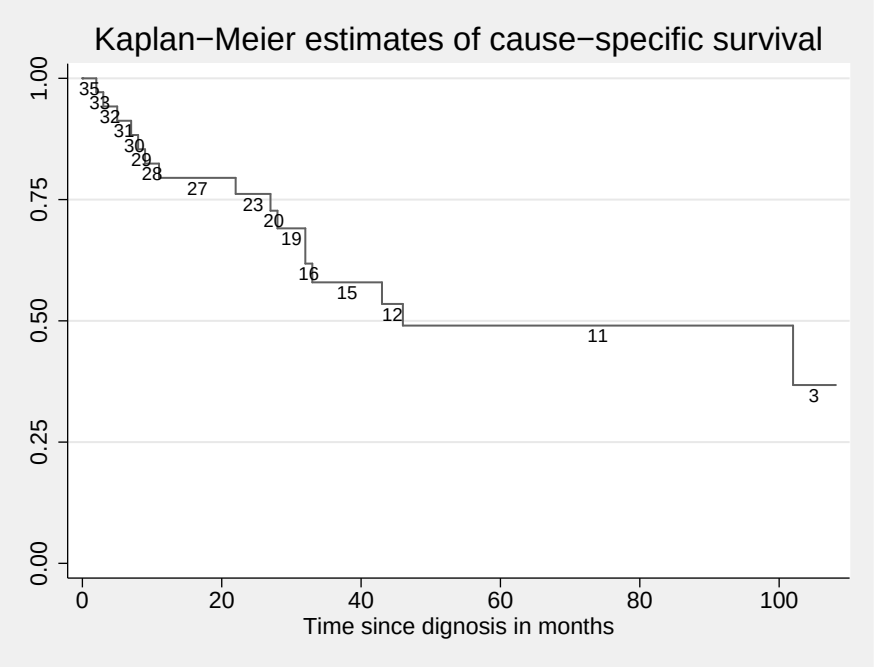

Figure 1: Kaplan-Meier plot of the cause-specific survivor function for sample of 35 patients diagnosed with colon carcinoma. The number at risk at each time point are shown on the curve.

2. **Comparing survival, proportions and mortality rates by stage for cause-specific and all-cause survival**

We start by reading the data and listing the first few observations to get an idea about the data.

```
. use melanoma, clear
(Skin melanoma, all stages, Finland 1975-94, follow-up to 1995)
. list age sex stage surv_mm surv_yy in 1/30

     +-----------------------------------------------+
     | age      sex       stage   surv_mm   surv_yy |
     |-----------------------------------------------|
  1. |  81   Female   Localised      26.5       2.5 |
  2. |  75   Female   Localised      55.5       4.5 |
  3. |  78   Female   Localised     177.5      14.5 |
  4. |  75   Female     Unknown      29.5       2.5 |
  5. |  81   Female     Unknown      57.5       4.5 |
     +-----------------------------------------------+
```

Now we define the data as survival time (st) data and look at the distribution of stage.

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

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

------------------------------------------------------------------------------
     7775  total obs.
        0  exclusions
------------------------------------------------------------------------------
     7775  obs. remaining, representing
     1913  failures in single record/single failure data
 615236.5  total analysis time at risk, at risk from t =         0
                              earliest observed entry t =         0
                                   last observed exit t =     251.5
. tab stage

   Clinical |
   stage at |
  diagnosis |      Freq.     Percent        Cum.
------------+-----------------------------------
    Unknown |      1,631       20.98       20.98
  Localised |      5,318       68.40       89.38
   Regional |        350        4.50       93.88
    Distant |        476        6.12      100.00
------------+-----------------------------------
      Total |      7,775      100.00
```

(a) Survival depends heavily on stage. It is interesting to note that patients with stage 0 (unknown) appear to have a similar survival to patients with stage 1 (localized).

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

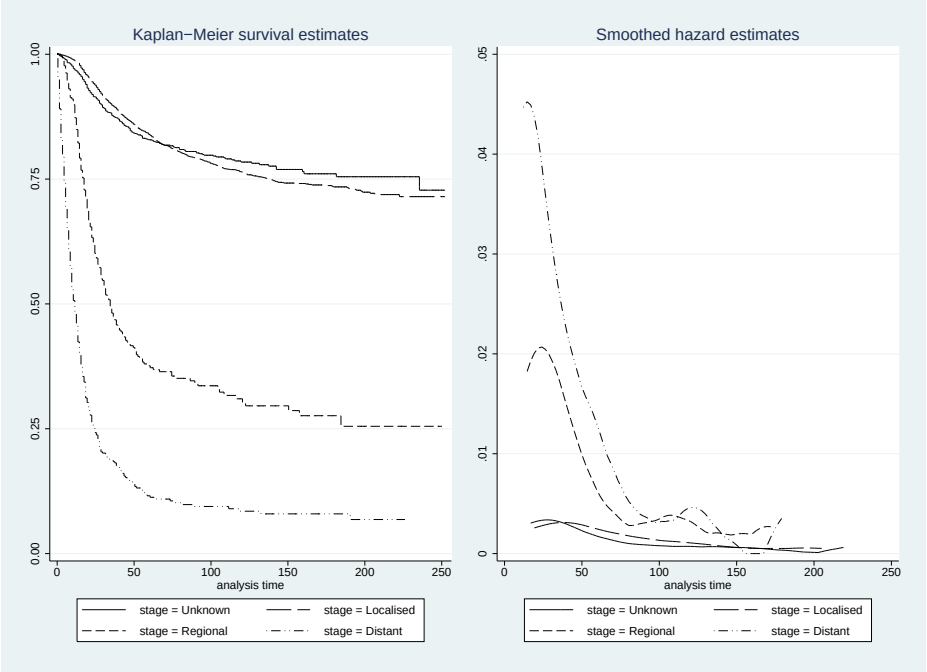

Figure 2: Skin melanoma. Kaplan-Meier estimates of cause-specific survival and mortality rate for each stage.

(b) `. strate stage`

```
      failure _d:  status == 1
analysis time _t:  surv_mm

Estimated rates and lower/upper bounds of 95% confidence intervals
(7775 records included in the analysis)
```

stage	D	Y	Rate	Lower	Upper
Unknown	274	1.2e+05	0.0022239	0.0019756	0.0025035
Localised	1013	4.6e+05	0.0021855	0.0020549	0.0023243
Regional	218	1.8e+04	0.0121091	0.0106038	0.0138281
Distant	408	1.1e+04	0.0388239	0.0352337	0.0427799

The time unit (defined when we `stset` the data) is months (since we specified `surv_mm` as the analysis time). Therefore, the units of the rates shown above are events/person-month. We could multiply these rates by 12 to obtain estimates with units events/person-year or we can change the default time unit by specifying the `scale()` option when we `stset` the data. For example,

```
. stset surv_mm, failure(status==1) scale(12)
. strate stage

        failure _d:  status == 1
  analysis time _t:  surv_mm/12

Estimated rates and lower/upper bounds of 95% confidence intervals
(7775 records included in the analysis)
  +-------------------------------------------------------------+
  |     stage      D          Y       Rate      Lower      Upper |
  |-------------------------------------------------------------|
  |   Unknown    274    1.0e+04   0.026687   0.023707   0.030042 |
  | Localised   1013    3.9e+04   0.026225   0.024659   0.027891 |
  |  Regional    218    1.5e+03   0.145309   0.127245   0.165937 |
  |   Distant    408   875.7500   0.465886   0.422804   0.513359 |
  +-------------------------------------------------------------+
```

(c) To obtain mortality rates per 1000 person years:

```
. strate stage,per(1000)

        failure _d:  status == 1
  analysis time _t:  surv_mm/12

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(7775 records included in the analysis)

  +-----------------------------------------------------------+
  |     stage      D         Y      Rate     Lower     Upper |
  |-----------------------------------------------------------|
  |   Unknown    274   10.2671    26.687    23.707    30.042 |
  | Localised   1013   38.6266    26.225    24.659    27.891 |
  |  Regional    218    1.5003   145.309   127.245   165.937 |
  |   Distant    408    0.8758   465.886   422.804   513.359 |
  +-----------------------------------------------------------+
```

(d) We see that the crude mortality rate is higher for males than females, a difference which is also reflected in the survival and hazard curves (Figure 3).

```
. strate sex, per(1000)

        failure _d:  status == 1
  analysis time _t:  surv_mm/12

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(7775 records included in the analysis)

+-------------------------------------------------------+
|    sex      D         Y     Rate    Lower    Upper |
|-------------------------------------------------------|
|   Male   1074   21.9689   48.887   46.049   51.900 |
| Female    839   29.3008   28.634   26.761   30.639 |
+-------------------------------------------------------+

. sts graph, by(sex)
```

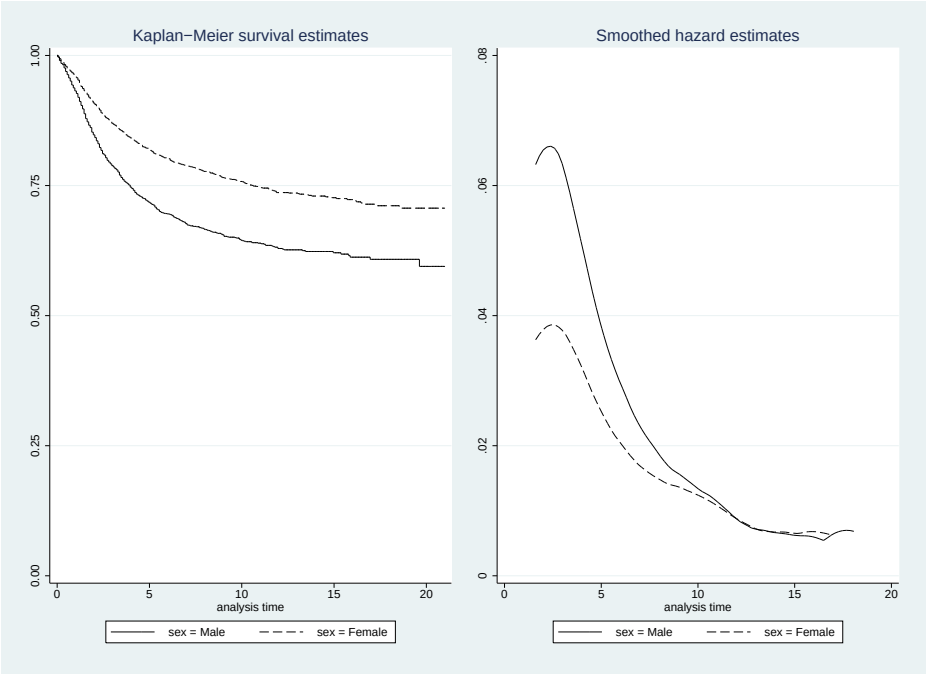

Figure 3: Skin melanoma (all stages). Kaplan-Meier estimates of cause-specific survival and mortality for each sex.

(e) The majority of patients are alive at end of study. 1,913 died from cancer while 1,134 died from another cause. The cause of death is highly depending of age, as young people die less from other causes.

```
. codebook status

-------------------------------------------------------------------------------
status                                          Vital status at last date of contact
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  status

                 range:  [0,4]                        units:  1
         unique values:  4                        missing .:  0/7775

            tabulation:  Freq.   Numeric  Label
                         4720         0  Alive
                         1913         1  Dead: cancer
                         1134         2  Dead: other
                            8         4  Lost to follow-up

. tab status agegrp

  Vital status at |
     last date of |             Age in 4 categories
          contact |      0-44      45-59      60-74        75+ |     Total
------------------+--------------------------------------------+----------
            Alive |     1,615      1,568      1,178        359 |     4,720
     Dead: cancer |       386        522        640        365 |     1,913
      Dead: other |        39        147        461        487 |     1,134
Lost to follow-up |         6          1          1          0 |         8
------------------+--------------------------------------------+----------
            Total |     2,046      2,238      2,280      1,211 |     7,775
```

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

     failure event:  status == 1 2
obs. time interval:  (0, surv_mm]
 exit on or before:  failure

-------------------------------------------------------------------------------
     7775  total obs.
        0  exclusions
-------------------------------------------------------------------------------
     7775  obs. remaining, representing
     3047  failures in single record/single failure data
 615236.5  total analysis time at risk, at risk from t =         0
                              earliest observed entry t =         0
                                   last observed exit t =     251.5
```

The survival is worse for all-cause survival than for cause-specific, since you now can die from other causes, and these deaths are incorporated in the Kaplan-Meier estimates. The "other cause" mortality is particularly present in patients with localised and unknown stage.

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

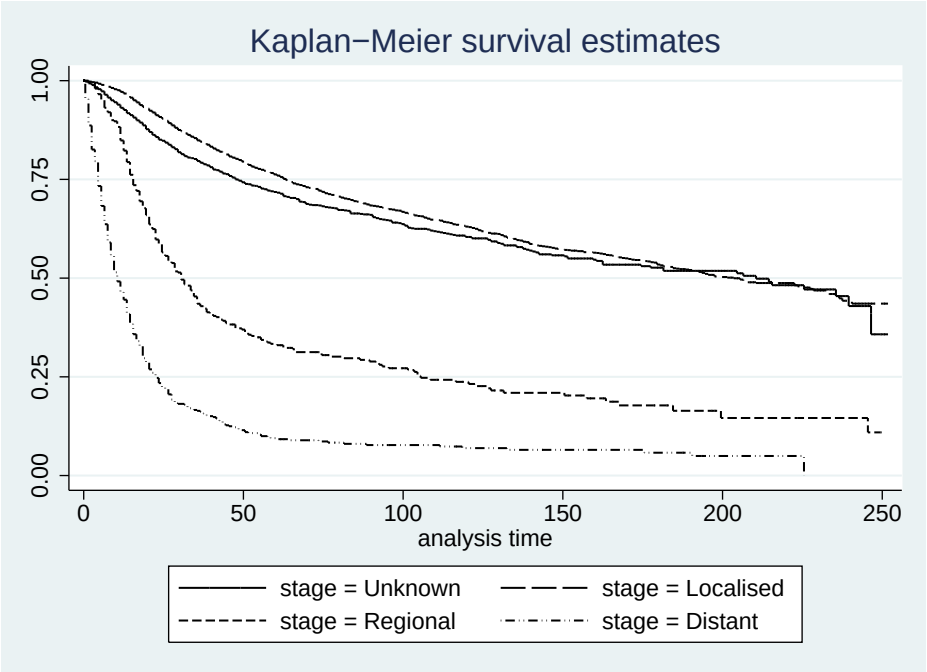

Figure 4: Skin melanoma (all stages). Kaplan-Meier estimates of all-cause survival for each stage.

(g) We see that the "other" cause mortality is particularly influential in patients with localised and unknown stage. Patients with localised disease, have a better prognosis (i.e. the cancer does not kill them), and are thus more likely to experience death from another cause. For regional and distant stage, the cancer is more aggressive and is the cause of death for most of these patients (i.e. it is the cancer that kills these patients before they have "the chance" to die from something else).

```
. 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, iscale(0.5)
```

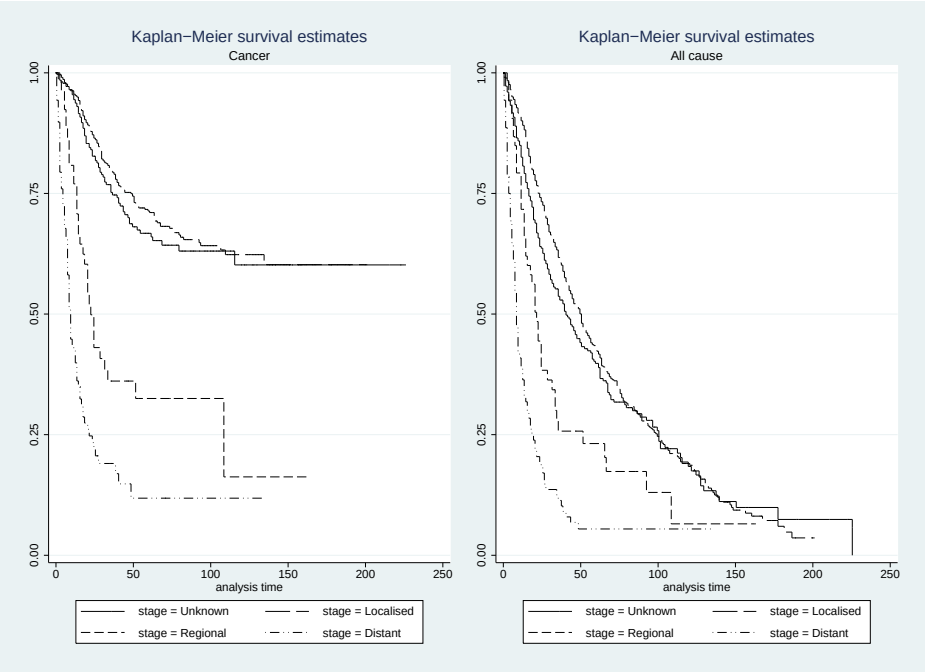

Figure 5: Skin melanoma (all stages). Kaplan-Meier estimates of all-cause survival versus cause-specific survival for each stage.

(h)
```
. 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")

[output omitted]
```

3. **Comparing estimates of cause-specific survival between periods**

```
. use melanoma if stage==1, clear
(Skin melanoma, all stages, Finland 1975-94, follow-up to 1995)

. stset surv_mm, failure(status==1)

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

------------------------------------------------------------------------------
     5318  total obs.
        0  exclusions
------------------------------------------------------------------------------
     5318  obs. remaining, representing
     1013  failures in single record/single failure data
   463519  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =     251.5

. sts graph, by(year8594)
```

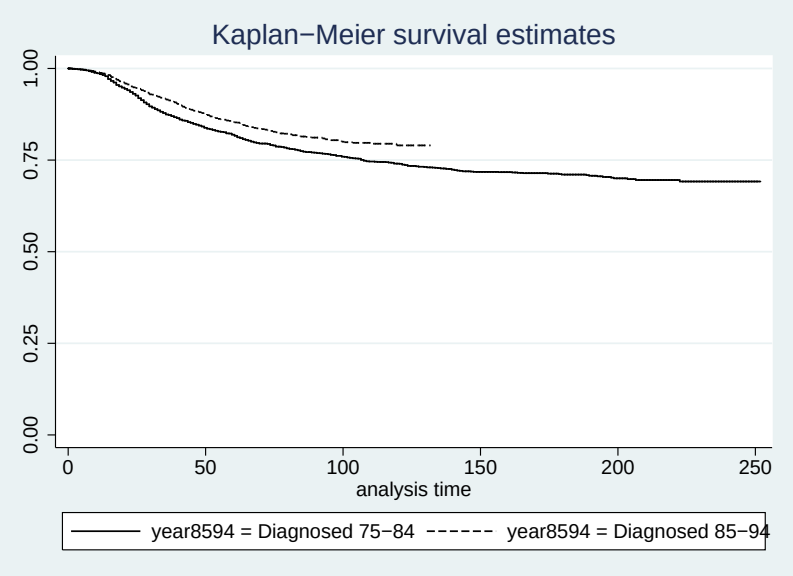

Figure 6: Skin melanoma. Kaplan-Meier plot of the cause-specific survivor function for each calendar period of diagnosis

(a) There seems to be a clear difference in survival between the two periods. Patients diagnosed during 1985–94 have superior survival to those diagnosed 1975–84.

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

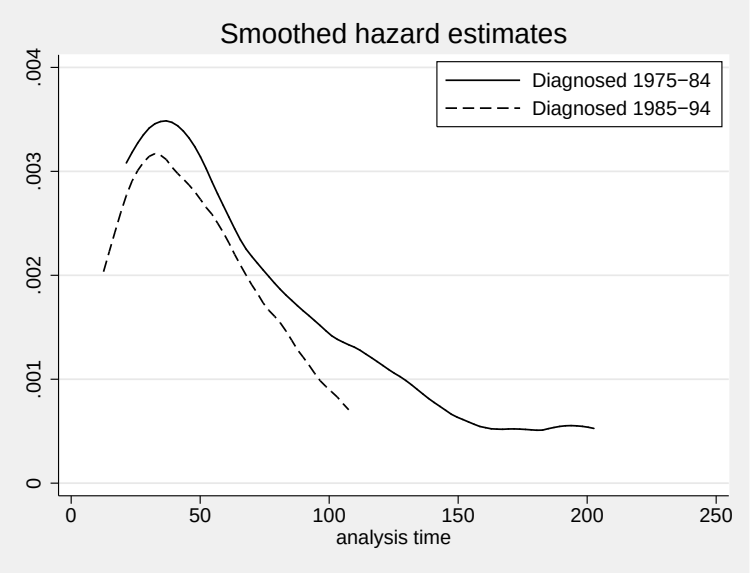

Figure 7: Skin melanoma. Plot of the cause-specific hazard for each calendar period of diagnosis

The plot shows the instantaneous cancer-specific mortality rate (the hazard) as a function of time. It appears that mortality is highest approximately 40 months following diagnosis. Remember that all patients were classified as having localised cancer at the time of diagnosis so we would not expect mortality to be high directly following diagnosis.

The plot of the hazard clearly illustrates the pattern of cancer-specific mortality as a function of time whereas this pattern is not obvious in the plot of the survivor function.

(c) `. sts test year8594`

```
Log-rank test for equality of survivor functions
------------------------------------------------
                |  Events
year8594        |  observed       expected
----------------+-------------------------
Diagnosed 75-84 |       572         512.02
Diagnosed 85-94 |       441         500.98
----------------+-------------------------
Total           |      1013        1013.00

                      chi2(1) =      15.50
                      Pr>chi2 =     0.0001
```

```
. sts test year8594, wilcoxon

Wilcoxon (Breslow) test for equality of survivor functions
------------------------------------------------------------
                |  Events                         Sum of
year8594        |  observed       expected         ranks
----------------+-------------------------------------------
Diagnosed 75-84 |       572         512.02        251185
Diagnosed 85-94 |       441         500.98       -251185
----------------+-------------------------------------------
Total           |      1013        1013.00             0

                     chi2(1) =      16.74
                     Pr>chi2 =     0.0000
```

There is strong evidence that survival differs between the two periods. The log-rank and the Wilcoxon tests give very similar results. The Wilcoxon test gives more weight to differences in survival in the early period of follow-up (where there are more individuals at risk) whereas the log rank test gives equal weight to all points in the follow-up. Both tests assume that, if there is a difference, a proportional hazards assumption is appropriate.

(d) We see that mortality increases with age at diagnosis (and survival decreases).

```
. strate agegrp, per(1000)

        failure _d:  status == 1
  analysis time _t:  surv_mm

Estimated rates (per 1000) and lower/upper bounds of 95\% confidence intervals
(5318 records included in the analysis)
  +-------------------------------------------------------+
  | agegrp     D          Y     Rate    Lower    Upper |
  |-------------------------------------------------------|
  |   0-44   217   157.1215   1.3811   1.2090   1.5776 |
  |  45-59   282   148.8215   1.8949   1.6861   2.1295 |
  |  60-74   333   121.3380   2.7444   2.4649   3.0556 |
  |    75+   181    36.2380   4.9948   4.3176   5.7781 |
  +-------------------------------------------------------+
```

The rates are (cause-specific) deaths per 1000 person-months. When we stset we defined time as time in months and then asked for rates per 1000 units of time.

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

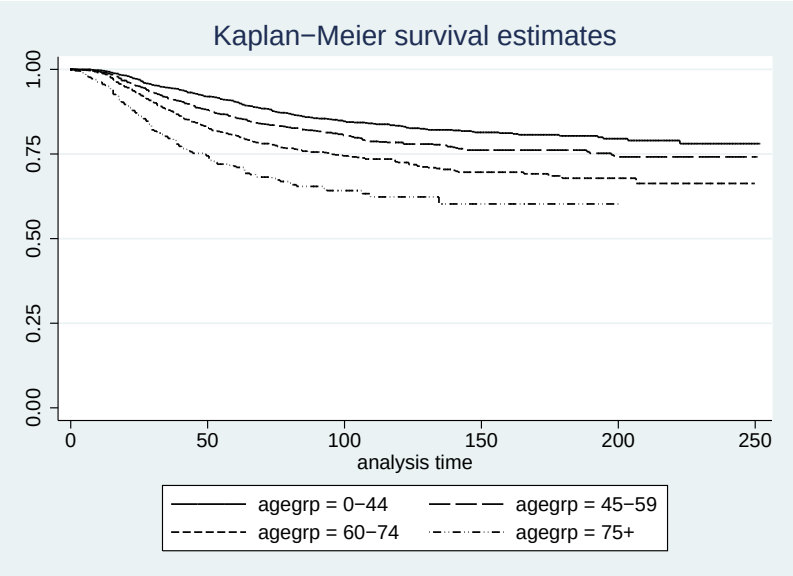

Figure 8: Skin melanoma. Plot of the cause-specific survival function for each age group

(e) No written solutions for this part.

(f) No written solutions for this part.

4. **Comparing various approaches to estimating the 10-year survival proportion**

```
. use melanoma if stage==1, clear
. generate csr_fail=0
. replace csr_fail=1 if status==1

. ltable surv_yy csr_fail
. ltable surv_mm csr_fail

. stset surv_yy, failure(status==1)
. sts list

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

	Actuarial	Kaplan-Meier
Years	0.7633	0.7729
Months	0.7637	0.7645

(a) The actuarial method is most appropriate because it deals with ties (events and censorings at the same time) in a more appropriate manner. The fact that there are a reasonably large number of ties in these data means that there is a difference between the estimates.

(b) The K-M estimate changes more. Because the actuarial method deals with ties in an appropriate manner it is not biased when data are heavily tied so is not heavily affected when we reduce the number of ties.

6. **Tabulating incidence rates and modelling with Poisson regression**

(a) We see that individuals with a high energy intake have a lower CHD incidence rate. The estimated crude incidence rate ratio is 0.52.

```
. strate hieng, per(1000)

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(337 records included in the analysis)
  +-------------------------------------------------+
  | hieng    D        Y      Rate    Lower     Upper |
  |-------------------------------------------------|
  |   low   28   2.0594   13.5960   9.3875   19.6912 |
  |  high   18   2.5442    7.0748   4.4574   11.2291 |
  +-------------------------------------------------+

. display 7.0748/13.596
.52035893
```

(b) The IRR calculated by the Poisson regression is the same as the IRR calculated in 6(a). A theoretical observation: If we consider the data as being cross classified solely by hieng then the Poisson regression model with one parameter is a saturated model so the IRR estimated from the model will be identical to the 'observed' IRR. That is, the model is a perfect fit.

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

Poisson regression                      Number of obs   =        337
                                        LR chi2(1)      =       4.82
                                        Prob > chi2     =     0.0282
Log likelihood =  -175.0016             Pseudo R2       =     0.0136
------------------------------------------------------------------------------
         chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |   .5203602   .1572055    -2.16   0.031     .2878382    .9407184
           y |  (exposure)
------------------------------------------------------------------------------
```

(c) A histogram (Figure 9) gives us an idea of the distribution of energy intake. We can also tabulate moments and percentiles of the distribution using the `summarize` command.

```
. histogram energy, normal
```

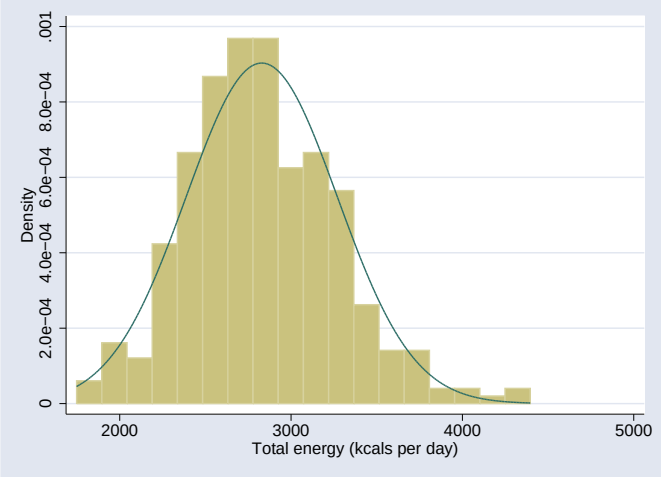

Figure 9: Histogram of energy with superimposed normal density curve (with the sample mean and variance).

```
. sum energy, detail

                Total energy (kcals per day)
-------------------------------------------------------------
      Percentiles      Smallest
 1%      1876.13        1748.43
 5%      2168.86        1854.02
10%      2311.24         1858.8       Obs                 337
25%      2536.69        1876.13       Sum of Wgt.         337

50%      2802.98                      Mean           2828.872
                        Largest       Std. Dev.      441.7528
75%      3109.66        4063.02
90%      3366.61        4234.06       Variance       195145.5
95%      3595.05        4256.81       Skewness       .4430434
99%      4063.02        4395.75       Kurtosis       3.506768
```

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

       eng3 |      Freq.     Percent        Cum.
------------+-----------------------------------
       1500 |         75       22.26       22.26
       2500 |        150       44.51       66.77
       3000 |        112       33.23      100.00
------------+-----------------------------------
      Total |        337      100.00
```

(e) We see that the CHD incidence rate decreases as the level of total energy intake increases.

```
. strate eng3,per(1000)

Estimated rates (per 1000) and lower/upper bounds of 95% Cis
(337 records included in the analysis)
  +-------------------------------------------------+
  | eng3    D        Y      Rate     Lower     Upper |
  |-------------------------------------------------|
  | 1500   16   0.9466   16.9020   10.3547   27.5892 |
  | 2500   22   2.0173   10.9059    7.1810   16.5629 |
  | 3000    8   1.6398    4.8787    2.4398    9.7555 |
  +-------------------------------------------------+

. display 10.9059/16.9020
.64524317

. display 4.8787/16.9020
.28864631
```

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

       eng3 |      Freq.     Percent        Cum.
------------+-----------------------------------
       1500 |         75       22.26       22.26
       2500 |        150       44.51       66.77
       3000 |        112       33.23      100.00
------------+-----------------------------------
      Total |        337      100.00
```

(g)
```
. set more off
. list eng3 X1 X2 X3 if eng3==1500 in 1/100
     +---------------------+
     | eng3   X1   X2   X3 |
     |---------------------|
  1. | 1500    1    0    0 |
  2. | 1500    1    0    0 |
  3. | 1500    1    0    0 |
  4. | 1500    1    0    0 |
  5. | 1500    1    0    0 |
     |---------------------|

. list eng3 X1 X2 X3 if eng3==2500 in 1/100
     +---------------------+
     | eng3   X1   X2   X3 |
     |---------------------|
 76. | 2500    0    1    0 |
 77. | 2500    0    1    0 |
 78. | 2500    0    1    0 |
 79. | 2500    0    1    0 |
 80. | 2500    0    1    0 |
     |---------------------|

. list eng3 X1 X2 X3 if eng3==3000 in 200/300
     +---------------------+
     | eng3   X1   X2   X3 |
     |---------------------|
226. | 3000    0    0    1 |
227. | 3000    0    0    1 |
228. | 3000    0    0    1 |
229. | 3000    0    0    1 |
230. | 3000    0    0    1 |
     |---------------------|

. set more on
```

(h) Level 1 of the categorized total energy is the reference category. The estimated rate ratio comparing level 2 to level 1 is 0.6452 and the estimated rate ratio comparing level 3 to level 1 is 0.2886.

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

Poisson regression                      Number of obs   =        337
                                        LR chi2(2)      =       9.20
                                        Prob > chi2     =     0.0100
Log likelihood = -172.81043             Pseudo R2       =     0.0259
------------------------------------------------------------------------------
         chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          X2 |   .6452416   .2120034    -1.33   0.182     .3388815    1.228561
          X3 |   .2886479   .1249882    -2.87   0.004     .1235342    .6744495
           y | (exposure)
------------------------------------------------------------------------------
```

(i) Now use level 2 as the reference (by omitting X2 but including X1 and X3). The estimated rate ratio comparing level 1 to level 2 is 1.5498 and the estimated rate ratio comparing level 3 to level 2 is 0.4473.

```
. poisson chd X1 X3, e(y) irr

Poisson regression                  Number of obs   =        337
                                    LR chi2(2)      =       9.20
                                    Prob > chi2     =     0.0100
Log likelihood = -172.81043         Pseudo R2       =     0.0259
------------------------------------------------------------------------------
chd |        IRR   Std. Err.      z    P>|z|  [95% Conf. Interval]
----+-------------------------------------------------------------------
 X1 |   1.549807   .5092114    1.33   0.182  .8139601    2.950884
 X3 |   .4473485   .1846929   -1.95   0.051  .1991671    1.004788
  y | (exposure)
------------------------------------------------------------------------------
```

(j) The estimates are identical (as we would hope) when we have Stata create indicator variables for us.

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

Poisson regression                                Number of obs   =        337
                                                  LR chi2(2)      =       9.20
                                                  Prob > chi2     =     0.0100
Log likelihood = -172.81043                       Pseudo R2       =     0.0259

------------------------------------------------------------------------------
         chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        eng3 |
       2500  |   .6452416   .2120034    -1.33   0.182     .3388815    1.228561
       3000  |   .2886479   .1249882    -2.87   0.004     .1235342    .6744495
           y | (exposure)
------------------------------------------------------------------------------
```

(k) Somehow (there are many different alternatives) you'll need to calculate the total number of events and the total person-time at risk and then calculate the incidence rate as events/person-time. For example,

```
. summarize y chd

Variable | Obs      Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------
       y | 337  13.66074    4.777274   .2874743   20.04107
     chd | 337  .1364985    .3438277          0          1

. display (337*0.1364985)/(337*13.66074)
.00999203
```

The estimated incidence rate is 0.00999 events per person-year (note that the two 337's cancel in the calculations are are only included for completeness). We get the same answer using `stptime`.

```
. stset dox, id(id) fail(chd) or(doe) scale(365.24)
. stptime
Cohort     |  person-time   failures        rate
-----------+-------------------------------------
     total |    4603.7948         46   .00999176
```

To give these estimates per 1000 person-years, they can simply be multiplied by 1000, or the `per(1000)` option of `stptime` can be used.

7. **Model cause-specific mortality with poisson regression**

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

(a) i. Survival is better during the latter period.

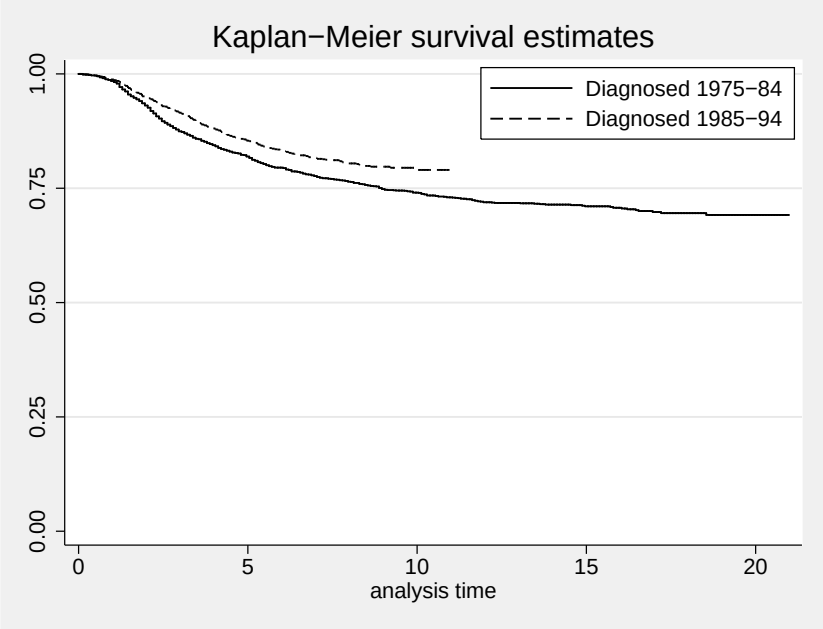

Figure 10: Localised melanoma. Kaplan-Meier estimates of cause-specific survival.

ii. Mortality is lower during the latter period.

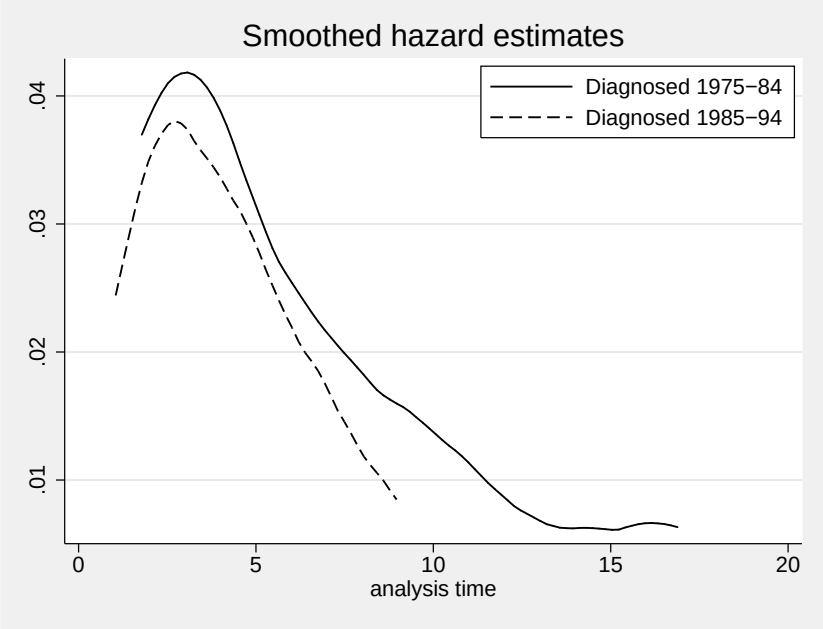

Figure 11: Localised melanoma. Smoothed cause-specific hazards (cause-specific mortality rates).

iii. The two graphs both show that prognosis is better during the latter period. Patients diagnosed during the latter period have lower mortality and higher survival.

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

```
      failure _d:  status == 1
analysis time _t:  surv_mm/12
              id:  id
```

```
Estimated rates (per 1000) and lower/upper bounds of 95% confidence
 intervals (5318 records included in the analysis)
  +-------------------------------------------------------------+
  |        year8594     D         Y     Rate    Lower    Upper |
  |-------------------------------------------------------------|
  | Diagnosed 75-84   572   22.6628   25.240   23.254   27.395 |
  | Diagnosed 85-94   441   15.9638   27.625   25.163   30.327 |
  +-------------------------------------------------------------+
```

The estimated mortality rate is lower for patients diagnosed during the early period. This is not consistent with what we saw in previous analyses. The inconsistency is due to the fact that we have not controlled for time since diagnosis. look at the graph of the estimated hazards (on the previous page) and try and estimate the overall average value for each group. We see that the average hazard for patients diagnosed in the early period is drawn down by the low mortality experienced by patients 10 years subsequent to diagnosis.

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

```
      failure _d:  status == 1
analysis time _t:  surv_mm/12
exit on or before:  time 120
              id:  id
```

```
Estimated rates (per 1000) and lower/upper bounds of 95% confidence
intervals (5318 records included in the analysis)
  +-------------------------------------------------------------+
  |        year8594     D         Y     Rate    Lower    Upper |
  |-------------------------------------------------------------|
  | Diagnosed 75-84   519   16.5010   31.453   28.860   34.278 |
  | Diagnosed 85-94   441   15.8756   27.778   25.303   30.496 |
  +-------------------------------------------------------------+
```

Now that we have restricted follow-up to a maximum of 10 years we see that the average mortality rate for patients diagnosed in the early period is higher than for the latter period. This is consistent with the graphs we examined in part (a).

ii. $27.778/31.453 = 0.883159$

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

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
year8594     |   .8831852   .0571985    -1.92   0.055     .7779016    1.002718
------------------------------------------------------------------------------
```

We see that Poisson regression is estimating the mortality rate ratio which, in this simple example, is the ratio of the two mortality rates.

(d)
```
. stsplit fu, at(0(1)10) trim
(no obs. trimmed because none out of range)
(28991 observations (episodes) created)
```

(e) It seems reasonable (at least to me) that melanoma-specific mortality is lower during the first year. These patients were classified as having localised skin melanoma at the time of diagnosis. That is, there was no evidence of metastases at the time of diagnosis although many of the patients who died would have had undetectable metastases or micrometastases at the time of diagnosis. It appears that it takes at least one year for these initially undetectable metastases to progress and cause the death of the patient.

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

        failure _d:  status == 1
  analysis time _t:  surv_mm/12
  exit on or before:  time 120
                id:  id

Estimated rates (per 1000) and lower/upper bounds of 95% confidence
intervals (34309 records included in the analysis)
  +-------------------------------------------------+
  | fu     D        Y      Rate     Lower     Upper |
  |-------------------------------------------------|
  |  0    71   5.2570   13.5058   10.7029   17.0427 |
  |  1   228   4.8579   46.9337   41.2204   53.4388 |
  |  2   202   4.2355   47.6926   41.5490   54.7446 |
  |  3   138   3.7116   37.1809   31.4674   43.9318 |
  |  4   100   3.2656   30.6224   25.1721   37.2528 |
  |-------------------------------------------------|
  |  5    80   2.8647   27.9265   22.4310   34.7683 |
  |  6    56   2.5248   22.1800   17.0693   28.8210 |
  |  7    35   2.1902   15.9799   11.4735   22.2563 |
  |  8    34   1.8864   18.0240   12.8787   25.2250 |
  |  9    16   1.5830   10.1071    6.1919   16.4979 |
  +-------------------------------------------------+
```

(f) The pattern is similar. The plot of the mortality rates (Figure 12) could be considered an approximation to the 'true' functional form depicted in Figure 13. By estimating the rates for each year of follow-up we are essentially approximating the curve in Figure 13 using a step function. It would probably be more informative to use narrower intervals (e.g., 6-month intervals) for the first 6 months of follow-up.

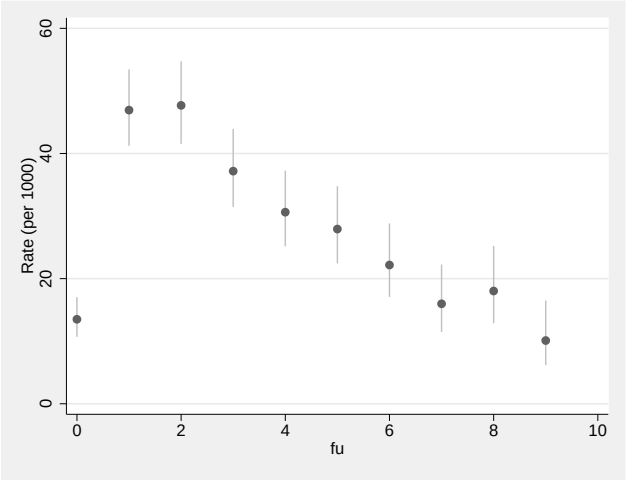

Figure 12: Localised melanoma. Disease-specific mortality rates as a function of time since diagnosis (annual intervals).

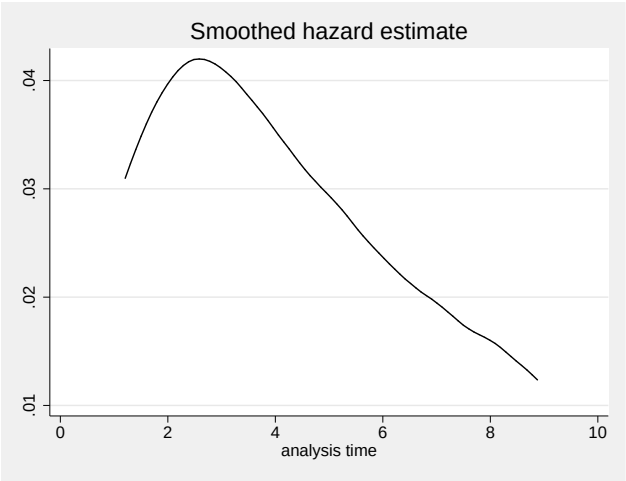

Figure 13: Localised melanoma. Disease-specific mortality rates as continuous function of time since diagnosis (using a smoother).

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

```
Exponential regression -- log relative-hazard form

No. of subjects =         5318                     Number of obs   =     34309
No. of failures =          960
Time at risk    =  32376.66667
                                                   LR chi2(9)      =    205.01
Log likelihood  =   -3264.6254                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          fu |
          1  |   3.475077   .4722842     9.17   0.000     2.662447    4.535737
          2  |   3.531267   .4871997     9.14   0.000     2.694589    4.627737
          3  |   2.752957   .4020721     6.93   0.000     2.067667    3.665374
          4  |   2.267352   .3518745     5.27   0.000     1.672705    3.073395
          5  |   2.067738   .3371396     4.46   0.000     1.502136    2.846308
          6  |   1.642261   .2935086     2.78   0.006     1.156947    2.331153
          7  |   1.183189   .2443677     0.81   0.415     .7893192    1.773598
          8  |   1.334537   .2783278     1.38   0.166     .8867597    2.008422
          9  |   .7483544   .2070989    -1.05   0.295     .4350575    1.287265
------------------------------------------------------------------------------
```

The pattern of the estimated mortality rate ratios mirrors the pattern we saw in the plot of the rates. Note that the first year of follow-up is the reference so the estimated rate ratio labelled `1` for `fu` is the rate ratio for the second year compared to the first year.

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

Exponential regression -- log relative-hazard form

No. of subjects =         5318                     Number of obs   =     34309
No. of failures =          960
Time at risk    =  32376.66667
                                                   LR chi2(10)     =    218.85
Log likelihood  =   -3257.7021                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          fu |
          1  |   3.467801   .4712995     9.15   0.000     2.656866    4.526251
          2  |   3.503269   .4833963     9.09   0.000     2.673136    4.591198
          3  |   2.711162   .3961271     6.83   0.000     2.036041    3.610141
          4  |   2.213063   .3437536     5.11   0.000     1.632214    3.000615
          5  |   1.998642   .3263829     4.24   0.000     1.451215    2.752569
          6  |   1.569936   .2812163     2.52   0.012     1.105121    2.230254
          7  |   1.114537   .2308644     0.52   0.601     .7426385    1.672676
          8  |   1.234277   .2586587     1.00   0.315      .818526      1.8612
          9  |   .6754363   .1877805    -1.41   0.158     .3916867    1.164743
             |
    year8594 |   .7831406   .0515257    -3.72   0.000     .6883924    .8909297
------------------------------------------------------------------------------
```

The estimated mortality rate ratio is 0.7831406 compared to 0.8831852 (part c) and a value greater than 1 in part (b). The estimate we obtained in part (b) was subject to confounding by time-since-diagnosis. In part (c) we restricted to the first 10 years of follow-up subsequent to diagnosis. This did not, however, completely remove the confounding effect of time since diagnosis. There was still some confounding within the first 10 years of follow-up (if this is not clear to you then look in the data to see if there are associations between the confounder and the exposure and the confounder and the outcome) so the estimate was subject to residual confounding. Now, when we adjust for time since diagnosis we see that the estimate changes further.

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

Exponential regression -- log relative-hazard form

No. of subjects =         5318                     Number of obs   =     34309
No. of failures =          960
Time at risk    =  32376.66667
                                                   LR chi2(14)     =    418.10
Log likelihood  =   -3158.0791                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          fu |
           1 |   3.554685   .4831685     9.33   0.000     2.723341     4.63981
           2 |   3.693498    .509924     9.46   0.000      2.81787    4.841218
           3 |   2.932197   .4288972     7.35   0.000     2.201337    3.905707
           4 |   2.447753   .3808518     5.75   0.000     1.804376    3.320536
           5 |   2.256233   .3693067     4.97   0.000      1.63703    3.109646
           6 |   1.797453   .3227726     3.27   0.001      1.26417    2.555699
           7 |   1.288667   .2675039     1.22   0.222     .8579195    1.935685
           8 |    1.43946   .3023764     1.73   0.083      .953661    2.172726
           9 |   .7961573   .2216843    -0.82   0.413     .4613046    1.374073
             |
      agegrp |
           1 |   1.327795    .125042     3.01   0.003     1.104005    1.596948
           2 |   1.862376    .169244     6.84   0.000     1.558527    2.225464
           3 |   3.400287   .3551404    11.72   0.000     2.770846    4.172715
             |
    year8594 |   .7224105   .0478125    -4.91   0.000     .6345233    .8224709
         sex |   .5875465   .0384565    -8.12   0.000     .5168076     .667968
------------------------------------------------------------------------------
```

i. For patients of the same sex diagnosed in the same calendar period, those aged 60–74 at diagnosis have an estimated 86% higher risk of death due to skin melanoma than those aged 0–44 at diagnosis. The difference is statistically significant.

ii. The parameter estimate for period changes from 0.78 to 0.72 when age and sex are added to the model. Whether this is 'strong confounding', or even 'confounding' is a matter of judgement. I would consider this confounding but not strong confounding but there is no correct answer.

iii. Age (modelled as a categorical variable with 4 levels) is highly significant in the model.

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

 ( 1)  [_t]1.agegrp = 0
 ( 2)  [_t]2.agegrp = 0
 ( 3)  [_t]3.agegrp = 0

           chi2(  3) =  155.82
         Prob > chi2 =    0.0000
```

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

Exponential regression -- log relative-hazard form

No. of subjects =         5318                     Number of obs   =     34309
No. of failures =          960
Time at risk    =  32376.66667
                                                   LR chi2(15)     =    418.29
Log likelihood  =   -3157.9807                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          fu |
           1 |   3.554795   .4831838     9.33   0.000     2.723425    4.639955
           2 |   3.693547   .5099324     9.46   0.000     2.817906    4.841287
           3 |   2.932013   .4288725     7.35   0.000     2.201195    3.905468
           4 |   2.447604   .3808316     5.75   0.000     1.804262    3.320341
           5 |    2.25602   .3692772     4.97   0.000     1.636868    3.109367
           6 |   1.797325   .3227558     3.26   0.001     1.264071    2.555534
           7 |   1.288401    .267454     1.22   0.222     .8577355    1.935301
           8 |   1.439152   .3023187     1.73   0.083     .9534478    2.172282
           9 |   .7958958    .221615    -0.82   0.412     .4611492    1.373634
             |
      agegrp |
           1 |   1.326709   .1249663     3.00   0.003     1.103059    1.595705
           2 |   1.861131   .1691561     6.83   0.000     1.557443    2.224035
           3 |   3.399539   .3550374    11.72   0.000     2.770277    4.171737
             |
  1.year8594 |   .7414351   .0655414    -3.38   0.001     .6234888    .8816936
       2.sex |   .6031338   .0531555    -5.74   0.000     .5074526     .716856
             |
year8594#sex |
        1 2  |   .9437245   .1232639    -0.44   0.657     .7305772    1.219058
------------------------------------------------------------------------------
```

The interaction term is not statistically significant indicating that there is no evidence that the effect of sex is modified by period.

(k) i. The effect of sex for patients diagnosed 1975–84 is 0.6031338 and the effect of sex for patients diagnosed 1985–94 is $0.6031338 \times 0.9437245 = 0.56919214$.

ii. We can use `lincom` to get the estimated effect for patients diagnosed 1985–94.

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

 ( 1)  [_t]2.sex + [_t]1.year8594#2.sex = 0

------------------------------------------------------------------------------
          _t |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .5691922    .055267    -5.80   0.000     .4705541    .6885069
------------------------------------------------------------------------------
```

The advantage of `lincom` is that we also get a confidence interval (not easy to calculate by hand since the SE is a function of variances and covariances).

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

Exponential regression -- log relative-hazard form

No. of subjects =         5318          Number of obs   =     34309
No. of failures =          960
Time at risk    =  32376.66667
                                        LR chi2(15)     =    418.29
Log likelihood  =   -3157.9807          Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Ifu_1 |   3.554795   .4831838    9.33   0.000     2.723425    4.639955
      _Ifu_2 |   3.693547   .5099324    9.46   0.000     2.817906    4.841287
      _Ifu_3 |   2.932013   .4288725    7.35   0.000     2.201195    3.905468
      _Ifu_4 |   2.447604   .3808316    5.75   0.000     1.804262    3.320341
      _Ifu_5 |    2.25602   .3692772    4.97   0.000     1.636868    3.109367
      _Ifu_6 |   1.797325   .3227558    3.26   0.001     1.264071    2.555534
      _Ifu_7 |   1.288401    .267454    1.22   0.222     .8577355    1.935301
      _Ifu_8 |   1.439152   .3023187    1.73   0.083     .9534478    2.172282
      _Ifu_9 |   .7958958    .221615   -0.82   0.412     .4611492    1.373634
_Iagegrp_1   |   1.326709   .1249663    3.00   0.003     1.103059    1.595705
_Iagegrp_2   |   1.861131   .1691561    6.83   0.000     1.557443    2.224035
_Iagegrp_3   |   3.399539   .3550374   11.72   0.000     2.770277    4.171737
  year8594   |   .7414351   .0655414   -3.38   0.001     .6234888    .8816936
 sex_early   |   .6031338   .0531555   -5.74   0.000     .5074526     .716856
sex_latter   |   .5691922    .055267   -5.80   0.000     .4705541    .6885069
------------------------------------------------------------------------------
```

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

```
Exponential regression -- log relative-hazard form

No. of subjects =         5318                     Number of obs   =     34309
No. of failures =          960
Time at risk    =  32376.66667
                                                   LR chi2(15)     =    418.29
Log likelihood  =   -3157.9807                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          fu |
           1 |   3.554795   .4831838     9.33   0.000     2.723425    4.639955
           2 |   3.693547   .5099324     9.46   0.000     2.817906    4.841287
           3 |   2.932013   .4288725     7.35   0.000     2.201195    3.905468
           4 |   2.447604   .3808316     5.75   0.000     1.804262    3.320341
           5 |    2.25602   .3692772     4.97   0.000     1.636868    3.109367
           6 |   1.797325   .3227558     3.26   0.001     1.264071    2.555534
           7 |   1.288401    .267454     1.22   0.222     .8577355    1.935301
           8 |   1.439152   .3023187     1.73   0.083     .9534478    2.172282
           9 |   .7958958    .221615    -0.82   0.412     .4611492    1.373634
             |
      agegrp |
           1 |   1.326709   .1249663     3.00   0.003     1.103059    1.595705
           2 |   1.861131   .1691561     6.83   0.000     1.557443    2.224035
           3 |   3.399539   .3550374    11.72   0.000     2.770277    4.171737
             |
  1.year8594 |   .7414351   .0655414    -3.38   0.001     .6234888    .8816936
             |
year8594#sex |
        0 2  |   .6031338   .0531555    -5.74   0.000     .5074526     .716856
        1 2  |   .5691922    .055267    -5.80   0.000     .4705541    .6885069
------------------------------------------------------------------------------
```

(l) If we fit stratified models we get slightly different estimates (0.6165815 and 0.5549737) since the models stratified by calendar period imply that all estimates are modified by calendar period. That is, we are actually estimating the following model:

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

8. **Using Poisson regression adjusting for confounders on two different time-scales**

   (a) The rates plotted on timescale attained age show a clear increasing trend as age increases, which is to be expected (older persons are more likely to suffer from CHD). The rates plotted on timescale time-since-entry are almost constant (if you have some imagination you can see that the rates are flat).

```
. use diet, clear

* Timescale: Attained age
. stset dox, id(id) fail(chd) origin(dob) entry(doe) scale(365.24)

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

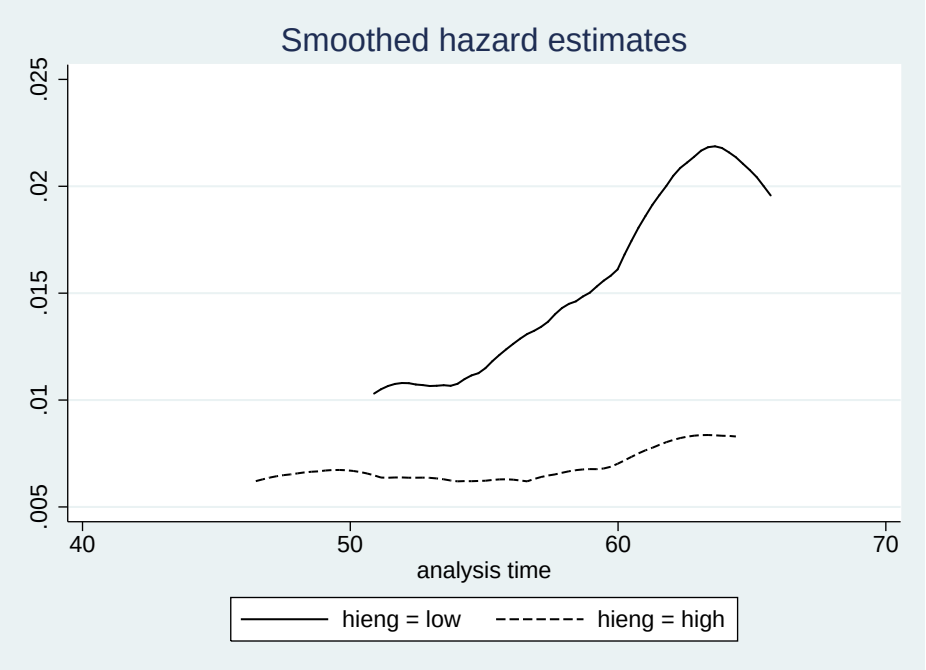

Figure 14: Diet data. Kaplan-Meier estimates of hazard rate for each energy intake level, with attained age as time scale.

```
* Timescale: Time since entry
. stset dox, id(id) fail(chd) origin(doe) enter(doe) scale(365.24)

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

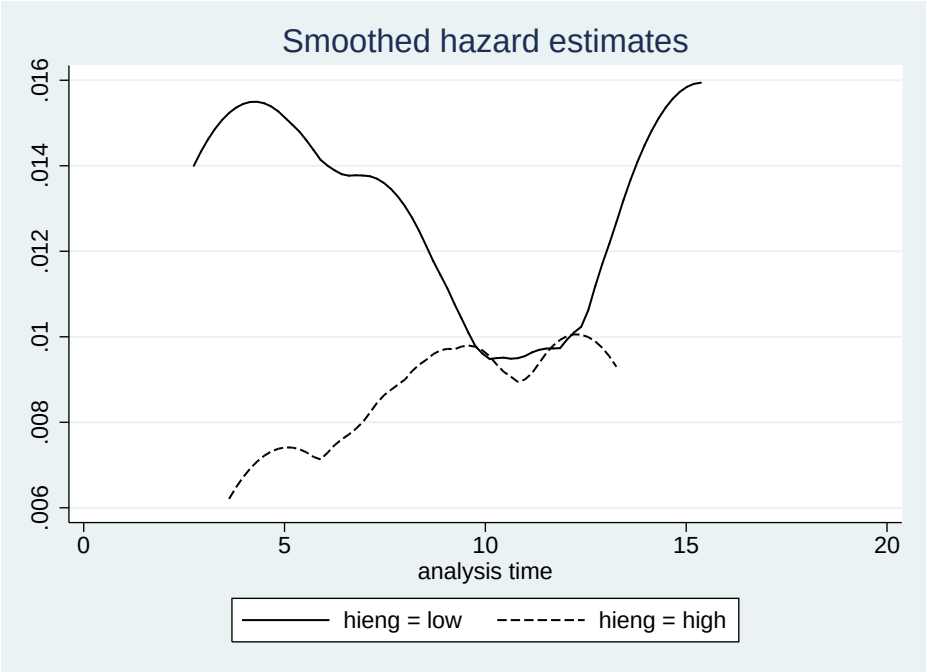

Figure 15: Diet data. Kaplan-Meier estimates of hazard rate for each energy intake level, with time since entry as time scale.

(b) Patients with high energy intake have 48% less CHD rate. The underlying shape of the rates is assumed to be constant (i.e. the baseline is flat) over time.

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

Poisson regression                              Number of obs   =        337
                                                LR chi2(1)      =       4.82
                                                Prob > chi2     =     0.0282
Log likelihood =  -175.0016                     Pseudo R2       =     0.0136

------------------------------------------------------------------------------
         chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |   .5203602   .1572055    -2.16   0.031     .2878382    .9407184
           y | (exposure)
------------------------------------------------------------------------------
```

(c) The effect of high energy intake is slightly confounded by bmi and job, since the point estimate changes a little.

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

Poisson regression                                Number of obs   =        332
                                                  LR chi2(3)      =       5.98
                                                  Prob > chi2     =     0.1127
Log likelihood =  -169.5164                       Pseudo R2       =     0.0173

------------------------------------------------------------------------------
         chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |   .4966098   .1538834    -2.26   0.024     .2705548     .911539
         job |   .9166234   .1573876    -0.51   0.612     .6546912    1.283351
         bmi |   1.052232   .0500593     1.07   0.285     .9585526    1.155066
           y | (exposure)
------------------------------------------------------------------------------
```

(d) The y variable is not correct since it is kept for all splitted records, and contains the complete follow-up rather than the risktime in that specific timeband.

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

     +-------------------------------------------------+
     |  id         _t0          _t   ageband          y |
     |-------------------------------------------------|
  1. | 127   49.389443          50        30   16.79124 |
  2. | 127          50          60        50   16.79124 |
  3. | 127          60   66.181141        60   16.79124 |
  4. | 200   47.497536          50        30   19.95893 |
  5. | 200          50          60        50   19.95893 |
     |-------------------------------------------------|
  6. | 200          60   67.457015        60   19.95893 |
  7. | 198   46.465338          50        30   19.95893 |
  8. | 198          50          60        50   19.95893 |
  9. | 198          60   66.424817        60   19.95893 |
 10. | 222   54.605191          60        50   15.39493 |
     +-------------------------------------------------+
```

The risktime variable contains the correct amount of risktime for each timeband.

```
. gen risktime=_t-t_0
. list id _t0 _t ageband y risktime in 1/10
```

```
     +-------------------------------------------------------------+
     |  id         _t0          _t   ageband          y   risktime |
     |-------------------------------------------------------------|
  1. | 127   49.389443          50        30   16.79124   .6105574 |
  2. | 127          50          60        50   16.79124         10 |
  3. | 127          60   66.181141        60   16.79124   6.181141 |
  4. | 200   47.497536          50        30   19.95893   2.502464 |
  5. | 200          50          60        50   19.95893         10 |
     |-------------------------------------------------------------|
  6. | 200          60   67.457015        60   19.95893   7.457015 |
  7. | 198   46.465338          50        30   19.95893   3.534662 |
  8. | 198          50          60        50   19.95893         10 |
  9. | 198          60   66.424817        60   19.95893   6.424817 |
 10. | 222   54.605191          60        50   15.39493   5.394809 |
     +-------------------------------------------------------------+
```

The event variable chd is not correct since it is kept constant for all splitted records, while it should only be 1 for the last record (if the person has the event). For all other records (timebands) for that person it should be 0.

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

```
           |   Failure: 1=chd, 0 otherwise
   ageband |         0          1          . |     Total
-----------+---------------------------------+----------
        30 |        10          6        180 |       196
        50 |        63         18        212 |       293
        60 |       218         22          0 |       240
-----------+---------------------------------+----------
     Total |       291         46        392 |       729
```

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

```
           |          _d
   ageband |         0          1 |     Total
-----------+----------------------+----------
        30 |       190          6 |       196
        50 |       275         18 |       293
        60 |       218         22 |       240
-----------+----------------------+----------
     Total |       683         46 |       729
```

The effect of high energy intake is somewhat confounded by age, but also confounded by job and bmi.

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

Poisson regression                                Number of obs   =        729
                                                  LR chi2(3)      =       9.64
                                                  Prob > chi2     =     0.0218
Log likelihood = -201.70224                       Pseudo R2       =     0.0234

------------------------------------------------------------------------------
          _d |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |   .5361689   .1622749    -2.06   0.039     .2962648    .9703384
             |
     ageband |
          50 |   1.353255   .6388848     0.64   0.522     .5364372    3.413816
          60 |   2.328214   1.074106     1.83   0.067      .942598     5.75068
    risktime | (exposure)
------------------------------------------------------------------------------

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

Poisson regression                                Number of obs   =        719
                                                  LR chi2(6)      =      14.47
                                                  Prob > chi2     =     0.0248
Log likelihood = -194.38638                       Pseudo R2       =     0.0359

------------------------------------------------------------------------------
          _d |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |   .4901577   .1538543    -2.27   0.023     .2649442     .906812
             |
         job |
           2 |   1.545112   .6284217     1.07   0.285     .6962464    3.428919
           3 |   .8711755   .3239507    -0.37   0.711     .4203222    1.805631
             |
         bmi |   1.076678   .0522368     1.52   0.128     .9790126    1.184086
             |
     ageband |
          50 |   1.710734   .8703232     1.06   0.291     .6311608     4.63687
          60 |   2.927686   1.454295     2.16   0.031     1.105859    7.750847
    risktime | (exposure)
------------------------------------------------------------------------------
```

(e)
```
. use diet, clear

. gen bmi=weight/(height/100*height/100)
. stset dox, id(id) fail(chd) origin(doe) enter(doe) scale(365.24)

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

	id	_t0	_t	fuband	y
1.	127	0	5	0	16.79124
2.	127	5	10	5	16.79124
3.	127	10	15	10	16.79124
4.	127	15	16.791699	15	16.79124
5.	200	0	5	0	19.95893
6.	200	5	10	5	19.95893
7.	200	10	15	10	19.95893
8.	200	15	19.959479	15	19.95893
9.	198	0	5	0	19.95893
10.	198	5	10	5	19.95893

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

	id	_t0	_t	fuband	y	risktime
1.	127	0	5	0	16.79124	5
2.	127	5	10	5	16.79124	5
3.	127	10	15	10	16.79124	5
4.	127	15	16.791699	15	16.79124	1.791699
5.	200	0	5	0	19.95893	5
6.	200	5	10	5	19.95893	5
7.	200	10	15	10	19.95893	5
8.	200	15	19.959479	15	19.95893	4.959479
9.	198	0	5	0	19.95893	5
10.	198	5	10	5	19.95893	5

```
. tab fuband chd, missing

           |   Failure: 1=chd, 0 otherwise
    fuband |         0          1          . |     Total
-----------+---------------------------------+----------
         0 |        13         17        307 |       337
         5 |        26         12        269 |       307
        10 |        69         13        187 |       269
        15 |       183          4          0 |       187
-----------+---------------------------------+----------
     Total |       291         46        763 |     1,100

. tab fuband _d, missing

           |          _d
    fuband |         0          1 |     Total
-----------+----------------------+----------
         0 |       320         17 |       337
         5 |       295         12 |       307
        10 |       256         13 |       269
        15 |       183          4 |       187
-----------+----------------------+----------
     Total |     1,054         46 |     1,100

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

Poisson regression                              Number of obs   =       1100
                                                LR chi2(4)      =       5.65
                                                Prob > chi2     =     0.2270
Log likelihood = -238.76022                     Pseudo R2       =     0.0117

------------------------------------------------------------------------------
          _d |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |    .522449   .1578565    -2.15   0.032      .288972    .9445654
             |
      fuband |
           5 |   .7916051   .2984822    -0.62   0.535      .378055    1.657533
          10 |     1.1292   .4160427     0.33   0.742     .5484711    2.324811
          15 |   .9511141   .5285699    -0.09   0.928      .320028    2.826684
    risktime | (exposure)
------------------------------------------------------------------------------
```

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

Poisson regression                                Number of obs   =       1084
                                                  LR chi2(7)      =       9.14
                                                  Prob > chi2     =     0.2429
Log likelihood = -232.10988                       Pseudo R2       =     0.0193

------------------------------------------------------------------------------
          _d |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |   .4895596   .1526123    -2.29   0.022     .2657402    .9018907
             |
         job |
          2  |   1.584205   .6439641     1.13   0.258     .7141775    3.514121
          3  |   .8711819   .3246359    -0.37   0.711     .4196801     1.80842
             |
         bmi |   1.071175   .0521887     1.41   0.158     .9736194    1.178506
             |
      fuband |
          5  |   .8451327   .3227979    -0.44   0.660      .399769    1.786655
         10  |   1.245226   .4667926     0.59   0.559     .5972581    2.596179
         15  |   1.142386   .6449991     0.24   0.814     .3777621    3.454675
    risktime | (exposure)
------------------------------------------------------------------------------
```

There seems to be no confounding by time-since-entry, but there is confounding by bmi and job.

(f) No written solutions for this part.

9. **Modelling cause-specific mortality using Cox regression**

```
. stcox year8594

Cox regression -- Breslow method for ties

No. of subjects =         5318          Number of obs   =      5318
No. of failures =          960
Time at risk    =       388520
                                        LR chi2(1)      =     14.78
Log likelihood  =   -7893.0592          Prob > chi2     =    0.0001
------------------------------------------------------------------------------
      _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
year8594 |   .7768217   .0511092    -3.84   0.000     .6828393    .8837392
------------------------------------------------------------------------------
```

(a) Patients diagnosed during 1985–94 experience only 77.7% of the cancer mortality experienced by those diagnosed 1975–84. That is, mortality due to skin melanoma has decreased by 22.3% in the latter period compared to the earlier period. This estimate is not adjusted for potential confounders. There is strong evidence of a statistically significant difference in survival between the two periods (based on the test statistic or the fact that the CI for the hazard ratio does not contain 1).

(b) The three test statistics are

log-rank 14.85 (from `sts test year8594`)

Wald $-3.84^2 = 14.75$ (from the z test above)

Likelihood ratio 14.78 (from the output above)

The three test statistics are very similar. We would expect each of these test statistics to be similar since they each test the same null hypothesis that survival is independent of calendar period. The null hypothesis in each case is that survival depends on calendar period in such a way that the hazard ratio between the two periods is constant over follow-up time (i.e. proportional hazards).

(c)
```
. stcox sex year8594 i.agegrp

Cox regression -- Breslow method for ties

No. of subjects =         5318                     Number of obs   =      5318
No. of failures =          960
Time at risk    =       388520
                                                   LR chi2(5)      =    211.94
Log likelihood  =   -7794.4811                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .5888144   .0385379    -8.09   0.000     .5179256    .6694059
    year8594 |   .7168836   .0474446    -5.03   0.000     .6296723    .8161739
             |
      agegrp |
           1 |   1.326397   .1249113     3.00   0.003     1.102841     1.59527
           2 |   1.857323   .1687866     6.81   0.000     1.554295     2.21943
           3 |   3.372652   .3522268    11.64   0.000     2.748371    4.138736
------------------------------------------------------------------------------
```

i. For patients of the same sex diagnosed in the same calendar period, those aged 60–74 at diagnosis have an estimated 86% higher risk of death due to skin melanoma than those aged 0–44 at diagnosis. The difference is statistically significant.
If this were an exam question the previous paragraph would be awarded full marks. It is worth noting, however, that the analysis is adjusted for the fact that mortality may depend on time since diagnosis (since this is the underlying time scale) and the mortality ratio between the two age groups is assumed to be the same at each point during the follow-up (i.e., proportional hazard).

ii. The parameter estimate for period changes from 0.78 to 0.72 when age and sex are added to the model. Whether this is 'strong confounding', or even 'confounding', is a matter of judgement. I would consider this confounding but not strong confounding but there is no correct answer to this question.

iii. Age (modelled as a categorical variable with 4 levels) is highly significant in the model.
```
. test 1.agegrp 2.agegrp 3.agegrp

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

           chi2(  3) =  153.78
         Prob > chi2 =    0.0000
```

(d) Age (modelled as a categorical variable with 4 levels) is highly significant in the model. The Wald test is an approximation to the LR test and we would expect the two to be similar (which they are).

```
. lrtest A

Likelihood-ratio test                                 LR chi2(3)  =    142.85
(Assumption: . nested in A)                           Prob > chi2 =    0.0000
```

(e) i. Both models adjust for the same factors. When fitting the Poisson regression model we split time since diagnosis into annual intervals and explicitly estimated the effect of this factor in the model. The Cox model does not estimate the effect of 'time' but the other estimates are adjusted for 'time'.

ii. Since the two models are conceptually similar we would expect the parameter estimates to be similar, which they are.

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Cox regression						
sex	.5888144	.0385379	-8.09	0.000	.5179256	.6694059
year8594	.7168836	.0474446	-5.03	0.000	.6296723	.8161739
agegrp						
1	1.326397	.1249113	3.00	0.003	1.102841	1.59527
2	1.857323	.1687866	6.81	0.000	1.554295	2.21943
3	3.372652	.3522268	11.64	0.000	2.748371	4.138736
Poisson regression						
sex	.5875465	.0384565	-8.12	0.000	.5168076	.667968
year8594	.7224105	.0478125	-4.91	0.000	.6345233	.8224709
agegrp						
1	1.327795	.125042	3.01	0.003	1.104005	1.596948
2	1.862376	.169244	6.84	0.000	1.558527	2.225464
3	3.400287	.3551404	11.72	0.000	2.770846	4.172715

iii. Yes, both models assume 'proportional hazards'. The proportional hazards assumption implies that the risk ratios for sex, period, and age are constant across all levels of follow-up time. In other words, the assumption is that there is no effect modification by follow-up time. This assumption is implicit in Poisson regression (as it is in logistic regression) where it is assumed that estimated risk ratios are constant across all combination of the other covariates. We can, of course, relax this assumption by fitting interaction terms.

(f) No written soultions for this part.

(g) No written solutions for this part.

10. **Examining the proportional hazards hypothesis**

(a) If we look at the hazard curves, at their peak the ratio is approximately $0.038/0.048 \approx 0.79$. The ratio is similar at other follow-up times.

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

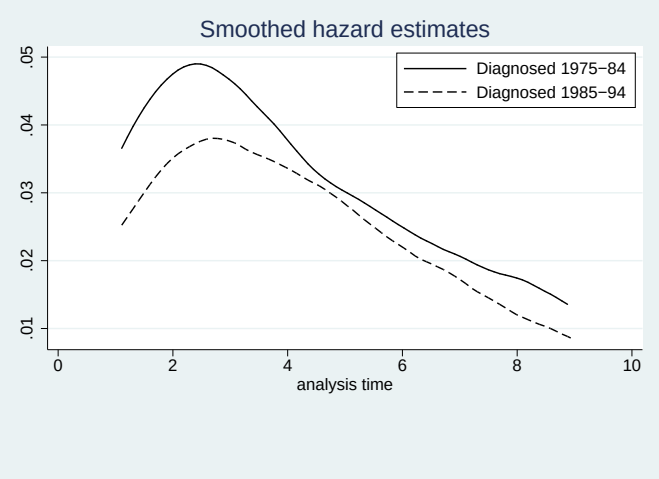

Figure 16: Localised skin melanoma. Plot of the estimated hazard function for each calendar period of diagnosis.

(b) There is no strong evidence against an assumption of proportional hazards since we see (close to) parallel curves when plotting the instantaneous cause-specific hazard on the log scale.

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

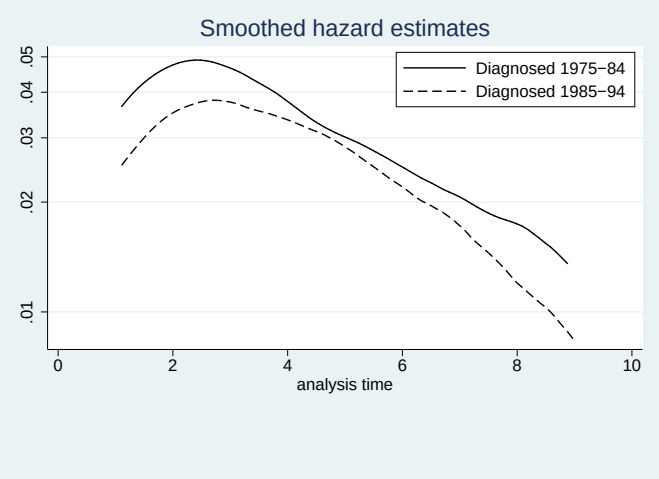

Figure 17: Localised skin melanoma. Plot of the estimated hazard function for each calendar period of diagnosis using a log scale for the y axis.

(c) If the proportional hazards assumption is appropriate then we should see parallel lines in Figure 18. This looks okay, we shouldn't put too much weight on the fact that the curves cross early in the follow-up since there are so few deaths there. The difference between the two log-cumulative hazard curves is similar during the part of the follow-up where we have the most information (most deaths). Note that these curves are not based on the estimated Cox model (i.e., they are unadjusted).

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

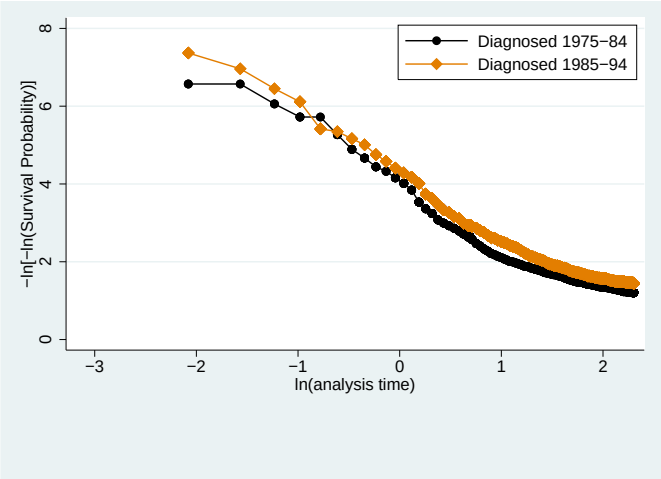

Figure 18: Localised skin melanoma. Plot of the log cumulative hazard function for each calendar period of diagnosis. Each plot symbol represents an event time. Note that the x axis is the natural logarithm of time in years, so a value of 0 corresponds to 1 year.

(d) The estimated hazard ratio from the Cox model is 0.78 which is similar (as it should be) to the estimate made by looking at the hazard function plot.

(e) The command `estat phtest, plot(1.year8594)` plots the scaled Schoenfeld residuals for the effect of period. Under proportional hazards, the smoother will be a horizontal line. The line is not, however, perfectly horizontal; it appears that the effect of period is greater earlier in the follow-up.

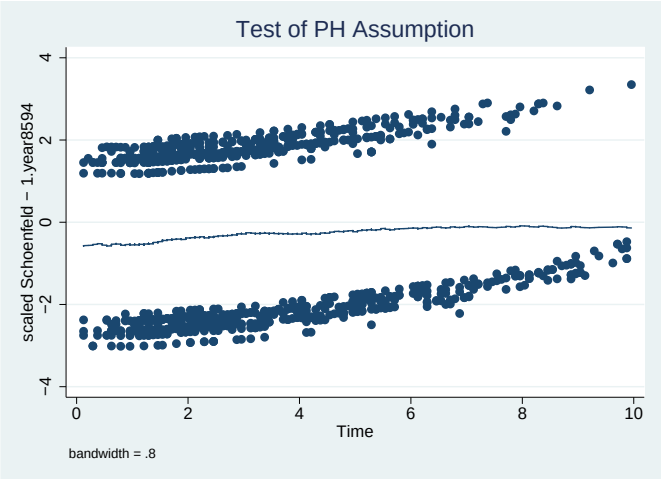

Figure 19: Localised skin melanoma. Plot of the scaled Schoenfeld residuals for calendar period 1985–94. The smooth line shows the estimated hazard ratio as a function of time.

(f) No written solutions for this part.

(g) It seems that there is evidence of non-proportional hazards by age (particularly for the comparison of the oldest to youngest) but not for calendar period. The plot of Schoenfeld residuals suggested non-proportionality for period but this was not statistically significant.

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

```
Test of proportional-hazards assumption

Time:  Time
------------------------------------------------------------------
             |       rho            chi2       df       Prob>chi2
-------------+----------------------------------------------------
1b.sex       |            .            .        1             .
2.sex        |      0.04705         2.09        1         0.1482
0b.year8594  |            .            .        1             .
1.year8594   |      0.04878         2.28        1         0.1308
0b.agegrp    |            .            .        1             .
1.agegrp     |     -0.04431         1.89        1         0.1690
2.agegrp     |     -0.08247         6.48        1         0.0109
3.agegrp     |     -0.12450        14.19        1         0.0002
-------------+----------------------------------------------------
global test  |                     18.29        5         0.0026
------------------------------------------------------------------
```

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

Cox regression -- Breslow method for ties

No. of subjects =         5318                     Number of obs   =      5318
No. of failures =          960
Time at risk    =  32376.66667
                                                   LR chi2(8)      =    221.75
Log likelihood  =   -7789.5752                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
         sex |   .5906795   .0386481    -8.05   0.000     .5195865    .6714998
    year8594 |   .7153885   .0473797    -5.06   0.000     .6283005    .8145476
     agegrp2 |   1.698848   .3335545     2.70   0.007     1.156187    2.496208
     agegrp3 |   2.457673   .4605845     4.80   0.000     1.702171    3.548502
     agegrp4 |   5.399496   1.035355     8.79   0.000      3.70796    7.862694
-------------+----------------------------------------------------------------
tvc          |
     agegrp2 |   .7257338   .1624357    -1.43   0.152     .4680143    1.125371
     agegrp3 |    .693004   .1487645    -1.71   0.088     .4550003    1.055504
     agegrp4 |   .4931264   .1144418    -3.05   0.002     .3129079    .7771414
------------------------------------------------------------------------------
Note: variables in tvc equation interacted with _t>=2
```

The hazard ratios for age in the top panel are for the first two years subsequent to diagnosis. To obtain the hazard ratios for the period two years or more following diagnosis we multiply the hazard ratios in the top and bottom panel. That is, during the first two years following diagnosis patients aged 75 years or more at diagnosis have 5.4 times higher cancer-specific mortality than patients aged 0–44 at diagnosis. During the period two years or more following diagnosis the corresponding hazard ratio is $5.4 \times 0.49 = 2.66$.

Using `stsplit` to split on time will give you the same results as above. We see that the age*follow up interaction is statistically significant.

```
. testparm i.agegrp#i.fuband
```

```
 ( 1)  1.agegrp#2.fuband = 0
 ( 2)  2.agegrp#2.fuband = 0
 ( 3)  3.agegrp#2.fuband = 0

           chi2(  3) =    9.55
         Prob > chi2 =    0.0228
```

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

Cox regression -- Breslow method for ties

No. of subjects =         5318                     Number of obs   =      9856
No. of failures =          960
Time at risk    =  32376.66667
                                                   LR chi2(8)      =    221.75
Log likelihood  =   -7789.5752                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .5906795   .0386481    -8.05   0.000     .5195865    .6714998
    year8594 |   .7153885   .0473797    -5.06   0.000     .6283005    .8145476
    2.fuband |   7.388391          .        .       .            .           .
             |
      fuband#|
      agegrp |
        0 1  |   1.698848   .3335545     2.70   0.007     1.156187    2.496208
        0 2  |   2.457673   .4605845     4.80   0.000     1.702171    3.548502
        0 3  |   5.399496   1.035355     8.79   0.000      3.70796    7.862694
        2 1  |   1.232911   .1328384     1.94   0.052     .9982062    1.522802
        2 2  |   1.703178   .1784726     5.08   0.000     1.386961    2.091489
        2 3  |   2.662634    .350343     7.44   0.000      2.05737    3.445963
------------------------------------------------------------------------------
```

	0-2 years	2+ years
Agegrp1	1.00	1.00
Agegrp2	1.70	1.23
Agegrp3	2.46	1.70
Agegrp4	5.40	2.66

(j) No written solutions for this part.

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

```
. stset surv_mm, failure(status==1,2) exit(time 120)

     failure event:  status == 1 2
obs. time interval:  (0, surv_mm]
 exit on or before:  time 120

------------------------------------------------------------------------------
     5318  total obs.
        0  exclusions
------------------------------------------------------------------------------
     5318  obs. remaining, representing
     1580  failures in single record/single failure data
   388520  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =       120


. stcox sex year8594 i.agegrp

Cox regression -- Breslow method for ties

No. of subjects =         5318             Number of obs   =      5318
No. of failures =         1580
Time at risk    =       388520
                                           LR chi2(5)      =    890.37
Log likelihood  =   -12506.145             Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .6101738   .0311091    -9.69   0.000     .5521485     .674297
    year8594 |    .753006   .0390759    -5.47   0.000     .6801847    .8336238
             |
      agegrp |
           1 |   1.502939   .1307488     4.68   0.000     1.267333    1.782346
           2 |   2.937808    .234755    13.49   0.000     2.511917    3.435907
           3 |   8.427357   .6966317    25.79   0.000     7.166851     9.90956
------------------------------------------------------------------------------
```

(a) For patients of the same sex diagnosed in the same period, those aged 60–74 at diagnosis have a 2.9 times higher risk of death *due to any causes* than those aged 0–44 at diagnosis. This difference is statistically significant.

(b) Note that the previous model estimated cause-specific hazard ratios whereas the current model estimates all-cause hazard ratios. The estimated hazard ratios for sex and period are similar, whereas the estimated hazard ratios for age are markedly different. This is because non-cancer mortality is heavily dependent on age, but only lightly dependent on sex and calendar period.

12. **Cox model for cause-specific mortality**

(a) `. stcox sex`

```
Cox regression -- Breslow method for ties

No. of subjects =         7775            Number of obs   =      7775
No. of failures =         1913
Time at risk    =    611349.29
                                          LR chi2(1)      =    103.25
Log likelihood  =   -16342.555            Prob > chi2     =    0.0000
------------------------------------------------------------------------------
 _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----+-------------------------------------------------------------------------
sex |   .6273066   .0289338   -10.11   0.000      .573085    .6866581
------------------------------------------------------------------------------
```

We see, without adjusting for potential confounders, that females have a 38% lower mortality than males.

(b) `. stcox sex year8594 i.agegrp i.subsite i.stage`

```
Cox regression -- Breslow method for ties

No. of subjects =         7775                  Number of obs   =      7775
No. of failures =         1913
Time at risk    =     615236.5
                                                LR chi2(11)     =   1835.82
Log likelihood  =   -15476.269                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .7490676    .036445    -5.94   0.000     .6809368    .8240153
             |
      agegrp |
           1 |   1.268542   .0855596     3.53   0.000     1.111459    1.447824
           2 |   1.730767   .1126805     8.43   0.000     1.523427    1.966326
           3 |   2.785848   .2128337    13.41   0.000     2.398431    3.235845
             |
       stage |
           1 |   1.038328   .0713262     0.55   0.584     .9075334    1.187972
           2 |   4.771515   .4363494    17.09   0.000     3.988549     5.70818
           3 |   13.48664   1.097917    31.96   0.000     11.49766     15.8197
             |
     subsite |
           2 |   1.393153   .0984179     4.69   0.000     1.213016    1.600041
           3 |   1.032021   .0767263     0.42   0.672     .8920829     1.19391
           4 |   1.305318    .133562     2.60   0.009      1.06812     1.59519
             |
    year8594 |   .7867739   .0376881    -5.01   0.000     .7162681    .8642199
------------------------------------------------------------------------------
```

After adjusting for a range of potential confounders we see that the estimated difference in cancer-specific mortality between males and females has decreased slightly but there is still quite a large difference.

(c) Let's first estimate the effect of gender for each age group without adjusting for confounders.

```
. stcox i.agegrp i.sex#i.agegrp

Cox regression -- Breslow method for ties

No. of subjects =         7775                     Number of obs   =      7775
No. of failures =         1913
Time at risk    =     615236.5
                                                   LR chi2(7)      =    331.08
Log likelihood  =   -16228.639                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |
           1 |   1.197101   .1017692     2.12   0.034     1.013369    1.414145
           2 |   1.497299   .1267028     4.77   0.000     1.268466    1.767412
           3 |   2.322161   .2401309     8.15   0.000     1.896142    2.843895
             |
  sex#agegrp |
         2 0 |   .4578165   .0478157    -7.48   0.000     .3730692    .5618151
         2 1 |   .5526258   .0504729    -6.49   0.000     .4620494     .660958
         2 2 |   .7132982   .0565997    -4.26   0.000     .6105607     .833323
         2 3 |   .6750958   .0713516    -3.72   0.000     .5487834    .8304813
------------------------------------------------------------------------------

. test 2.sex#0.agegrp = 2.sex#1.agegrp = 2.sex#2.agegrp = 2.sex#3.agegrp

 ( 1)  2.sex#0b.agegrp - 2.sex#1.agegrp = 0
 ( 2)  2.sex#0b.agegrp - 2.sex#2.agegrp = 0
 ( 3)  2.sex#0b.agegrp - 2.sex#3.agegrp = 0

           chi2(  3) =   13.50
         Prob > chi2 =    0.0037
```

We see that there is some evidence that the survival advantage experienced by females depends on age. The hazard ratio for males/females in the youngest age group is 0.46, while in the highest age group the hazard ratio is 0.68. There is evidence that the hazard ratios for gender differ across the age groups (p=0.0037). However, after adjusting for stage, subsite, and period there is no longer evidence of an interaction. See the following.

```
. stcox year8594 i.subsite i.stage i.agegrp i.sex#i.agegrp

Cox regression -- Breslow method for ties

No. of subjects =         7775                     Number of obs   =      7775
No. of failures =         1913
Time at risk    =     615236.5
                                                   LR chi2(14)     =   1840.42
Log likelihood  =   -15473.971                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    year8594 |   .7868595   .0376845    -5.01   0.000     .7163599    .8642973
             |
     subsite |
           2 |   1.401988   .0992064     4.78   0.000     1.220428    1.610558
           3 |   1.039415   .0773326     0.52   0.603     .8983792    1.202593
           4 |   1.315538   .1349198     2.67   0.007     1.075983    1.608428
             |
       stage |
           1 |   1.036942   .0712433     0.53   0.598     .9063011    1.186414
           2 |   4.702828   .4312718    16.88   0.000     3.929161    5.628833
           3 |   13.38869   1.091144    31.83   0.000     11.41215    15.70757
             |
      agegrp |
           1 |   1.188947   .1014449     2.03   0.043     1.005855    1.405367
           2 |     1.5508   .1318113     5.16   0.000     1.312827    1.831911
           3 |   2.485421   .2605605     8.68   0.000     2.023782    3.052363
             |
  sex#agegrp |
        2 0  |   .6251314   .0662091    -4.44   0.000     .5079472    .7693502
        2 1  |   .7300673   .0678894    -3.38   0.001      .608428    .8760252
        2 2  |   .8120201   .0653462    -2.59   0.010     .6935337    .9507494
        2 3  |   .8068979    .086154    -2.01   0.044      .654537    .9947249
------------------------------------------------------------------------------

. test 2.sex#0.agegrp = 2.sex#1.agegrp = 2.sex#2.agegrp = 2.sex#3.agegrp

 ( 1)  2.sex#0b.agegrp - 2.sex#1.agegrp = 0
 ( 2)  2.sex#0b.agegrp - 2.sex#2.agegrp = 0
 ( 3)  2.sex#0b.agegrp - 2.sex#3.agegrp = 0

           chi2(  3) =    4.56
         Prob > chi2 =    0.2067
```

That is, there is not strong evidence in support of the hypothesis (although some may consider that there is weak evidence).

(d) After having fitted a main effects model we can check the proportional hazards assumption by fitting a regression line through the model-based Schoenfeld residulas and check if the slope is statistically different from zero.

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

```
 Test of proportional-hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      sex         |      0.03157         1.93        1          0.1644
      year8594    |     -0.00805         0.13        1          0.7229
      0b.agegrp   |            .            .        1               .
      1.agegrp    |     -0.00847         0.14        1          0.7096
      2.agegrp    |     -0.00901         0.16        1          0.6918
      3.agegrp    |     -0.02301         1.04        1          0.3078
      1b.subsite  |            .            .        1               .
      2.subsite   |      0.01695         0.58        1          0.4477
      3.subsite   |      0.00398         0.03        1          0.8587
      4.subsite   |     -0.00694         0.09        1          0.7641
      0b.stage    |            .            .        1               .
      1.stage     |      0.08211        12.85        1          0.0003
      2.stage     |     -0.01781         0.60        1          0.4373
      3.stage     |     -0.06603         7.95        1          0.0048
      ------------+---------------------------------------------------
      global test |                     82.21       11          0.0000
      ----------------------------------------------------------------
```

There is strong evidence that the proportional hazard assumption is not satisfied for the effect of stage. Unless our primary interest is in the stage effect we can fit a stratified Cox model where we stratify on stage (i.e. estimate a separate baseline hazard function for each stage group).

```
stcox sex year8594 i.agegrp i.subsite, strata(stage)
```

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |    .741208   .0361298    -6.14   0.000     .6736723    .8155141
    year8594 |   .7877028   .0376795    -4.99   0.000     .7172086    .8651258
             |
      agegrp |
           1 |   1.263398   .0852288     3.47   0.001     1.106925     1.44199
           2 |   1.734631    .112968     8.46   0.000     1.526766    1.970796
           3 |   2.756441    .210658    13.27   0.000     2.372994     3.20185
             |
     subsite |
           2 |    1.33654   .0943198     4.11   0.000     1.163892    1.534799
           3 |   .9950338   .0738293    -0.07   0.947     .8603607    1.150787
           4 |   1.250443   .1282923     2.18   0.029     1.022664    1.528956
------------------------------------------------------------------------------
                                                        Stratified by stage
```

If we re-do a test for non-proportional hazards we find that there is no longer evidence that any of the remaining covariates effects seem to depend on time since diagnosis.

Having accounted for the time-dependent effect of stage, there is still no evidence that the effect of sex is modified by age at diagnosis.

```
stcox i.sex#i.agegrp year8594 i.agegrp i.subsite, strata(stage)
test 2.sex#0.agegrp = 2.sex#1.agegrp = 2.sex#2.agegrp = 2.sex#3.agegrp
```

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  sex#agegrp |
       2 0   |   .6115151   .0647711    -4.64   0.000     .4968768    .7526024
       2 1   |   .7330985   .0682897    -3.33   0.001     .6107606    .8799411
       2 2   |   .8004243   .0644649    -2.76   0.006     .6835429    .9372916
       2 3   |   .7982689   .0852012    -2.11   0.035     .6475874    .9840111
             |
    year8594 |    .788275   .0376984    -4.97   0.000     .7177446    .8657361
             |
     agegrp  |
          1  |   1.171996   .1000088     1.86   0.063     .9914973    1.385355
          2  |   1.549262   .1316249     5.15   0.000     1.311617    1.829964
          3  |   2.447562    .256747     8.53   0.000     1.992707    3.006242
             |
     subsite |
          2  |   1.345398   .0950902     4.20   0.000     1.171357    1.545297
          3  |   1.002342   .0744343     0.03   0.975     .8665735    1.159382
          4  |   1.260847   .1296178     2.25   0.024     1.030758    1.542296
------------------------------------------------------------------------------
                                                          Stratified by stage
```

```
 ( 1)  2.sex#0b.agegrp - 2.sex#1.agegrp = 0
 ( 2)  2.sex#0b.agegrp - 2.sex#2.agegrp = 0
 ( 3)  2.sex#0b.agegrp - 2.sex#3.agegrp = 0

           chi2(  3) =    4.79
         Prob > chi2 =    0.1878
```

If you have time make sure you check for additional interaction terms between the remaining covariates, i.e. between age at diagnosis and stage.

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

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

Poisson regression                                Number of obs   =        337
                                                  LR chi2(1)      =       4.82
                                                  Prob > chi2     =     0.0282
Log likelihood =  -175.0016                       Pseudo R2       =     0.0136

------------------------------------------------------------------------------
         chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |   .5203602   .1572055    -2.16   0.031     .2878382    .9407184
           y | (exposure)
------------------------------------------------------------------------------

. stset dox, id(id) fail(chd) origin(doe) scale(365.25)
. stcox hieng

Cox regression -- no ties

No. of subjects =          337                    Number of obs   =       337
No. of failures =           46
Time at risk    =  4603.794765
                                                  LR chi2(1)      =      4.73
Log likelihood  =   -253.32253                    Prob > chi2     =    0.0296

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |   .5233587     .15814    -2.14   0.032     .2894658    .9462409
------------------------------------------------------------------------------
```

These two models are conceptually different since the Cox model adjusts for 'time' even though this is not explicit in the `stcox` command. In this example, 'time' refers to 'time on study' (time since entry) which we do not expect to be a strong confounder. That is, we would expect the estimates of the effect of high energy to be similar for the two models, which they are.

(b) If we use a different timescale then this amounts to adjusting for a different factor. As such, we would not expect the estimates to be identical. Attained age, unlike time since entry, is expected to be a confounder but we see that it is not a strong confounder.

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

Cox regression -- Breslow method for ties

No. of subjects =          337                    Number of obs   =       337
No. of failures =           46
Time at risk    =  4603.794765
                                                  LR chi2(1)      =      4.20
Log likelihood  =   -234.78217                    Prob > chi2     =    0.0405

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

22. **Estimating the effect of a time-varying exposure**

(a) `. use brv, clear`

```
. list id sex doe dosp dox fail if couple==3
     +-----------------------------------------------------+
     | id   sex         doe        dosp         dox   fail |
     |-----------------------------------------------------|
168. | 60     1   20jan1981   31dec1981   03aug1981      1 |
384. | 63     2   20jan1981   03aug1981   31dec1981      1 |
     +-----------------------------------------------------+

. list id sex doe dosp dox fail if couple==4
     +-----------------------------------------------------+
     |  id   sex         doe        dosp         dox   fail |
     |-----------------------------------------------------|
 12. | 156     1   20jan1981   23nov1988   01jan1991      0 |
300. | 220     2   20jan1981   01jan2000   23nov1988      1 |
     +-----------------------------------------------------+

. list id sex doe dosp dox fail if couple==19
     +-----------------------------------------------------+
     |   id   sex         doe        dosp         dox   fail |
     |-----------------------------------------------------|
167. | 2122     1   06may1981   01jan2000   01jan1991      0 |
298. | 2128     2   06may1981   01jan2000   01jan1991      0 |
     +-----------------------------------------------------+
```

(b) `. stset dox, fail(fail) origin(dob) entry(doe) scale(365.24) id(id) noshow`

```
                id:  id
     failure event:  fail != 0 & fail < .
obs. time interval:  (dox[_n-1], dox]
 enter on or after:  time doe
 exit on or before:  failure
    t for analysis:  (time-origin)/365.24
            origin:  time dob
------------------------------------------------------------------------------
      399  total obs.
        0  exclusions
------------------------------------------------------------------------------
      399  obs. remaining, representing
      399  subjects
      278  failures in single failure-per-subject data
 2435.708  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =  75.13963
                                  last observed exit t =  96.50641

. strate sex, per(1000)

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(399 records included in the analysis)
  +-------------------------------------------------+
  | sex     D        Y      Rate     Lower     Upper |
  |-------------------------------------------------|
  |   1   181   1.3405   135.022   116.717   156.198 |
  |   2    97   1.0952    88.569    72.587   108.071 |
  +-------------------------------------------------+
```

i. The timescale is attained age, which would seem to be a reasonable choice.

ii. Males have the higher mortality which is to be expected.

iii. Age could potentially be a confounder.

```
. tabstat _t0, by(sex)

Summary for variables: _t0
     by categories of: sex (1=M, 2=F)

     sex |      mean
---------+----------
       1 |  79.06936
       2 |   78.6578
---------+----------
   Total |  78.90123
--------------------
```

Males are slightly older at diagnosis (although we haven't studied pairwise differences).

```
. streg sex, dist(exp) nolog
Exponential regression -- log relative-hazard form
No. of subjects =          399              Number of obs  =  399
No. of failures =          278
Time at risk    =  2435.641342
                                            LR chi2(1)     =  11.64
Log likelihood  =    355.79411              Prob > chi2    =  0.0006
------------------------------------------------------------------------------
 _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----+-------------------------------------------------------------------------
sex |   .6559621   .0825422    -3.35   0.001     .5125885     .839438
------------------------------------------------------------------------------
```

(c)
```
. stsplit brv, after(time=dosp) at(0)
. recode brv -1=0 0=1
(brv: 555 changes made)
```

(d)
```
. streg brv, distribution(exponential) nolog
Exponential regression -- log relative-hazard form
No. of subjects =          399          Number of obs   =       555
No. of failures =          278
Time at risk    =  2435.641342
                                        LR chi2(1)      =      0.81
Log likelihood  =    350.37937          Prob > chi2     =    0.3686
------------------------------------------------------------------------------
 _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----+-------------------------------------------------------------------------
brv |   1.127154    .148775     0.91   0.364      .870225    1.459939
------------------------------------------------------------------------------
```

(e)
```
. streg brv if sex==1, nolog
Exponential regression -- log relative-hazard form
No. of subjects =          236        Number of obs   =       295
No. of failures =          181
Time at risk    =    1340.4846
                                      LR chi2(1)      =      0.00
Log likelihood  =    258.40461        Prob > chi2     =    0.9548
------------------------------------------------------------------------------
         _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        brv |   1.010863   .1923683     0.06   0.955     .6961579    1.467834
------------------------------------------------------------------------------

. streg brv if sex==2, nolog
Exponential regression -- log relative-hazard form
No. of subjects =          163        Number of obs   =       260
No. of failures =           97
Time at risk    =  1095.156742
                                      LR chi2(1)      =      5.62
Log likelihood  =    100.20223        Prob > chi2     =    0.0177
------------------------------------------------------------------------------
         _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        brv |   1.624613   .3300669     2.39   0.017     1.090974    2.419277
------------------------------------------------------------------------------
```

Now we create indicator variables (`brv_m` and `brv_f`) to allow us to estimate the effect of bereavement separately for each sex.

```
. streg i.sex i.brv#i.sex, dist(exp)

Iteration 0:   log likelihood =  349.97514
Iteration 1:   log likelihood =  358.42347
Iteration 2:   log likelihood =  358.60677
Iteration 3:   log likelihood =  358.60684
Iteration 4:   log likelihood =  358.60684

Exponential regression -- log relative-hazard form

No. of subjects =          399                     Number of obs   =       555
No. of failures =          278
Time at risk    =  2435.708028
                                                   LR chi2(3)      =     17.26
Log likelihood  =    358.60684                     Prob > chi2     =    0.0006

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |   .5348431    .087562    -3.82   0.000     .3880357     .737193
             |
     brv#sex |
        1 1  |   1.010863   .1923683     0.06   0.955     .6961579    1.467834
        1 2  |   1.624613   .3300669     2.39   0.017     1.090974    2.419277
------------------------------------------------------------------------------
```

(f)
```
. /* Split by attained age */
. stsplit age, at(70(5)100)
(481 observations (episodes) created)

. strate age

Estimated rates and lower/upper bounds of 95% confidence intervals
(1036 records included in the analysis)
  +-------------------------------------------------------+
  | age     D          Y       Rate      Lower      Upper |
  |-------------------------------------------------------|
  |  75    45   703.6124   0.063956   0.047752   0.085658 |
  |  80   123    1.2e+03   0.103825   0.087007   0.123895 |
  |  85    95   490.0214   0.193869   0.158554   0.237050 |
  |  90    12    55.0904   0.217824   0.123704   0.383554 |
  |  95     3     2.2999   1.304429   0.420706   4.044471 |
  +-------------------------------------------------------+

. /* Poisson regression: effect of bereavement
                                    controlled for attained age */
. streg brv i.age, nolog
                                                LR chi2(5)      =     56.61
Log likelihood  =    378.28189                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         brv |   .8594122   .1178685    -1.10   0.269     .6568393     1.12446
             |
         age |
         80  |    1.66633    .292713     2.91   0.004     1.180962     2.35118
         85  |   3.198481    .597915     6.22   0.000      2.21729    4.613866
         90  |   3.613713   1.188938     3.90   0.000     1.896279    6.886607
         95  |   20.97061   12.51454     5.10   0.000     6.510932    67.54276
------------------------------------------------------------------------------
```

```
. /* Poisson regression: effect of bereavement
                               controlled for attained age and sex */
. streg brv i.age sex, nolog
                                                LR chi2(6)      =     71.38
Log likelihood  =    385.66573                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         brv |   .9735923   .1364956    -0.19   0.849     .7396742    1.281486
             |
         age |
         80  |   1.675997   .2944392     2.94   0.003     1.187774    2.364897
         85  |   3.171938   .5908462     6.20   0.000     2.201754    4.569624
         90  |    3.65729   1.203318     3.94   0.000     1.919102     6.96981
         95  |   27.80767   16.74873     5.52   0.000     8.540449    90.54167
             |
         sex |    .611474   .0798274    -3.77   0.000     .4734285    .7897718
------------------------------------------------------------------------------
```

(g)
```
. /* Poisson regression: effect of bereavement for each
                                 gender (controlled for attained age) */
. streg i.age i.sex i.brv#i.sex, nolog dist(exp)

Exponential regression -- log relative-hazard form

No. of subjects =          399                  Number of obs   =      1036
No. of failures =          278
Time at risk    =  2435.708028
                                                LR chi2(7)      =     73.22
Log likelihood  =    386.58403                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |
         80  |   1.677943   .2948222     2.95   0.003     1.189097    2.367757
         85  |   3.129915   .5842027     6.11   0.000     2.170974    4.512429
         90  |   3.655497   1.203045     3.94   0.000     1.917834    6.967575
         95  |   28.74863   17.34039     5.57   0.000     8.814459    93.76454
             |
       2.sex |   .5368135   .0889125    -3.76   0.000     .3880064    .7426907
             |
     brv#sex |
        1 1  |    .823687   .1585562    -1.01   0.314     .5648194    1.201199
        1 2  |   1.199917   .2501707     0.87   0.382     .7974142    1.805586
------------------------------------------------------------------------------
```

(h) We could split the post bereavement period into multiple categories (e.g., within one year and subsequent to one year following bereavement) and compare the risks between these categories.

(i)
```
. /* Cox regression: effect of brv controlled for attained age */
. stcox brv, nolog

Cox regression -- Breslow method for ties

No. of subjects =          399         Number of obs   =      1036
No. of failures =          278
Time at risk    =  2435.641342
                                       LR chi2(1)      =      2.25
Log likelihood  =   -1379.1483         Prob > chi2     =    0.1333
------------------------------------------------------------------------------
 _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----+-------------------------------------------------------------------------
brv |   .8134514   .1131032    -1.48   0.138     .6194119    1.068276
------------------------------------------------------------------------------

. /* Cox: effect of brv controlled for attained age and sex */
. stcox brv sex, nolog

Cox regression -- Breslow method for ties

No. of subjects =          399       Number of obs   =      1036
No. of failures =          278
Time at risk    =  2435.641342
                                       LR chi2(2)      =     15.82
Log likelihood  =   -1372.3656         Prob > chi2     =    0.0004
------------------------------------------------------------------------------
 _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----+-------------------------------------------------------------------------
brv |   .9249887   .1317637    -0.55   0.584     .6996545    1.222895
sex |   .6233905   .0815085    -3.61   0.000     .4824643    .8054806
------------------------------------------------------------------------------
```

(j)
```
. /* Cox regression estimating the effect of brv for each sex,
        controlling for age */
. stcox i.sex i.sex#i.brv, nolog

Cox regression -- Breslow method for ties

No. of subjects =          399                     Number of obs   =      1036
No. of failures =          278
Time at risk    =  2435.708028
                                                   LR chi2(3)      =     17.08
Log likelihood  =   -1371.7342                     Prob > chi2     =    0.0007

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |   .5592749   .0925961    -3.51   0.000     .4042933     .773667
             |
     sex#brv |
        1 1  |   .8055967    .155495    -1.12   0.263     .5518488    1.176022
        2 1  |   1.103135   .2337666     0.46   0.643      .728198     1.67112
------------------------------------------------------------------------------
```

23. **Calculating SMRs/SIRs**

(a)
```
. use melanoma, clear
(Skin melanoma, all stages, Finland 1975-94, follow-up to 1995)

. stset exit, fail(status == 1 2) origin(bdate) entry(dx) scale(365.25) id(id)

                id:  id
     failure event:  status == 1 2
obs. time interval:  (exit[_n-1], exit]
 enter on or after:  time dx
 exit on or before:  failure
    t for analysis:  (time-origin)/365.25
            origin:  time bdate

--------------------------------------------------------------------------------
     7775  total obs.
        0  exclusions
--------------------------------------------------------------------------------
     7775  obs. remaining, representing
     7775  subjects
     3047  failures in single failure-per-subject data
 51377.85  total analysis time at risk, at risk from t =         0
                              earliest observed entry t =         0
                                   last observed exit t =  101.4593
. stsplit _age, at(0(1)110) trim
(no obs. trimmed because none out of range)
(50508 observations (episodes) created)
```

(b)
```
. stsplit _year, after(time=d(1/1/1900)) at(70(1)100) trim
(no obs. trimmed because none out of range)
(48743 observations (episodes) created)
. tab _year

      _year |      Freq.     Percent        Cum.
------------+-----------------------------------
         75 |        423        0.40        0.40
         76 |        711        0.66        1.06
         77 |      1,102        1.03        2.09
          :
        output omitted
          :
         92 |      8,841        8.26       72.64
         93 |      9,408        8.79       81.43
         94 |     10,070        9.41       90.84
         95 |      9,808        9.16      100.00
------------+-----------------------------------
      Total |    107,026      100.00

. replace _year=1900+_year
_year was byte now int
(107026 real changes made)

. tab _year

      _year |      Freq.     Percent        Cum.
------------+-----------------------------------
       1975 |        423        0.40        0.40
       1976 |        711        0.66        1.06
       1977 |      1,102        1.03        2.09
          :
        output omitted
          :
       1992 |      8,841        8.26       72.64
       1993 |      9,408        8.79       81.43
       1994 |     10,070        9.41       90.84
       1995 |      9,808        9.16      100.00
------------+-----------------------------------
      Total |    107,026      100.00
```

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

----------------------------------------
ageband_1 |           period_5
0         | 1975  1980  1985  1990  1995
----------+-----------------------------
        0 |    0     0     0     1
       10 |    2     1     0     0     0
       20 |    8    10    11     8     1
       30 |   18    45    49    30     4
       40 |   42    56    80   102    29
       50 |   45    96   105   139    32
       60 |   74   128   178   185    48
       70 |   53   153   226   279    54
       80 |   29    81   158   298    63
       90 |    1    12    33    64    12
      100 |          1           3
----------------------------------------

. table ageband_10 period_5, c(sum _y) format(%5.3f)

------------------------------------------------------------
ageband_1 |                    period_5
0         |     1975      1980      1985      1990      1995
----------+-------------------------------------------------
        0 |    0.824    17.713    13.021     1.318
       10 |   25.785    36.949    67.034    82.780    11.478
       20 |  153.075   357.187   580.864   726.161   123.450
       30 |  318.515  1061.527  1646.500  1913.121   390.853
       40 |  465.988  1376.970  2708.173  4069.910   850.004
       50 |  570.943  1683.505  3004.309  4487.079  1025.253
       60 |  562.719  1563.439  3037.539  4666.730  1061.766
       70 |  378.820  1307.218  2407.125  3723.926   865.351
       80 |   92.853   378.127   961.110  1806.420   435.370
       90 |   10.068    29.621    87.367   184.157    44.507
      100 |              0.624              2.705
------------------------------------------------------------
```

(d)
```
. gen obsrate=_d/_y

. table ageband_10 period_5 [iw=_y] , c(mean obsrate) format(%5.3f)

-------------------------------------------
ageband_1 |             period_5
0         |  1975   1980   1985   1990   1995
----------+--------------------------------
        0 | 0.000  0.000  0.000  0.759
       10 | 0.078  0.027  0.000  0.000  0.000
       20 | 0.052  0.028  0.019  0.011  0.008
       30 | 0.057  0.042  0.030  0.016  0.010
       40 | 0.090  0.041  0.030  0.025  0.034
       50 | 0.079  0.057  0.035  0.031  0.031
       60 | 0.132  0.082  0.059  0.040  0.045
       70 | 0.140  0.117  0.094  0.075  0.062
       80 | 0.312  0.214  0.164  0.165  0.145
       90 | 0.099  0.405  0.378  0.348  0.270
      100 |        1.602         1.109
-------------------------------------------
```

(e)
```
. sort _year sex _age

. merge _year sex _age using popmort
variables _year sex _age do not uniquely identify observations in the master data
_year was int now float
_age was int now float

. tab _merge

     _merge |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |      7,180        6.29        6.29
          3 |    107,026       93.71      100.00
------------+-----------------------------------
      Total |    114,206      100.00

. drop if _merge==2
(7180 observations deleted)

. drop _merge
```

(f)
```
. gen mortrate=(-ln(prob))
. gen e=_y*mortrate
. list e _d in 1/10
     +---------------+
     |        e   _d |
     |---------------|
  1. | .0005001    0 |
  2. | 3.21e-06    0 |
  3. | .0011804    0 |
  4. | 9.73e-07    0 |
  5. | .0012089    0 |
     |---------------|
  6. | 3.43e-06    0 |
  7. | .0005415    0 |
  8. | 1.34e-06    0 |
  9. | .0007909    0 |
 10. | 2.59e-06    0 |
     +---------------+
```

(g)
```
. egen obs=total(_d)
. egen exp=total(e)
.
. 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
. restore
.
. display "SMR(95%CI)=" round(SMR,.001) " (" round(LL,.001) ":" round(UL,.001) ")"
SMR(95%CI)=2.411 (2.327:2.499)

. strate, smr(mortrate)

Estimated SMRs and lower/upper bounds of 95% confidence intervals
(107026 records included in the analysis)
  +---------------------------------------+
  |    D         E     SMR   Lower   Upper |
  |---------------------------------------|
  | 3047   1263.57   2.411   2.327   2.499 |
  +---------------------------------------+
```

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

Estimated SMRs and lower/upper bounds of 95% confidence intervals
(107026 records included in the analysis)
  +-----------------------------------------------------+
  |     stage      D        E      SMR    Lower    Upper |
  |-----------------------------------------------------|
  |   Unknown    557   284.42    1.958    1.802    2.128 |
  | Localised   1795   915.27    1.961    1.873    2.054 |
  |  Regional    260    37.55    6.925    6.132    7.820 |
  |   Distant    435    26.33   16.520   15.039   18.148 |
  +-----------------------------------------------------+
```

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

First, fit a Cox model to the full cohort

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

 Cox regression -- Breslow method for ties

No. of subjects =         5318                     Number of obs   =      5318
No. of failures =          960
Time at risk    =       388520
                                                   LR chi2(5)      =    211.94
Log likelihood  =   -7794.4811                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |   .5888144   .0385379    -8.09   0.000     .5179256    .6694059
  1.year8594 |   .7168836   .0474446    -5.03   0.000     .6296723    .8161739
             |
      agegrp |
          1  |   1.326397   .1249113     3.00   0.003     1.102841     1.59527
          2  |   1.857323   .1687866     6.81   0.000     1.554295     2.21943
          3  |   3.372652   .3522268    11.64   0.000     2.748371    4.138736
------------------------------------------------------------------------------
```

(a) There are 5318 individuals in the study that we would need to collect data for if we were to use the complete cohort of patients.

(b) 960 cancer patients die from melanoma during the first 10 years of follow-up.

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

(d) We knew that 960 patients in the cohort experienced the event of interest within 10 years. The sttoc command confirms this by stating that there are 960 cases with 1 control sampled per case. The cases are all unique, but the 960 controls are not necessarily unique so there will be fewer than 1920 individuals in our study. An individual may be both a case and a control and/or may be a control to several cases. Some epidemiologists struggle with this concept, but it is easy to understand when we think in terms of sampling from risk sets. In standard Cox regression analysis, for example, every individual is a control for every case until they themselves become a case. Restricting a nested case-control study to consist of unique individuals will result in biased estimates.

Your results will differ, but our nested case-control study contained 1700 unique individuals.

```
. duplicates report id
--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |         1497             0
        2 |          374           187
        3 |           45            30
        4 |            4             3
--------------------------------------
```

(e)
```
est table Complete_Cox Nested_CC, eform equations(1) ///
b(%9.6f) se modelwidth(10) title("Hazard ratio")

Hazard ratios and standard errors from the two models
---------------------------------------
    Variable | Complete~x   Nested_CC
-------------+-------------------------
         sex |
           2 |   0.588814     0.611493
             |   0.038538     0.058835
             |
    year8594 |
           1 |   0.716884     0.724686
             |   0.047445     0.072384
             |
      agegrp |
           1 |   1.326397     1.342203
             |   0.124911     0.174505
           2 |   1.857323     1.966943
             |   0.168787     0.257653
           3 |   3.372652     4.404641
             |   0.352227     0.787285
---------------------------------------
                            legend: b/se
```

Note that, since every nested case-control study is different, the parameter estimates you obtain will not be identical to those above. However, the hazard ratios from the two models should be very similar. The standard errors are slightly larger for the nested case-control study since the estimates are based on a sample from the full cohort. Loss of precision is the trade-off we have to make when designing a nested case-control study. The precision can be improved by adding more controls to each case.

(f) Following is some code to estimate and analyse the nested case-control study 5 times.

```
/* Repeatedly generate nested case-control studies */
set more off
use melanoma, clear
keep if stage == 1
stset surv_mm, failure(status==1) id(id) exit(time 120)

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

Here are the results. We see that there is sampling variation in the parameter estimates from the five nested case-control studies but they are centered on the full cohort estimate. We see that the standard errors of the estimates from the nested case-control studies are larger than for the full cohort but there is some sampling variation.

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

```
-------------------------------------------------------------------------------
Variable |   Complete     ncc1        ncc2        ncc3        ncc4       ncc5
---------+---------------------------------------------------------------------
     sex |
       2 |   0.588814    0.616907    0.602383    0.544285    0.574463   0.599772
         |   0.038538    0.060836    0.057810    0.051935    0.057257   0.059603
         |
year8594 |
       1 |   0.716884    0.699482    0.762841    0.747950    0.811977   0.715201
         |   0.047445    0.069447    0.076288    0.074391    0.083310   0.069803
         |
  agegrp |
       1 |   1.326397    1.272060    1.350298    1.208072    1.321977   1.398562
         |   0.124911    0.163739    0.178126    0.155366    0.169123   0.180422
       2 |   1.857323    1.931832    1.841300    1.890836    1.700583   2.157252
         |   0.168787    0.250121    0.239062    0.242986    0.216667   0.286852
       3 |   3.372652    3.678843    3.248771    3.359871    3.763965   2.996758
         |   0.352227    0.618735    0.549156    0.568002    0.648790   0.486675
-------------------------------------------------------------------------------
                                                                legend: b/se
```

28. **Royston-Parmar Models**

(a) The `stpm2` output can be seen below.

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

Log likelihood = -3249.8427                     Number of obs   =       5318
------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
    year8594 |   .7840079   .0516192    -3.70   0.000      .689092    .8919975
       _rcs1 |    2.47851   .0741825    30.33   0.000     2.337297    2.628254
       _rcs2 |   1.318875   .0481845     7.58   0.000     1.227737    1.416778
       _rcs3 |   1.071562   .0207102     3.58   0.000      1.03173    1.112932
       _rcs4 |   1.000146   .0077078     0.02   0.985     .9851525    1.015368
------------------------------------------------------------------------------
```

The hazard ratio, 95% confidence interval and statistical significance are very similar to the Cox model.

(b) The predicted survival and hazard functions are shown in Figures 20 and 21.

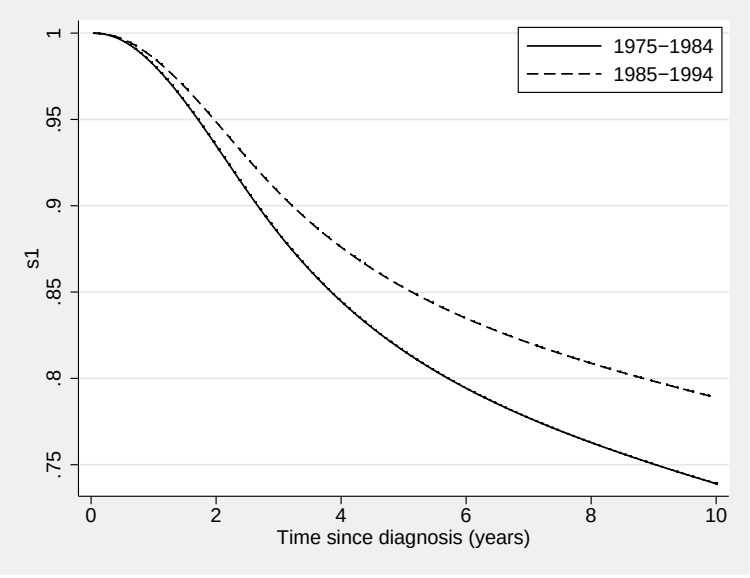

Figure 20: Localised skin melanoma. Predicted survival functions from a flexible parametric model.

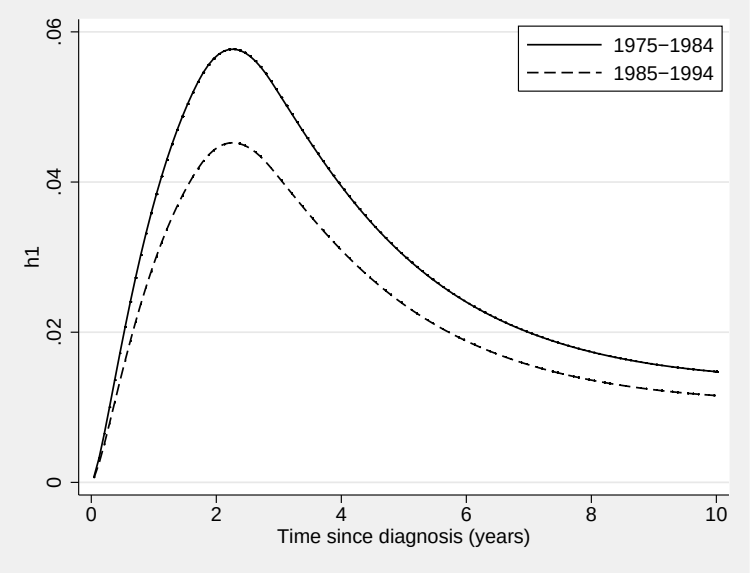

Figure 21: Localised skin melanoma. Predicted hazard functions from a flexible parametric model.

(c) Plotting on the log scale is shown in Figure 22. There is a constant difference as the predictions are from a proportional hazards model and a multiplicative effect becomes additive on the log scale.

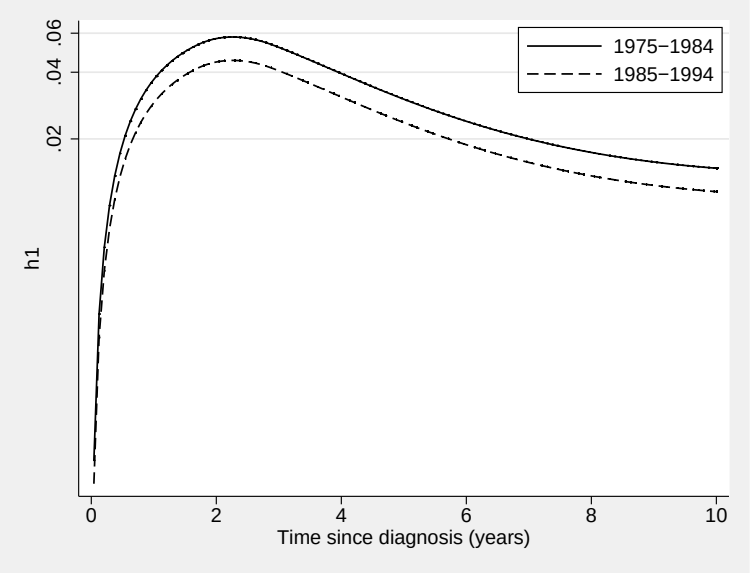

Figure 22: Localised skin melanoma. Predicted hazard functions on log scale from a flexible parametric model.

(d) The estimates are shown below

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

-------------------------------------------------------------------------------
Variable |    df1          df2          df3          df4          df5          df6
---------+---------------------------------------------------------------------
year8594 | -.11573888   -.23960014   -.24389406    -.2433362   -.24554507   -.24591639
         |  .06575724    .06584093    .06581084     .0658401    .06580441    .06580344
---------+---------------------------------------------------------------------
     AIC |  6736.0089    6512.0253    6511.6302    6511.6853    6506.6137    6507.6984
     BIC |  6750.6097     6531.493    6535.9648    6540.8869    6540.6822    6546.6339
-------------------------------------------------------------------------------
```

The log hazard ratios (and hence the hazard ratios) from 2 df and up are similar and for 3 df they are very similar. The main difference is for 1 df, which is equivalent to a Weibull model. The Weibull model enforces a monotonic hazard function and as the hazard function in the melanoma data has a turning point it is clearly inappropriate.
The lowest AIC is for the model with 5 df and for the BIC it is the model with 2 df. The penalty term in the AIC is twice the number of parameters ($2 \times k$) whereas in the BIC it is $\ln(D) \times k$ where D is the number of events. Since $\ln(D) > k$ the BIC penalizes extra parameters much more strongly than AIC. Since we have a large data set and there are no disadvantages to including extra parameters we would use 5df for the baseline hazard.

(e) The baseline survival functions are shown in Figure 23 and the hazard functions in Figure 24.

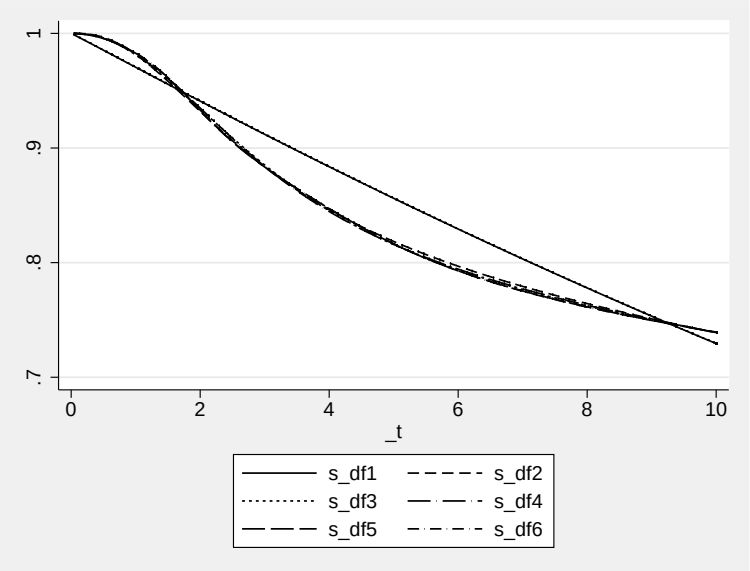

Figure 23: Localised skin melanoma. Predicted survival functions for various df from a flexible parametric model.

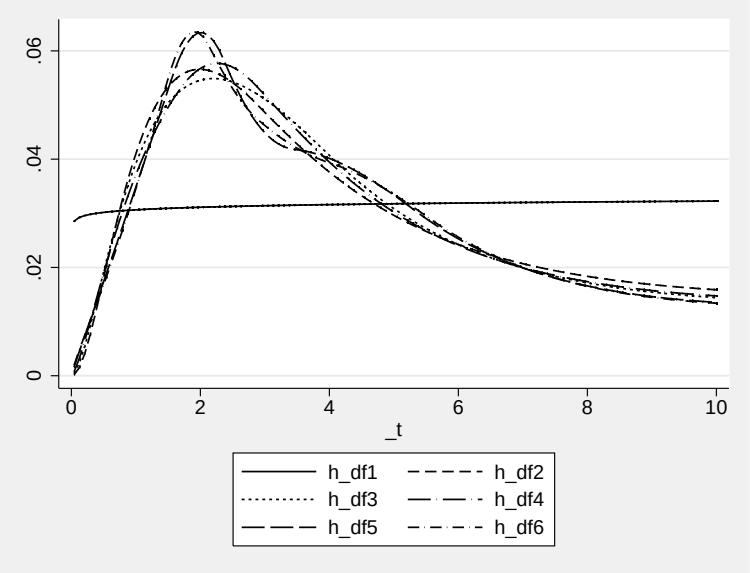

Figure 24: Localised skin melanoma. Predicted hazard functions for various df from a flexible parametric model.

With the exception of 1 df (the Weibull model), the survival and hazard functions show similar shapes, so as long we have enough knots our conclusions would be very similar.

(f) The model is shown below

```
. stpm2 i.sex year8594 i.agegrp, df(4) scale(hazard) nolog eform

Log likelihood = -3150.2605                     Number of obs   =       5318

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
         sex |
           1 |          1  (base)
           2 |   .5876853   .0384659    -8.12   0.000      .516929    .6681266
             |
    year8594 |   .7231987   .0479007    -4.89   0.000     .6351536    .8234488
             |
      agegrp |
           0 |          1  (base)
           1 |   1.327816   .1250451     3.01   0.003     1.104021    1.596976
           2 |   1.862093   .1692228     6.84   0.000     1.558282    2.225136
           3 |   3.400181   .3551542    11.72   0.000      2.77072    4.172644
             |
       _rcs1 |   2.545712       .077    30.89   0.000     2.399181    2.701192
       _rcs2 |   1.310266   .0479852     7.38   0.000     1.219513    1.407773
       _rcs3 |   1.073125   .0209887     3.61   0.000     1.032767    1.115061
       _rcs4 |   1.000181   .0080248     0.02   0.982     .9845751    1.016033
       _cons |   .1371494   .0115726   -23.54   0.000     .1162439    .1618147
------------------------------------------------------------------------------

. estimates store ph
```

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

 ( 1)  [xb]1.agegrp = 0
 ( 2)  [xb]2.agegrp = 0
 ( 3)  [xb]3.agegrp = 0

           chi2(  3) =  155.78
         Prob > chi2 =    0.0000
```

The estimates are similar to those obtained from the Cox model. The Wald test yields a very highly significant result, which is similar to that obtained from the comparable test for the Cox model.

(g) The estimates are so similar because very similar models are being fitted with exactly the same covariates. The two models differ only in the manner in which they account for the baseline hazard. In the Cox model it is assumed arbitrary and not directly estimated. In the flexible parametric model the baseline hazard is modelled using splines. The 5 df spline allows sufficient flexibility to make the model estimates virtually identical.

(h) The model including a time-varying effect for age is shown below

```
. stpm2 i.sex year8594 agegrp2-agegrp4, df(4) scale(hazard) ///
>         tvc(agegrp2 agegrp3 agegrp4) dftvc(2) nolog eform

Log likelihood =  -3135.625                     Number of obs = 5318
------------------------------------------------------------------------------
             |   exp(b)   Std. Err.      z    P>|z|   [95% Conf. Interval]
-------------+----------------------------------------------------------------
  x          |
         sex |
           1 |          1  (base)
           2 |  .5911324   .0386738    -8.04   0.000   .5199916     .672006
             |
    year8594 |  .7201623   .0477596    -4.95   0.000   .6323836    .8201252
     agegrp2 |  1.465683   .1595012     3.51   0.000   1.184157    1.814141
     agegrp3 |  2.081347   .2167703     7.04   0.000   1.697042     2.55268
     agegrp4 |  3.958773   .4523931    12.04   0.000    3.16438    4.952592
       _rcs1 |   3.24619   .3119881    12.25   0.000   2.688845    3.919061
       _rcs2 |  1.472048   .1259346     4.52   0.000   1.244804    1.740775
       _rcs3 |  1.069076   .0216773     3.29   0.001   1.027423    1.112418
       _rcs4 |  1.001397   .0080042     0.17   0.861   .9858309    1.017208
_rcs_agegrp21 |  .8390881   .0957004    -1.54   0.124   .6710048    1.049275
_rcs_agegrp22 |   .934734   .0872182    -0.72   0.469   .7785099    1.122308
_rcs_agegrp31 |  .8321523   .0912377    -1.68   0.094   .6712381    1.031642
_rcs_agegrp32 |  .9934604   .0893837    -0.07   0.942   .8328488    1.185045
_rcs_agegrp41 |   .643227    .069256    -4.10   0.000   .5208537    .7943515
_rcs_agegrp42 |  .8421151   .0763089    -1.90   0.058   .7050812    1.005782
        _cons |  .1206609   .0114631   -22.26   0.000   .1001614    .1453559
------------------------------------------------------------------------------

. estimates store nonph
```

The likelihood ratio test is shown below.

```
. lrtest ph nonph

Likelihood-ratio test                                 LR chi2(6)  =     29.27
(Assumption: ph nested in nonph)                      Prob > chi2 =    0.0001
```

There is strong evidence of a non-proportional effect of age.

(i) The baseline hazard is shown in Figure 25. This baseline is for the youngest age group who are male and diagnosed in 1975–1984, i.e, when all the covariates are equal to zero.

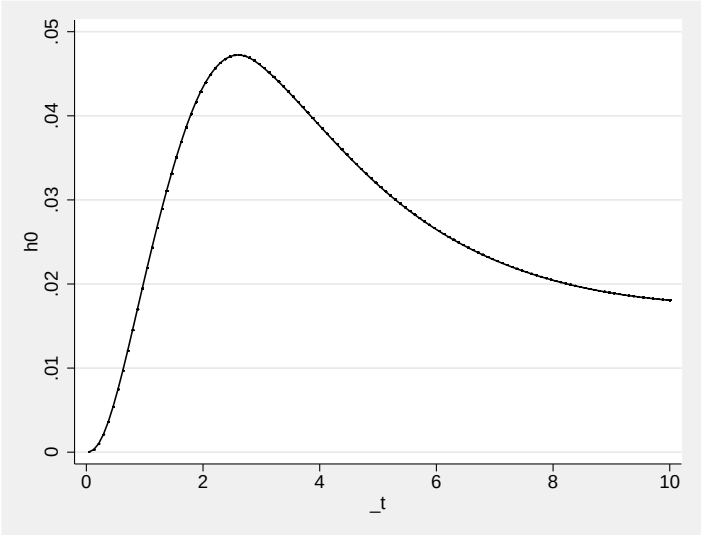

Figure 25: Localised skin melanoma. Predicted baseline hazard functions from a flexible parametric model.

(j) The hazard ratios are plotted in Figure 26.

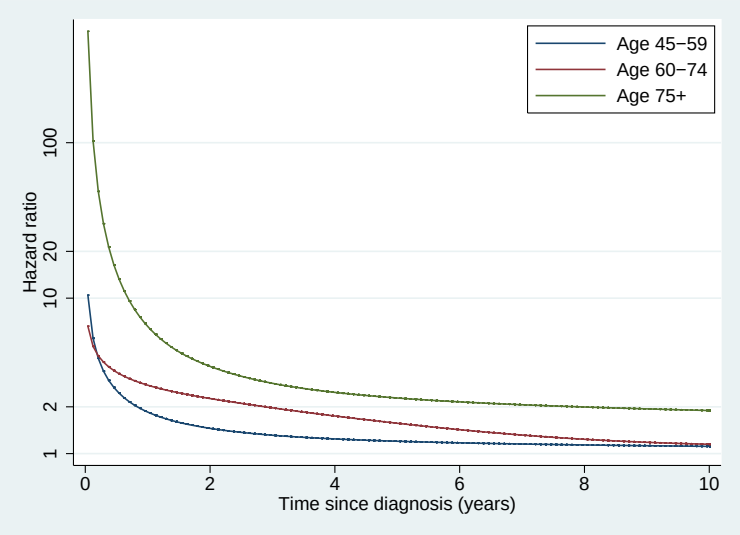

Figure 26: Localised skin melanoma. Predicted time-dependent hazard ratios for age from a flexible parametric model.

The hazard ratios decrease as a function of follow-up time. The hazard ratio is so high during the early years of follow-up because the hazard in the reference group is close to zero

(Figure 25). The hazard ratio for the oldest age group with 95% confidence intervals is shown in Figure 27.

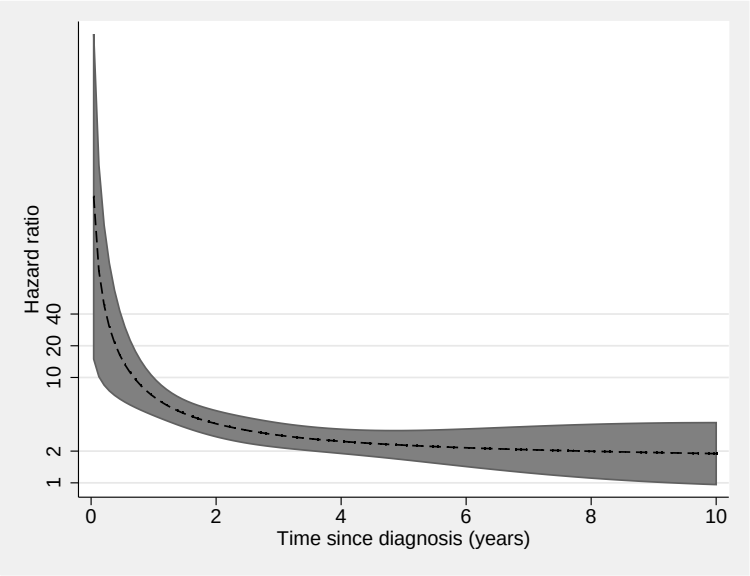

Figure 27: Localised skin melanoma. Predicted time-dependent hazard ratio (with a 95% confidence interval) for oldest age group from a flexible parametric model.

(k) The hazard difference for the oldest age group (with other covariates set to zero) is shown in Figure 28.

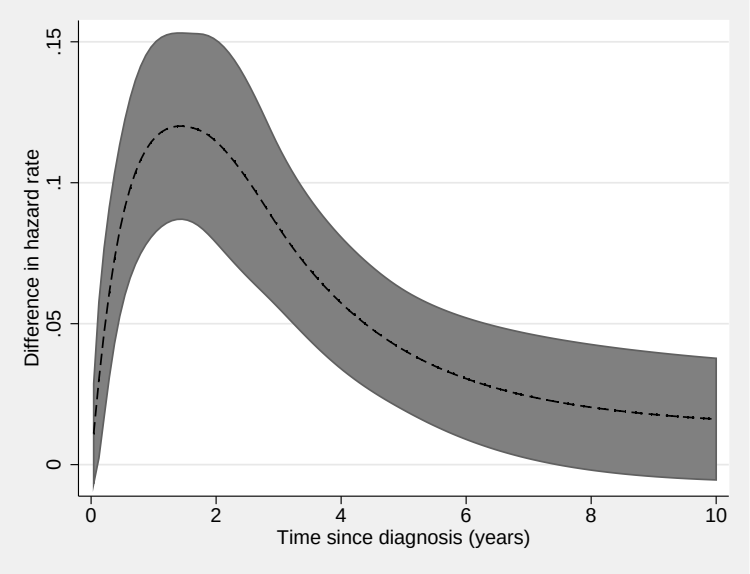

Figure 28: Localised skin melanoma. Predicted difference in hazard rates (with a 95% confidence interval) for oldest age group from a flexible parametric model (other covariates are set to zero).

The hazard difference is small early on, despite the hazard ratio being large, because the underlying hazard is so low.

(l) The survival difference, with 95% confidence interval, is shown in Figure 29.

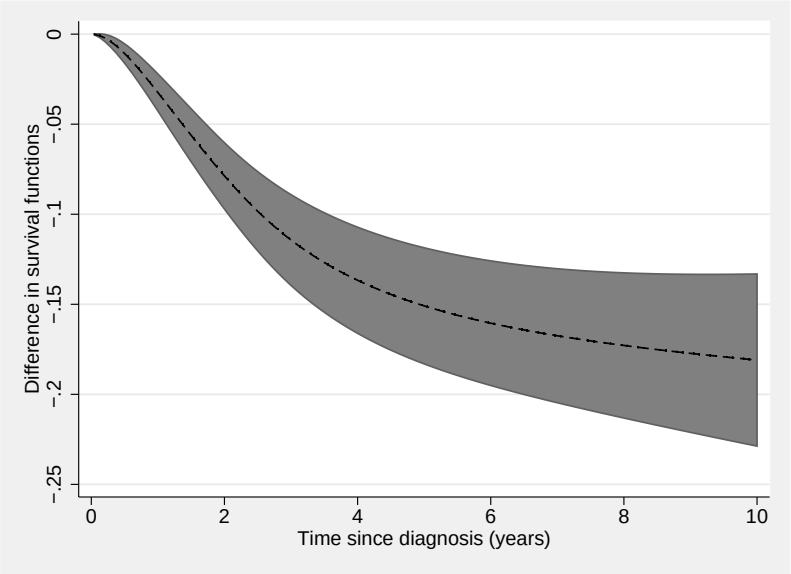

Figure 29: Localised skin melanoma. Predicted difference in survival functions (with a 95% confidence interval) for oldest age group from a flexible parametric model (for females diagnosed in 1984-1994).

(m) The AIC and BIC for the different models are

```
. count if _d==1
  960

. estimates stats dftvc*, n(‘r(N)’)

-------------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC          BIC
-------------+-----------------------------------------------------------------
      dftvc1 |    960           .   -3139.579     13     6305.158     6368.428
      dftvc2 |    960           .   -3135.625     16      6303.25     6381.121
      dftvc3 |    960           .    -3134.91     19      6307.82     6400.292
-------------------------------------------------------------------------------
               Note:  N=960 used in calculating BIC
```

AIC selects 2 df and BIC selects 1 df. As discussed in the previous part, the BIC imposes a stronger penalty on aditional parameters. The fitted time-dependent effects are similar. We would suggest 2 df for the time-varying effect of age.

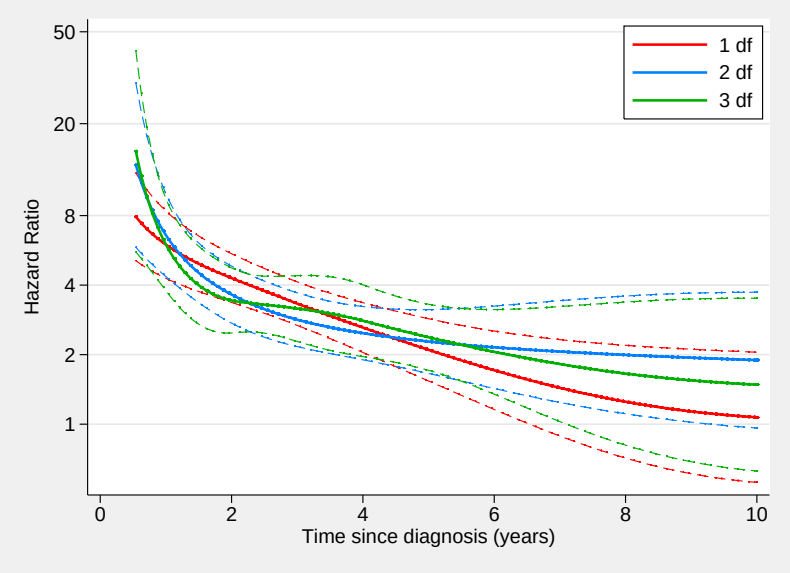

Figure 30: Localised skin melanoma. Sensitivity to using different df for time dependent effects.

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

(a) Load the melanoma data. Plot the complement of the Kaplan-Meier survival estimate.

```
. * Cancer
. stset surv_mm, failure(status==1) scale(12)

     failure event:  status == 1
obs. time interval:  (0, surv_mm]
 exit on or before:  failure
    t for analysis:  time/12

--------------------------------------------------------------------------------
     7775  total obs.
        0  exclusions
--------------------------------------------------------------------------------
     7775  obs. remaining, representing
     1913  failures in single record/single failure data
 51269.71  total analysis time at risk, at risk from t =         0
                              earliest observed entry t =         0
                                   last observed exit t =  20.95833

. sts generate surv1 = s

. * Other Causes
. stset surv_mm, failure(status==2) scale(12)

     failure event:  status == 2
obs. time interval:  (0, surv_mm]
 exit on or before:  failure
    t for analysis:  time/12

--------------------------------------------------------------------------------
     7775  total obs.
        0  exclusions
--------------------------------------------------------------------------------
     7775  obs. remaining, representing
     1134  failures in single record/single failure data
 51269.71  total analysis time at risk, at risk from t =         0
                              earliest observed entry t =         0
                                   last observed exit t =  20.95833

. sts generate surv2 = s

. * Plot the complement of the Kaplan-Meier survival estimate for each cause

. gen prob1=1-surv1

. gen prob2=1-surv2

. twoway (line prob1 _t, sort lcolor(black) lpattern(solid)) ///
> (line prob2 _t, sort lcolor(black) lpattern(dash)), ///
> ytitle("Probability of Death") ///
> xtitle("Time Since Diagnosis (Years)") ///
> legend(order(1 "Cancer" 2 "Other Causes")) ///
> ylabel(0(0.1)0.5, angle(0) format(%3.1f)) name(KM, replace)
```

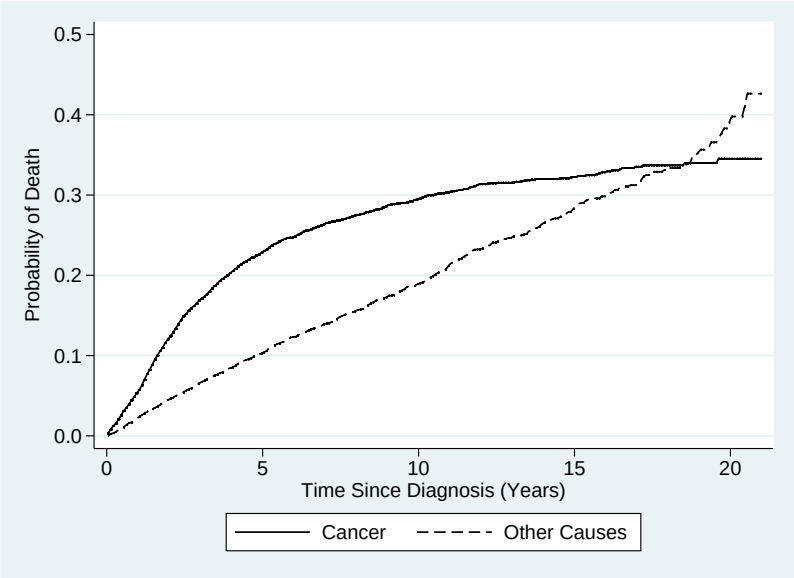

Figure 31: Complement of the Kaplan-Meier survival estimate for both cancer and other causes.

(b) Use the `stcompet` command to estimate the cumulative incidence function for both cancer and other causes. Plot the two functions along with the complement of the Kaplan-Meier function.

```
. stset surv_mm, failure(status==1) scale(12)

     failure event:  status == 1
obs. time interval:  (0, surv_mm]
 exit on or before:  failure
    t for analysis:  time/12

--------------------------------------------------------------------------------
     7775  total obs.
        0  exclusions
--------------------------------------------------------------------------------
     7775  obs. remaining, representing
     1913  failures in single record/single failure data
 51269.71  total analysis time at risk, at risk from t =         0
                              earliest observed entry t =         0
                                   last observed exit t =  20.95833

. stcompet CIF=ci, compet1(2)

. gen CIFcancer=CIF if status==1
(5862 missing values generated)

. gen CIFother=CIF if status==2
(6641 missing values generated)

* Plot the cumulative incidence function along with the complement of the K-M estimates

. twoway (line prob1 _t, sort lcolor(black) lpattern(solid)) ///
       (line prob2 _t, sort lcolor(black) lpattern(dash)) ///
       (line CIFcancer _t, sort lcolor(gs10) lpattern(solid)) ///
   (line CIFother _t, sort lcolor(gs10) lpattern(dash)), ///
       ytitle("Probability of Death") ///
   xtitle("Time Since Diagnosis (Years)") ///
   legend(order(1 "Cancer K-M" 2 "Other K-M" 3 "Cancer CIF" 4 "Other CIF")) ///
   ylabel(0(0.1)0.5, angle(0) format(%3.1f)) name(KMandCIF, replace)
```

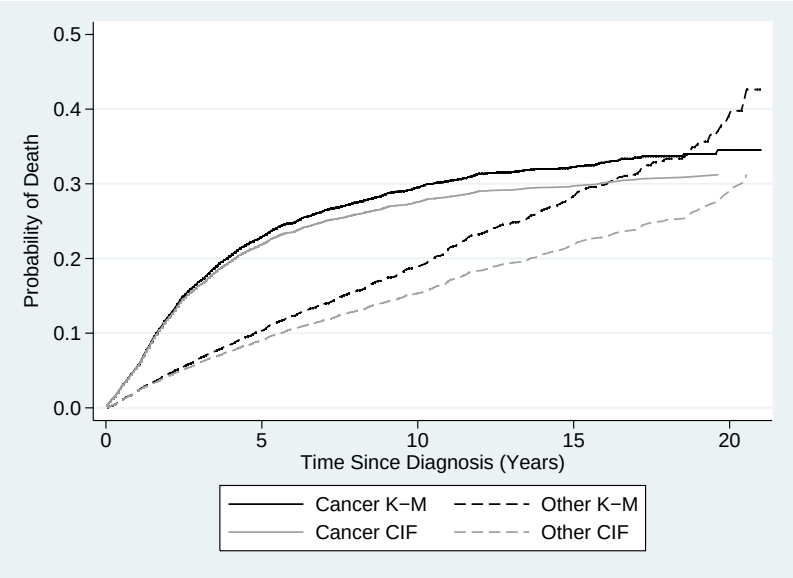

Figure 32: Comparison of Kaplan-Meier complement and cumulative incidence function for cancer and other causes.

The complement of the Kaplan-Meier estimate is giving higher values for the probability of death from cancer and other causes than the cumulative incidence function. The cumulative incidence function is taking into account the fact that patients may die from other causes and so the probabilities for both cancer and other causes are not as high.

(c) Use the `stcompadj` command to estimate the cumulative incidence function for cancer and other causes taking into account the effect of sex. We shall assume that the effect of sex is the same for both cancer and other causes.

```
. gen female=sex==2

. stset surv_mm, failure(status==1) scale(12)

     failure event:  status == 1
obs. time interval:  (0, surv_mm]
 exit on or before:  failure
    t for analysis:  time/12

--------------------------------------------------------------------------------
     7775  total obs.
        0  exclusions
--------------------------------------------------------------------------------
     7775  obs. remaining, representing
     1913  failures in single record/single failure data
 51269.71  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  20.95833

. stcompadj female=0, compet(2) gen(CIFcancermale CIFothermale)

. stcompadj female=1, compet(2) gen(CIFcancerfemale CIFotherfemale)

. * Plot the cumulative incidence functions against time _t
. twoway (line CIFcancermale _t, sort lcolor(navy) lpattern(solid)) ///
> (line CIFothermale _t, sort lcolor(navy) lpattern(dash)), ///
> ytitle("Probability of Death") title("Males") ///
> xtitle("Time Since Diagnosis (Years)") ///
```

```
> legend(order(1 "Cancer" 2 "Other")) ///
> ylabel(0(0.1)0.5, angle(0) format(%3.1f)) name(males, replace)

. twoway (line CIFcancerfemale _t, sort lcolor(maroon) lpattern(solid)) ///
> (line CIFotherfemale _t, sort lcolor(maroon) lpattern(dash)), ///
> ytitle("Probability of Death") title("Females") ///
> xtitle("Time Since Diagnosis (Years)") ///
> legend(order(1 "Cancer" 2 "Other")) ///
> ylabel(0(0.1)0.5, angle(0) format(%3.1f)) name(females, replace)

. graph combine males females, ycommon xcommon name(Cox, replace)
```

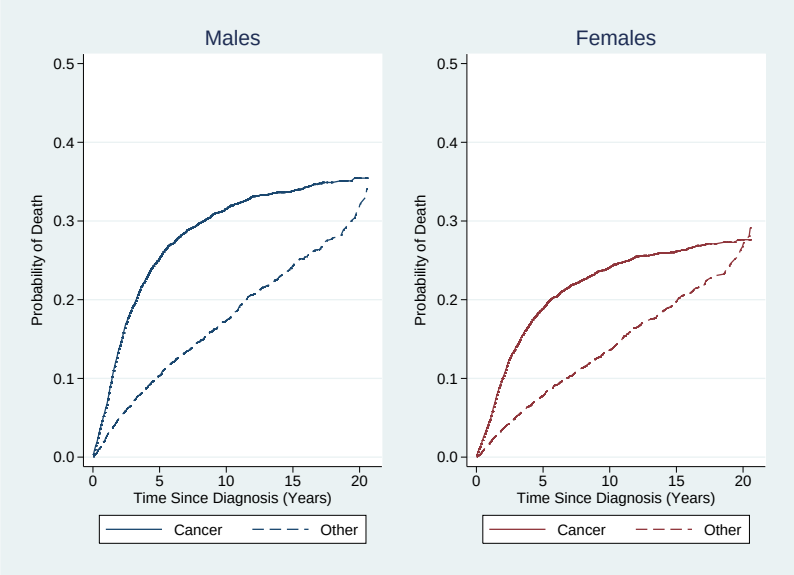

Figure 33: Cumulative incidence functions for cancer and other causes by sex. Obtained through competing risks analogue of the Cox model.

The probability of death from cancer and other causes is higher for males than for females.

(d) Consider the same model but using the flexible parametric approach.

i. Expand the data set.

```
. expand 2
(7775 observations created)

. * Recode and set up data for competing risk analysis
. * cause =1 for cause 1, cause =2 for cause 2
. bysort id: gen cause=_n

. * cancer is a dummy for cause 1
. gen cancer=(cause==1)

. * other is a dummy for cause 2
. gen other=(cause==2)

. * Event is the event indicator, coded like this:
. * For the first row (death due to cancer):
. * event is 1 if person died from cancer
. * For the second row (death due to other):
. * event is 1 if person died from other

. * status=1 death due to cancer, =2 death due to other
. gen event=(cause==status)

. * Look at the created data
. list id sex status cause sex event in 21/26, nolabel

     +---------------------------------------+
     | id   sex   status   cause   sex   event |
     |---------------------------------------|
 21. | 11     2        1       1     2       1 |
 22. | 11     2        1       2     2       0 |
 23. | 12     1        2       1     1       0 |
 24. | 12     1        2       2     1       1 |
 25. | 13     2        2       1     2       0 |
     |---------------------------------------|
 26. | 13     2        2       2     2       1 |
     +---------------------------------------+
```

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.

```
. stset surv_mm, failure(event) scale(12)

     failure event:  event != 0 & event < .
obs. time interval:  (0, surv_mm]
 exit on or before:  failure
    t for analysis:  time/12

--------------------------------------------------------------------------------
    15550  total obs.
        0  exclusions
--------------------------------------------------------------------------------
    15550  obs. remaining, representing
     3047  failures in single record/single failure data
 102539.4  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  20.95833

. * Fit the stpm2 model assuming the effect of sex is the same
. * for both cancer and other causes
. stpm2 cancer other female, scale(hazard) ///
>       rcsbaseoff dftvc(3) nocons tvc(cancer other) eform nolog

Log likelihood =  -10585.48                     Number of obs   =      15550

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      cancer |   .2350793   .0069002   -49.33   0.000     .2219369        .249
       other |   .1139254   .0044081   -56.14   0.000     .1056051    .1229013
      female |   .7050712   .0256194    -9.62   0.000     .6566045    .7571155
_rcs_cancer1 |   2.274767   .0402761    46.42   0.000     2.197181    2.355092
_rcs_cancer2 |   1.249747   .0204284    13.64   0.000     1.210343    1.290434
_rcs_cancer3 |   1.066001   .0095287     7.15   0.000     1.047488    1.084842
 _rcs_other1 |   3.033144   .0934966    36.00   0.000      2.85532    3.222043
 _rcs_other2 |   1.085711   .0317998     2.81   0.005      1.02514    1.149861
 _rcs_other3 |   .9533664   .0155478    -2.93   0.003     .9233751    .9843318
------------------------------------------------------------------------------
```

The hazard ratio for sex is 0.70. This means that the hazard of death from both cancer and other causes is 1.42 (1/0.7) times higher for males compared to females.

iii. Use the `stpm2cif` postestimation command to obtain the cumulative incidence functions for cancer and other causes for each sex.

```
. * predict the CIF for males (generates CIF_cancermale, CIF_othermale, _newt)
. stpm2cif cancermale othermale, cause1(cancer 1) ///
>    cause2(other 1)

. * predict the CIF for males (generates CIF_cancerfemale, CIF_otherfemale, _newt)
. stpm2cif cancerfemale otherfemale, cause1(cancer 1 female 1) ///
>    cause2(other 1 female 1)

. * Plot CIF's for cancer and other causes for males and females
. * It is important that we use the _newt variable (which is outputted from stpm2cif),
. * and not _t variable from stset.

. twoway (line CIF_cancermale _newt, sort lcolor(navy) lpattern(solid)) ///
> (line CIF_othermale _newt, sort lcolor(navy) lpattern(dash)), ///
> ytitle("Probability of Death") title("Males") ///
> xtitle("Time Since Diagnosis (Years)") ///
> legend(order(1 "Cancer" 2 "Other")) ///
> ylabel(0(0.1)0.5, angle(0) format(%3.1f)) name(malesfpm, replace)

. twoway (line CIF_cancerfemale _newt, sort lcolor(maroon) lpattern(solid)) ///
> (line CIF_otherfemale _newt, sort lcolor(maroon) lpattern(dash)), ///
> ytitle("Probability of Death") title("Females") ///
> xtitle("Time Since Diagnosis (Years)") ///
> legend(order(1 "Cancer" 2 "Other")) ///
> ylabel(0(0.1)0.5, angle(0) format(%3.1f)) name(femalesfpm, replace
> )

. graph combine malesfpm femalesfpm, ycommon xcommon name(FPM, replace)
```

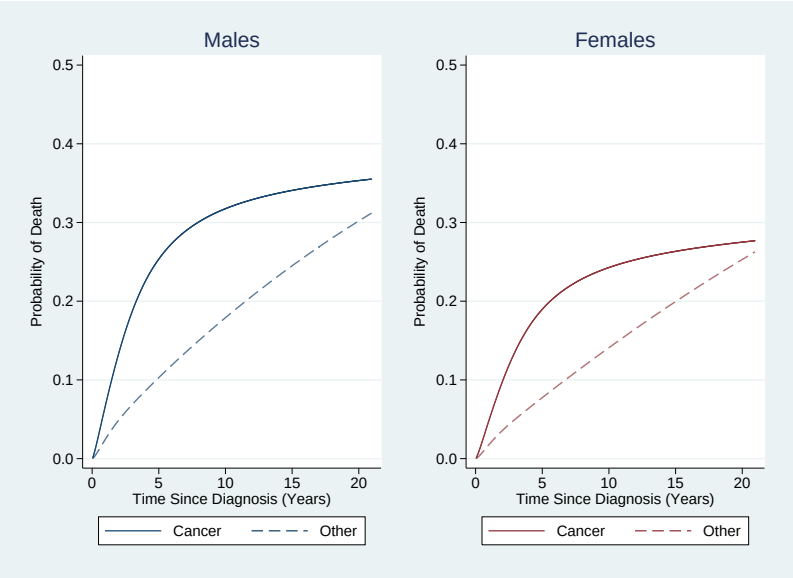

Figure 34: Cumulative incidence functions for cancer and other causes by sex. Obtained through postestimation after fitting flexible parametric model.

iv. Stack the cumulative incidence functions for cancer and other causes.

```
. gen male_total1=CIF_cancermale
(14550 missing values generated)

. gen male_total2=male_total1 + CIF_othermale
(14550 missing values generated)

. gen female_total1=CIF_cancerfemale
(14550 missing values generated)

. gen female_total2=female_total1 + CIF_otherfemale
(14550 missing values generated)

. twoway (area male_total2 _newt, sort fintensity(100)) ///
> (area male_total1 _newt, sort fintensity(100)), ///
> ylabel(0(0.2)1, angle(0) format(%3.1f)) ///
> ytitle("Probability of Death") xtitle("Time Since Diagnosis (Years)") ///
> legend(order(2 "Cancer" 1 "Other") size(small)) ///
> title("Males") plotregion(margin(zero)) name(malestack, replace)

. twoway (area female_total2 _newt, sort fintensity(100)) ///
> (area female_total1 _newt, sort fintensity(100)), ///
> ylabel(0(0.2)1, angle(0) format(%3.1f)) ///
> ytitle("Probability of Death") xtitle("Time Since Diagnosis (Years)") ///
> legend(order(2 "Cancer" 1 "Other") size(small)) ///
> title("Female") plotregion(margin(zero)) name(femalestack, replace)

. graph combine malestack femalestack, ycommon xcommon name(FPM, replace)
```

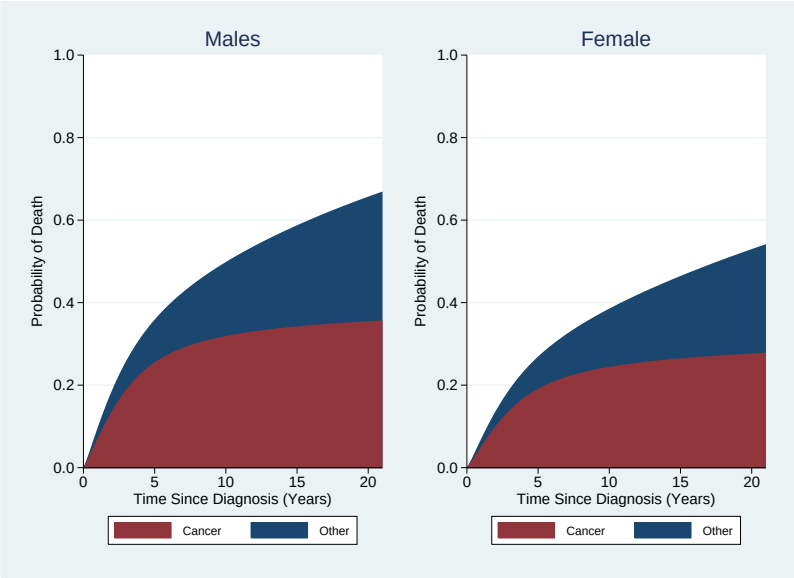

Figure 35: Stacked cumulative incidence functions for cancer and other causes by sex. Obtained through postestimation after fitting flexible parametric model.

(e) Adjust the model for age assuming that the effect of age is different for cancer and other causes of death.

i. Categorise age into 4 groups. To allow the effect of age to vary for the two causes create interaction terms between age group and the causes of death.

```
* To allow for different effects of age for the different causes
*   we must create dummies for age and each cause
gen age0ca=(agegrp==0 & cancer==1)
gen age1ca=(agegrp==1 & cancer==1)
gen age2ca=(agegrp==2 & cancer==1)
gen age3ca=(agegrp==3 & cancer==1)

gen age0oth=(agegrp==0 & other==1)
gen age1oth=(agegrp==1 & other==1)
gen age2oth=(agegrp==2 & other==1)
gen age3oth=(agegrp==3 & other==1)
```

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

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

Log likelihood = -9647.6016                     Number of obs   =      15550

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      cancer |   .1610153   .0087898   -33.45   0.000     .1446772    .1791985
       other |   .0097935   .0016017   -28.29   0.000     .0071076    .0134944
      female |   .6368877   .0233689   -12.30   0.000     .5926938     .684377
      age1ca |   1.309543   .0880211     4.01   0.000     1.147906    1.493941
      age2ca |   1.820119   .1176145     9.27   0.000       1.6036    2.065874
      age3ca |   2.787821   .2058777    13.88   0.000     2.412151    3.221997
     age1oth |   3.996579   .7200584     7.69   0.000     2.807557    5.689162
     age2oth |   16.37298   2.735226    16.73   0.000     11.80125    22.71575
     age3oth |   63.72519   10.73031    24.67   0.000     45.81233    88.64209
 _rcs_cancer1 |   2.333724   .0421174    46.96   0.000     2.252618     2.41775
 _rcs_cancer2 |   1.239049   .0203928    13.02   0.000     1.199718     1.27967
 _rcs_cancer3 |   1.067022   .0099092     6.99   0.000     1.047776    1.086622
  _rcs_other1 |   3.648889   .1178379    40.08   0.000     3.425089    3.887313
  _rcs_other2 |   .9942078   .0302275    -0.19   0.848     .9366937    1.055253
  _rcs_other3 |   .9110387   .0166227    -5.11   0.000     .8790344    .9442082
------------------------------------------------------------------------------
```

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

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

Log likelihood = -9643.2669                     Number of obs   =      15550

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      cancer |   .1616374   .0088594   -33.25   0.000     .1451735    .1799685
       other |   .0098492    .001612   -28.23   0.000     .0071463    .0135743
      female |   .6288809   .0245906   -11.86   0.000     .5824848    .6789727
      age1ca |   1.308775   .0879738     4.00   0.000     1.147225    1.493074
      age2ca |   1.817989   .1174813     9.25   0.000     1.601715    2.063466
      age3ca |   2.793172   .2062493    13.91   0.000     2.416821    3.228129
     age1oth |   3.997847   .7202928     7.69   0.000      2.80844    5.690983
     age2oth |   16.35058   2.731512    16.73   0.000     11.78507    22.68476
     age3oth |   63.64521   10.71583    24.67   0.000     45.75624    88.52808
_rcs_cancer1 |   2.314656   .0520792    37.30   0.000     2.214801    2.419014
_rcs_cancer2 |   1.271691   .0267621    11.42   0.000     1.220306    1.325241
_rcs_cancer3 |   1.069834   .0126225     5.72   0.000     1.045379    1.094862
 _rcs_other1 |   3.619297   .1302917    35.73   0.000     3.372731    3.883888
 _rcs_other2 |   1.023829   .0347366     0.69   0.488     .9579611    1.094226
 _rcs_other3 |   .9128674   .0184401    -4.51   0.000     .8774317    .9497343
_rcs_female1 |   1.015445   .0319292     0.49   0.626     .9547538    1.079993
_rcs_female2 |   .9443859   .0272901    -1.98   0.048     .8923849    .9994172
_rcs_female3 |   .9973606   .0165087    -0.16   0.873     .9655233    1.030248
------------------------------------------------------------------------------
```

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

```
. * predict the CIF for males aged 0-44
. * (generates cancersex1age0, othersex1age0, _newt)
. stpm2cif cancermaleage0 othermaleage0, cause1(cancer 1) cause2(other 1)

. * predict the CIF for males aged 75+
. * (generates cancersex1age3, othersex1age3, _newt)
. stpm2cif cancermaleage3 othermaleage3,
> cause1(cancer 1 age3ca 1) cause2(other 1 age3oth 1)

. * predict the CIF for females aged 0-44
. * (generates cancersex2age0, othersex2age0, _newt)
. stpm2cif cancerfemaleage0 otherfemaleage0,
> cause1(cancer 1 female 1) cause2(other 1 female 1)

. * predict the CIF for females aged 75+
. * (generates cancersex2age3, othersex2age3, _newt)
. stpm2cif cancerfemaleage3 otherfemaleage3,
> cause1(cancer 1 age3ca 1 female 1) cause2(other 1 age3oth 1 female 1)

. * plot against _newt (which is outputted from stpm2cif)

. twoway (line CIF_cancermaleage0 _newt, sort lcolor(navy)) ///
> (line CIF_cancerfemaleage0 _newt, sort lcolor(maroon) lpattern(dash)), ///
> ytitle("Probability of Death") title("Ages 0-44") ///
> xtitle("Time Since Diagnosis (Years)") ///
> legend(order(1 "Males" 2 "Females")) ///
> ylabel(0(0.1)0.5, angle(0) format(%3.1f)) name(age0, replace)
```

```
. twoway (line CIF_cancermaleage3 _newt, sort lcolor(navy)) ///
> (line CIF_cancerfemaleage3 _newt, sort lcolor(maroon) lpattern(dash)), ///
> ytitle("Probability of Death") title("Ages 75+") ///
> xtitle("Time Since Diagnosis (Years)") ///
> legend(order(1 "Males" 2 "Females")) ///
> ylabel(0(0.1)0.5, angle(0) format(%3.1f)) name(age3, replace)

. graph combine age0 age3, ycommon xcommon title("Cancer") name(ages, replace)
```

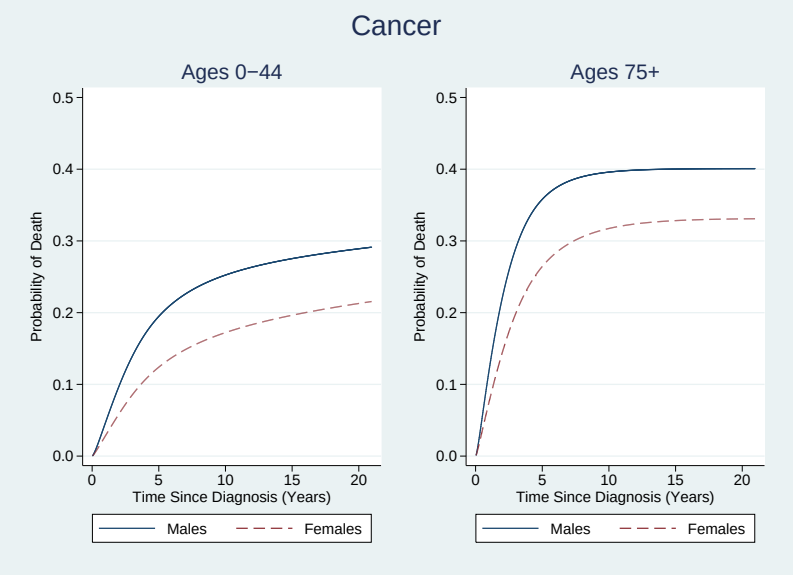

Figure 36: Cumulative incidence functions for cancer and other causes by sex and age group. Obtained through postestimation after fitting flexible parametric model.