Statistical methods for population-based cancer survival analysis

Solutions to exercises

Paul W. Dickman[1], Paul C. Lambert[1,2], Sandra Eloranta[1],
Therese Andersson[1], Mark J Rutherford[2], Anna Johansson[1],
Caroline E. Weibull[1], Sally Hinchliffe[2], Hannah Bower[1], Michael Crowther[2]

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

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

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

February 2018

Exercise solutions

100. **Life table and Kaplan-Meier estimates of survival**

The results are contained in the Excel file `\solutions\exercise100.xls` and in the Stata output for exercise 101.

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

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

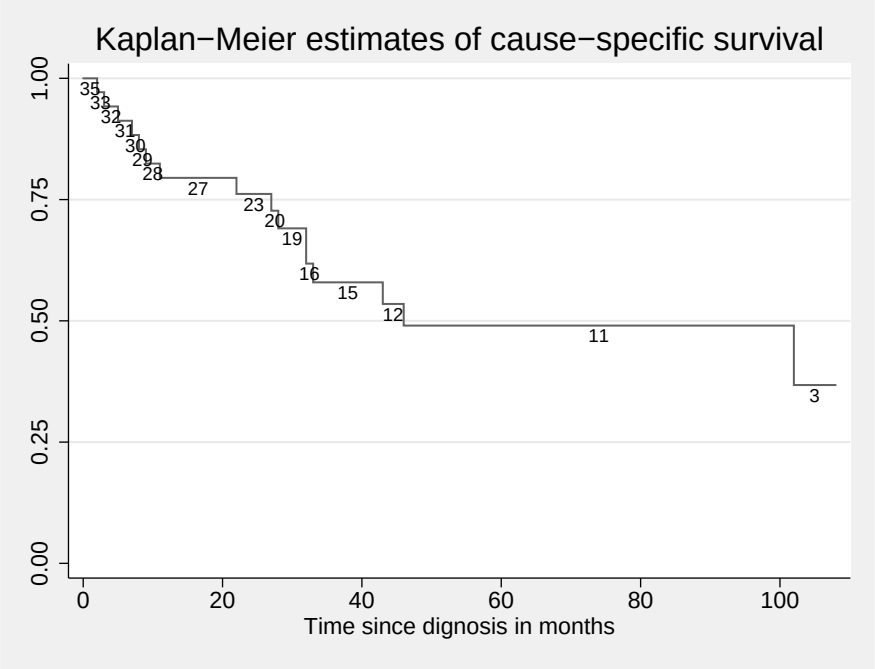


Figure 1: Kaplan-Meier plot of the cause-specific survivor function for sample of 35 patients diagnosed with colon carcinoma. The number at risk at each time point are shown on the curve.

102. **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.

103. **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, diagnosed 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)
```

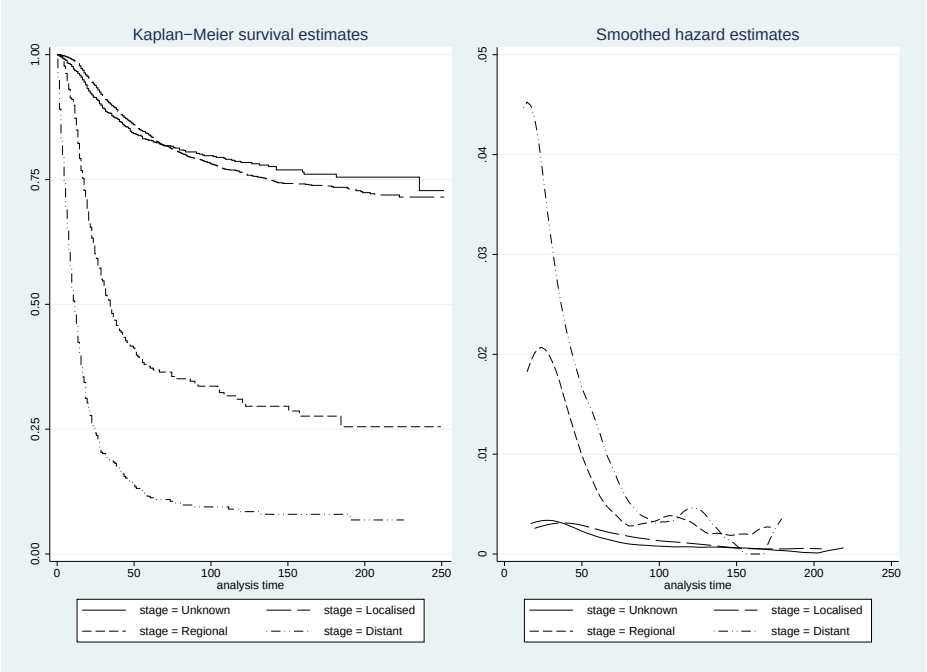


Figure 2: Skin melanoma. Kaplan-Meier estimates of cause-specific survival and mortality rate for each stage.

(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)
```

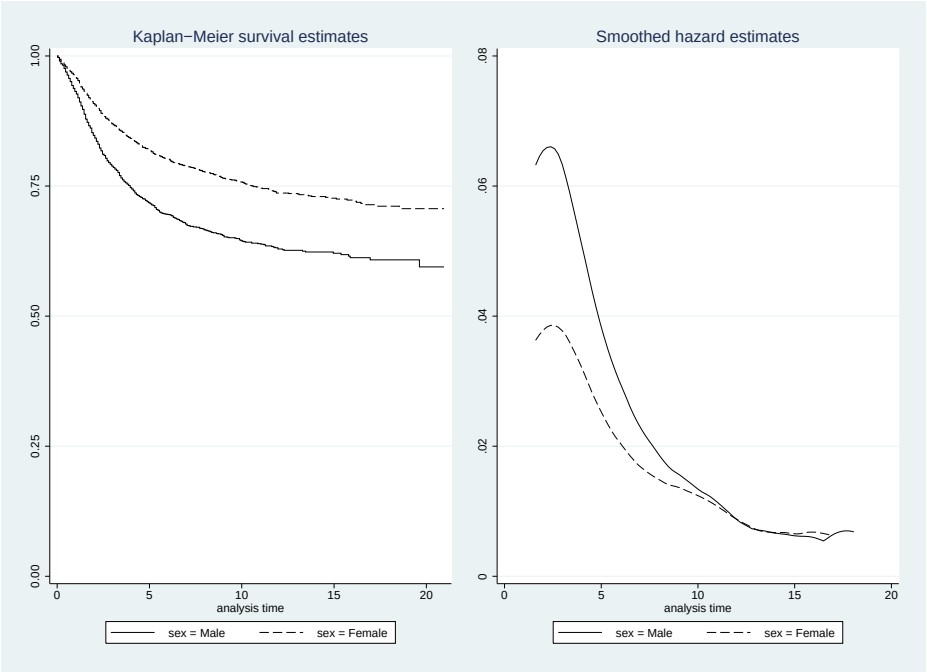


Figure 3: Skin melanoma (all stages). Kaplan-Meier estimates of cause-specific survival and mortality for each sex.

(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 exit
-------------------------------------------------------------------------------

                  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 |            Age in 4 categories
             exit |      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)
```

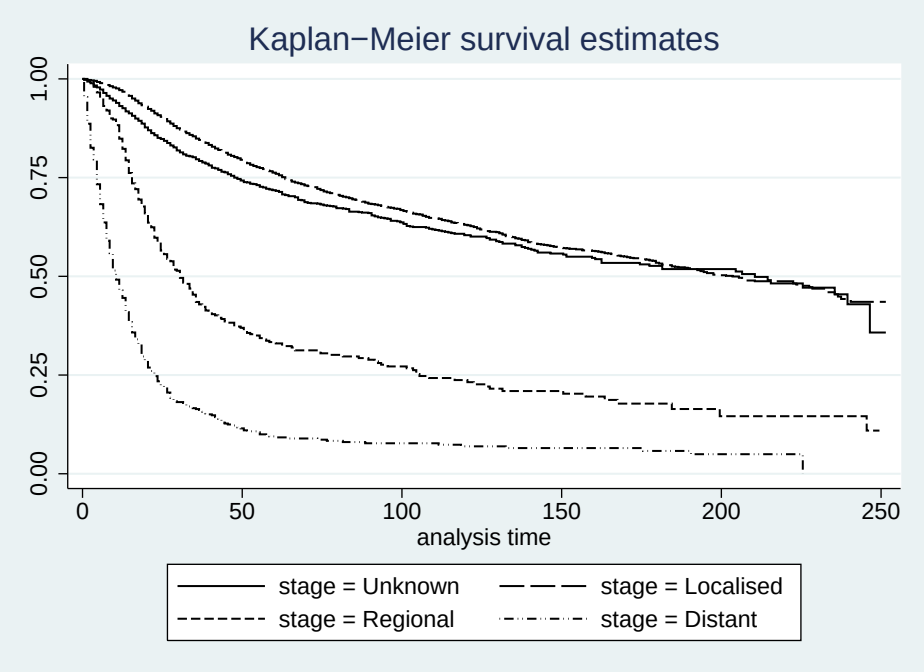


Figure 4: Skin melanoma (all stages). Kaplan-Meier estimates of all-cause survival for each stage.

(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)
```

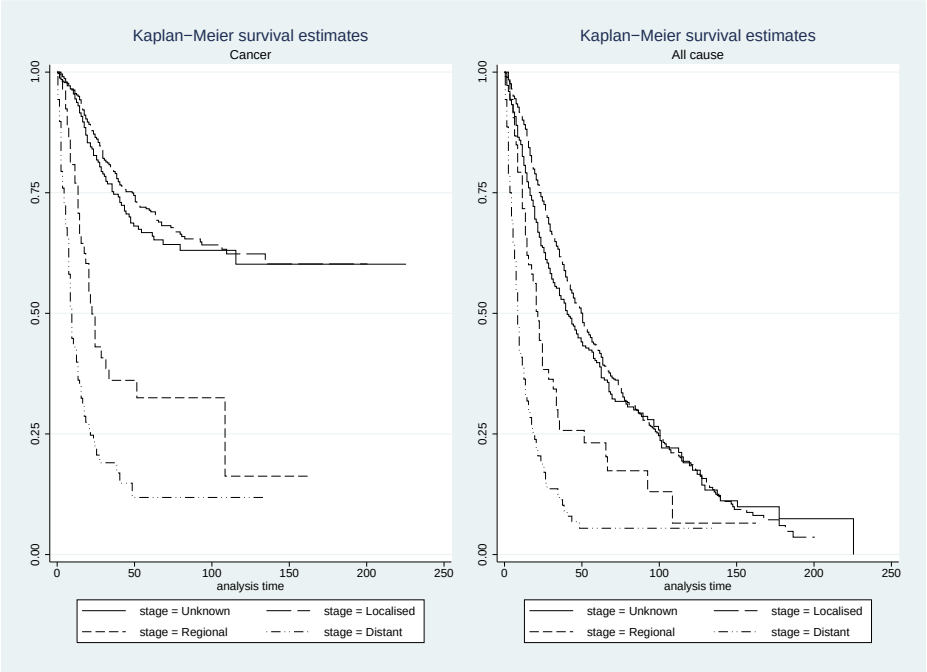


Figure 5: Skin melanoma (all stages). Kaplan-Meier estimates of all-cause survival versus cause-specific survival for each stage.

(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]
```

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

```
. use melanoma if stage==1, clear
(Skin melanoma, diagnosed 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)
```

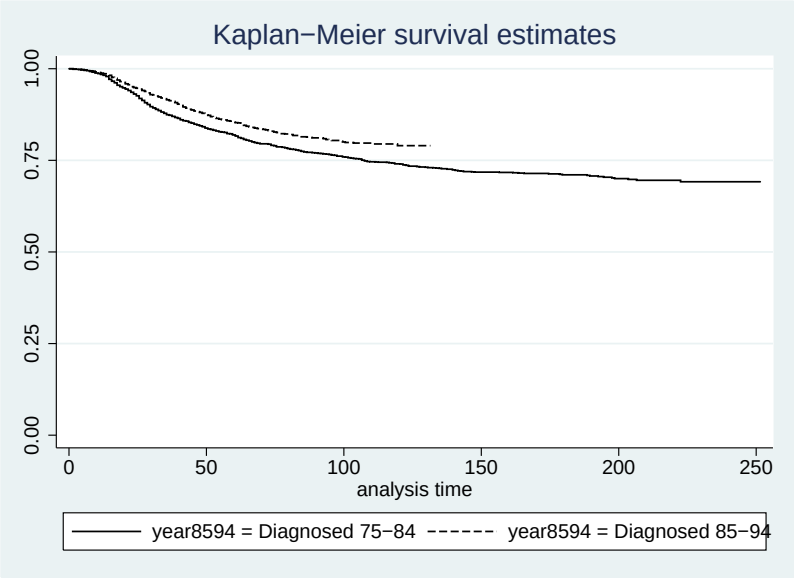


Figure 6: Skin melanoma. Kaplan-Meier plot of the cause-specific survivor function for each calendar period of diagnosis

(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)`

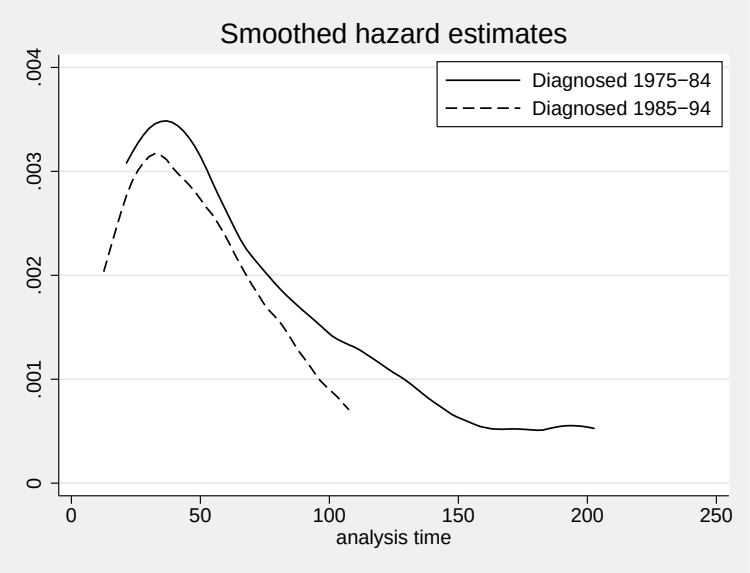


Figure 7: Skin melanoma. Plot of the cause-specific hazard for each calendar period of diagnosis

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)
```

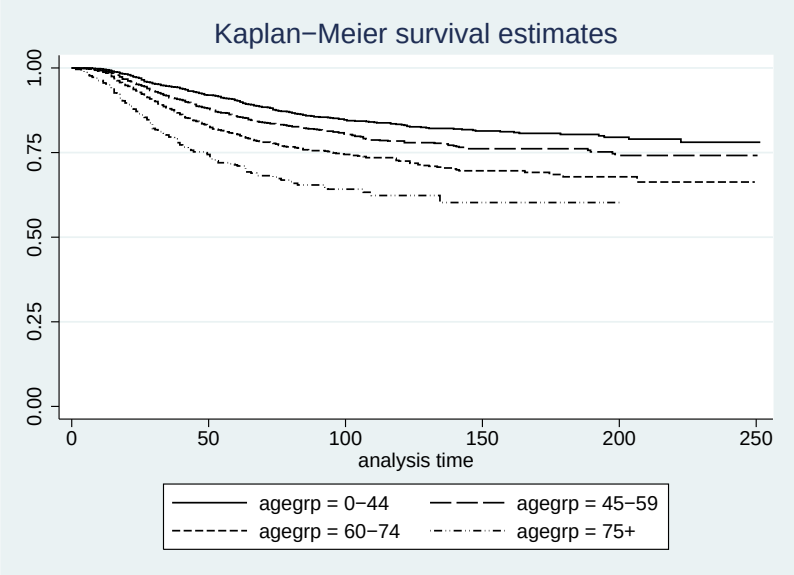


Figure 8: Skin melanoma. Plot of the cause-specific survival function for each age group

(e)
```
. 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

--------------------------------------------------------------------------------
     5318  total observations
        0  exclusions
--------------------------------------------------------------------------------
     5318  observations remaining, representing
     1013  failures in single-record/single-failure data
 38626.58  total analysis time at risk and under observation
                                             at risk from t =         0
                                  earliest observed entry t =         0
                                       last observed exit t =  20.95833

. sts graph, by(agegrp)
[output omitted]

. strate agegrp, per(1000)

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

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   13.0935   16.573   14.508   18.932 |
  |  45-59   282   12.4018   22.739   20.234   25.554 |
  |  60-74   333   10.1115   32.933   29.579   36.667 |
  |    75+   181    3.0198   59.937   51.812   69.337 |
  +-------------------------------------------------+
```

(f)
```
. sts graph, by(sex)
. sts graph, hazard by(sex) noshow
[output omitted]

. 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
(5318 records included in the analysis)

  +-------------------------------------------------+
  |    sex     D         Y     Rate    Lower    Upper |
  |-------------------------------------------------|
  |   Male   542   16.0974   33.670   30.952   36.627 |
  | Female   471   22.5292   20.906   19.101   22.882 |
  +-------------------------------------------------+
```

Males seem to have a higher mortality rate compared to females. This difference is also statistically significant according to the log-rank test below.

```
. sts test sex

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

Log-rank test for equality of survivor functions

       |   Events         Events
sex    |  observed       expected
-------+-------------------------
Male   |       542         432.55
Female |       471         580.45
-------+-------------------------
Total  |      1013        1013.00

             chi2(1) =      48.55
             Pr>chi2 =     0.0000
```

110. **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
       _cons |    .013596   .0025694   -22.74   0.000     .0093875    .0196912
       ln(y) |          1  (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
```

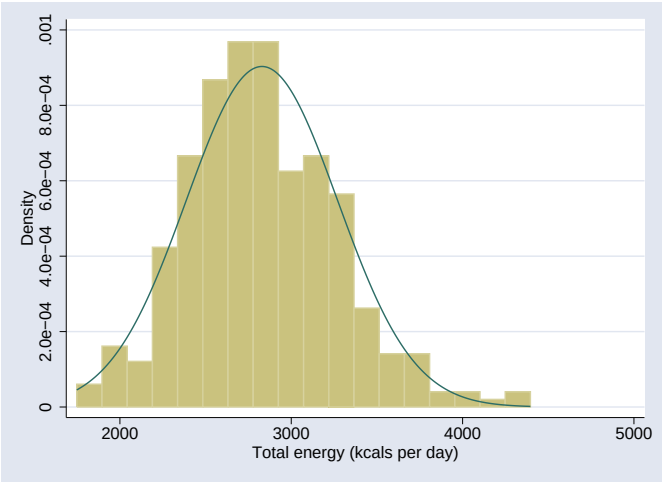


Figure 9: Histogram of energy with superimposed normal density curve (with the sample mean and variance).

```
. 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
       _cons |    .016902   .0042255   -16.32   0.000     .0103547    .0275892
      ln(y) |          1  (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
       _cons |   .0109059   .0023251   -21.19   0.000      .007181    .0165629
       ln(y) |          1  (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
             |
       _cons |    .016902   .0042255   -16.32   0.000     .0103547    .0275892
       ln(y) |          1  (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.

111. **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.

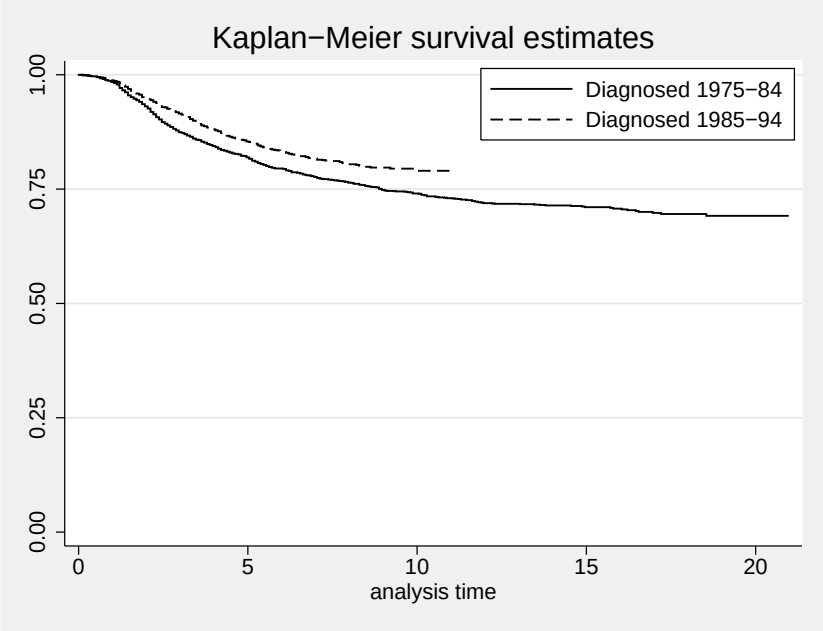


Figure 10: Localised melanoma. Kaplan-Meier estimates of cause-specific survival.

ii. Mortality is lower during the latter period.

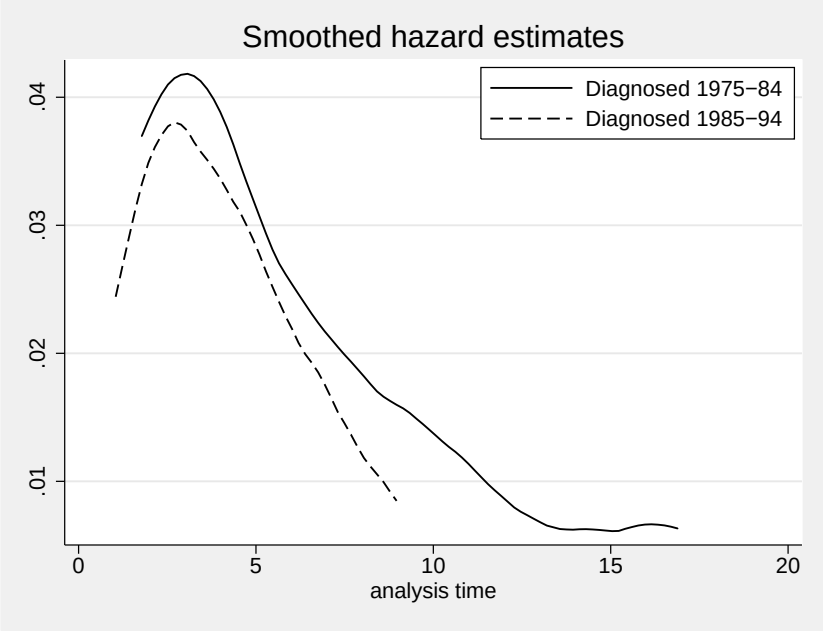


Figure 11: Localised melanoma. Smoothed cause-specific hazards (cause-specific mortality rates).

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.
```
. stset surv_mm, failure(status==1) scale(12) id(id) exit(time 120)

                id:  id
     failure event:  status == 1
obs. time interval:  (surv_mm[_n-1], surv_mm]
 exit on or before:  time 120
    t for analysis:  time/12

--------------------------------------------------------------------------------
     5318  total observations
        0  exclusions
--------------------------------------------------------------------------------
     5318  observations remaining, representing
     5318  subjects
      960  failures in single-failure-per-subject data
 32376.67  total analysis time at risk and under observation
                                             at risk from t =         0
                                  earliest observed entry t =         0
                                       last observed exit t =        10
. 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
       _cons |   .0314526   .0013806   -78.81   0.000     .0288597    .0342783
------------------------------------------------------------------------------
```

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.

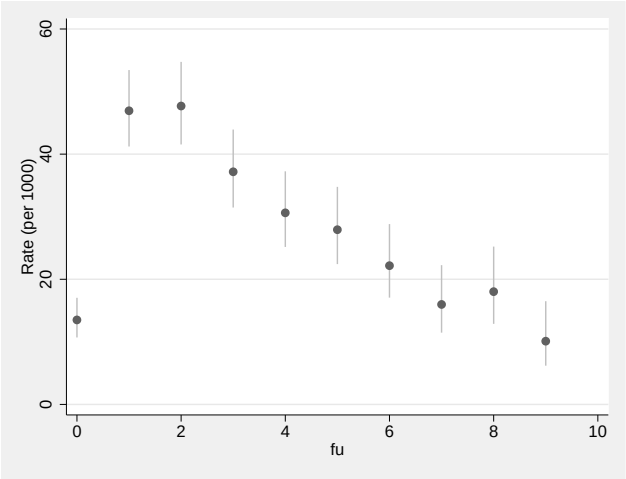


Figure 12: Localised melanoma. Disease-specific mortality rates as a function of time since diagnosis (annual intervals).

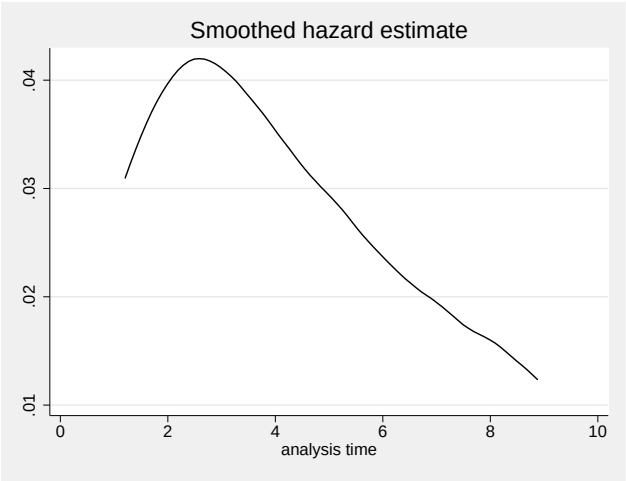


Figure 13: Localised melanoma. Disease-specific mortality rates as continuous function of time since diagnosis (using a smoother).

(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
             |
       _cons |   .0135058   .0016028   -36.27   0.000     .0107029    .0170427
------------------------------------------------------------------------------
```

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
       _cons |   .0155123   .0019207   -33.65   0.000     .0121698    .0197728
------------------------------------------------------------------------------
```

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 |
      45-59  |   1.327795    .125042     3.01   0.003     1.104005    1.596948
      60-74  |   1.862376    .169244     6.84   0.000     1.558527    2.225464
        75+  |   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
       _cons |   .0216012   .0036626   -22.62   0.000     .0154936    .0301163
------------------------------------------------------------------------------
```

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 |
       45-59 |   1.326709   .1249663     3.00   0.003     1.103059    1.595705
       60-74 |   1.861131   .1691561     6.83   0.000     1.557443    2.224035
         75+ |   3.399539   .3550374    11.72   0.000     2.770277    4.171737
             |
    year8594 |
Diagnosed 85-94 |   .7414351   .0655414    -3.38   0.001     .6234888    .8816936
             |
         sex |
      Female |   .6031338   .0531555    -5.74   0.000     .5074526     .716856
             |
 year8594#sex |
Diagnosed 85-94#Female |   .9437245   .1232639    -0.44   0.657     .7305772    1.219058
             |
       _cons |   .0125379     .00183   -30.00   0.000     .0094185    .0166904
------------------------------------------------------------------------------
```

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)
```

```
. 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]
-------------+----------------------------------------------------------------
          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 |
      45-59  |   1.326709   .1249663     3.00   0.003     1.103059    1.595705
      60-74  |   1.861131   .1691561     6.83   0.000     1.557443    2.224035
        75+  |   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
       _cons |   .0125379     .00183   -30.00   0.000     .0094185    .0166904
------------------------------------------------------------------------------
```

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 |
       45-59 |   1.326709   .1249663    3.00   0.003     1.103059    1.595705
       60-74 |   1.861131   .1691561    6.83   0.000     1.557443    2.224035
         75+ |   3.399539   .3550374   11.72   0.000     2.770277    4.171737
             |
   year8594 |
 Diagnosed 85-94 |   .7414351   .0655414   -3.38   0.001     .6234888    .8816936
             |
 year8594#sex |
Diagnosed 75-84#Female |   .6031338   .0531555   -5.74   0.000     .5074526     .716856
Diagnosed 85-94#Female |   .5691922    .055267   -5.80   0.000     .4705541    .6885069
             |
       _cons |   .0125379     .00183  -30.00   0.000     .0094185    .0166904
------------------------------------------------------------------------------
```

(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)
```

112. **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)
```

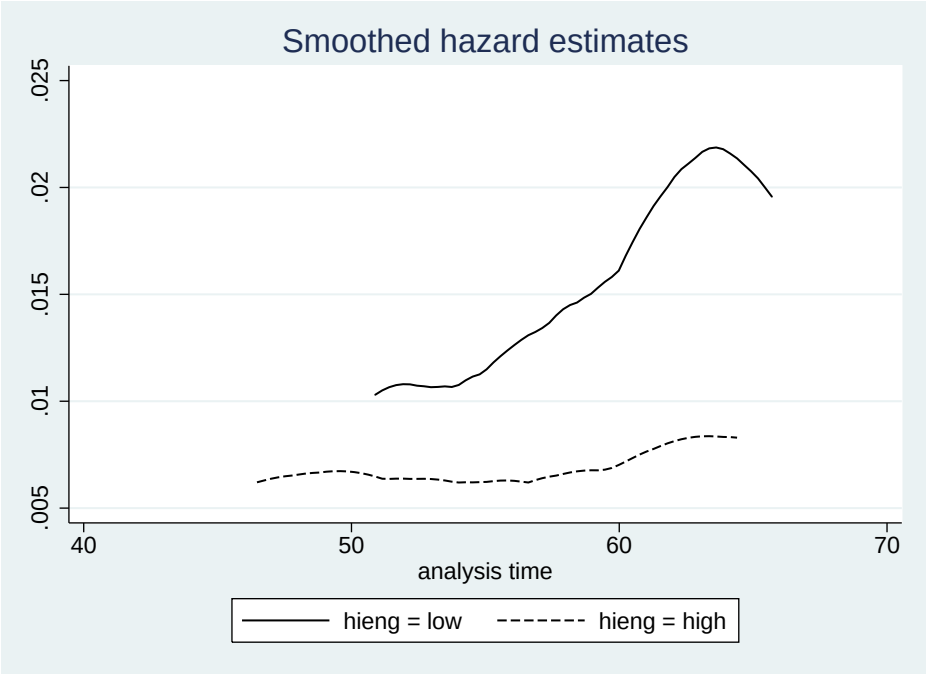


Figure 14: Diet data. Kaplan-Meier estimates of hazard rate for each energy intake level, with attained age as time scale.

```
* 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)
```

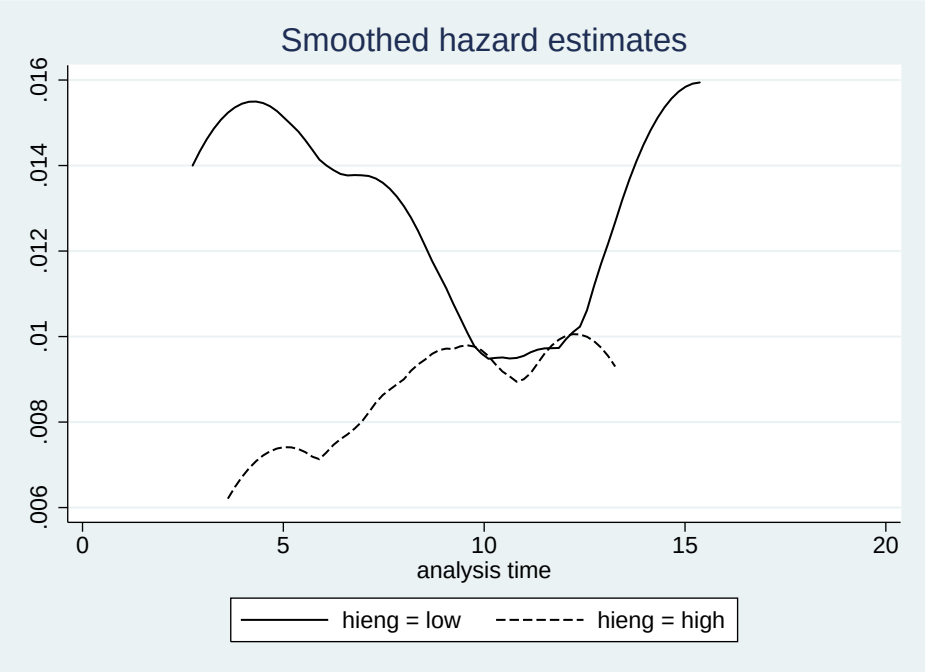


Figure 15: Diet data. Kaplan-Meier estimates of hazard rate for each energy intake level, with time since entry as time scale.

(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
       _cons |    .013596   .0025694   -22.74   0.000     .0093875    .0196912
       ln(y) |          1  (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
       _cons |   .0048706   .0059874    -4.33   0.000     .0004377    .0541948
      ln(y) |          1  (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
             |
       _cons |   .0083976   .0036279   -11.06   0.000      .003601    .0195835
ln(risktime) |          1  (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 |
  conductor  |   1.545112   .6284217     1.07   0.285     .6962464    3.428919
        bank |   .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
             |
       _cons |   .0011229   .0014748    -5.17   0.000     .0000856    .0147317
ln(risktime) |          1  (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
             |
       _cons |   .0141283   .0038053   -15.82   0.000     .0083335    .0239524
ln(risktime) |          1  (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 |
  conductor  |   1.584205   .6439641     1.13   0.258     .7141775    3.514121
       bank  |   .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
             |
       _cons |   .0024216   .0030584    -4.77   0.000     .0002038    .0287817
ln(risktime) |          1  (exposure)
------------------------------------------------------------------------------
```

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

(f) Using `streg` will give you the same results as using `poisson`. The advantage using `streg` is that this command understands and respects the internal st variables (`_st`, `_t`, `_t0`, and `_d`).

120. **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) `. est table Cox Poisson, eform equations(1)`

```
Hazard ratios and standard errors for Cox and Poisson models

--------------------------------------
    Variable |    Cox       Poisson
-------------+------------------------
         sex |  0.588814    0.587547
             |  0.038538    0.038456
    year8594 |  0.716884    0.722411
             |  0.047445    0.047813
             |
      agegrp |
      45-59  |  1.326397    1.327795
             |  0.124911    0.125042
      60-74  |  1.857323    1.862376
             |  0.168787    0.169244
        75+  |  3.372652    3.400287
             |  0.352227    0.355140
--------------------------------------
                           legend: b/se
```

The table shows hazard ratios and standard errors for Cox regression and Poisson regression with annual intervals. We see that the estimates are very similar.

(g) `. est table Cox Poisson_fine Poisson, eform equations(1)`

```
Hazard ratios and standard errors for various models

------------------------------------------------------------
    Variable |     Cox        Poisson_fine     Poisson
-------------+----------------------------------------------
         sex |     0.588814       0.588814       0.587547
             |     0.038538       0.038538       0.038456
    year8594 |     0.716884       0.716884       0.722411
             |     0.047445       0.047445       0.047813
             |
      agegrp |
      45-59  |     1.326397       1.326397       1.327795
             |     0.124911       0.124911       0.125042
      60-74  |     1.857323       1.857323       1.862376
             |     0.168787       0.168787       0.169244
        75+  |     3.372652       3.372652       3.400287
             |     0.352227       0.352227       0.355140
------------------------------------------------------------
                                                legend: b/se
```

The table shows hazard ratios and standard errors for Cox regression, Poisson regression after splitting at each failure time (`Poisson_fine`), and Poisson regression with annual intervals. Both the estimates and standard errors are identical for the first two.

(h) No written solutions for this part.

121. **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)
```

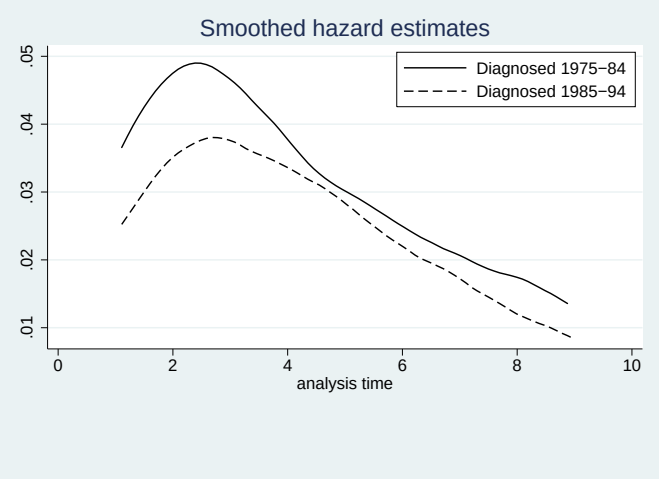


Figure 16: Localised skin melanoma. Plot of the estimated hazard function for each calendar period of diagnosis.

(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)
```

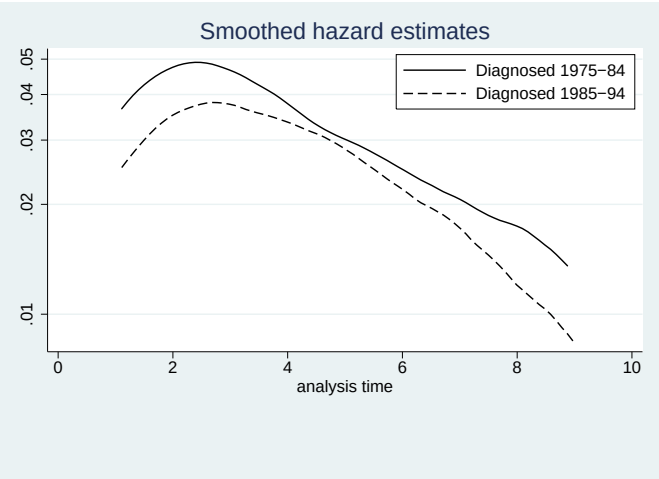


Figure 17: Localised skin melanoma. Plot of the estimated hazard function for each calendar period of diagnosis using a log scale for the y axis.

(c) If the proportional hazards assumption is appropriate then we should see parallel lines in Figure 17. 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)
```

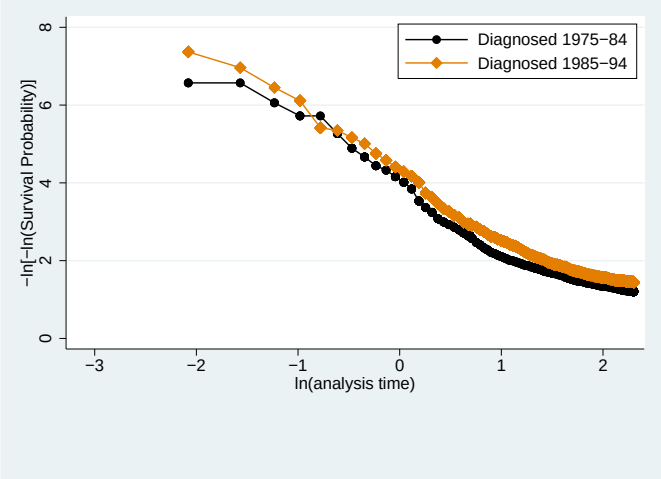


Figure 18: Localised skin melanoma. Plot of the log cumulative hazard function for each calendar period of diagnosis. Each plot symbol represents an event time. Note that the x axis is the natural logarithm of time in years, so a value of 0 corresponds to 1 year.

(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.

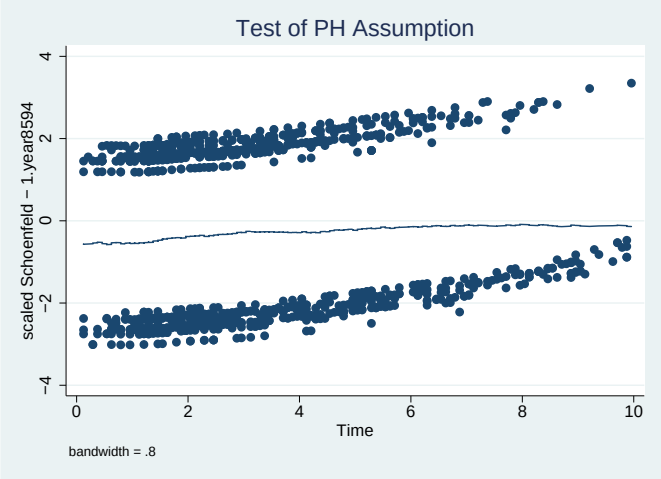


Figure 19: Localised skin melanoma. Plot of the scaled Schoenfeld residuals for calendar period 1985–94. The smooth line shows the estimated hazard ratio as a function of time.

(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) Splitting time since diagnosis into yearly intervals and estimating the effect of age separate for 0–2 years and 2+ years after diagnosis gives similar estimates to those obtained from the Cox model.

122. **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.

123. **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.

124. **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
------------------------------------------------------------------------------
```

125. **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)
```
. 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 |
  +-------------------------------------------------------+

. 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
------------------------------------------------------------------------------

. 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)
```
. 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)
```
. 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
------------------------------------------------------------------------------

. 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) `. 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
------------------------------------------------------------------------------
```

130. **Melanoma: Understanding splines**

```
. use melanoma
(Skin melanoma, diagnosed 1975-94, follow-up to 1995)

. gen female = sex == 2

. stset surv_mm, failure(status=1,2) scale(12) exit(time 120) id(id)

                id:  id
     failure event:  status == 1 2
obs. time interval:  (surv_mm[_n-1], surv_mm]
 exit on or before:  time 120
    t for analysis:  time/12

------------------------------------------------------------------------------
       7775  total observations
          0  exclusions
------------------------------------------------------------------------------
       7775  observations remaining, representing
       7775  subjects
       2773  failures in single-failure-per-subject data
  43306.833  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =        10
```

(a)
```
. stsplit fu, every(‘=1/12’)
(514,861 observations (episodes) created)

. gen risktime = _t - _t0

. collapse (sum) d = _d risktime (min) start=_t0 (max) end=_t, ///
>  by(fu female year8594 agegrp)

. // Fit a model with a parameter for each interval
. egen interval = group(start)
. gen midtime = (start + end)/2

. glm d ibn.interval, family(poisson) link(log) lnoffset(risktime) nocons

Generalized linear models                         No. of obs      =      1,920
Optimization     : ML                             Residual df     =      1,800
                                                  Scale parameter =          1
Deviance         =  3108.787038                   (1/df) Deviance =   1.727104
Pearson          =  4379.789968                   (1/df) Pearson  =   2.433217

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   3.324284
Log likelihood   = -3071.312939                   BIC             =  -10499.36
------------------------------------------------------------------------------
             |                 OIM
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    interval |
           1 |    -3.1046   .1856953   -16.72   0.000    -3.468556   -2.740643
           2 |  -2.534902    .140028   -18.10   0.000    -2.809352   -2.260452
           3 |  -2.699421   .1524986   -17.70   0.000    -2.998313    -2.40053
           4 |  -2.929231   .1714986   -17.08   0.000    -3.265362     -2.5931
```

5	-2.38904	.1313064	-18.19	0.000	-2.646395	-2.131684
6	-2.453025	.1360828	-18.03	0.000	-2.719743	-2.186308
7	-2.464522	.1373606	-17.94	0.000	-2.733744	-2.1953
8	-2.457342	.1373606	-17.89	0.000	-2.726564	-2.18812
9	-2.528921	.1428571	-17.70	0.000	-2.808916	-2.248926
10	-2.564062	.145865	-17.58	0.000	-2.849953	-2.278172
11	-2.744761	.1601282	-17.14	0.000	-3.058607	-2.430916
12	-2.29056	.1280369	-17.89	0.000	-2.541507	-2.039612
13	-2.500236	.1428571	-17.50	0.000	-2.780231	-2.220242
14	-2.301949	.1301889	-17.68	0.000	-2.557115	-2.046784
15	-2.160058	.1221694	-17.68	0.000	-2.399506	-1.92061
16	-2.160067	.1230915	-17.55	0.000	-2.401322	-1.918812
17	-2.384106	.138675	-17.19	0.000	-2.655904	-2.112308
18	-2.244205	.1301889	-17.24	0.000	-2.49937	-1.989039
19	-2.264819	.1324532	-17.10	0.000	-2.524423	-2.005216
20	-2.486988	.1490712	-16.68	0.000	-2.779162	-2.194814
21	-2.253717	.1336306	-16.87	0.000	-2.515628	-1.991806
22	-2.527711	.1543033	-16.38	0.000	-2.83014	-2.225282
23	-2.208612	.1324532	-16.67	0.000	-2.468215	-1.949008
24	-2.476555	.1524986	-16.24	0.000	-2.775446	-2.177663
25	-2.614548	.164399	-15.90	0.000	-2.936764	-2.292332
26	-2.550046	.1601282	-15.93	0.000	-2.863891	-2.236201
27	-2.350446	.145865	-16.11	0.000	-2.636336	-2.064556
28	-2.38006	.1490712	-15.97	0.000	-2.672235	-2.087886
29	-2.300847	.1443376	-15.94	0.000	-2.583744	-2.017951
30	-2.469775	.1581139	-15.62	0.000	-2.779673	-2.159878
31	-2.745043	.1825742	-15.04	0.000	-3.102881	-2.387204
32	-2.548794	.1666667	-15.29	0.000	-2.875455	-2.222133
33	-2.752635	.1856953	-14.82	0.000	-3.116591	-2.388679
34	-2.813133	.1924501	-14.62	0.000	-3.190328	-2.435938
35	-2.802705	.1924501	-14.56	0.000	-3.179901	-2.42551
36	-2.374244	.1561738	-15.20	0.000	-2.680339	-2.068149
37	-2.858575	.2	-14.29	0.000	-3.250568	-2.466582
38	-2.890082	.2041241	-14.16	0.000	-3.290158	-2.490006
39	-2.689391	.1856953	-14.48	0.000	-3.053347	-2.325434
40	-2.609536	.1796053	-14.53	0.000	-2.961556	-2.257516
41	-2.56525	.1767767	-14.51	0.000	-2.911726	-2.218774
42	-2.800731	.2	-14.00	0.000	-3.192723	-2.408738
43	-2.748872	.1961161	-14.02	0.000	-3.133253	-2.364492
44	-2.62625	.1856953	-14.14	0.000	-2.990206	-2.262294
45	-3.091989	.2357023	-13.12	0.000	-3.553957	-2.630021
46	-2.570596	.1825742	-14.08	0.000	-2.928435	-2.212757
47	-3.015384	.2294157	-13.14	0.000	-3.465031	-2.565738
48	-2.857754	.2132007	-13.40	0.000	-3.27562	-2.439888
49	-2.994306	.2294157	-13.05	0.000	-3.443952	-2.544659
50	-2.750205	.2041241	-13.47	0.000	-3.150281	-2.350129
51	-2.548682	.1856953	-13.73	0.000	-2.912638	-2.184725
52	-2.859817	.2182179	-13.11	0.000	-3.287516	-2.432118
53	-2.802901	.2132007	-13.15	0.000	-3.220767	-2.385035
54	-3.173995	.2581989	-12.29	0.000	-3.680055	-2.667934
55	-3.097767	.25	-12.39	0.000	-3.587758	-2.607776
56	-2.969108	.2357023	-12.60	0.000	-3.431076	-2.50714
57	-3.210027	.2672612	-12.01	0.000	-3.73385	-2.686205
58	-2.794058	.2182179	-12.80	0.000	-3.221757	-2.366359
59	-3.430805	.3015113	-11.38	0.000	-4.021757	-2.839854
60	-2.984889	.2425356	-12.31	0.000	-3.46025	-2.509528
61	-3.035178	.25	-12.14	0.000	-3.525169	-2.545187
62	-2.907331	.2357023	-12.33	0.000	-3.369299	-2.445363
63	-2.452518	.1889822	-12.98	0.000	-2.822916	-2.082119
64	-2.726789	.2182179	-12.50	0.000	-3.154488	-2.29909

```
          65 |  -3.050457   .2581989   -11.81   0.000    -3.556518   -2.544397
          66 |  -3.037887   .2581989   -11.77   0.000    -3.543947   -2.531826
          67 |  -3.095093   .2672612   -11.58   0.000    -3.618915    -2.57127
          68 |  -3.083438   .2672612   -11.54   0.000     -3.60726   -2.559615
          69 |  -3.409634   .3162278   -10.78   0.000    -4.029429   -2.789839
          70 |  -2.868901   .2425356   -11.83   0.000    -3.344262    -2.39354
          71 |  -3.611481   .3535534   -10.21   0.000    -4.304433   -2.918529
          72 |  -3.888555   .4082483    -9.52   0.000    -4.688707   -3.088403
          73 |  -4.062166   .4472136    -9.08   0.000    -4.938688   -3.185643
          74 |  -2.770561   .2357023   -11.75   0.000    -3.232529   -2.308593
          75 |  -2.940631   .2581989   -11.39   0.000    -3.446691    -2.43457
          76 |  -2.929563   .2581989   -11.35   0.000    -3.435623   -2.423502
          77 |  -3.323086   .3162278   -10.51   0.000    -3.942881   -2.703291
          78 |  -3.417423   .3333333   -10.25   0.000    -4.070744   -2.764102
          79 |  -3.300609   .3162278   -10.44   0.000    -3.920404   -2.680814
          80 |  -3.289179   .3162278   -10.40   0.000    -3.908974   -2.669384
          81 |  -3.384233   .3333333   -10.15   0.000    -4.037555   -2.730912
          82 |  -3.171403   .3015113   -10.52   0.000    -3.762354   -2.580452
          83 |  -3.764908   .4082483    -9.22   0.000     -4.56506   -2.964756
          84 |  -2.905795   .2672612   -10.87   0.000    -3.429617   -2.381972
          85 |  -3.231298   .3162278   -10.22   0.000    -3.851093   -2.611503
          86 |  -4.136665         .5    -8.27   0.000    -5.116647   -3.156683
          87 |  -3.208825   .3162278   -10.15   0.000    -3.828621    -2.58903
          88 |  -3.420285   .3535534    -9.67   0.000    -4.113237   -2.727333
          89 |  -3.290335   .3333333    -9.87   0.000    -3.943656   -2.637013
          90 |   -3.07525   .3015113   -10.20   0.000    -3.666202   -2.484299
          91 |   -3.37588   .3535534    -9.55   0.000    -4.068831   -2.682928
          92 |  -3.493075   .3779645    -9.24   0.000    -4.233871   -2.752278
          93 |  -3.347159   .3535534    -9.47   0.000    -4.040111   -2.654207
          94 |  -3.336288   .3535534    -9.44   0.000     -4.02924   -2.643337
          95 |  -3.458455   .3779645    -9.15   0.000    -4.199252   -2.717658
          96 |  -3.447339   .3779645    -9.12   0.000    -4.188135   -2.706542
          97 |  -3.437246   .3779645    -9.09   0.000    -4.178043   -2.696449
          98 |  -3.581588   .4082483    -8.77   0.000     -4.38174   -2.781436
          99 |     -4.266   .5773503    -7.39   0.000    -5.397586   -3.134414
         100 |  -2.955541   .3015113    -9.80   0.000    -3.546493    -2.36459
         101 |  -3.034552   .3162278    -9.60   0.000    -3.654347   -2.414757
         102 |  -2.923487   .3015113    -9.70   0.000    -3.514439   -2.332536
         103 |  -3.357809   .3779645    -8.88   0.000    -4.098606   -2.617012
         104 |  -3.086825   .3333333    -9.26   0.000    -3.740146   -2.433503
         105 |  -3.475669   .4082483    -8.51   0.000    -4.275821   -2.675517
         106 |  -4.154533   .5773503    -7.20   0.000    -5.286119   -3.022948
         107 |  -3.041873   .3333333    -9.13   0.000    -3.695195   -2.388552
         108 |  -3.145184   .3535534    -8.90   0.000    -3.838136   -2.452233
         109 |  -2.907356   .3162278    -9.19   0.000    -3.527151   -2.287561
         110 |  -4.096194   .5773502    -7.09   0.000     -5.22778   -2.964609
         111 |  -4.488385   .7071007    -6.35   0.000    -5.874277   -3.102493
         112 |  -3.558201   .4472136    -7.96   0.000    -4.434724   -2.681679
         113 |  -2.954862   .3333333    -8.86   0.000    -3.608183   -2.301541
         114 |  -3.750729         .5    -7.50   0.000    -4.730711   -2.770747
         115 |  -3.513037   .4472136    -7.86   0.000    -4.389559   -2.636514
         116 |  -2.910235   .3333333    -8.73   0.000    -3.563556   -2.256914
         117 |  -3.481496   .4472136    -7.78   0.000    -4.358019   -2.604974
         118 |  -4.384297   .7070817    -6.20   0.000    -5.770151   -2.998442
         119 |  -3.455265   .4472136    -7.73   0.000    -4.331787   -2.578742
         120 |  -3.106077   .3779645    -8.22   0.000    -3.846874    -2.36528
ln(risktime) |          1  (exposure)
------------------------------------------------------------------------------

. // predict the baseline (one parameter for each interval)
```

```
. predict haz_grp, nooffset
(option mu assumed; predicted mean d)

. replace haz_grp = haz_grp*1000
(1,920 real changes made)

. twoway  (scatter haz_grp midtime)  ///
>                 , xtitle("Years from diagnosis") ///
>                 ytitle("Baseline hazard (1000 pys)") ///
>                 ylabel(5 10 20 50 100 150, angle(h)) ///
>                 name(piecewise, replace)

. di "Total number of parameters is `e(k)'"
Total number of parameters is 120
```

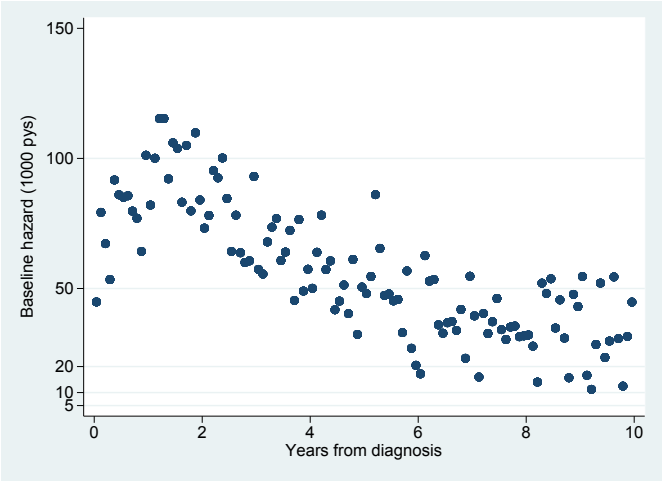


Figure 20: Localised skin melanoma. Plot of the estimated baseline hazard function for the piecewise model.

(b) The log hazard function before the knot at 1.5 year, $t \leq 1.5$, is:

$$\ln h(t) = \beta_0 + \beta_1 t$$

The log hazard function after the knot at 1.5 year, $t > 1.5$, is:

$$\ln h(t) = \beta_0 + \beta_1 t + \beta_2 + \beta_3 (t - 1)$$

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

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

Generalized linear models                         No. of obs      =      1,920
Optimization     : ML                             Residual df     =      1,916
                                                  Scale parameter =          1
Deviance         =  3241.142594                   (1/df) Deviance =   1.691619
Pearson          =  4714.038396                   (1/df) Pearson  =   2.460354

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   3.272386
Log likelihood   = -3137.490717                   BIC             =  -11243.97
------------------------------------------------------------------------------
             |                 OIM
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lin_s1 |   .3833764   .0767377     5.00   0.000     .2329733    .5337795
    lin_int2 |  -.2135571   .0730092    -2.93   0.003    -.3566525   -.0704617
      lin_s2 |  -.5338942   .0775133    -6.89   0.000    -.6858175   -.3819709
       _cons |   -2.76861   .0698084   -39.66   0.000    -2.905432   -2.631788
ln(risktime) |          1  (exposure)
------------------------------------------------------------------------------

. predict haz_lin1, nooffset
(option mu assumed; predicted mean d)

. replace haz_lin1 = haz_lin1*1000
(1,920 real changes made)

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

. di "the gradient up to 1.5 years is: " _b[lin_s1]
the gradient up to 1.5 years is: .38337637

. di "the gradient after 1.5 years is: " _b[lin_s1] + _b[lin_s2]
the gradient after 1.5 years is: -.15051783
```

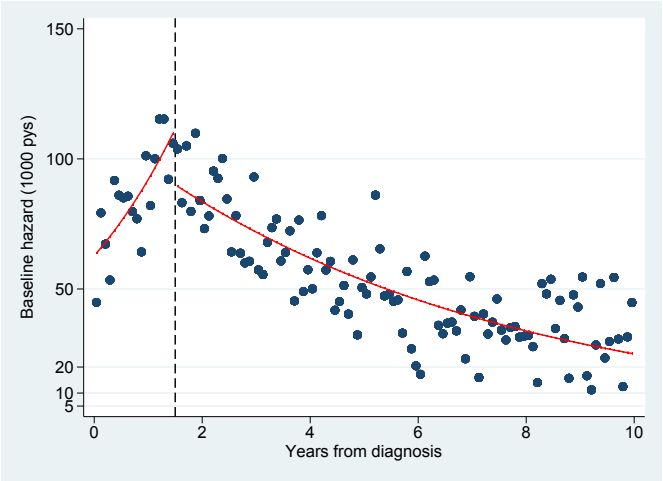


Figure 21: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and linear spline model.

Comparing the piecewise fitted function and the linear spline function, shown in Figure 21, we observe that the linear spline model fits the data very well.

```
. di "the gradient up to 1 year is: " _b[lin_s1]
the gradient up to 1 year is: .24828023

. di "the gradient after 1 year is: " _b[lin_s1] + _b[lin_s2]
the gradient after 1 year is: -.271407
```

(c)
```
. glm d lin_s1 lin_s2 , family(poisson) link(log) lnoffset(risktime)

Iteration 0:   log likelihood = -3325.6269
Iteration 1:   log likelihood =   -3143.98
Iteration 2:   log likelihood = -3141.6801
Iteration 3:   log likelihood = -3141.6762
Iteration 4:   log likelihood = -3141.6762

Generalized linear models                         No. of obs      =      1,920
Optimization     : ML                             Residual df     =      1,917
                                                  Scale parameter =          1
Deviance         =  3249.513617                   (1/df) Deviance =   1.695104
Pearson          =  4756.012765                   (1/df) Pearson  =   2.480966

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   3.275704
Log likelihood   = -3141.676229                   BIC             =  -11243.16

------------------------------------------------------------------------------
             |                 OIM
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lin_s1 |   .2178297   .0513656     4.24   0.000     .1171549    .3185045
      lin_s2 |   -.380508   .0567922    -6.70   0.000    -.4918187   -.2691973
       _cons |  -2.681235   .0619486   -43.28   0.000    -2.802652   -2.559818
ln(risktime) |          1  (exposure)
------------------------------------------------------------------------------
```

```
. predict haz_lin2, nooffset
(option mu assumed; predicted mean d)

. replace haz_lin2 = haz_lin2*1000
(1,920 real changes made)

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

. di "the gradient up to 1.5 years is: " _b[lin_s1]
the gradient up to 1.5 years is: .21782972

. di "the gradient after to 1.5 years is: " _b[lin_s1] + _b[lin_s2]
the gradient after to 1.5 years is: -.16267827
```

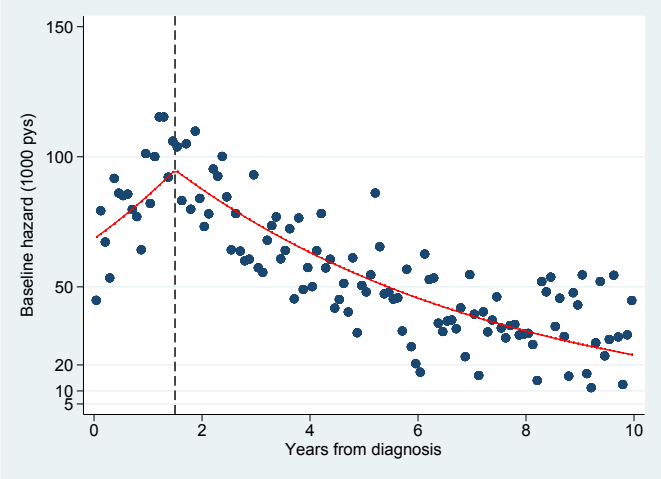


Figure 22: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and linear spline model.

```
. di "the gradient up to 1 year is: " _b[lin_s1]
the gradient up to 1 year is: .6310592

. di "the gradient after to 1 year is: " _b[lin_s1] + _b[lin_s2]
the gradient after to 1 year is: -.24886701
```

(d)
```
. gen cubic_s1 = midtime
. gen cubic_s2 = midtime^2
. gen cubic_s3 = midtime^3
. gen cubic_int = midtime>2
. gen cubic_lin = (midtime - 2)*(midtime>2)
. gen cubic_quad = ((midtime - 2)^2)*(midtime>2)
. gen cubic_s4 = ((midtime - 2)^3)*(midtime>2)
. glm d cubic* , family(poisson) link(log) lnoffset(risktime)

Iteration 0:   log likelihood = -3314.3924
Iteration 1:   log likelihood = -3136.0859
Iteration 2:   log likelihood = -3133.1534
Iteration 3:   log likelihood = -3133.1501
Iteration 4:   log likelihood = -3133.1501

Generalized linear models                         No. of obs      =      1,920
Optimization     : ML                             Residual df     =      1,912
                                                  Scale parameter =          1
Deviance         =  3232.461336                   (1/df) Deviance =   1.690618
Pearson          =  4648.482544                   (1/df) Pearson  =   2.431215

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   3.272031
Log likelihood   = -3133.150088                   BIC             =  -11222.41
------------------------------------------------------------------------------
             |                 OIM
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cubic_s1 |   .6523493   .5301936     1.23   0.219     -.386811     1.69151
    cubic_s2 |  -.1244914    .604615    -0.21   0.837    -1.309515    1.060532
    cubic_s3 |  -.0480855   .1971288    -0.24   0.807    -.4344508    .3382799
   cubic_int |  -.0358033   .1387985    -0.26   0.796    -.3078434    .2362367
   cubic_lin |   .2325272   .5186172     0.45   0.654    -.7839438    1.248998
  cubic_quad |   .4106761   .5955855     0.69   0.490      -.75665    1.578002
    cubic_s4 |   .0495792   .1971493     0.25   0.801    -.3368264    .4359847
       _cons |  -2.841688   .1277767   -22.24   0.000    -3.092126    -2.59125
ln(risktime) |          1  (exposure)
------------------------------------------------------------------------------

. predict haz_cubic1, nooffset
(option mu assumed; predicted mean d)

. replace haz_cubic1 = haz_cubic1*1000
(1,920 real changes made)

. twoway  (scatter haz_grp midtime)  ///
>                 (line haz_cubic1 midtime if midtime<=2, lcolor(red)) ///
>                 (line haz_cubic1 midtime if midtime>2, lcolor(red)) ///
>                 , xtitle("Years from diagnosis") ///
>                 ytitle("Baseline hazard (1000 pys)") ///
>                 xline(2, lcolor(black) lpattern(dash)) ///
>                 ylabel(5 10 20 50 100 150, angle(h)) ///
>                 legend(off) ///
>                 name(cubic1, replace)
```

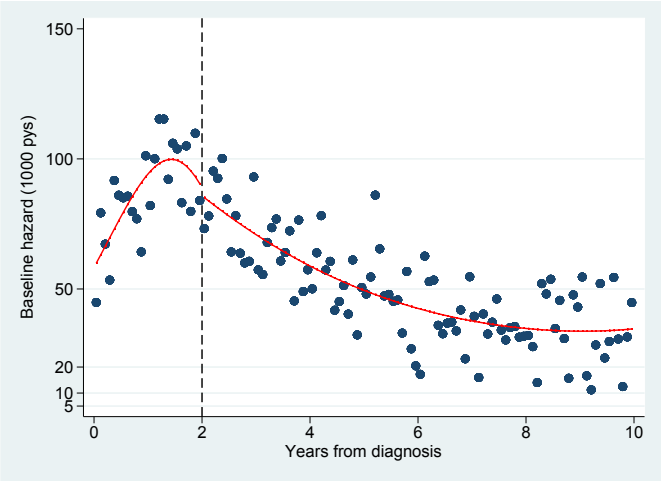


Figure 23: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and cubic spline model.

(e)
```
. glm d cubic_s* cubic_lin cubic_quad, family(poisson) link(log) lnoffset(risktime)

Iteration 0:   log likelihood = -3314.4284
Iteration 1:   log likelihood = -3136.1237
Iteration 2:   log likelihood = -3133.1865
Iteration 3:   log likelihood = -3133.1833
Iteration 4:   log likelihood = -3133.1833

Generalized linear models                         No. of obs      =      1,920
Optimization     : ML                             Residual df     =      1,913
                                                  Scale parameter =          1
Deviance         =  3232.527663                   (1/df) Deviance =   1.689769
Pearson          =  4648.358616                   (1/df) Pearson  =   2.429879

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   3.271024
Log likelihood   = -3133.183252                   BIC             =  -11229.91
------------------------------------------------------------------------------
             |                 OIM
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cubic_s1 |   .5997222   .4889988     1.23   0.220    -.3586977    1.558142
    cubic_s2 |  -.0478583   .5263989    -0.09   0.928    -1.079581    .9838645
    cubic_s3 |  -.0774854   .1608245    -0.48   0.630    -.3926957    .2377248
    cubic_s4 |   .0787461   .1614884     0.49   0.626    -.2377654    .3952575
   cubic_lin |    .320885   .3899094     0.82   0.411    -.4433234    1.085093
  cubic_quad |    .513397   .4429728     1.16   0.246    -.3548136    1.381608
       _cons |  -2.834161    .124225   -22.81   0.000    -3.077638   -2.590685
ln(risktime) |          1  (exposure)
------------------------------------------------------------------------------

. predict haz_cubic2, nooffset
(option mu assumed; predicted mean d)

. replace haz_cubic2 = haz_cubic2*1000
(1,920 real changes made)

. twoway  (scatter haz_grp midtime)  ///
```

```
>                (line haz_cubic2 midtime, lcolor(red)) ///
>                , xtitle("Years from diagnosis") ///
>                ytitle("Baseline hazard (1000 pys)") ///
>                xline(2, lcolor(black) lpattern(dash)) ///
>                ylabel(5 10 20 50 100 150, angle(h)) ///
>                legend(off) ///
>                name(cubic2, replace)
```

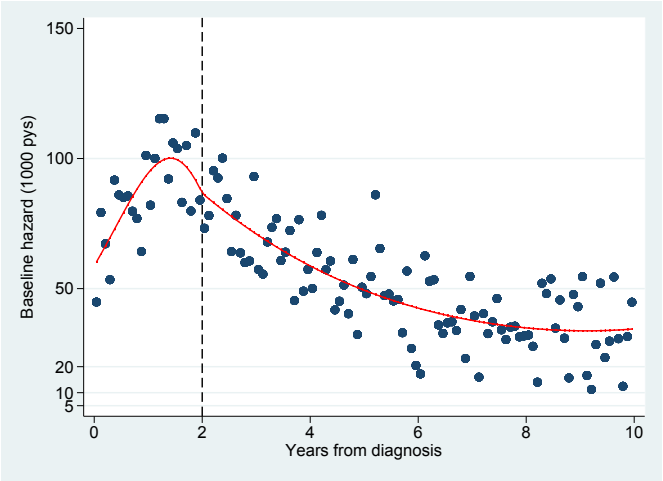


Figure 24: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and cubic spline model.

The fitted cubic spline function appears over-parameterised.

(f)
```
. glm d cubic_s* cubic_quad, family(poisson) link(log) lnoffset(risktime)

Generalized linear models                         No. of obs      =      1,920
Optimization     : ML                             Residual df     =      1,914
                                                  Scale parameter =          1
Deviance         =  3233.205488                   (1/df) Deviance =    1.68924
Pearson          =  4648.130991                   (1/df) Pearson  =   2.428491

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   3.270336
Log likelihood   = -3133.522164                   BIC             =  -11236.79
------------------------------------------------------------------------------
             |                 OIM
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cubic_s1 |   .8568882   .3786741     2.26   0.024     .1147007    1.599076
    cubic_s2 |  -.3818574   .3374689    -1.13   0.258    -1.043284    .2795696
    cubic_s3 |   .0351165   .0851876     0.41   0.680    -.1318482    .2020812
    cubic_s4 |  -.0350218   .0841447    -0.42   0.677    -.1999424    .1298989
  cubic_quad |   .1861311   .1969974     0.94   0.345    -.1999767    .5722389
       _cons |  -2.875102   .1148165   -25.04   0.000    -3.100138   -2.650066
ln(risktime) |          1  (exposure)
------------------------------------------------------------------------------

. predict haz_cubic3, nooffset
(option mu assumed; predicted mean d)

. replace haz_cubic3 = haz_cubic3*1000
```

```
(1,920 real changes made)

. twoway  (scatter haz_grp midtime)  ///
>                 (line haz_cubic3 midtime, lcolor(red)) ///
>                 , xtitle("Years from diagnosis") ///
>                 ytitle("Baseline hazard (1000 pys)") ///
>                 xline(2, lcolor(black) lpattern(dash)) ///
>                 ylabel(5 10 20 50 100 150, angle(h)) ///
>                 legend(off) ///
>                 name(cubic3, replace)
```

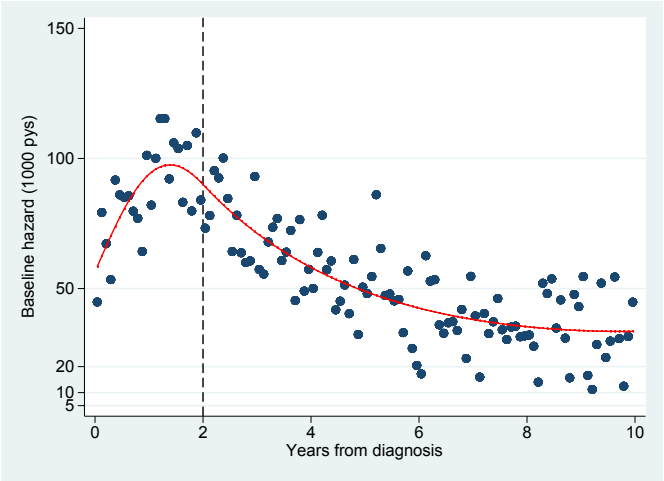


Figure 25: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and cubic spline model with continuous first derivatives.

If you brought your magnifying glass, you can see an ever so slight improvement in the stability and smoothness of the fitted function.

(g)
```
glm d cubic_s*, family(poisson) link(log) lnoffset(risktime)
predict haz_cubic4, nooffset
replace haz_cubic4 = haz_cubic4*1000
twoway (scatter haz_grp midtime)  ///
(line haz_cubic4 midtime, lcolor(red)) ///
, xtitle("Years from diagnosis") ///
ytitle("Baseline hazard (1000 pys)") ///
xline(2, lcolor(black) lpattern(dash)) ///
ylabel(5 10 20 50 100 150, angle(h)) ///
legend(off) ///
name(cubic4, replace)
```

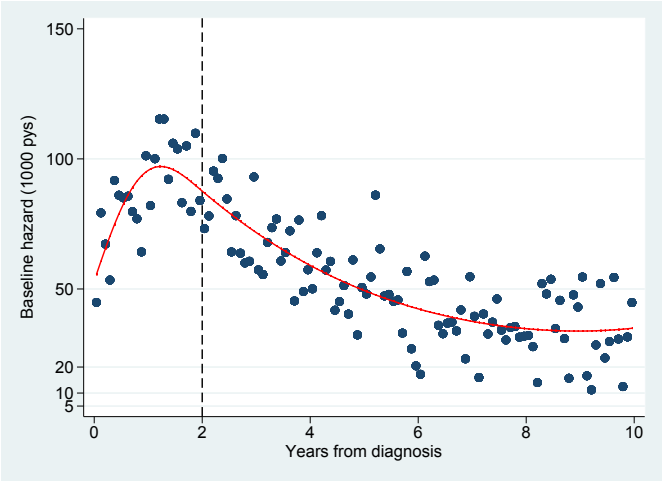


Figure 26: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and cubic spline model with continuous first and second derivatives.

The model fit appears to improve as the constraints are added, providing a more plausible fit to the data.

(h)
```
. rcsgen midtime, gen(rcs) df(4) fw(d)
Variables rcs1 to rcs4 were created

. global knots 'r(knots)'
```

(i)
```
. glm d rcs1, family(poisson) link(log) lnoffset(risktime)

Generalized linear models                         No. of obs      =      1,920
Optimization     : ML                             Residual df     =      1,918
                                                  Scale parameter =          1
Deviance         =  3296.146807                   (1/df) Deviance =   1.718533
Pearson          =   4685.68724                   (1/df) Pearson  =   2.443007

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   3.298951
Log likelihood   = -3164.992824                   BIC             =  -11204.09
------------------------------------------------------------------------------
             |                 OIM
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rcs1 |  -.1200737   .0077061   -15.58   0.000    -.1351773   -.1049701
       _cons |  -2.336551   .0301252   -77.56   0.000    -2.395595   -2.277506
ln(risktime) |          1  (exposure)
------------------------------------------------------------------------------

. estimates store rcs1

. predict haz_rcs1, nooffset
(option mu assumed; predicted mean d)

. replace haz_rcs1 = haz_rcs1*1000
(1,920 real changes made)

. twoway  (scatter haz_grp midtime)  ///
>                 (line haz_rcs1 midtime, lcolor(red)) ///
>                 , xtitle("Years from diagnosis") ///
```

```
>                 ytitle("Baseline hazard (1000 pys)") ///
>                 ylabel(5 10 20 50 100 150, angle(h)) ///
>                 legend(off) ///
>                 name(rcs1, replace)
```

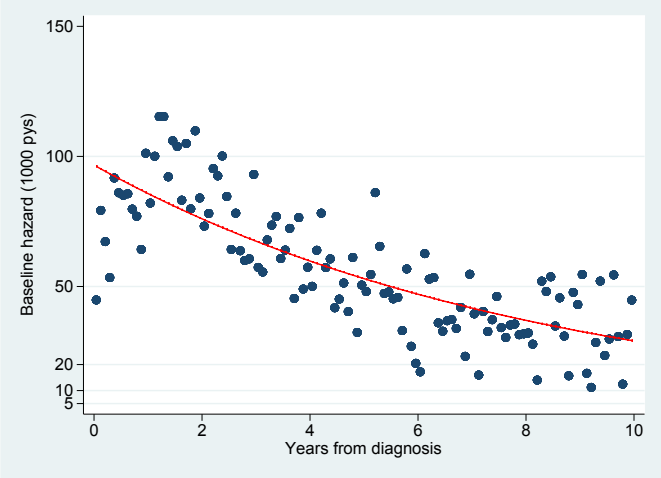


Figure 27: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and linear model.

The linear model appears to fit very poorly.

(j)
```
. glm d rcs*, family(poisson) link(log) lnoffset(risktime)

Generalized linear models                         No. of obs      =      1,920
Optimization     : ML                             Residual df     =      1,915
                                                  Scale parameter =          1
Deviance         =  3233.589355                   (1/df) Deviance =   1.688558
Pearson          =  4648.401252                   (1/df) Pearson  =   2.427364

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   3.269494
Log likelihood   = -3133.714098                   BIC             =  -11243.96
------------------------------------------------------------------------------
             |                 OIM
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rcs1 |   .5594366   .1069501     5.23   0.000     .3498183     .769055
        rcs2 |   .2341777   .0568007     4.12   0.000     .1228503    .3455051
        rcs3 |  -.1274038   .0418432    -3.04   0.002     -.209415   -.0453926
        rcs4 |   .0005971   .0084695     0.07   0.944    -.0160029    .0171971
       _cons |  -2.825642   .0782389   -36.12   0.000    -2.978988   -2.672297
ln(risktime) |          1  (exposure)
------------------------------------------------------------------------------

. estimates store rcs2
. lrtest rcs1 rcs2

Likelihood-ratio test                                 LR chi2(3)  =     62.56
(Assumption: rcs1 nested in rcs2)                     Prob > chi2 =    0.0000

. predict haz_rcs2, nooffset
(option mu assumed; predicted mean d)
```

```
. replace haz_rcs2 = haz_rcs2*1000
(1,920 real changes made)
```

The likelihood ratio test gave a p-value of <0.0001, indicating evidence against the null hypothesis that the effect is linear.

```
. predict haz_rcs2, nooffset
(option mu assumed; predicted mean d)

. replace haz_rcs2 = haz_rcs2*1000
(72 real changes made)

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

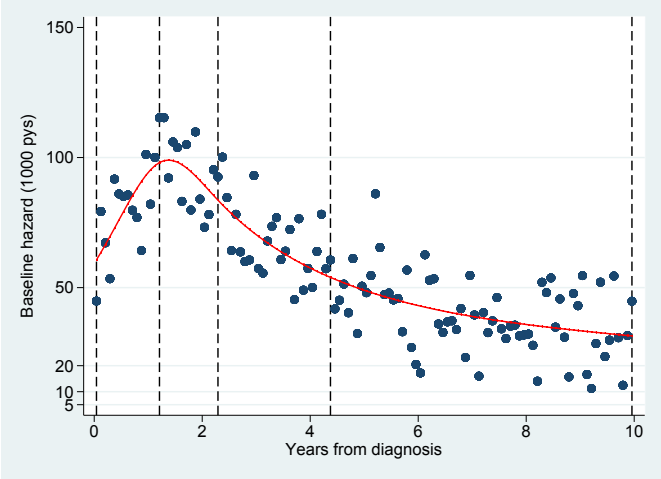


Figure 28: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and restricted cubic spline model.

(k)

```
. drop rcs*
. rcsgen midtime, gen(rcs) knots(1 2 3) fw(d)
Variables rcs1 to rcs2 were created

. global knots `r(knots)'

. glm d rcs*, family(poisson) link(log) lnoffset(risktime)

Generalized linear models                         No. of obs      =      1,920
Optimization     : ML                             Residual df     =      1,917
                                                  Scale parameter =          1
Deviance         =  3265.098545                   (1/df) Deviance =   1.703233
Pearson          =  4774.278604                   (1/df) Pearson  =   2.490495

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   3.283822
Log likelihood   = -3149.468693                   BIC             =  -11227.58
```

```
------------------------------------------------------------------------------
             |                 OIM
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rcs1 |   .0756425   .0364661     2.07   0.038      .0041702    .1471148
        rcs2 |   .0804797   .0145799     5.52   0.000      .0519036    .1090557
       _cons |  -2.568201   .0532653   -48.22   0.000     -2.672599   -2.463803
ln(risktime) |          1  (exposure)
------------------------------------------------------------------------------

. predict haz_rcs3, nooffset
(option mu assumed; predicted mean d)

. replace haz_rcs3 = haz_rcs3*1000
(1,920 real changes made)

. twoway  (scatter haz_grp midtime)  ///
>                 (line haz_rcs3 midtime, lcolor(red)) ///
>                 , xtitle("Years from diagnosis") ///
>                 ytitle("Baseline hazard (1000 pys)") ///
>                 xline($knots , lcolor(black) lpattern(dash)) ///
>                 ylabel(5 10 20 50 100 150, angle(h)) ///
>                 legend(off) ///
>                 name(rcs3, replace)
```

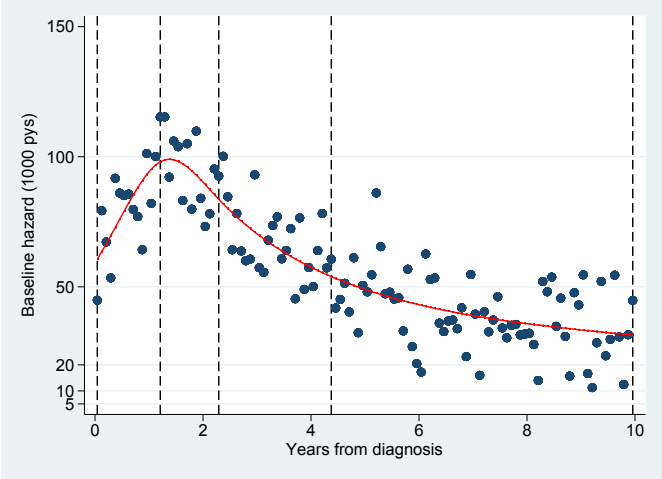


Figure 29: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and restricted cubic spline model with knots at 1, 2, and 3 years.

131. Flexible Parametric Survival (Royston-Parmar) Models

Load the Melanoma data and refit the Cox model to use as a comparison.

```
. // Load the Melanoma data, keep those with localized stage
. use melanoma, clear
(Skin melanoma, diagnosed 1975-94, follow-up to 1995)

. keep if stage == 1
(2,457 observations deleted)

. gen female = sex == 2

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

     failure event:  status == 1
obs. time interval:  (0, surv_mm]
 exit on or before:  time 120.5
    t for analysis:  time/12

------------------------------------------------------------------------------
       5318  total observations
          0  exclusions
------------------------------------------------------------------------------
       5318  observations remaining, representing
        961  failures in single-record/single-failure data
  32437.667  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  10.04167
```

(a) Kaplan-Meier curve.

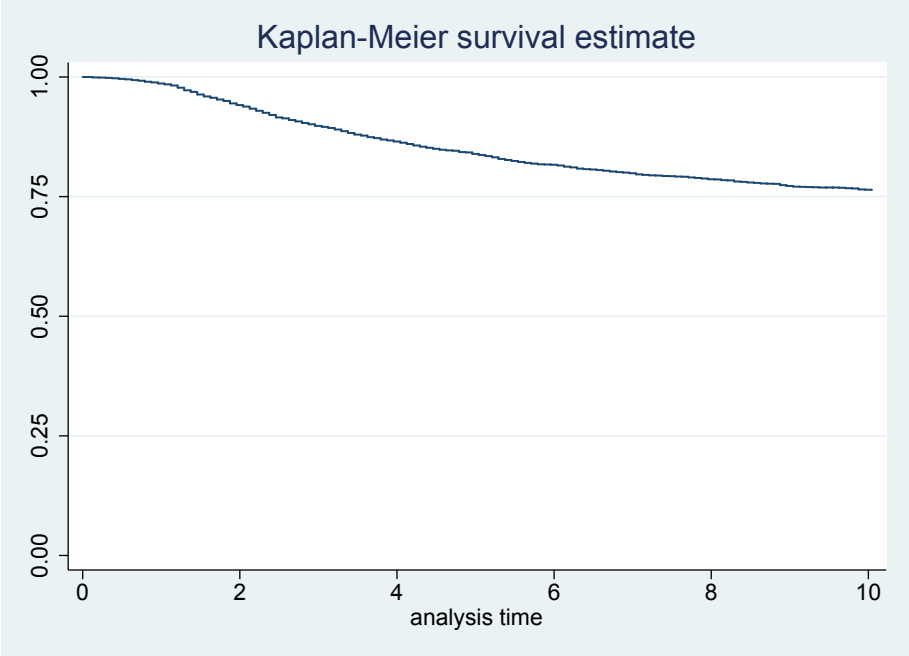
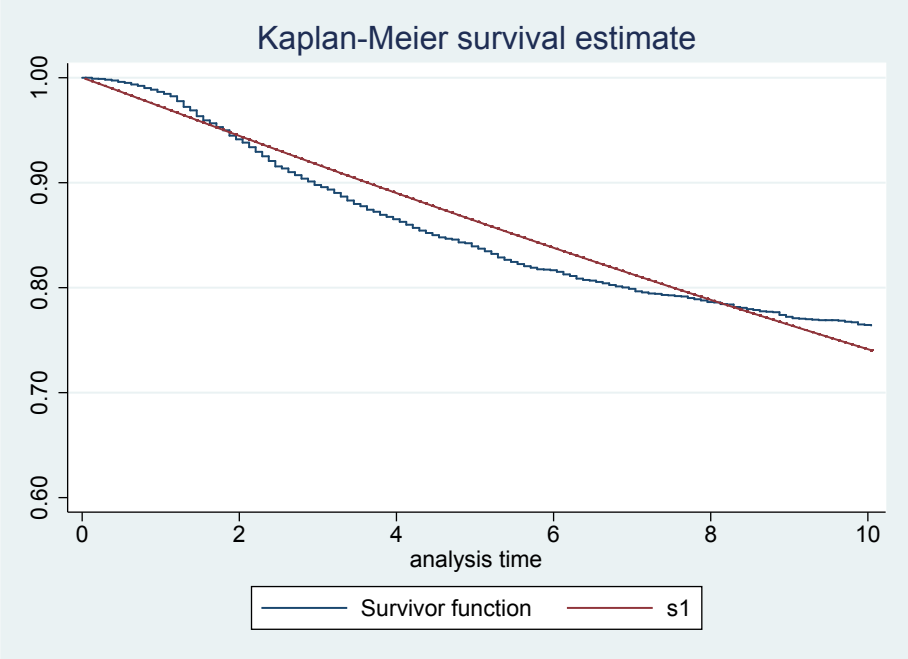


Figure 30: Localised skin melanoma. Plot of the estimated survival function.

(b) Weibull model using `stpm2`.

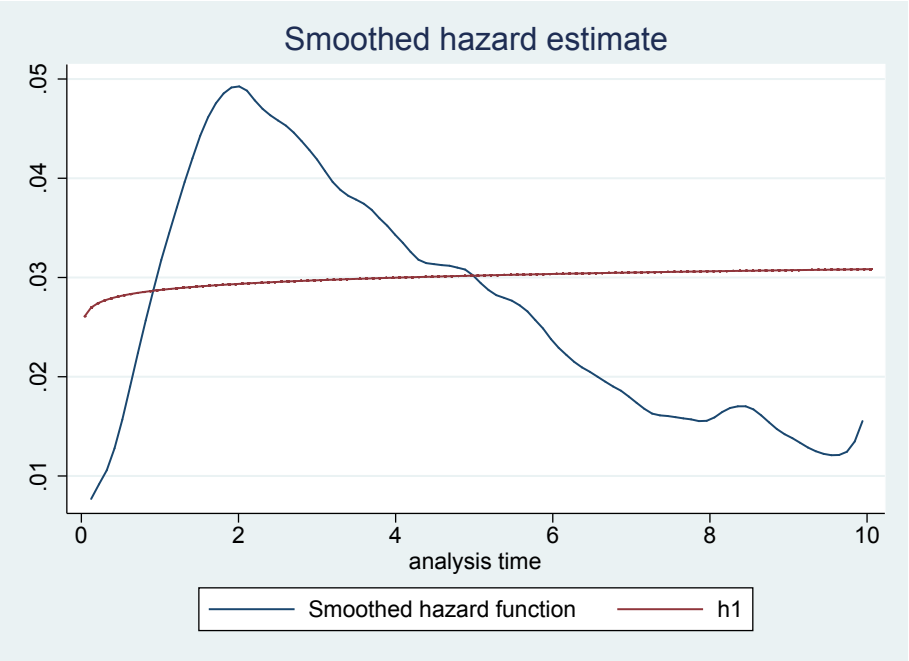
```
. stpm2, scale(hazard) df(1)

Iteration 0:   log likelihood = -3493.7327
Iteration 1:   log likelihood = -3374.1674
Iteration 2:   log likelihood = -3369.6234
Iteration 3:   log likelihood = -3369.6113
Iteration 4:   log likelihood = -3369.6113

Log likelihood = -3369.6113                     Number of obs     =      5,318

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
       _rcs1 |   .7948519    .022936    34.66   0.000     .7498981    .8398056
       _cons |  -1.947946   .0343742   -56.67   0.000    -2.015318   -1.880574
------------------------------------------------------------------------------

. predict s1, surv

. predict h1, hazard
```



(c) Obtain hazard kernel density estimate of hazard function and compare to Weibull model.

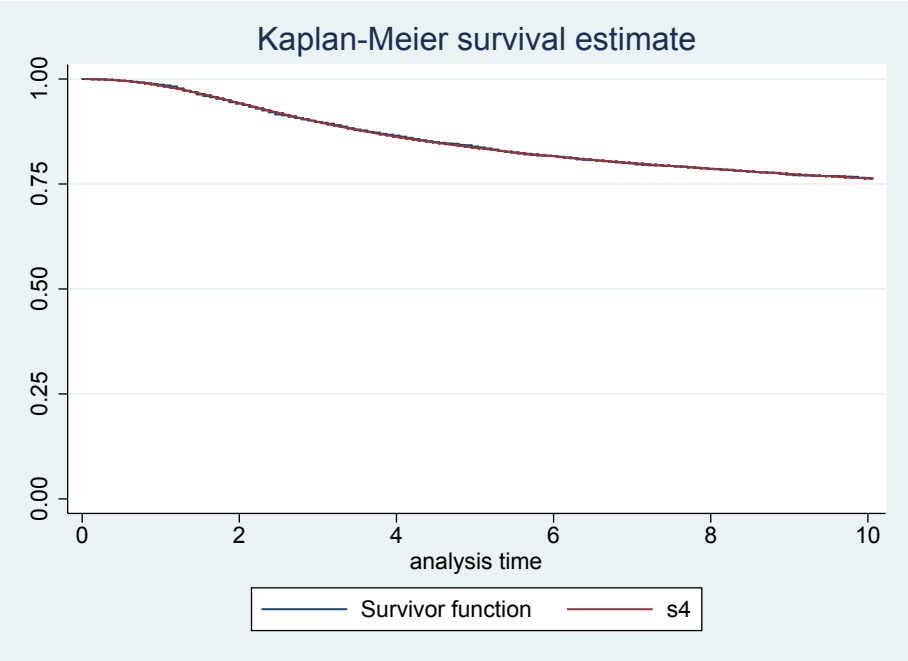
```
sts graph, hazard kernel(epan2) addplot(line h1 _t, sort) name(hazard1, replace)
```

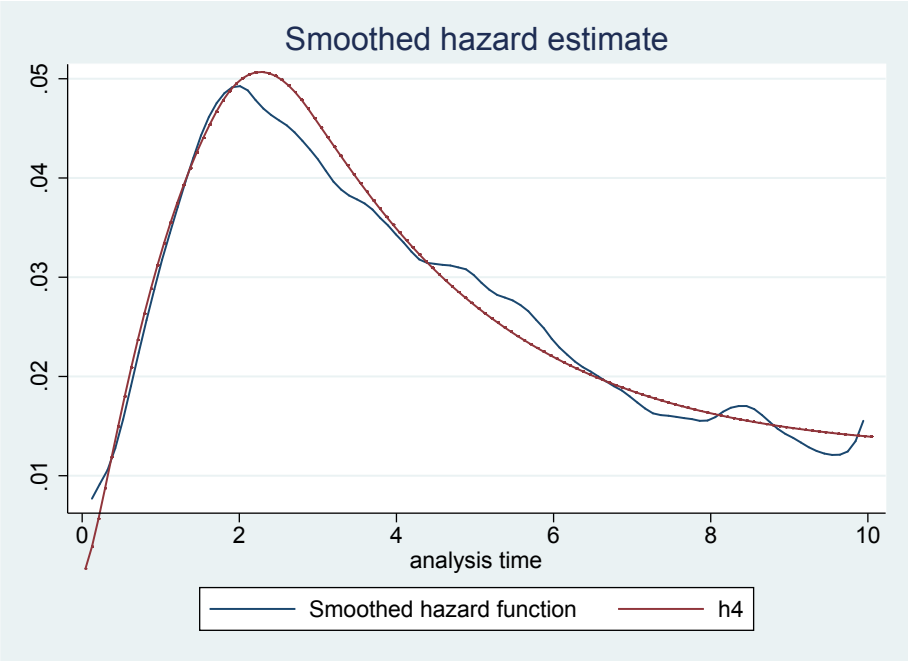
The Weibull model does not fit well as the hazard function appears to have a turning point. A Weibull model has either a increasing or decreasing hazard function.



(d) Fit flexible parametric model with 4df (5 knots) for the baseline.

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

Iteration 0:   log likelihood = -3277.5698
Iteration 1:   log likelihood = -3260.2601
Iteration 2:   log likelihood = -3259.4927
Iteration 3:   log likelihood =  -3259.491
Iteration 4:   log likelihood =  -3259.491

Log likelihood =  -3259.491                     Number of obs     =      5,318

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
       _rcs1 |   .9169168   .0299303    30.64   0.000     .8582546     .975579
       _rcs2 |   .2730108   .0365061     7.48   0.000       .20146    .3445615
       _rcs3 |   .0676424   .0194169     3.48   0.000     .0295859    .1056988
       _rcs4 |  -.0011682   .0078443    -0.15   0.882    -.0165428    .0142064
       _cons |  -1.965909   .0344635   -57.04   0.000    -2.033457   -1.898362
------------------------------------------------------------------------------

. predict s4, surv

. predict h4, hazard

. sts graph, addplot(line s4 _t, sort) name(km4, replace)

         failure _d:  status == 1
   analysis time _t:  surv_mm/12
  exit on or before:  time 120.5

. sts graph, hazard kernel(epan2) addplot(line h4 _t, sort) name(hazard4, replace)

         failure _d:  status == 1
   analysis time _t:  surv_mm/12
  exit on or before:  time 120.5
```





A much better fit than the Weibull model.

(e) Fit a Cox model.

```
. stcox year8594

        failure _d:  status == 1
  analysis time _t:  surv_mm/12
  exit on or before:  time 120.5

Iteration 0:   log likelihood =  -7907.738
Iteration 1:   log likelihood = -7900.3231
Iteration 2:   log likelihood = -7900.3231
Refining estimates:
Iteration 0:   log likelihood = -7900.3231
```

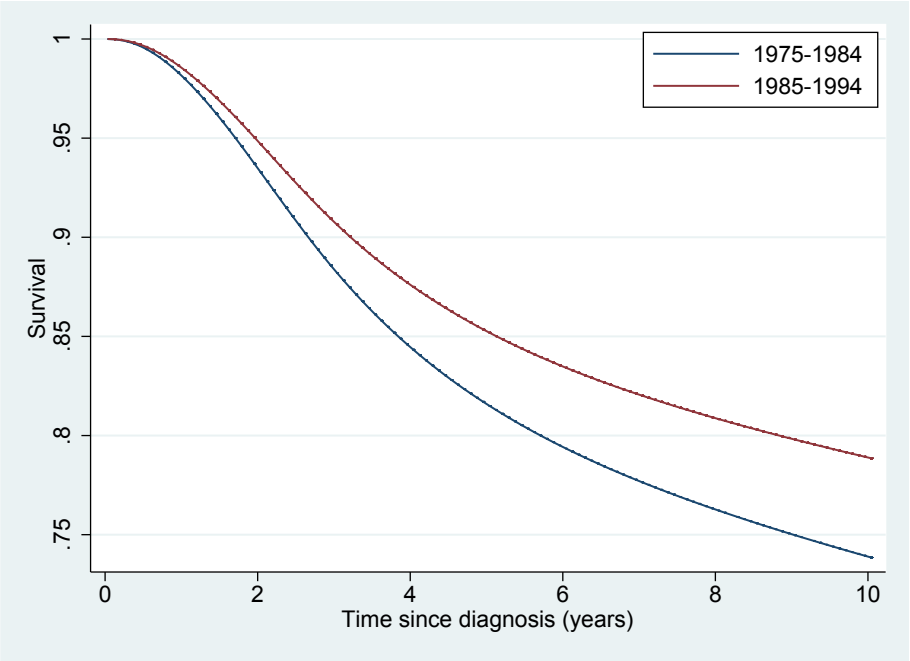
```
Cox regression -- Breslow method for ties

No. of subjects =        5,318                 Number of obs    =       5,318
No. of failures =          961
Time at risk    =  32437.66667
                                               LR chi2(1)       =       14.83
Log likelihood  =   -7900.3231                 Prob > chi2      =      0.0001

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    year8594 |   .7765254   .0510814    -3.84   0.000     .6825931    .8833839
------------------------------------------------------------------------------
```

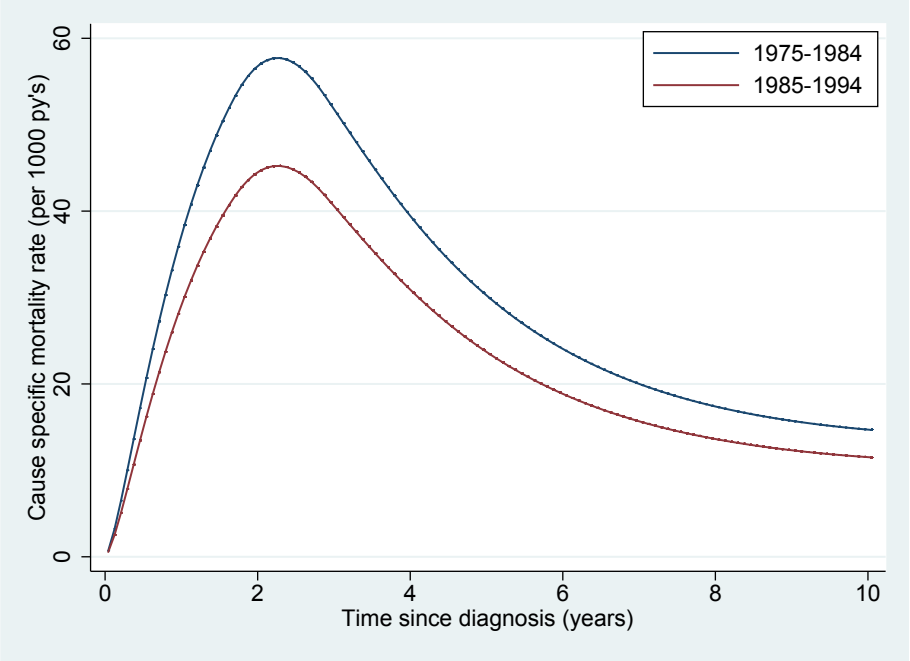
(f) Equivalent flexible parametric model.

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

Iteration 0:   log likelihood = -3272.2998
Iteration 1:   log likelihood = -3253.6208
Iteration 2:   log likelihood = -3252.6109
Iteration 3:   log likelihood = -3252.6073
Iteration 4:   log likelihood = -3252.6073

Log likelihood = -3252.6073                    Number of obs     =      5,318

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
    year8594 |   .7836011   .0515816    -3.70   0.000     .6887531    .8915105
       _rcs1 |   2.479199   .0741692    30.35   0.000     2.338009    2.628914
       _rcs2 |    1.31958   .0481939     7.59   0.000     1.228423    1.417501
       _rcs3 |   1.071416   .0207502     3.56   0.000     1.031508    1.112867
       _rcs4 |   .9999275   .0077227    -0.01   0.993     .9849053    1.015179
       _cons |   .1585156   .0074182   -39.36   0.000     .1446231    .1737427
------------------------------------------------------------------------------
```

(g) Predicted survival and hazard functions by period of diagnosis.

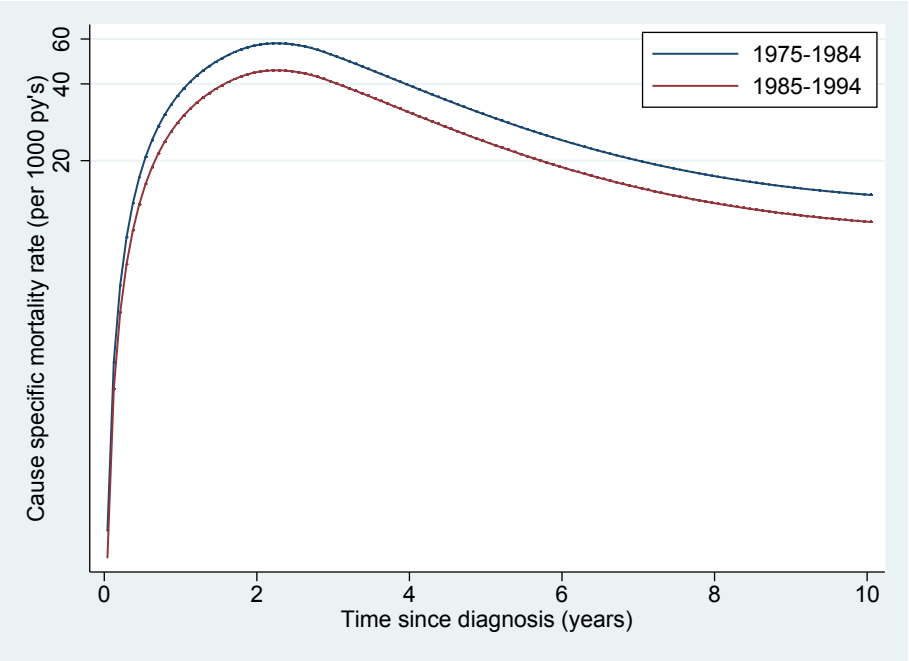
```
. predict s1ph, survival

. predict h1ph, hazard per(1000)

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

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





(h) Plot hazard functions on log scale.

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



A constant difference on the log scale means that the effect is proportional. The model is a proportional hazards model and so and predictions will have perfect proportional hazards.

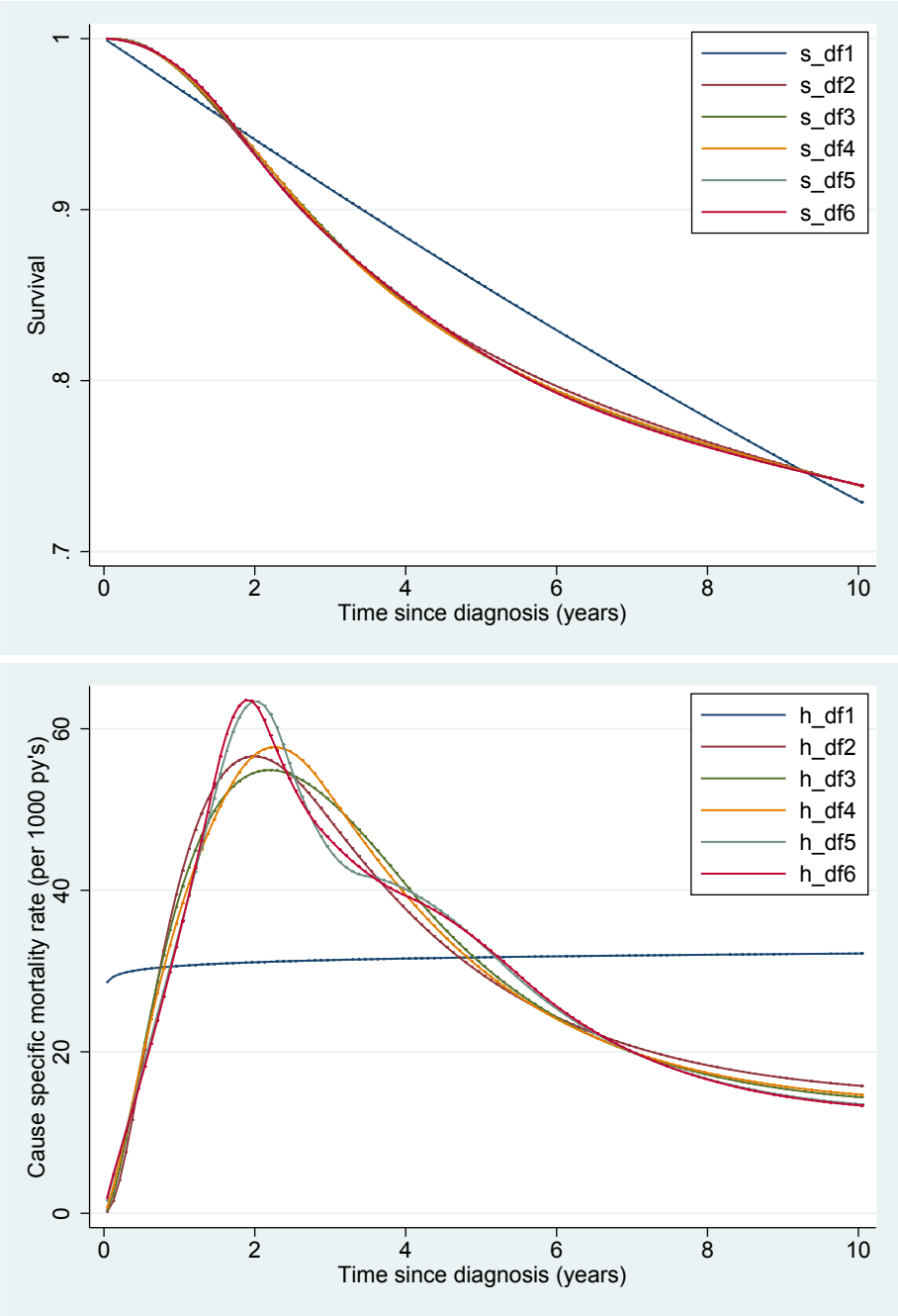
(i) Compare the number of knots.

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

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

-------------------------------------------------------------------------------------------
    Variable |    df1          df2          df3          df4          df5          df6
-------------+-----------------------------------------------------------------------------
    year8594 | -.11512481   -.24019646   -.24444962   -.24385523   -.24606124   -.24642169
             |  .06574271    .06582554      .065796    .06582631    .06579035    .06578964
-------------+-----------------------------------------------------------------------------
         AIC |  6742.1488    6517.4684    6517.1701    6517.2146    6512.2044    6513.2999
         BIC |  6756.7527    6536.9403      6541.51    6546.4225    6546.2802    6552.2437
-------------------------------------------------------------------------------------------
                                                                              legend: b/se
```

The AIC selects 5 df and the BIC 2 df. The hazards ratios are very similar with 2 or more df.

(j) Compare baseline hazard and survival functions with different degrees of freedom.

```
. line s_df* _t if year8594 == 0, sort  ///
        legend(ring(0) cols(1) pos(1)) ///
        xtitle("Time since diagnosis (years)") ///
        ytitle("Survival")

. line h_df* _t if year8594 == 0, sort ///
        legend(ring(0) cols(1) pos(1)) ///
        xtitle("Time since diagnosis (years)") ///
        ytitle("Cause specific mortality rate (per 1000 py’s)")
```



Having two or more df lead to similar fits, particularly for the survival function.

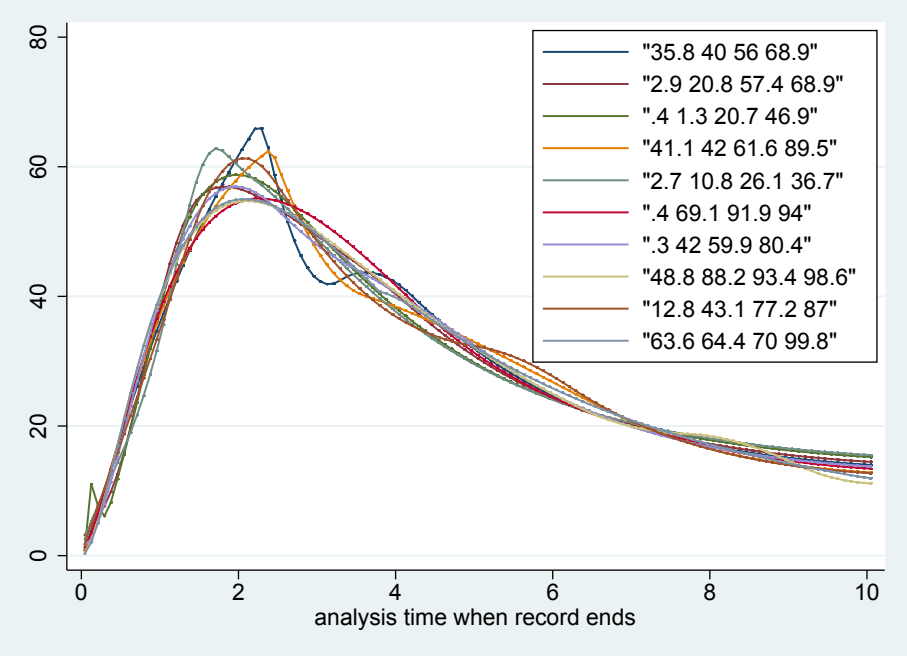
(k) Random knot locations.

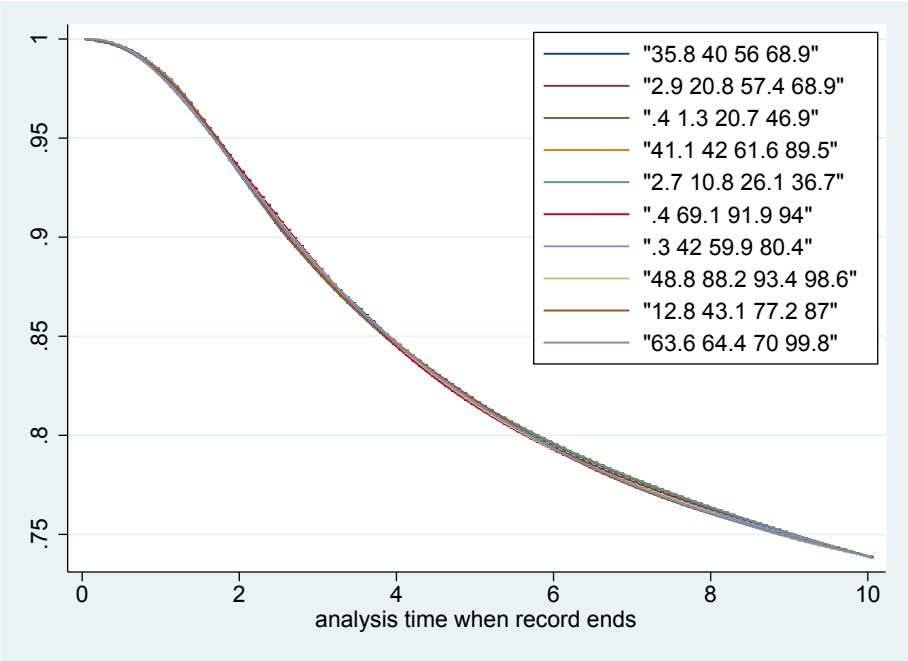
```
. replace _t = _t + runiform()*0.001
(5,318 real changes made)

. set seed 12345

. global legorder

. forvalues i = 1/10 {
  2.         local plist
  3.         forvalues j = 1/4 {
  4.                 local z`j': display %3.1f runiform()*100
  5.                 local plist  `plist' `z`j''
```

```
 6.         }
 7.         numlist "`plist'", sort
 8.         local plist `r(numlist)'
 9.         stpm2 year8594, scale(hazard) knots(`plist') knscale(centile) failconvlininit
10.         predict sp`i', surv zeros
11.         predict hp`i', hazard per(1000) zeros
12.         estimates store mp`i'
13.         global legorder ${legorder} `i' `""`plist'""'
14. }
```

```
. estimates table mp*, keep(year8594) se(%5.4f) b(%5.4f)

-------------------------------------------------------------------------------------------------
Variable |   mp1      mp2      mp3      mp4      mp5      mp6      mp7      mp8      mp9      mp10
---------+---------------------------------------------------------------------------------------
year8594 | -0.2450  -0.2448  -0.2428  -0.2466  -0.2416  -0.2461  -0.2459  -0.2470  -0.2469  -0.2468
         |  0.0658   0.0658   0.0658   0.0658   0.0658   0.0658   0.0658   0.0658   0.0658   0.0658
-------------------------------------------------------------------------------------------------
```

```
. // compare baseline hazard curves
. twoway (line hp* _t, sort), legend(order(${legorder}) ring(0) pos(1) cols(1)) ///
         name(hp,replace)

. // compare baseline survival curves
. twoway (line sp* _t, sort), legend(order(${legorder}) ring(0) pos(1) cols(1)) ///
         name(sp,replace)
```





(l) Add sex and age to the model and compare to a Cox model.

```
. stcox female year8594 i.agegrp

        failure _d:  status == 1
  analysis time _t:  surv_mm/12
  exit on or before:  time 120.5

Iteration 0:   log likelihood = -7902.3323
Iteration 1:   log likelihood = -7801.8606
Iteration 2:   log likelihood = -7796.3403
Iteration 3:   log likelihood =  -7796.318
Refining estimates:
Iteration 0:   log likelihood =  -7796.318

Cox regression -- no ties

No. of subjects =        5,318                 Number of obs    =       5,318
No. of failures =          961
Time at risk    =  32440.30996
                                               LR chi2(5)       =      212.03
Log likelihood  =    -7796.318                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .5891682   .0385376    -8.09   0.000     .5182772    .6697559
    year8594 |   .7204093   .0476836    -4.95   0.000     .6327594    .8202005
             |
      agegrp |
      45-59  |   1.321244   .1242452     2.96   0.003     1.098852    1.588646
      60-74  |   1.853307   .1681591     6.80   0.000     1.551365    2.214017
        75+  |   3.382446   .3528557    11.68   0.000     2.756981    4.149807
------------------------------------------------------------------------------

. estimate store cox

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

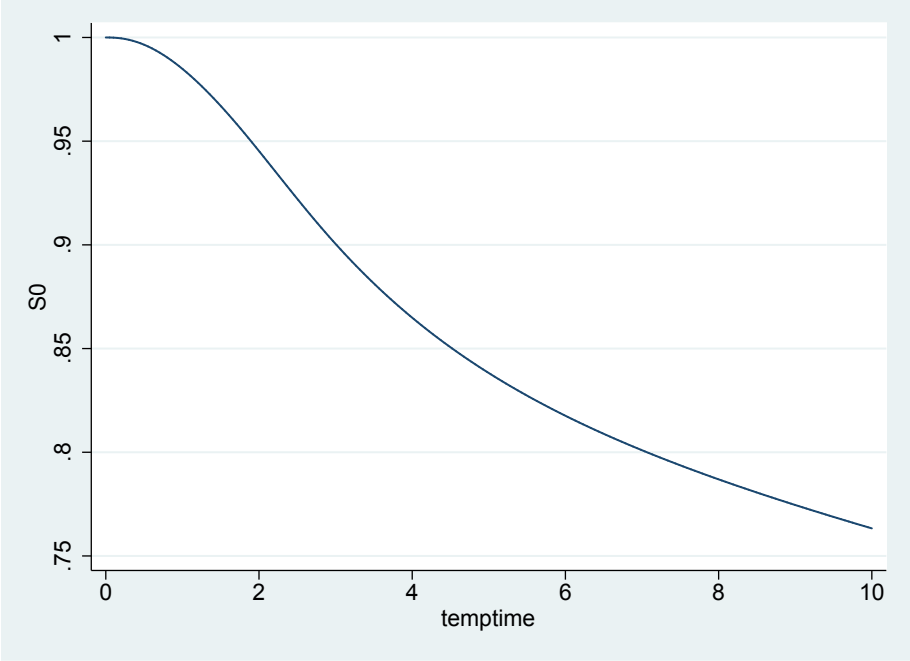
```
Iteration 0:   log likelihood = -3167.3947
Iteration 1:   log likelihood = -3153.8864
Iteration 2:   log likelihood = -3153.3628
Iteration 3:   log likelihood = -3153.3615
Iteration 4:   log likelihood = -3153.3615

Log likelihood = -3153.3615                     Number of obs     =      5,318

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      female |   .5888884   .0385204    -8.10   0.000      .518029    .6694404
    year8594 |   .7230319   .0478795    -4.90   0.000     .6350245    .8232361
             |
      agegrp |
      45-59  |   1.321555   .1242752     2.96   0.003     1.099109    1.589022
      60-74  |   1.853521   .1681828     6.80   0.000     1.551537    2.214282
        75+  |   3.385528   .3532167    11.69   0.000     2.759431    4.153684
             |
       _rcs1 |   2.546199   .0769614    30.92   0.000     2.399739    2.701599
       _rcs2 |   1.311274   .0479802     7.41   0.000     1.220528    1.408768
       _rcs3 |    1.07278   .0210209     3.59   0.000      1.03236    1.114781
       _rcs4 |   .9999819   .0080385    -0.00   0.998     .9843503    1.015862
       _cons |   .1376381   .0115929   -23.54   0.000     .1166929    .1623429
------------------------------------------------------------------------------

. estimates store stpm2_ph


. estimates table cox stpm2_ph,  equation(1) keep(#1:)  se

----------------------------------------
    Variable |    cox        stpm2_ph
-------------+--------------------------
      female | -.52904354   -.52951857
             |  .06541015    .06541214
    year8594 |  -.3279357   -.32430197
             |  .06618964    .06622047
             |
      agegrp |
      45-59  |  .27857398    .27880943
             |  .09403648    .09403706
      60-74  |  .61697173      .617087
             |  .09073462    .09073693
        75+  |  1.2185991      1.21951
             |  .10431967    .10433135
             |
       _rcs1 |               .93460183
             |               .03022597
       _rcs2 |               .27099947
             |               .03659051
       _rcs3 |               .07025301
             |               .01959482
       _rcs4 |              -.00001808
             |                .0080386
       _cons |              -1.9831271
             |                .08422746
----------------------------------------
                           legend: b/se
```

(m) Estimate are very similar as both models assume proportional hazards and we are using spline functions to model the hazard function flexibly.

(n) Using the predict command.

i. Creating and using the `temptime` option

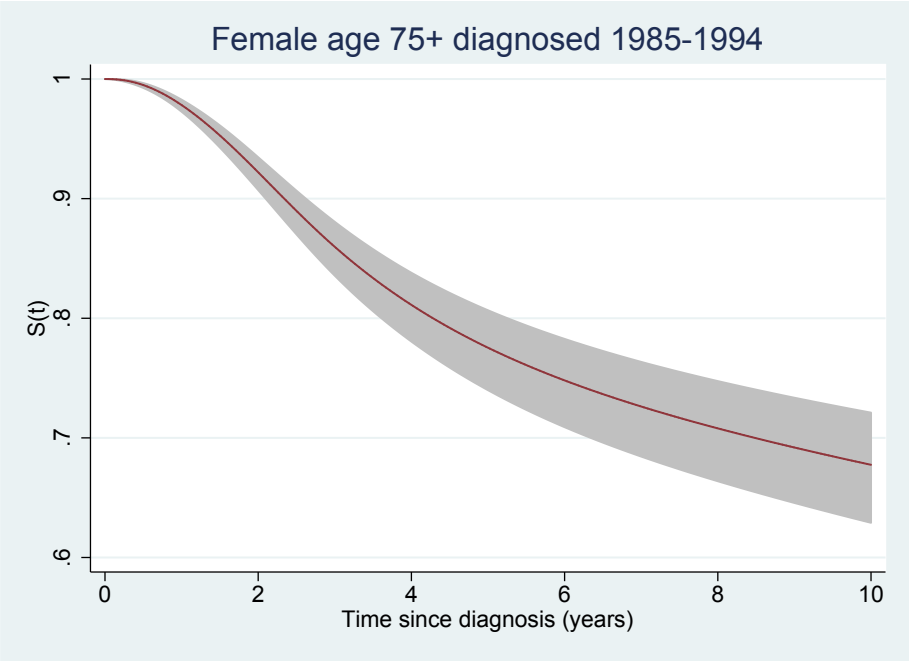
```
 . range temptime 0 10 200
(5,118 missing values generated)

. predict S0, survival zeros timevar(temptime)

. line S0 temptime, sort
```



The baseline represents males, aged ¡45 and diagnosed in 1975-1984.

ii. Using the `at()` and `zeros` options

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

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


Female age 75+ diagnosed 1985-1994
S(t)
1
.9
.8
.7
.6
0
2
4
6
8
10
Time since diagnosis (years)

132. **Modelling time-dependent effects using flexible parametric models**

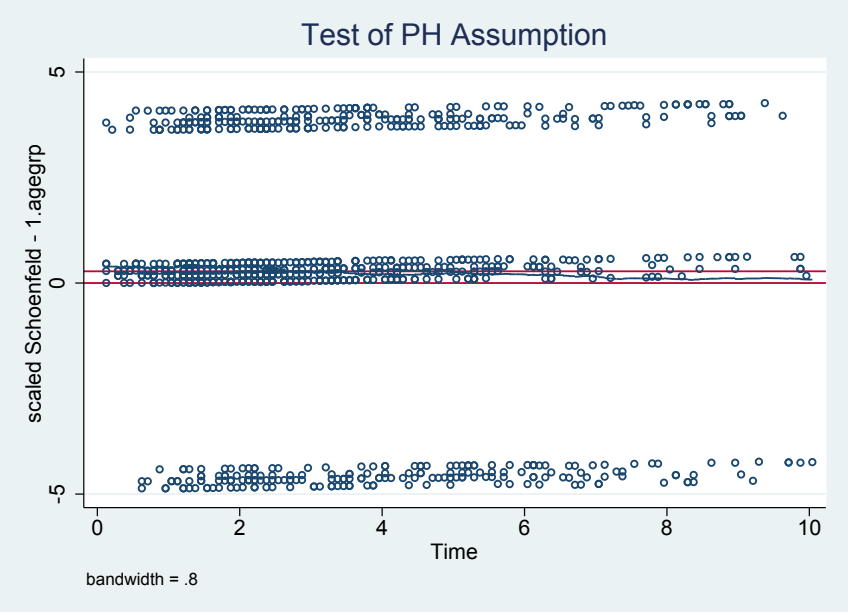
Load and `stset` the data

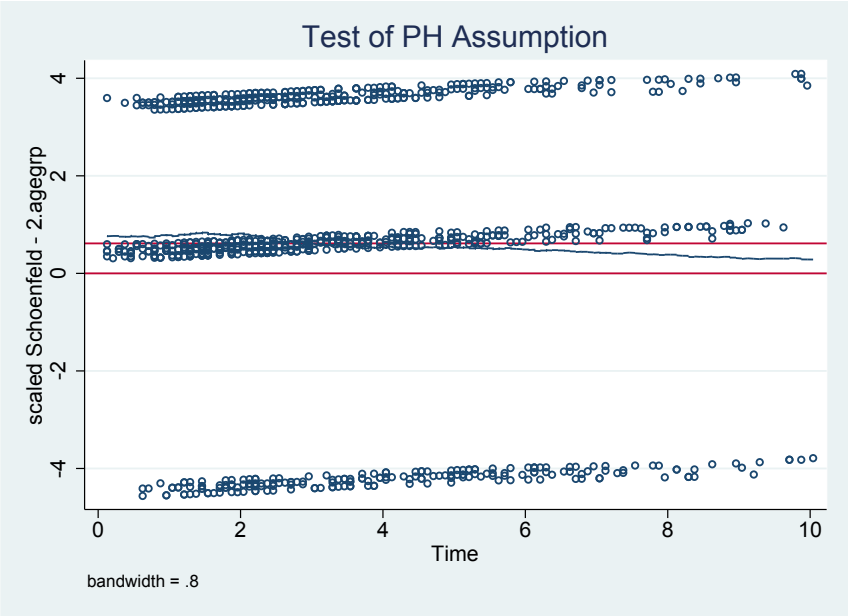
```
. use melanoma, clear
(Skin melanoma, diagnosed 1975-94, follow-up to 1995)

. keep if stage == 1
(2,457 observations deleted)

. gen female = sex == 2

. stset surv_mm, failure(status==1) exit(time 60.5) scale(12)

     failure event:  status == 1
obs. time interval:  (0, surv_mm]
 exit on or before:  time 60.5
    t for analysis:  time/12

--------------------------------------------------------------------------------
       5318  total observations
          0  exclusions
--------------------------------------------------------------------------------
       5318  observations remaining, representing
        747  failures in single-record/single-failure data
  21455.083  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  5.041667
```

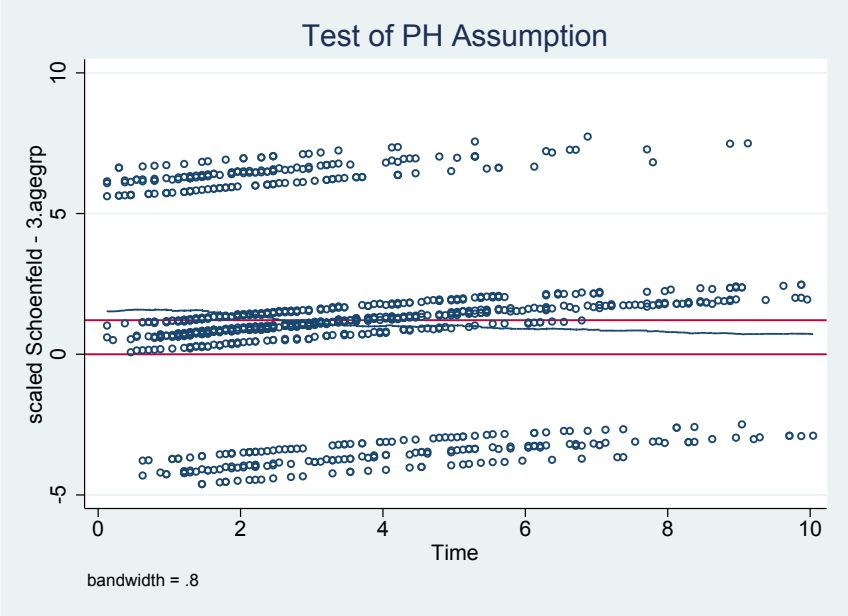
(a) First we will fit a Cox model and assess the proportional hazards assumption using Schoenfeld residuals.

```
. stcox female year8594 i.agegrp,

          failure _d:  status == 1
    analysis time _t:  surv_mm/12
   exit on or before:  time 60.5

Iteration 0:   log likelihood = -6243.0448
Iteration 1:   log likelihood = -6143.0805
Iteration 2:   log likelihood = -6137.2191
Iteration 3:   log likelihood = -6137.2003
Refining estimates:
Iteration 0:   log likelihood = -6137.2003

Cox regression -- Breslow method for ties

No. of subjects =        5,318                 Number of obs    =       5,318
No. of failures =          747
Time at risk    =  21455.08333
                                               LR chi2(5)       =      211.69
Log likelihood  =   -6137.2003                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .5592375   .0416501    -7.80   0.000     .4832833     .647129
    year8594 |   .6974691   .0514699    -4.88   0.000     .6035459    .8060085
             |
      agegrp |
```

```
      45-59  |   1.484577   .1677801     3.50   0.000     1.189608    1.852686
      60-74  |   2.149352   .2324899     7.07   0.000     1.738743    2.656929
        75+  |   3.976596   .4729993    11.61   0.000     3.149667    5.020631
------------------------------------------------------------------------------

. forvalue i = 1/3 {
  2.         local beta = _b[‘i’.agegrp]
  3.         estat phtest, plot(‘i’.agegrp) name(sch_age‘i’, replace) ///
>                yline(0 ‘beta’) msize(small) msymbol(Oh) bw(0.4)
  4. }

. estat phtest, detail

      Test of proportional-hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |       rho             chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      female      |      0.00207          0.00        1          0.9551
      year8594    |      0.08080          4.90        1          0.0269
      0b.agegrp   |            .             .        1               .
      1.agegrp    |     -0.02259          0.38        1          0.5356
      2.agegrp    |     -0.04408          1.45        1          0.2285
      3.agegrp    |     -0.11654          9.78        1          0.0018
      ------------+---------------------------------------------------
      global test |                      15.77        5          0.0075
      ----------------------------------------------------------------
```







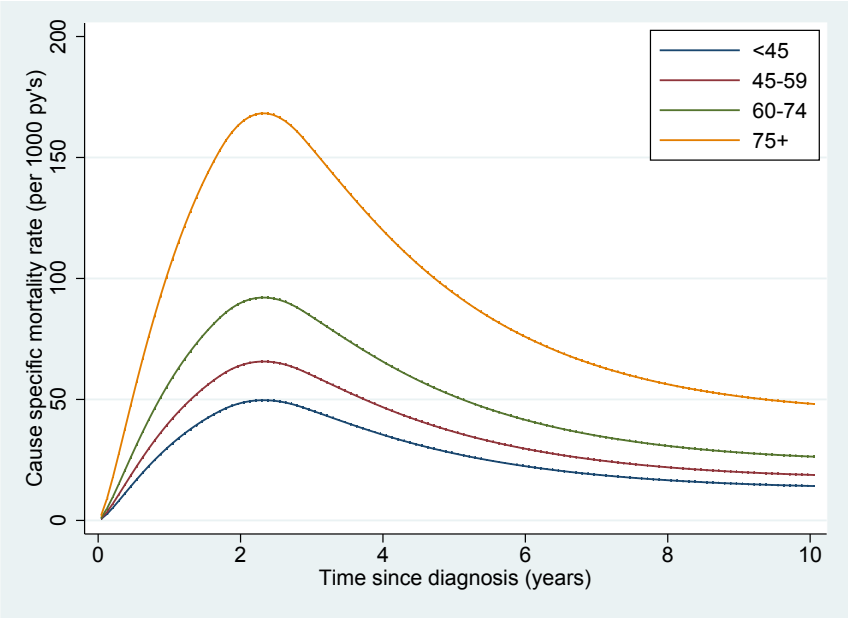
(b) Now fit a flexible parametric proportional hazards model with 4 df for the baseline.

```
. tab agegrp, gen(agegrp)

   Age in 4 |
 categories |      Freq.     Percent        Cum.
------------+-----------------------------------
       0-44 |      1,463       27.51       27.51
      45-59 |      1,575       29.62       57.13
      60-74 |      1,536       28.88       86.01
        75+ |        744       13.99      100.00
------------+-----------------------------------
      Total |      5,318      100.00

. tab agegrp, gen(agegrp)

   Age in 4 |
 categories |      Freq.     Percent        Cum.
------------+-----------------------------------
```

```
       0-44 |      1,463       27.51       27.51
      45-59 |      1,575       29.62       57.13
      60-74 |      1,536       28.88       86.01
        75+ |        744       13.99      100.00
------------+-----------------------------------
      Total |      5,318      100.00

. stpm2 female year8594 agegrp2-agegrp4, df(4) scale(hazard) eform

Iteration 0:   log likelihood = -2515.3648
Iteration 1:   log likelihood = -2508.7748
Iteration 2:   log likelihood = -2508.5979
Iteration 3:   log likelihood = -2508.5977
Iteration 4:   log likelihood = -2508.5977

Log likelihood = -2508.5977                     Number of obs     =      5,318

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      female |   .5580161   .0415611    -7.83   0.000     .4822244      .64572
    year8594 |   .7007966   .0517153    -4.82   0.000     .6064257    .8098533
     agegrp2 |   1.486106   .1679523     3.51   0.000     1.190834    1.854592
     agegrp3 |   2.154906   .2330888     7.10   0.000     1.743238    2.663789
     agegrp4 |    4.01077   .4770695    11.68   0.000     3.176727    5.063791
       _rcs1 |   2.315969   .0753367    25.82   0.000      2.17292    2.468435
       _rcs2 |   1.130169   .0396051     3.49   0.000      1.05515    1.210521
       _rcs3 |   1.076565   .0172889     4.59   0.000     1.043207    1.110989
       _rcs4 |   .9953895   .0065813    -0.70   0.485     .9825736    1.008373
       _cons |   .1050015   .0106141   -22.30   0.000     .0861294    .1280086
------------------------------------------------------------------------------

. estimates store ph
```

Predict and plot the hazard function for each age group for males diagnosed in 1975-1994.

```
. predict h_age1, hazard zeros per(1000)

. predict h_age2, hazard at(agegrp2 1) zeros per(1000)

. predict h_age3, hazard at(agegrp3 1) zeros per(1000)

. predict h_age4, hazard at(agegrp4 1) zeros per(1000)

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



(c) Now fit a model with time-dependent effects for age group.

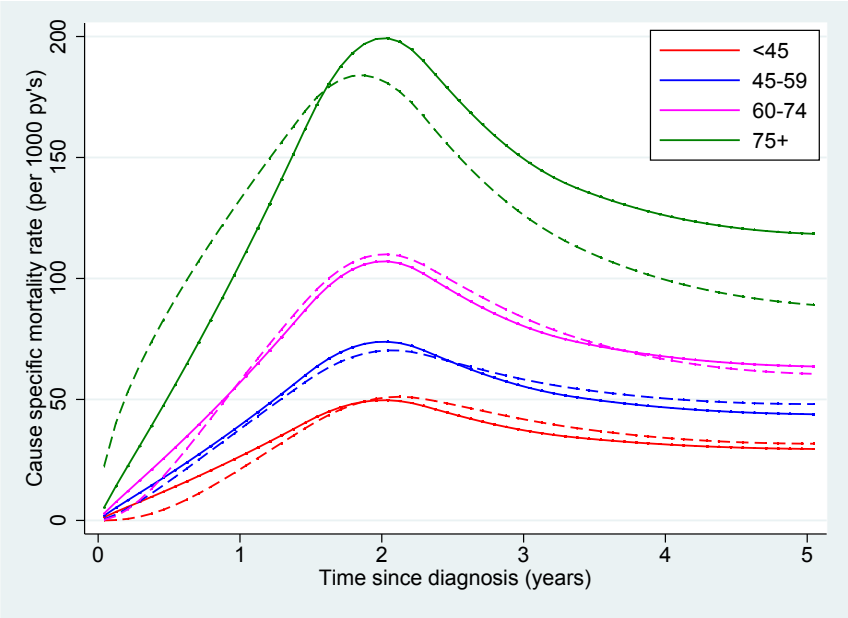
```
. stpm2 female year8594 agegrp2-agegrp4, df(4) scale(hazard) ///
>         tvc(agegrp2 agegrp3 agegrp4) dftvc(2)

Iteration 0:   log likelihood = -2515.8286
Iteration 1:   log likelihood = -2499.4895
Iteration 2:   log likelihood = -2498.5514
Iteration 3:   log likelihood = -2498.5494
Iteration 4:   log likelihood = -2498.5494

Log likelihood = -2498.5494                     Number of obs     =      5,318

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      female |  -.5803191   .0744504    -7.79   0.000    -.7262392    -.434399
    year8594 |  -.3577455   .0738423    -4.84   0.000    -.5024737   -.2130172
     agegrp2 |   .4584775   .1231253     3.72   0.000     .2171563    .6997986
     agegrp3 |   .8298068   .1176129     7.06   0.000     .5992898    1.060324
     agegrp4 |   1.499992   .1261885    11.89   0.000     1.252667    1.747317
       _rcs1 |   1.101495    .125085     8.81   0.000     .8563334    1.346658
       _rcs2 |   .2978602   .1086354     2.74   0.006     .0849387    .5107817
       _rcs3 |   .0714558   .0173555     4.12   0.000     .0374397     .105472
       _rcs4 |  -.0021103   .0066186    -0.32   0.750    -.0150826     .010862
_rcs_agegrp21 |  -.1883751   .1437494    -1.31   0.190    -.4701187    .0933686
_rcs_agegrp22 |  -.1341995   .1179674    -1.14   0.255    -.3654114    .0970124
_rcs_agegrp31 |  -.1597332   .1397683    -1.14   0.253     -.433674    .1142077
_rcs_agegrp32 |  -.0688189   .1150518    -0.60   0.550    -.2943163    .1566785
_rcs_agegrp41 |  -.4332123   .1341468    -3.23   0.001    -.6961352   -.1702894
_rcs_agegrp42 |   -.201846   .1116387    -1.81   0.071    -.4206539    .0169619
        _cons |  -2.341008   .1087981   -21.52   0.000    -2.554249   -2.127768
------------------------------------------------------------------------------

. estimates store nonph
```

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

```
. lrtest ph nonph

Likelihood-ratio test                                 LR chi2(6)  =     20.10
(Assumption: ph nested in nonph)                      Prob > chi2 =    0.0027
```

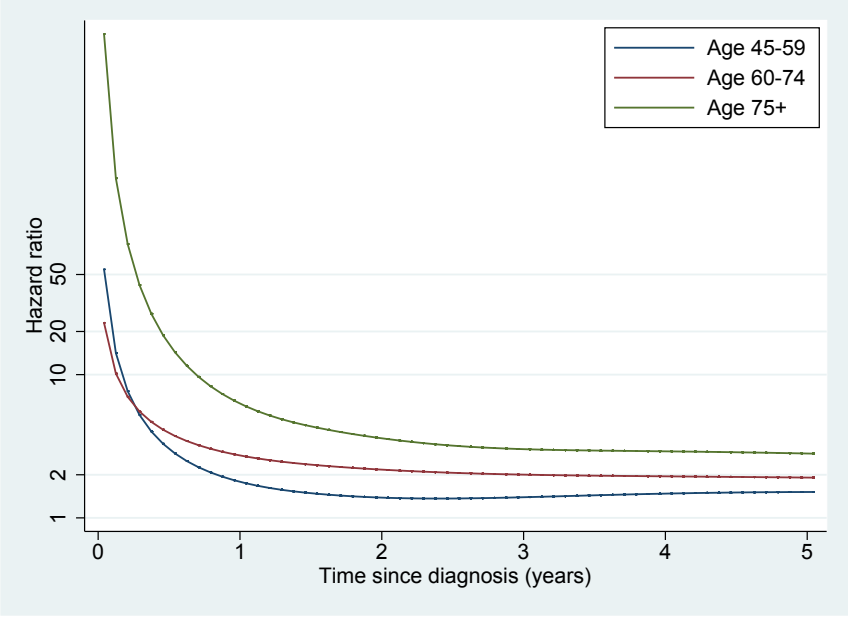
(d) Now predict the hazard function for each age group.

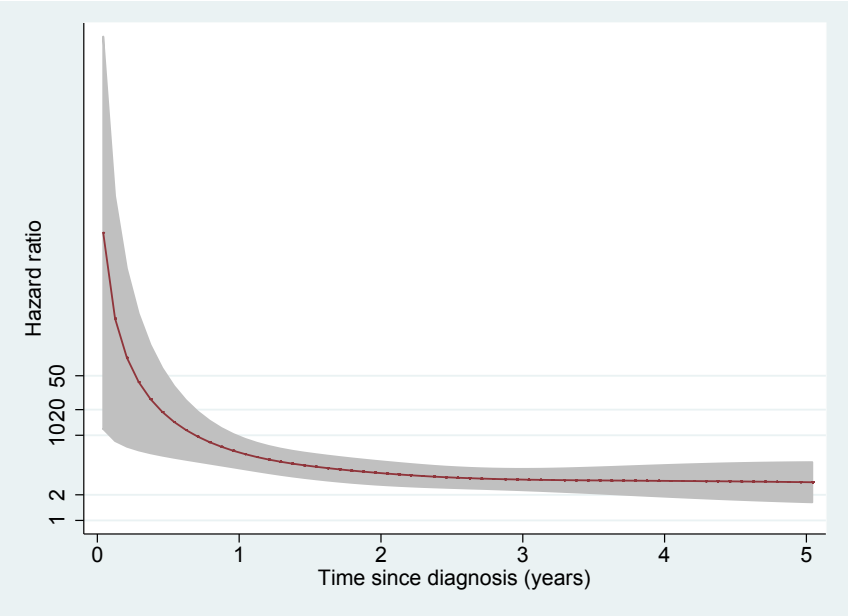
```
. predict h_age1_tvc, hazard zeros per(1000)
. predict h_age2_tvc, hazard at(agegrp2 1) zeros per(1000)
. predict h_age3_tvc, hazard at(agegrp3 1) zeros per(1000)
. predict h_age4_tvc, hazard at(agegrp4 1) zeros per(1000)

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



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

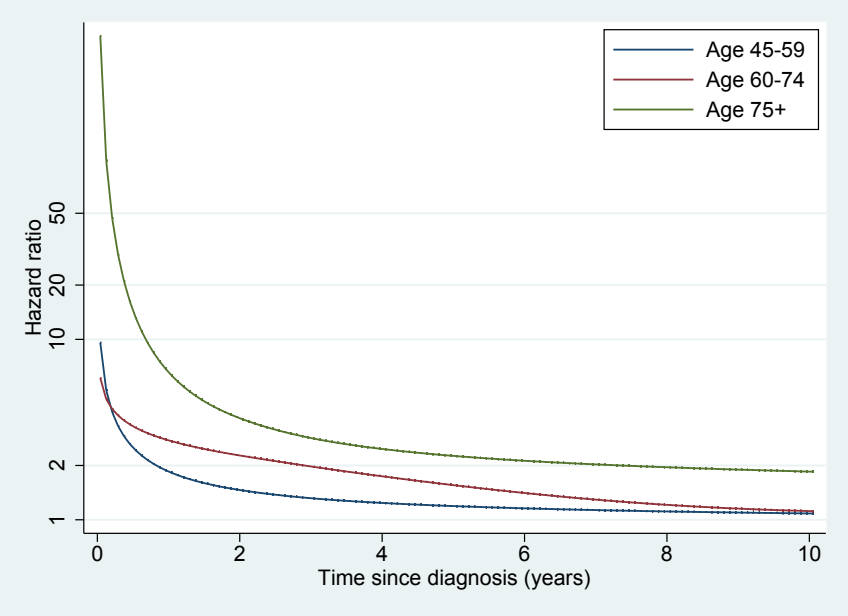
```
. predict hr2, hrnumerator(agegrp2 1) ci
. predict hr3, hrnumerator(agegrp3 1) ci
. predict hr4, hrnumerator(agegrp4 1) ci
```

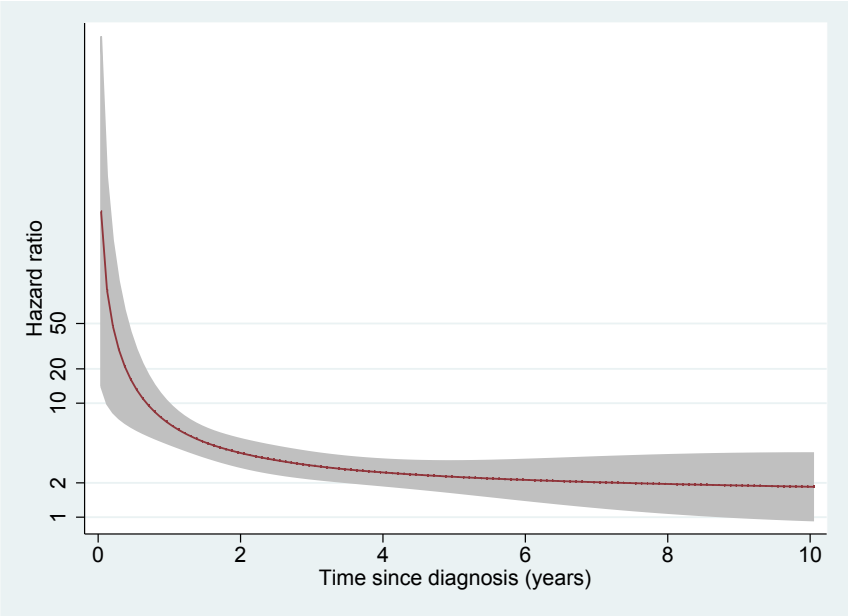




Note that by default the `hrdenominator` option sets all covariates to zero. As we only have one covariate with a time-dependent effect we can leave this unspecified.

Plot these hazard ratios versus follow-up time on the same graph. Also plot the hazard ratio for the oldest group with a 95% confidence interval.

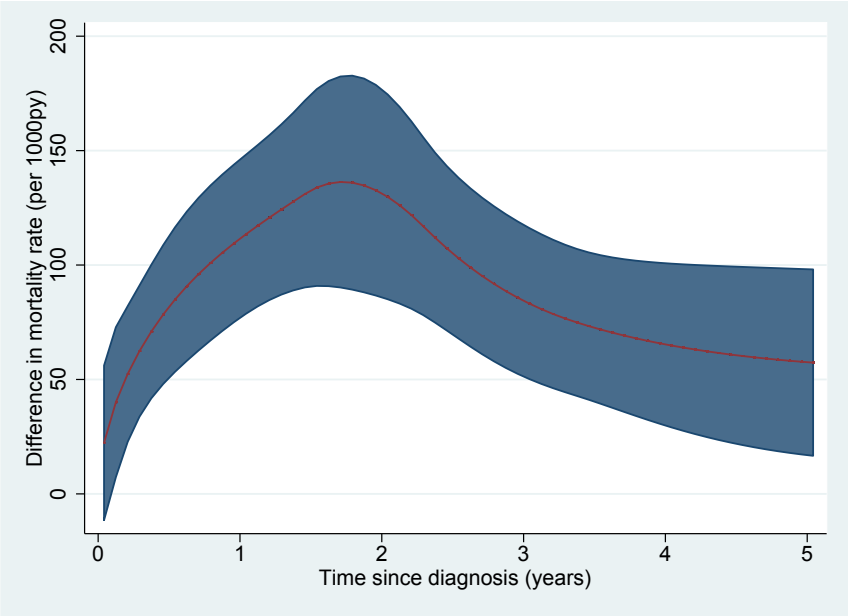




The hazard ratio is so high earlier on as there are very few early deaths in the youngest group. The means that the denominator of the hazard ratio is very small.

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

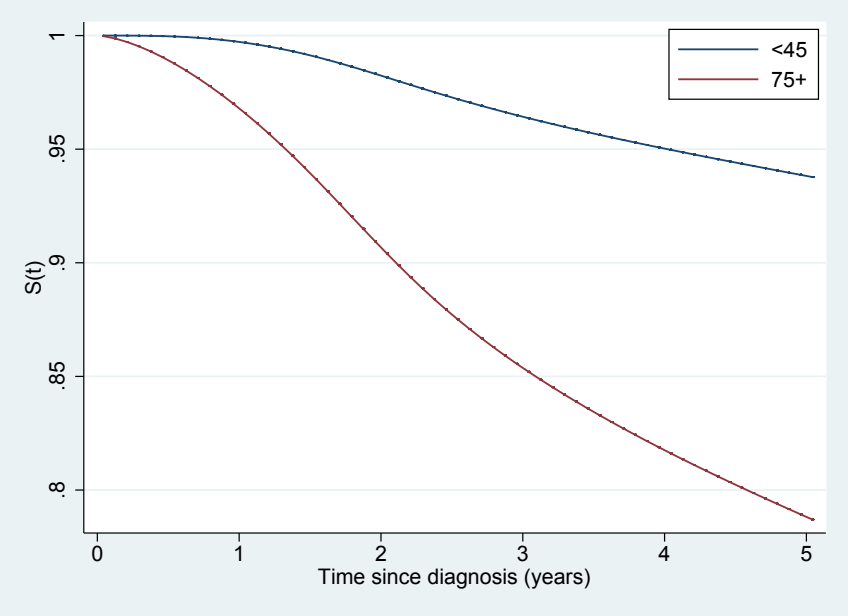
```
predict hdiff4, hdiff1(agegrp4 1) ci per(1000)
twoway (rarea hdiff4_lci hdiff4_uci _t, sort) ///
(line hdiff4 _t, sort) ///
,legend(off)  ///
xtitle("Time since diagnosis (years)") ///
ytitle("Difference in mortality rate (per 1000py)") ///
name(hdiff, replace)
```



The hazard difference is small early on in the time scale as each hazard rate is fairly low. Thus the large hazard ratio applied when the underlying rate is very low.

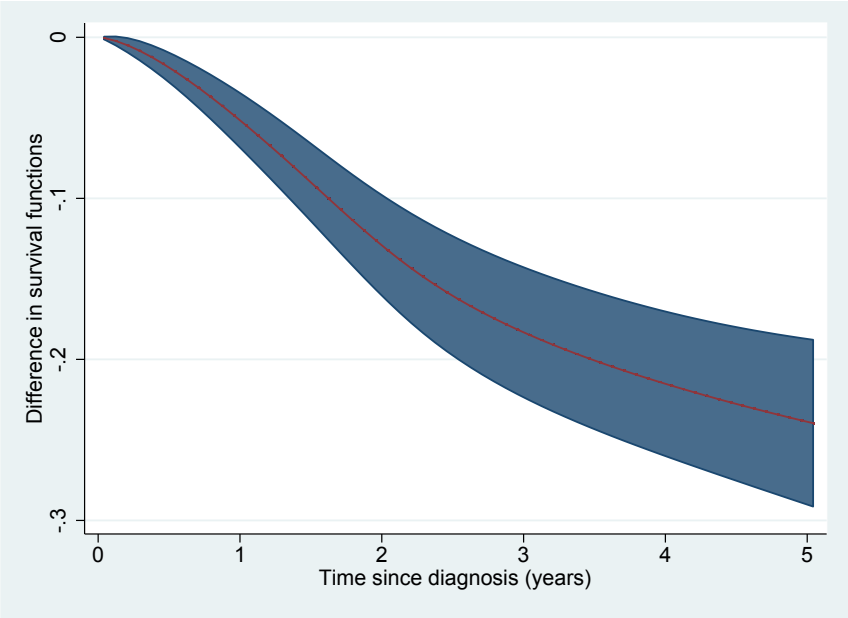
(g) Predict and plot the survival function for the youngest and oldest age groups for females diagnosed in 1985-1994.

```
predict s1, surv at(female 1 year8594 1) zeros
predict s2, surv at(agegrp4 1 female 1 year8594 1) zeros
twoway line s1 s2 _t, sort ///
xtitle("Time since diagnosis (years)") ///
ytitle("S(t)") ///
legend(order(1 "<45" 2 "75+") ring(0) pos(1) cols(1)) ///
name(surv_old_young, replace)
```



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

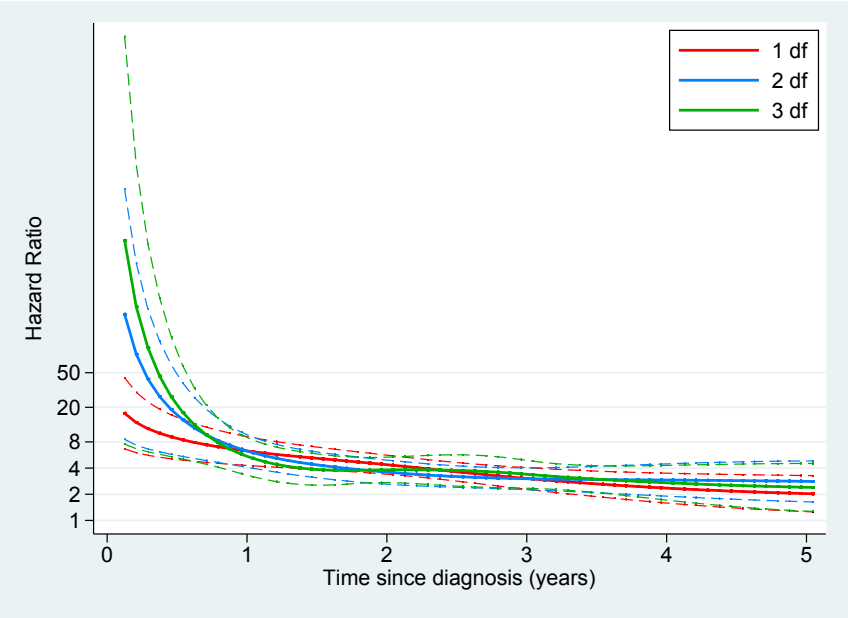
```
predict sdiff4, sdiff1(agegrp4 1 sex 2 year8594 1) ///
sdiff2(agegrp4 0 sex 2 year8594 1) ci
twoway (rarea sdiff4_lci sdiff4_uci _t, sort) ///
(line sdiff4 _t, sort) ///
,legend(off)  ///
xtitle("Time since diagnosis (years)") ///
ytitle("Difference in survival functions") ///
name(sdiff, replace)
```



(h) Fit models with 1, 2 and 3 df for the time-dependent effect of age. Use the AIC and BIC to compare models.

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

. count if _d==1
  747

. estimates stats dftvc*, n(‘r(N)’)

Akaike’s information criterion and Bayesian information criterion

------------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+----------------------------------------------------------------
      dftvc1 |        747         .  -2501.374      13    5028.747   5088.756
      dftvc2 |        747         .  -2498.549      16    5029.099   5102.956
      dftvc3 |        747         .  -2497.961      19    5033.922   5121.627
------------------------------------------------------------------------------
               Note: N=747 used in calculating BIC.

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

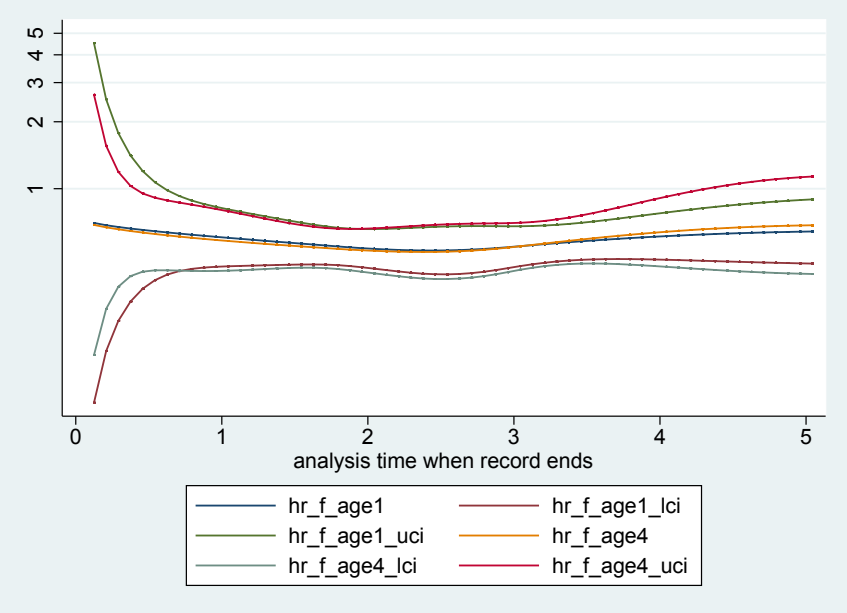
AIC selects 2df for the baseline and BIC selects 1 df (i.e. log(time))



(i) Now let effect of sex be time-dependent.

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

Iteration 0:   log likelihood = -2526.7407
Iteration 1:   log likelihood =  -2510.456
Iteration 2:   log likelihood = -2509.2177
Iteration 3:   log likelihood = -2509.2123
Iteration 4:   log likelihood = -2509.2123

Log likelihood = -2509.2123                     Number of obs     =      5,318

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      female |  -.5513793   .0766374    -7.19   0.000    -.7015859   -.4011727
    agegrp2 |   .4523901   .1238906     3.65   0.000      .209569    .6952111
    agegrp3 |   .8233369   .1184618     6.95   0.000     .5911561    1.055518
    agegrp4 |   1.455916   .1266905    11.49   0.000     1.207607    1.704225
      _rcs1 |   1.187959   .1564131     7.60   0.000     .8813948    1.494523
      _rcs2 |   .4407121   .1843816     2.39   0.017     .0793308    .8020935
      _rcs3 |   .0382407   .0408244     0.94   0.349    -.0417737     .118255
      _rcs4 |  -.0071244    .009576    -0.74   0.457     -.025893    .0116441
_rcs_agegrp21 |  -.2629583    .170383    -1.54   0.123    -.5969029    .0709862
_rcs_agegrp22 |  -.2735475    .193834    -1.41   0.158    -.6534551    .1063602
_rcs_agegrp23 |   .0376251   .0477674     0.79   0.431    -.0559973    .1312475
_rcs_agegrp31 |  -.2247332   .1675543    -1.34   0.180    -.5531335    .1036672
_rcs_agegrp32 |  -.1892325   .1915556    -0.99   0.323    -.5646746    .1862096
_rcs_agegrp33 |   .0338753   .0460845     0.74   0.462    -.0564487    .1241993
_rcs_agegrp41 |  -.5026386   .1635986    -3.07   0.002     -.823286   -.1819913
_rcs_agegrp42 |  -.3391512   .1870168    -1.81   0.070    -.7056973    .0273949
_rcs_agegrp43 |   .0467822   .0469483     1.00   0.319    -.0452347    .1387991
 _rcs_female1 |  -.0198806   .0654797    -0.30   0.761    -.1482185    .1084573
 _rcs_female2 |  -.0150768   .0651503    -0.23   0.817     -.142769    .1126154
 _rcs_female3 |  -.0171383   .0250381    -0.68   0.494     -.066212    .0319354
        _cons |   -2.53778   .1045078   -24.28   0.000    -2.742611   -2.332948
------------------------------------------------------------------------------
```

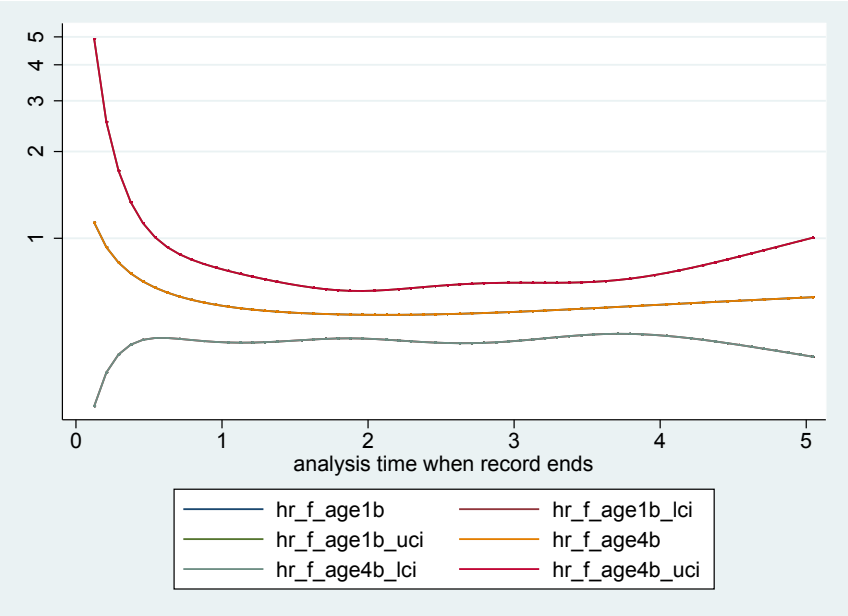
```
. predict hr_f_age1, hrnum(female 1) ci

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



(j) Use `strcs` command to fir model on the log hazard scale rather than the log cumulative hazard scale.

```
. strcs female  agegrp2-agegrp4, df(4) ///
>         tvc(agegrp2 agegrp3 agegrp4 female) dftvc(3) nodes(50)

Iteration 0:   log likelihood = -2509.3785  (not concave)
Iteration 1:   log likelihood = -2509.3785  (backed up)
Iteration 2:   log likelihood = -2508.7846
Iteration 3:   log likelihood = -2508.7785
Iteration 4:   log likelihood = -2508.7785

Log likelihood = -2508.7785                     Number of obs     =      5,318

------------------------------------------------------------------------------
             | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      female |    .602414   .0933718    -3.27   0.001     .4445925    .8162589
     agegrp2 |   1.363656   .2858438     1.48   0.139     .9042314    2.056506
     agegrp3 |   1.730756   .3613235     2.63   0.009     1.149566    2.605782
     agegrp4 |   2.406743   .6610432     3.20   0.001      1.40487    4.123094
-------------+----------------------------------------------------------------
rcs          |
        __s1 |   .3424356   .1674138     2.05   0.041     .0143105    .6705607
        __s2 |   .5330863   .1830909     2.91   0.004     .1742347     .891938
        __s3 |  -.0828143   .1081724    -0.77   0.444    -.2948282    .1291997
        __s4 |  -.0712097   .0383507    -1.86   0.063    -.1463757    .0039564
__s_agegrp21 |   -.229028   .1838626    -1.25   0.213     -.589392    .1313359
__s_agegrp22 |  -.2509745   .1888844    -1.33   0.184    -.6211812    .1192322
__s_agegrp23 |   .1323658   .1267311     1.04   0.296    -.1160226    .3807542
__s_agegrp31 |  -.3086703   .1827007    -1.69   0.091    -.6667571    .0494165
__s_agegrp32 |  -.1487228    .186751    -0.80   0.426     -.514748    .2173024
__s_agegrp33 |   .1533382   .1239147     1.24   0.216    -.0895301    .3962066
```

```
__s_agegrp41 |  -.5568512   .2105551    -2.64   0.008    -.9695315   -.1441708
__s_agegrp42 |   -.250374   .1921576    -1.30   0.193     -.626996    .126248
__s_agegrp43 |   .2092821   .1356667     1.54   0.123    -.0566198    .475184
 __s_female1 |   .0128572    .100218     0.13   0.898    -.1835665   .2092808
 __s_female2 |  -.0688224   .0715353    -0.96   0.336    -.2090289   .0713842
 __s_female3 |   -.010989   .0685836    -0.16   0.873    -.1454105   .1234324
       _cons |   -3.50583   .1814834   -19.32   0.000    -3.861531   -3.150129
------------------------------------------------------------------------------
 Quadrature method: Gauss-Legendre with 50 nodes

. predict hr_f_age1b, hrnum(female 1) ci

. predict hr_f_age4b, hrnum(female 1 agegrp4 1) hrdenom(agegrp4 1) ci

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



133. **Modelling on other scales (proportional odds and Aranda-Ordaz link function) non-linear effects using** `stpm2`

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

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

. gen female = sex == 2

. stset surv_mm, failure(status=1) scale(12) exit(time 60.5)

     failure event:  status == 1
obs. time interval:  (0, surv_mm]
 exit on or before:  time 60.5
    t for analysis:  time/12

------------------------------------------------------------------------------
     7775  total observations
        0  exclusions
------------------------------------------------------------------------------
     7775  observations remaining, representing
     1580  failures in single-record/single-failure data
 29159.46  total analysis time at risk and under observation
                                             at risk from t =         0
                                  earliest observed entry t =         0
                                       last observed exit t =  5.041667
```

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

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

Log likelihood = -5368.5831                     Number of obs   =       7775
------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      female |   .5605415   .0288516   -11.25   0.000     .5067522    .6200403
             |
      agegrp |
      45-59  |    1.39757   .1082142     4.32   0.000     1.200784    1.626606
      60-74  |   1.991024   .1466182     9.35   0.000     1.723433    2.300163
        75+  |   3.208854   .2612294    14.32   0.000     2.735612    3.763963
             |
    year8594 |   .7103591    .036085    -6.73   0.000     .6430406    .7847249
       _rcs1 |   2.154641    .040703    40.63   0.000     2.076323    2.235913
       _rcs2 |   1.075653   .0158898     4.94   0.000     1.044956    1.107252
       _rcs3 |   1.052009    .008968     5.95   0.000     1.034578    1.069734
       _rcs4 |   1.009169   .0048186     1.91   0.056     .9997686    1.018658
       _cons |   .1642373     .01134   -26.16   0.000     .1434497    .1880374
------------------------------------------------------------------------------

. forvalues i = 0/3 {
  2.         predict s_age`i'_ph, surv at(agegrp `i') zeros
  3.         predict h_age`i'_ph, hazard at(agegrp `i') zeros
  4. }

.         estimates store ph
```

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

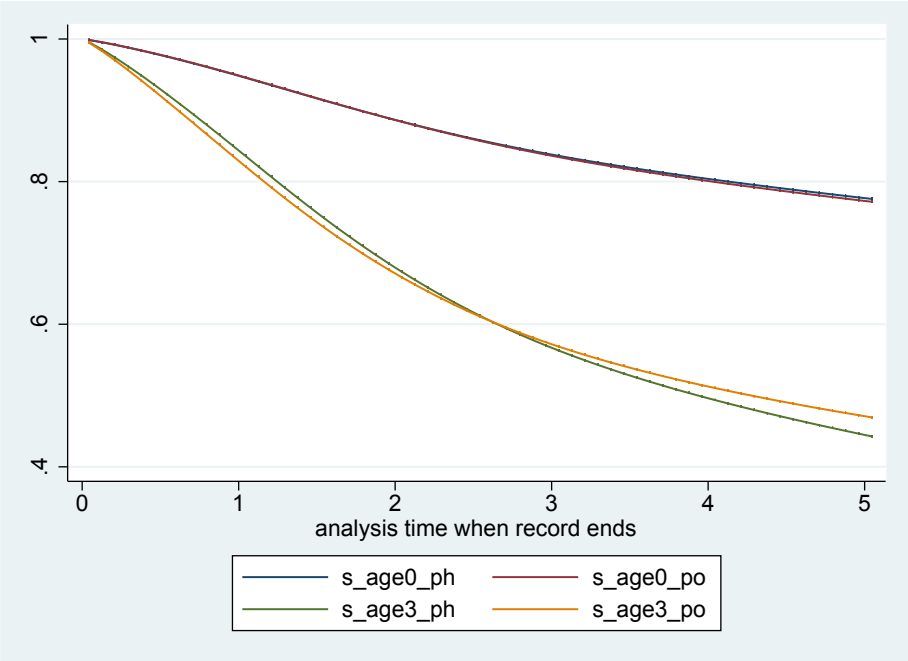
```
. stpm2 female i.agegrp year8594, scale(odds) df(4) eform

Log likelihood = -5366.4798                     Number of obs   =       7775
------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      female |   .5142867   .0300324   -11.39   0.000     .4586678      .57665
             |
      agegrp |
      45-59  |    1.44996   .1231269     4.38   0.000     1.227648    1.712528
      60-74  |   2.166539   .1768975     9.47   0.000     1.846146    2.542535
        75+  |   3.822862   .3560074    14.40   0.000     3.185075     4.58836
             |
    year8594 |   .6764815   .0392622    -6.73   0.000     .6037444    .7579817
       _rcs1 |   2.265334   .0441201    41.99   0.000      2.18049    2.353479
       _rcs2 |   1.057624   .0159041     3.73   0.000     1.026908     1.08926
       _rcs3 |    1.04979   .0093204     5.47   0.000      1.03168    1.068217
       _rcs4 |   1.010293   .0052687     1.96   0.050     1.000019    1.020672
       _cons |   .1838976   .0139394   -22.34   0.000     .1585094    .2133522
------------------------------------------------------------------------------
```

```
. forvalues i = 0/3 {
  2.         predict s_age‘i’_po, surv at(agegrp ‘i’) zeros
  3.         predict h_age‘i’_po, hazard at(agegrp ‘i’) zeros
  4. }

. estimates store po
```

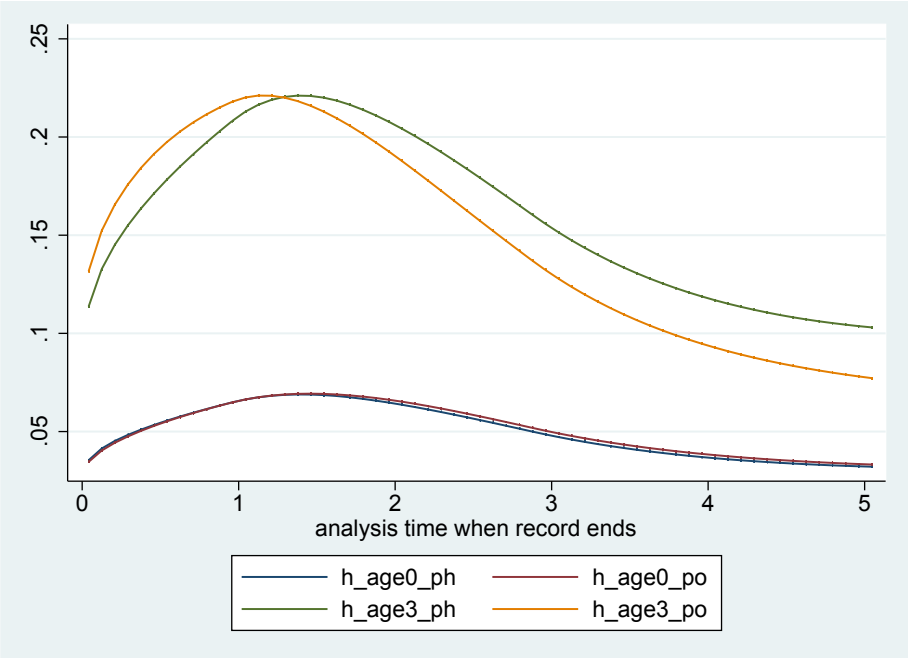
At each point in time the odds of an event for females are 0.51 that of males.

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

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




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

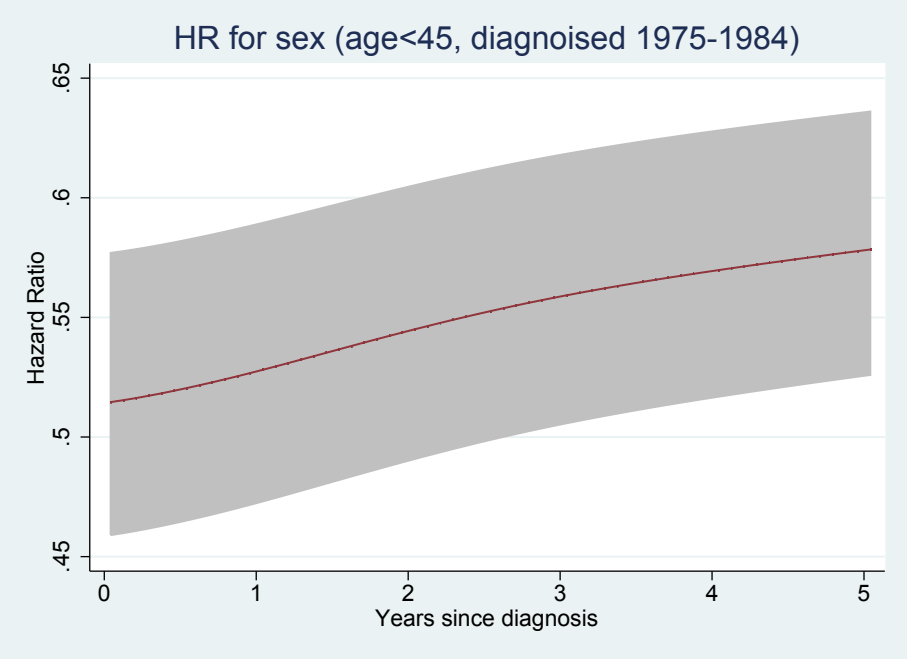


```
. count if _d == 1
 1580

. estimates stats ph po, n(‘r(N)’)

Akaike’s information criterion and Bayesian information criterion
------------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC          BIC
-------------+----------------------------------------------------------------
          ph |   1580           .   -5368.583     10     10757.17     10810.82
          po |   1580           .    -5366.48     10     10752.96     10806.61
------------------------------------------------------------------------------
```

According the both the AIC and BIC the proportional odds model gives the better fit.

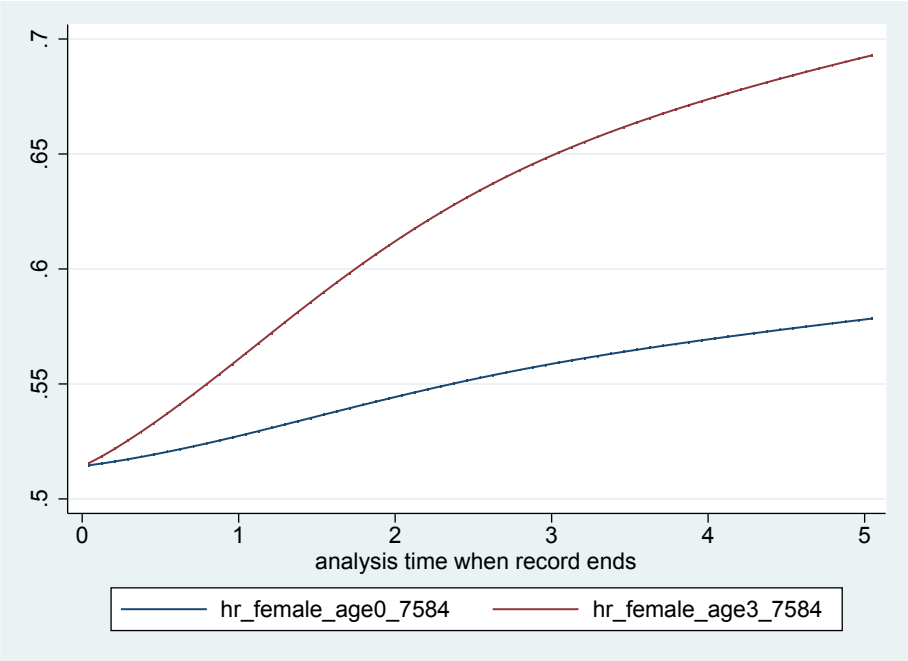
(d) For the proportional odds model the hazards will not be proportional. Predict and plot the hazard ratio for females in the youngest age group diagnosed in 1975-1984.

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



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

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



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

```
. stpm2 female i.agegrp year8594, scale(theta) df(4)

Iteration 0:   log likelihood = -5375.6278
Iteration 1:   log likelihood = -5366.5965
Iteration 2:   log likelihood = -5366.5186
Iteration 3:   log likelihood = -5366.4516
Iteration 4:   log likelihood = -5366.4491
Iteration 5:   log likelihood = -5366.4491

Log likelihood = -5366.4491                     Number of obs   =       7775

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      female |  -.6766366   .0757127    -8.94   0.000    -.8250307   -.5282425
             |
      agegrp |
      45-59  |   .3769008   .0887246     4.25   0.000     .2030038    .5507978
      60-74  |   .7849535   .0956142     8.21   0.000     .5975531    .9723539
        75+  |   1.364776   .1347484    10.13   0.000     1.100674    1.628878
             |
    year8594 |  -.3971262   .0641625    -6.19   0.000    -.5228824   -.2713699
       _rcs1 |    .825106   .0358258    23.03   0.000     .7548887    .8953233
       _rcs2 |   .0535342   .0181234     2.95   0.003     .0180129    .0890555
       _rcs3 |   .0482685    .009019     5.35   0.000     .0305915    .0659455
       _rcs4 |   .0104036    .005317     1.96   0.050    -.0000176    .0208248
       _cons |  -1.677235   .1009246   -16.62   0.000    -1.875044   -1.479426
-------------+----------------------------------------------------------------
ln_theta     |
       _cons |   .1348622   .5124562     0.26   0.792    -.8695336    1.139258
------------------------------------------------------------------------------

. estimates store ao
```

```
. count if _d == 1
 1580

. estimates stats ph po ao, n(‘r(N)’)

Akaike’s information criterion and Bayesian information criterion
------------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+----------------------------------------------------------------
          ph |   1580           .   -5368.583     10     10757.17    10810.82
          po |   1580           .    -5366.48     10     10752.96    10806.61
          ao |   1580           .   -5366.449     11      10754.9    10813.92
------------------------------------------------------------------------------
               Note:  N=1580 used in calculating BIC
```

The proportional odds model still gives the better fit.

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

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

 ( 1)  [ln_theta]_cons = 0

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.144379   .5864442     0.26   0.792      .419147    3.124449
------------------------------------------------------------------------------
```

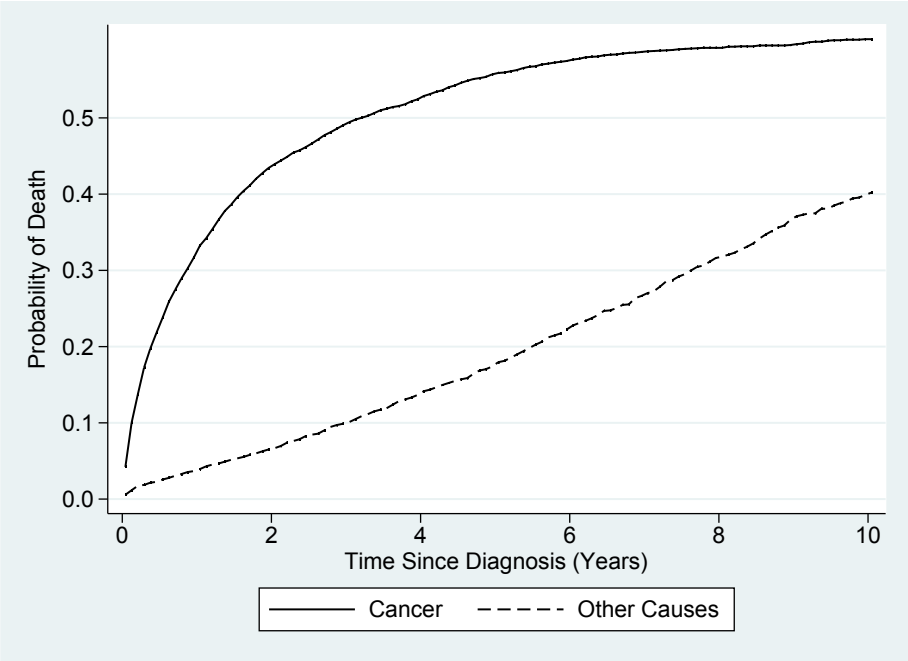
The estimated value of θ is close to one, which would equate to a proportional odds model.

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

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

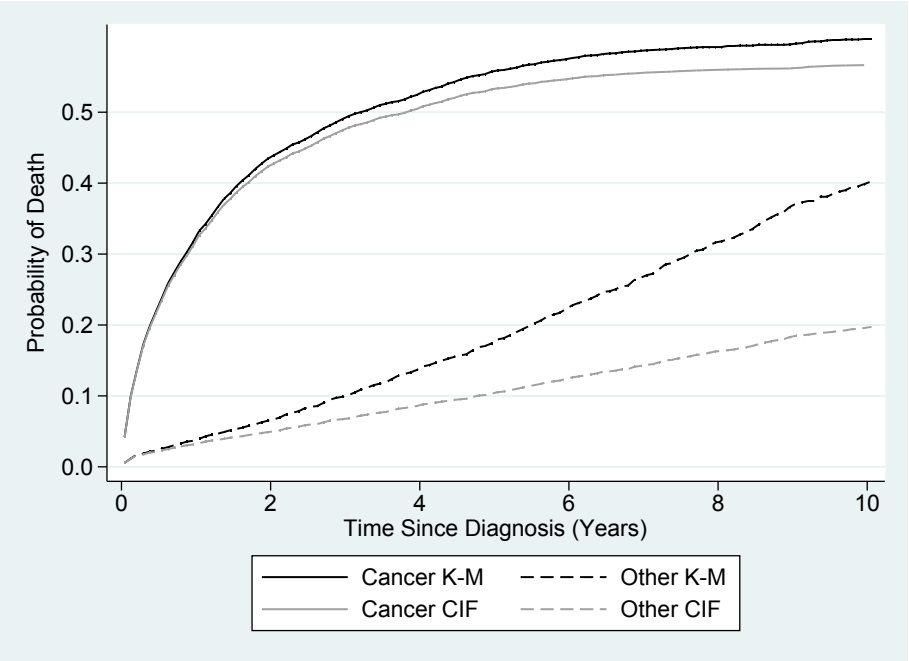
```
use colon, clear
drop if stage ==0
gen female = sex==2
```

Plot the complement of the Kaplan-Meier estimate for males (i.e. 1 minus Kaplan-Meier survival estimate) for both cancer and other causes. Describe what you see.



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

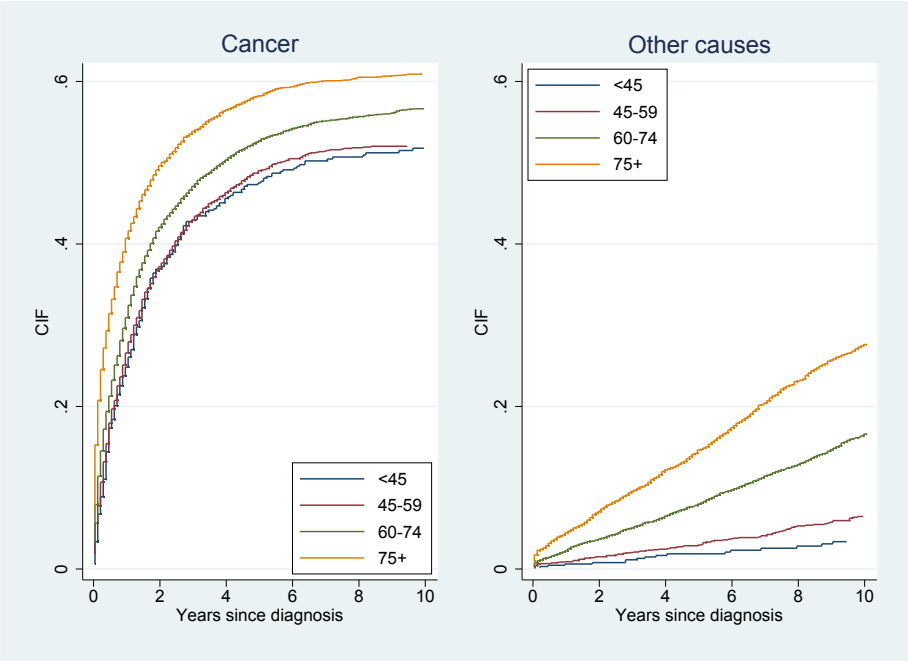
```
stset surv_mm, failure(status==1) scale(12) exit(time 120.5)
stcompet CIF_sex=ci, compet1(2) by(sex)
gen CIF_sex_cancer=CIF_sex if status==1
gen CIF_sex_other=CIF_sex if status==2
```



The cumulative incidence functions are lower than the cause-specific survival functions.

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

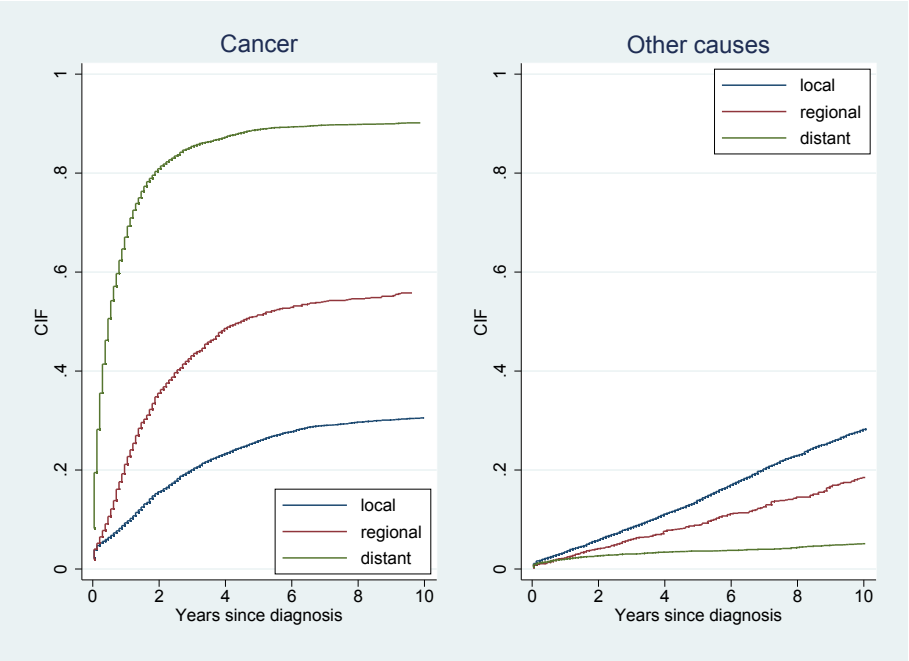
```
stset surv_mm, failure(status==1) scale(12) exit(time 120.5)
stcompet CIF_age=ci, compet1(2) by(agegrp)

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

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

graph combine CIF_age1 CIF_age2, nocopies ycommon
```



Being old increases the probability of both dying from cancer and from other causes. Younger people have a much lower probability of dying from other causes.

(d) Now obtain the CIF for cancer and other causes by stage group. Plot the results.

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

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

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

graph combine CIF_stage1 CIF_stage2, nocopies ycommon
```



Those diagonosed with regional and distant stage are more likely to die from their cancer and thus reducing their chance of dying from other causes.

(e) When fitting a Fine and Gray model the event of interest is indicated in the stset command and the competing events are indicated in the `stcrreg` command.

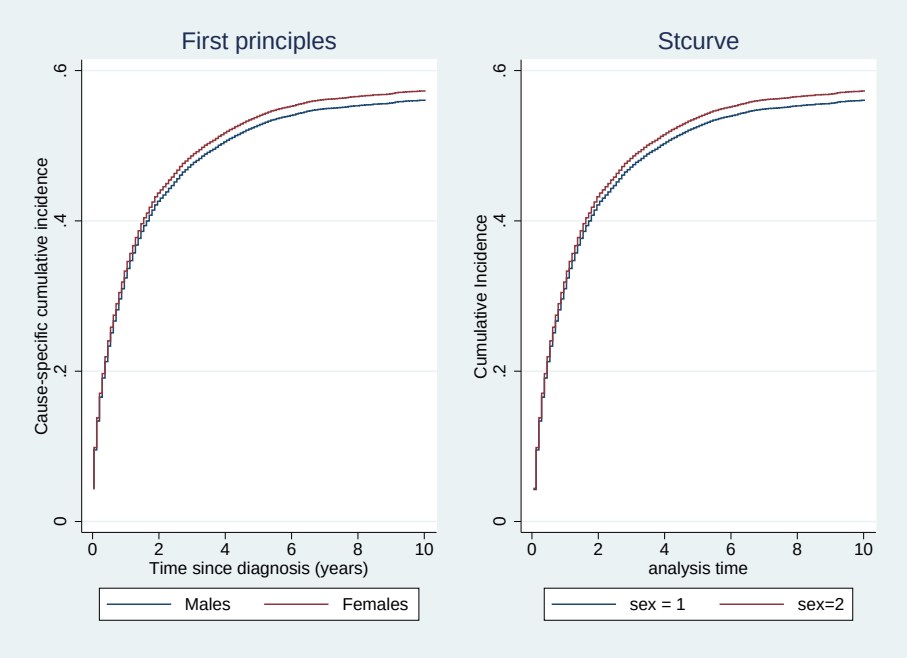
i.

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

         failure _d:  status == 1
   analysis time _t:  surv_mm/12
  exit on or before:  time 120.5

Competing-risks regression                    No. of obs        =      13208
                                              No. of subjects   =      13208
Failure event  : status == 1                  No. failed        =       7122
Competing event: status == 2                  No. competing     =       2062
                                              No. censored      =       4024

                                              Wald chi2(1)      =       2.06
Log pseudolikelihood = -64858.508             Prob > chi2       =     0.1515

------------------------------------------------------------------------------
             |               Robust
          _t |        SHR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |   1.034678   .0245912     1.43   0.151     .9875856    1.084016
------------------------------------------------------------------------------
```

The subhazard rate associated with cancer is 3% higher for females compared to males. However, this result is not statistically significant (p-value = .0151). This means that there is no evidence that the cause-specific CIFs for males and females that can be derived from this model are statistically different. The subhazard is conceptually different from the hazard that is estimated using cause-specific models (e.g. Cox regression or flexible parametric models). The difference is in how the risk set is defined. Individuals who are censored due to a competing event still contribute to the risk set for the event of main interest. This makes the interpretation of the subhazard per se complicated when the

competing events are absorbing (as is the case when the competing event is death due to some cause).

ii. Combined answer for ii.) and iii.)



The CIFs produced via calculation from first principles are identical to those produced by `stcurve` (as expected). We already know from the estimated regression parameter (SHR) that there is no evidence of a difference between the two CIFs. We can also verify this result using the Pepi-Mori test.

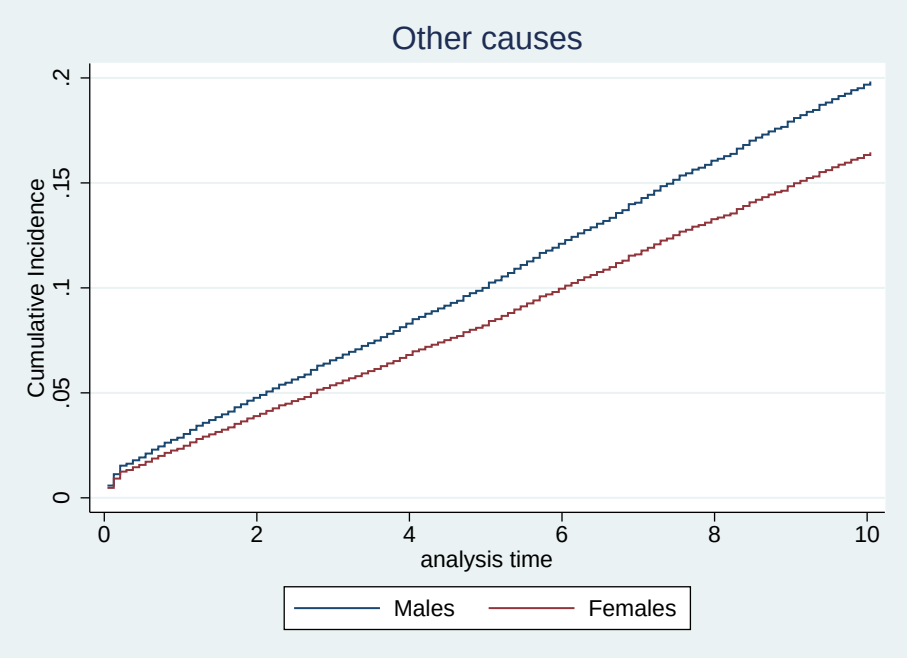
```
. stpepemori sex, compet(2)

Pepe and Mori test comparing the cumulative incidence of two groups of sex

     Main event failure:  status == 1
Chi2(1) = 1.8196  -  p =  0.17736

Competing event failure:  status == 2
Chi2(1) = 20.942  -  p <  0.00001
```

The test results shows that there is no evidence of statistical difference of the cancer-specific CIFs for males and females (p=0.17736). The p-value is not identical to that observed in the regression output. The reason is that a different test statistic is used. The test statistic for the Pepe-Mori test is based on cumulative weighted differences for the CIFs (with more weight given at the start of follow-up). Note that we also get a test for differences between the CIFs for males and females that are associated with the probability of death due to other causes.

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

         failure _d:  status == 2
   analysis time _t:  surv_mm/12
  exit on or before:  time 120.5

Competing-risks regression                    No. of obs       =     13208
                                              No. of subjects  =     13208
Failure event  : status == 2                  No. failed       =      1752
```

```
Competing event: status == 1                    No. competing    =      7186
                                                No. censored     =      4270

                                                Wald chi2(1)     =     18.79
Log pseudolikelihood = -16008.013               Prob > chi2      =    0.0000

------------------------------------------------------------------------------
             |               Robust
          _t |        SHR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |   .8134855    .038738    -4.33   0.000     .7409958    .8930666
------------------------------------------------------------------------------
```

The subhazard rate associated with death from other causes than cancer is 19% lower for females compared to males. This difference is statistically significant (p-value = 0.000). In other words, the CIFs associated with the estimated subhazards are significantly different (with women having a lower probability of dying from other causes during follow-up).



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

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

```
. expand 2
(13208 observations created)
. bysort id: gen cause=_n // cause =1 for cause 1, cause =2 for cause 2
. gen cancer=(cause==1)  // cancer is a dummy for cause 1
. gen other=(cause==2)   // other is a dummy for cause 2
. gen event=(cause==status)  // status=1 death due to cancer, =2 death due to other
. list id status cause sex event in 1/8, sepby(id)

     +-----------------------------------------+
     | id         status   cause      sex   event |
     |-----------------------------------------|
  1. |  1   Dead: cancer       1   Female       1 |
  2. |  1   Dead: cancer       2   Female       0 |
     |-----------------------------------------|
  3. |  2    Dead: other       1   Female       0 |
  4. |  2    Dead: other       2   Female       1 |
     |-----------------------------------------|
  5. |  3   Dead: cancer       1     Male       1 |
  6. |  3   Dead: cancer       2     Male       0 |
     |-----------------------------------------|
  7. |  4   Dead: cancer       1     Male       1 |
  8. |  4   Dead: cancer       2     Male       0 |
     +-----------------------------------------+
```

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

```
. stset surv_mm, failure(event) scale(12) exit(time 120.5)

     failure event:  event != 0 & event < .
obs. time interval:  (0, surv_mm]
 exit on or before:  time 120.5
    t for analysis:  time/12

--------------------------------------------------------------------------------
    26416  total observations
        0  exclusions
--------------------------------------------------------------------------------
    26416  observations remaining, representing
     8874  failures in single-record/single-failure data
 88021.33  total analysis time at risk and under observation
                                             at risk from t =         0
                                  earliest observed entry t =         0
                                       last observed exit t =  10.04167

. stpm2 cancer other female, scale(hazard) ///
>       rcsbaseoff dftvc(4) nocons tvc(cancer other) eform nolog

Log likelihood = -26104.379                     Number of obs   =      26416

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      cancer |   .3982399   .0072068   -50.88   0.000     .3843625    .4126185
       other |   .0586242   .0022495   -73.93   0.000     .0543771    .0632031
```

```
        female |   .9717029   .0209206    -1.33   0.182     .9315524    1.013584
  _rcs_cancer1 |   2.651499    .028922    89.40   0.000     2.595414    2.708795
  _rcs_cancer2 |   1.334031   .0112592    34.15   0.000     1.312144    1.356282
  _rcs_cancer3 |   .9966756    .004847    -0.68   0.494     .9872207    1.006221
  _rcs_cancer4 |   1.031213   .0028379    11.17   0.000     1.025666     1.03679
   _rcs_other1 |   4.195679   .1450027    41.49   0.000     3.920891    4.489726
   _rcs_other2 |    .877522   .0217591    -5.27   0.000     .8358946    .9212224
   _rcs_other3 |   .8458815   .0126606   -11.18   0.000     .8214275    .8710635
   _rcs_other4 |   .9983704   .0104806    -0.16   0.877     .9780387    1.019125
------------------------------------------------------------------------------
```

By including the two cause indicators (`cancer` and `other`) as both main effects and time-dependent effects (using `tvc` option) we have fitted a stratified model with two separate baselines, one for each cause. For this reason we have used the `rcsbaseoff` option together with the `nocons` option which excludes the baseline hazard from the model.
This model assume that the effect of being female on the cancer-specific and other cause-specific mortality rate is the same, i.e., being female reduces the mortality rate by 3%.

iii. . Now fit a model where the effect of sex is allowed to be different for cancer and other causes.

```
. gen fem_can = female*cancer

. gen fem_other = female*other

. stpm2 cancer other fem_can fem_other, scale(hazard) ///
>       rcsbaseoff dftvc(4) nocons tvc(cancer other) eform nolog

Log likelihood = -26094.579                     Number of obs   =      26416

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      cancer |   .3873045   .0075015   -48.97   0.000     .3728774    .4022897
       other |   .0652291   .0028992   -61.42   0.000     .0597872    .0711664
      fem_can |   1.018965    .024574     0.78   0.436     .9719214    1.068285
    fem_other |   .8029223   .0385147    -4.58   0.000     .7308747     .882072
  _rcs_cancer1 |   2.651391   .0289201    89.40   0.000      2.59531    2.708684
  _rcs_cancer2 |   1.334074   .0112591    34.15   0.000     1.312188    1.356325
  _rcs_cancer3 |   .9967069    .004847    -0.68   0.498     .9872521    1.006252
  _rcs_cancer4 |   1.031234   .0028375    11.18   0.000     1.025688    1.036811
   _rcs_other1 |   4.198961   .1451248    41.51   0.000     3.923942    4.493256
   _rcs_other2 |   .8769804   .0217488    -5.29   0.000     .8353729    .9206604
   _rcs_other3 |   .8455053   .0126566   -11.21   0.000     .8210591    .8706793
   _rcs_other4 |   .9980563   .0104807    -0.19   0.853     .9777245    1.018811
------------------------------------------------------------------------------
```

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

```
. test fem_can = fem_other

 ( 1)  [xb]fem_can - [xb]fem_other = 0

           chi2(  1) =   19.70
         Prob > chi2 =    0.0000
```

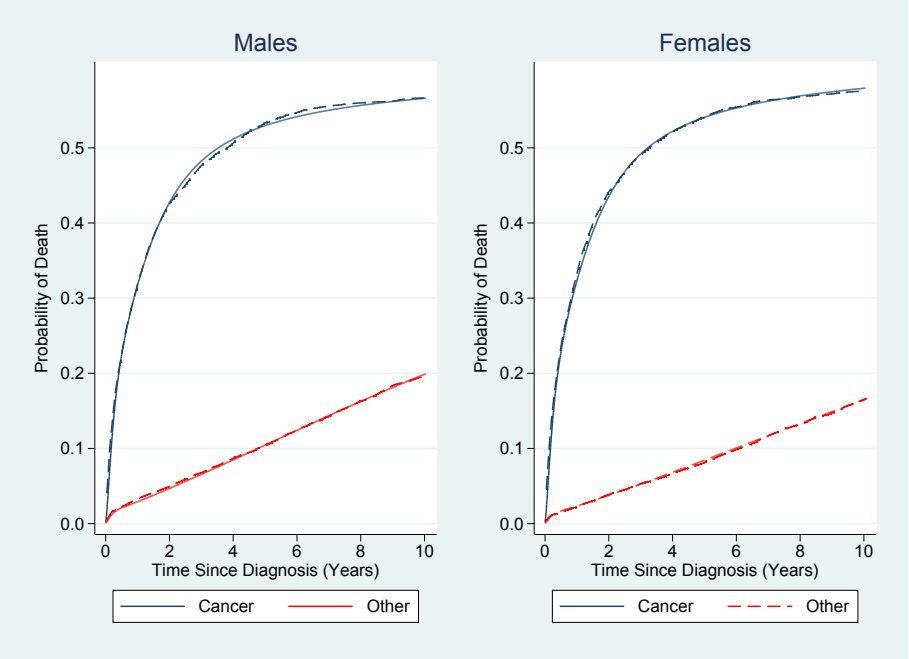
There is strong evidence that the effect of sex is different.

iv. The hazard ratios in the flexible parametric model are estimated using a riskset that reflects the actual number of people still at risk. That is, individuals who are censored due to a competing event no longer contributes to this risk set. Therefore, the hazards (and hazard ratios) estimated from the flexible parametric model can be interpreted as the hazard rate actually observed during follow up.

(h) Use the `stpm2cif` postestimation command to obtain the cumulative incidence functions for cancer and other causes for each sex. You will need to run this command twice - once for each sex. Do the results look the same as the empirical estimates (overlay these to make the comparison easier).

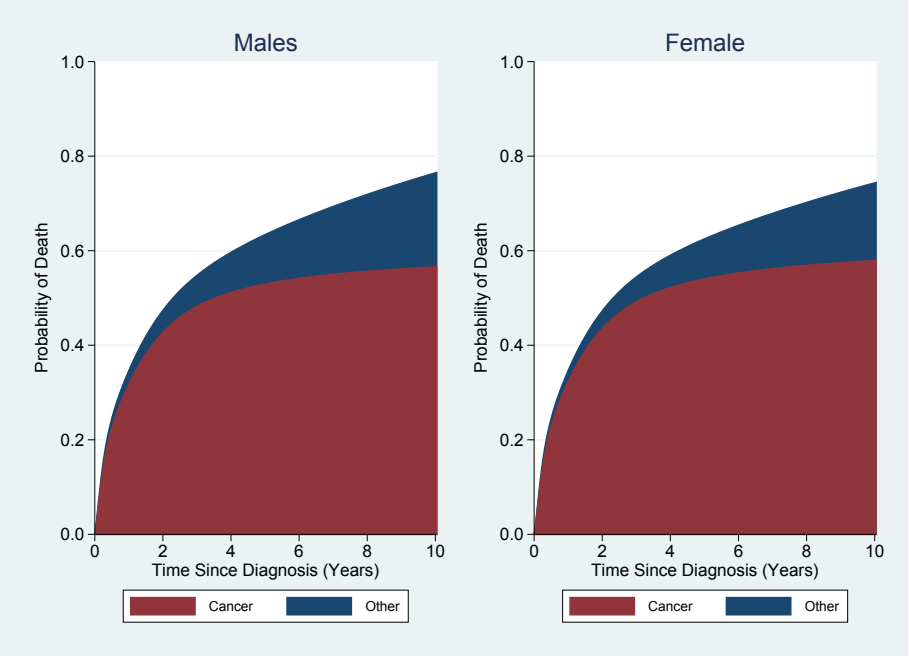
```
. stpm2cif cancermale othermale, cause1(cancer 1) ///
     cause2(other 1)
. stpm2cif cancerfemale otherfemale, cause1(cancer 1 fem_can 1) ///
     cause2(other 1 fem_other 1)
```

See the do file for the relevant Stata code.



Actually very good agreement here, but both models assume proportional hazards. If not a reasonable assumption then CIFs may differ.

(i) Try stacking the cumulative incidence functions for cancer and other causes (see do file for code to plot this).



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

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

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

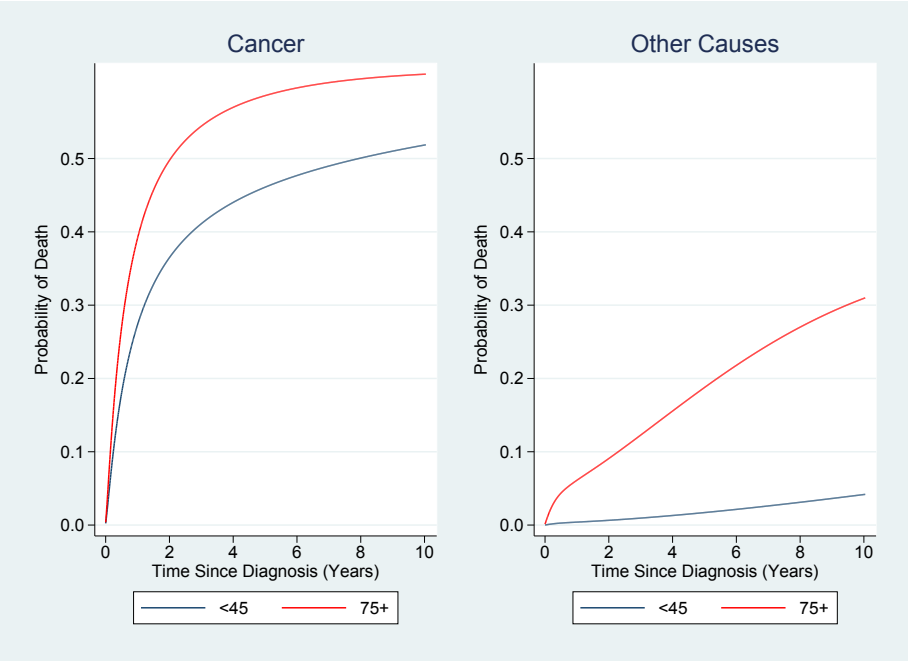
ii. Fit a flexible parametric model including sex and the interaction terms between age group and cause.

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

Log likelihood =  -25617.36                     Number of obs   =      26416

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      cancer |   .3042225   .0174918   -20.70   0.000     .2718004    .3405122
       other |   .0085571   .0020986   -19.41   0.000     .0052915    .0138382
     fem_can |   .9686434   .0236055    -1.31   0.191      .923465    1.016032
   fem_other |   .6501318   .0314869    -8.89   0.000     .5912571     .714869
     age1can |   1.039729   .0666383     0.61   0.543     .9169911    1.178896
     age2can |   1.235898    .072807     3.60   0.000     1.101129    1.387161
     age3can |   1.628803   .0968411     8.21   0.000      1.44964    1.830109
     age1oth |   2.063892   .5414761     2.76   0.006     1.234152    3.451481
     age2oth |   6.701714   1.645813     7.75   0.000     4.141399    10.84488
     age3oth |   17.13799   4.201882    11.59   0.000     10.59895    27.71131
_rcs_cancer1 |   2.656109    .028713    90.36   0.000     2.600424    2.712985
_rcs_cancer2 |   1.311476   .0101543    35.02   0.000     1.291725     1.33153
_rcs_cancer3 |   1.016337   .0042735     3.85   0.000     1.007996    1.024748
 _rcs_other1 |   4.469786   .1551555    43.14   0.000     4.175801    4.784468
 _rcs_other2 |   .8414316    .020001    -7.26   0.000     .8031294    .8815604
 _rcs_other3 |   .8541649    .012144   -11.09   0.000     .8306917    .8783014
------------------------------------------------------------------------------
```

All hazard ratios are cause-specific hazard ratios with separate effects for mortality due to cancer and due to other causes.

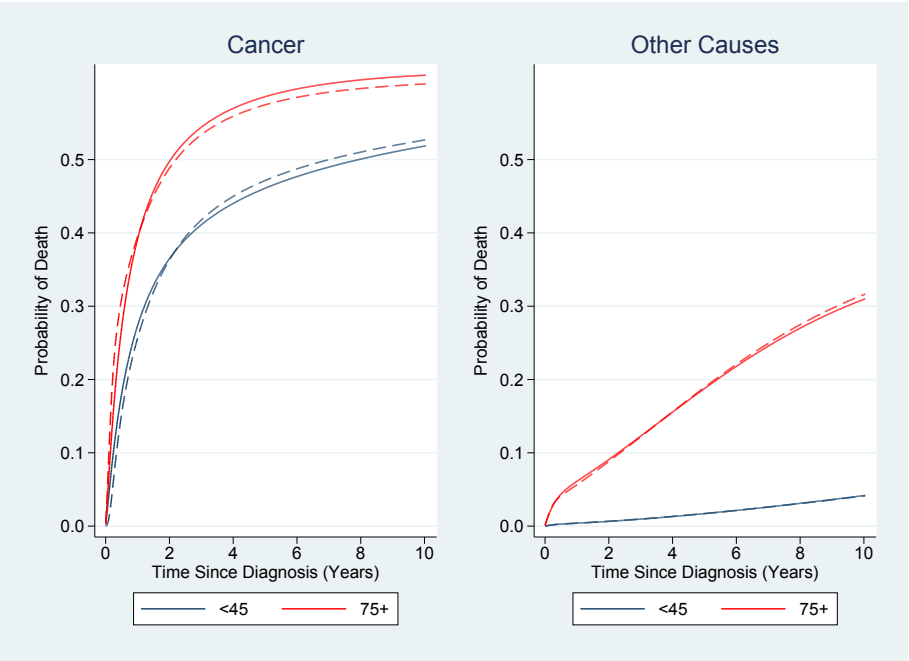
(k) Predict the cause-specific CIFs for males in the youngest and oldest age groups.

```
stpm2cif cancermale_age0 othermale_age0, cause1(cancer 1) ///
    cause2(other 1) mint(0.001)
stpm2cif cancermale_age3 othermale_age3, cause1(cancer 1 age3can 1) ///
    cause2(other 1 age3oth 1) mint(0.001)
```



(l) Now incorporate sex as a time-dependent effect in the model (use 3 df). Esimate the CIFs and compare to the model that assumes proportional hazards.

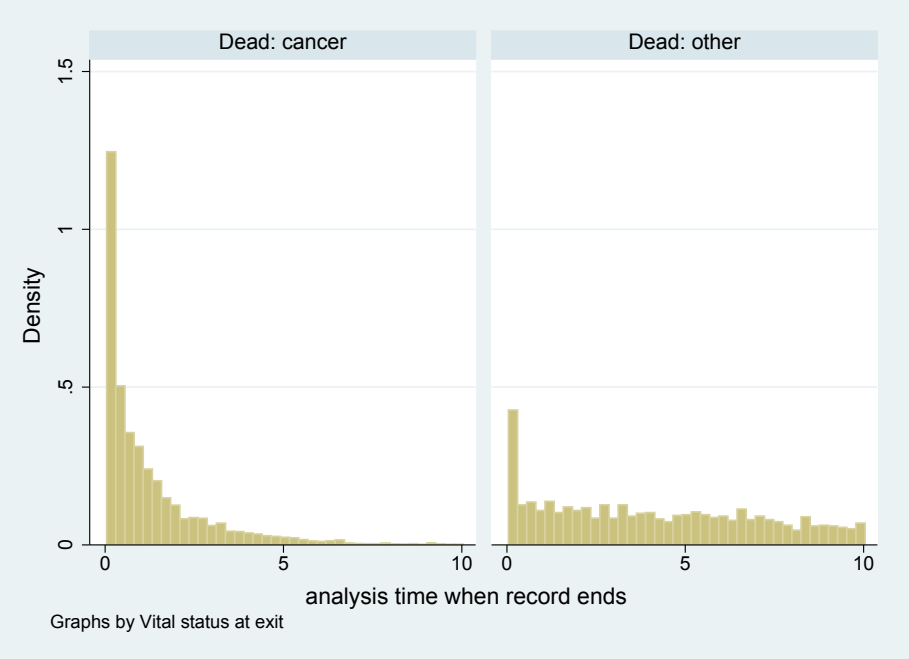
```
stpm2 cancer other fem_can fem_oth ///
    age1can age2can age3can age1oth age2oth age3oth , scale(hazard) ///
    rcsbaseoff dftvc(cancer:4 other:4 3) nocons ///
    tvc(cancer other fem_can age1can age2can age3can) eform nolog

stpm2cif cancermale_age0_tvc othermale_age0_tvc, cause1(cancer 1) ///
    cause2(other 1) mint(0.001)
stpm2cif cancermale_age3_tvc othermale_age3_tvc, cause1(cancer 1 age3can 1) ///
    cause2(other 1 age3oth 1) mint(0.001)
```



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

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

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



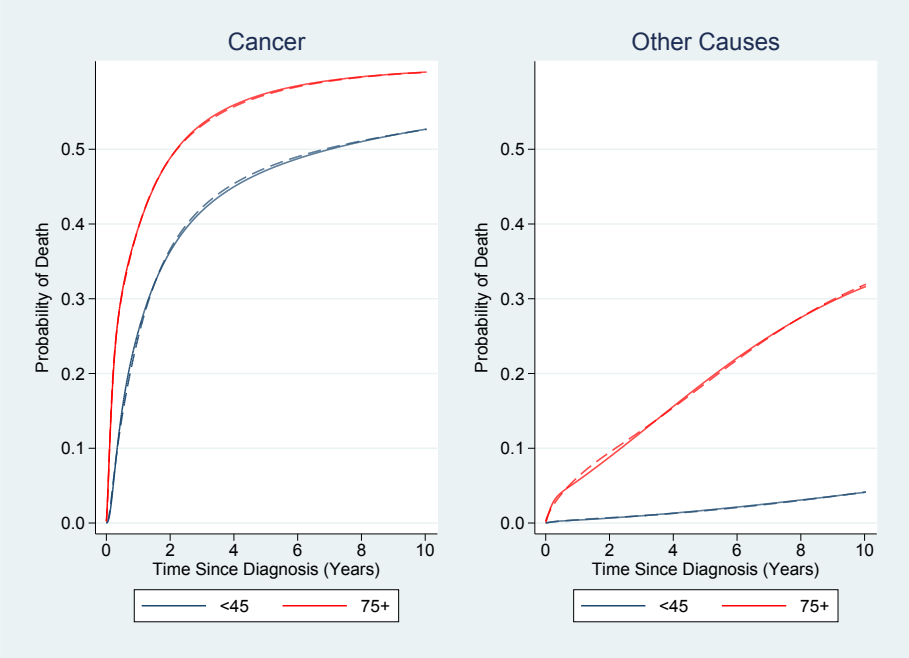
there are more early events for deaths due to cancer.

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

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

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

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

```
stpm2 cancer other fem_can fem_oth ///
    age1can age2can age3can age1oth age2oth age3oth , scale(hazard) ///
    rcsbaseoff nocons ///
    tvc(cancer other fem_can age1can age2can age3can) eform nolog ///
    knotstvc(cancer $knots_cancer other $knots_other ///
    fem_can $knots_cancer_tvc ///
    age1can $knots_cancer_tvc ///
    age2can $knots_cancer_tvc ///
    age3can $knots_cancer_tvc)
stpm2cif cancermale_age0_tvc2 othermale_age0_tvc2, cause1(cancer 1) ///
    cause2(other 1) mint(0.001)
stpm2cif cancermale_age3_tvc2 othermale_age3_tvc2, cause1(cancer 1 age3can 1) ///
    cause2(other 1 age3oth 1) mint(0.001)
```



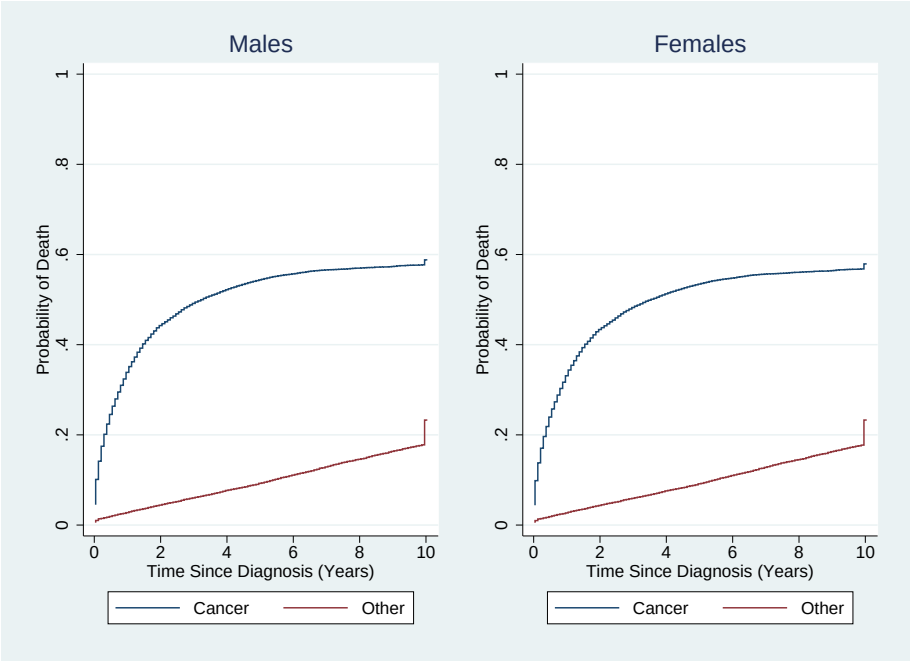
Changing the knot positions make little difference in this case.

(n) Estimatin CIFs in a Cox regression framework.

i. The data set is read in once again and stset with death due to cancer being coded the event of main interest.

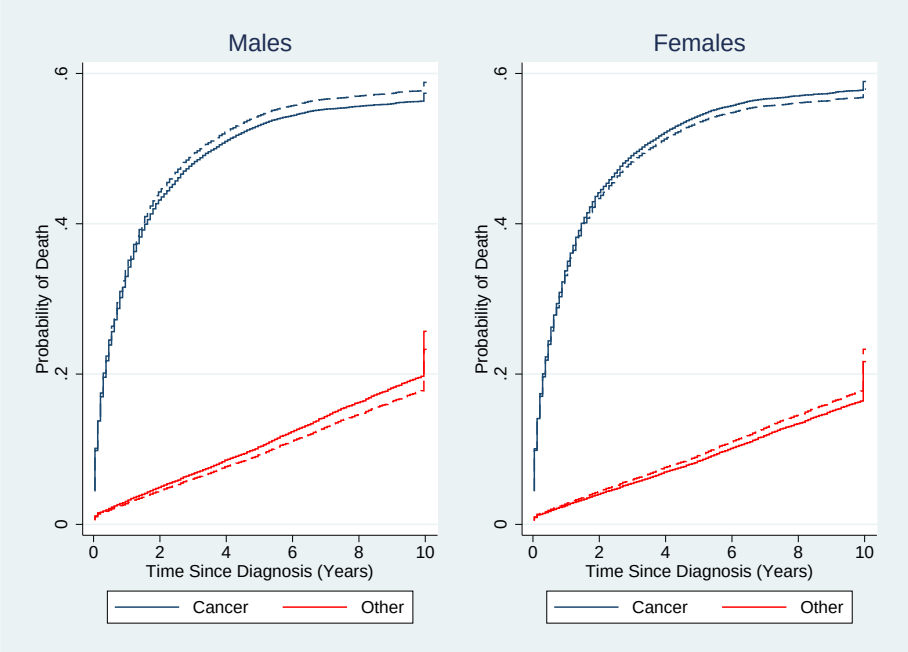
```
. stset surv_mm, failure(status==1) scale(12) exit(time 120.5)

     failure event:  status == 1
obs. time interval:  (0, surv_mm]
 exit on or before:  time 120.5
    t for analysis:  time/12

--------------------------------------------------------------------------------
    13208  total obs.
        0  exclusions
--------------------------------------------------------------------------------
    13208  obs. remaining, representing
     7122  failures in single record/single failure data
 44010.67  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  10.04167
```

ii.



The estimates look similar for males and females. In this model the association between sex and the two outcomes is constrained to be the same. This is not often no a realistic assumption in practise although when it is a reasonable assumption, it is nice to have the opportunity to save some power by not having to fit separate effects for each outcome under study.

(o)



The dashed line correspond to the estimates where the parameter estimate of sex is assumed to be the same for both outcomes, and the solid lines are estimated from the model where this assumption is relaxed. In this case, it doesn't make a huge difference on the CIfs what model is being fitted.

(p) Using the expanded data set, retrieved via `stcompadj`, we will now formally test if the effect of sex on the cause-specific hazards (*note: not on the subhazards and (equivalently) the cause-specific CIFs*) differs for our two outcomes.

```
. preserve

. stcompadj sex=1 , compet(2) savexp(silong,replace)

. use silong,clear
(Colon carcinoma, diagnosed 1975-94, follow-up to 1995)

. xi:stcox i.sex*i.stratum, strata(stratum) nohr nolog
i.sex             _Isex_1-2           (naturally coded; _Isex_1 omitted)
i.stratum         _Istratum_1-2       (naturally coded; _Istratum_1 omitted)
i.sex*i.stratum   _IsexXstr_#_#       (coded as above)

         failure _d:  __000003
   analysis time _t:  _t

Stratified Cox regr. -- Breslow method for ties

No. of subjects =        26416                     Number of obs   =     26416
No. of failures =         9248
Time at risk    =  88021.33333
                                                   LR chi2(2)      =     22.22
Log likelihood  =   -81956.045                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
           _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
      _Isex_2 |   .0237156   .0240223     0.99   0.324    -.0233672    .0707983
  _Istratum_2 |          0  (omitted)
_IsexXstr_2_2 |  -.2286699   .0503716    -4.54   0.000    -.3273965   -.1299434
------------------------------------------------------------------------------
                                                        Stratified by stratum

. restore
```

The p-value for the 2-way interaction effect is statistically significant (p = 0.000), indicating that there is strong evidence that the effect of sex on cancer-specific and other-case mortality is different.

(q) In this exercise we will find out what the two hazard ratios are by fitting a Cox model where the two outcomes are modelled simultaneously using the expanded data set.

```
. preserve

. stcompadj sex=1 , compet(2) maineffect(sex) competeffect(sex) savexp(silong,replace)

. use silong,clear
(Colon carcinoma, diagnosed 1975-94, follow-up to 1995)

. xi: stcox Main_sex Compet_sex stratum,  nolog

         failure _d:  __000003
   analysis time _t:  _t

Cox regression -- Breslow method for ties

No. of subjects =        26416                     Number of obs   =     26416
No. of failures =         9248
Time at risk    =  88021.33333
                                                   LR chi2(3)      =   3020.46
Log likelihood  =   -86867.151                     Prob > chi2     =    0.0000
```

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Main_sex |    1.01339   .0243455     0.55   0.580     .9667801    1.062248
  Compet_sex |   .8449906   .0374015    -3.81   0.000     .7747748    .9215699
     stratum |   .3818171   .0314702   -11.68   0.000     .3248609    .4487591
------------------------------------------------------------------------------

. restore
```

The hazard ratio comparing the cancer-specific mortality rate for females to males is 1.01 and not statistically significant. The corresponding hazard ratio for other-cause mortality is 0.85, suggesting that females have approximately 15 % lower mortality from other causes than cancer as compared to males. This difference is statistically significant (p-value = 0.000). However, although the reduction in the (other-cause) mortality rate associated with being female is quite large in relative terms, in absolute terms this reduction translates to quite small differences in the cause-specific CIfs. Can you think of one possible explanation for why this might be the case? Please talk to us in the labs if you would like to discuss this further or if you don't understand since this is a central learning outcome of this course.

150. **Adjusted/standardized survival curves**

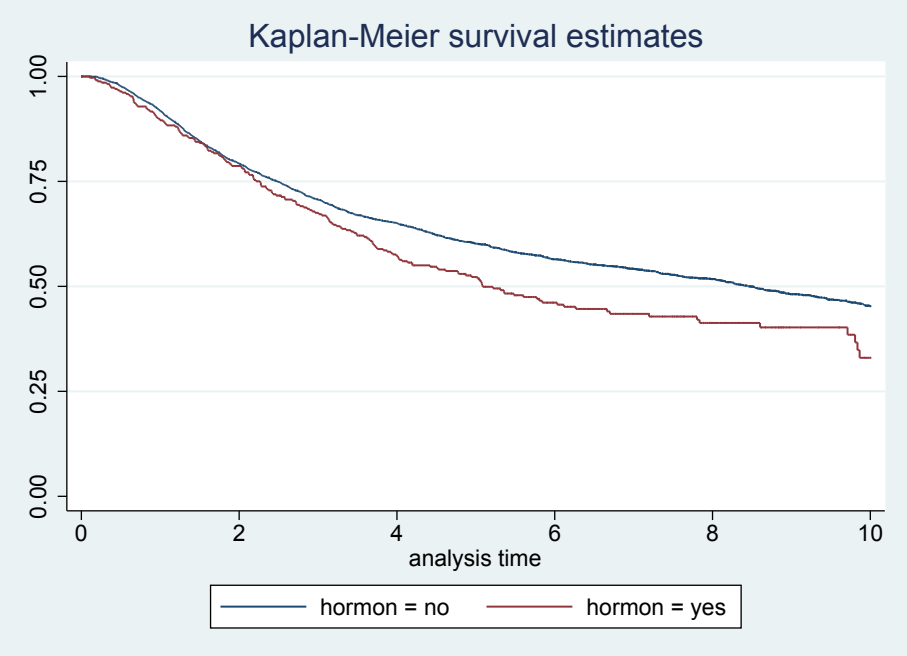
Stata addon required! This exercise requires the Stata user-written command `stpm2`

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

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

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

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



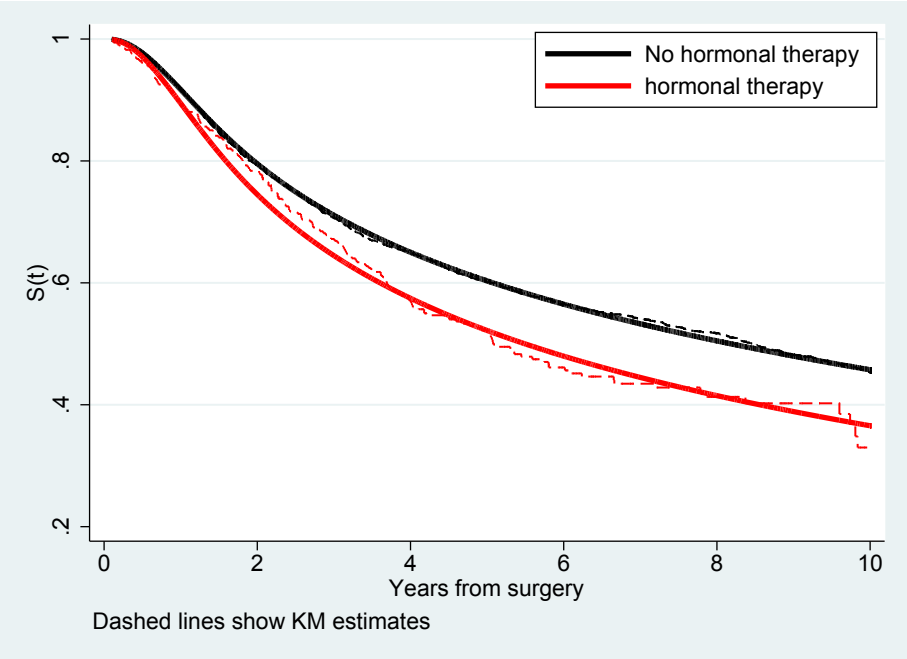
The hazard ratio will be greater than 1 as the survival is worse for the hormonal therapy group.

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

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

Iteration 0:   log likelihood = -3668.9419
Iteration 1:   log likelihood = -3668.8198
Iteration 2:   log likelihood = -3668.8197

Log likelihood = -3668.8197                     Number of obs   =       2982

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      hormon |   1.286309   .1020782     3.17   0.002     1.101023    1.502777
       _rcs1 |   2.667733   .0664965    39.37   0.000     2.540534    2.801299
       _rcs2 |   1.309768   .0283726    12.46   0.000     1.255323    1.366575
       _rcs3 |   .9909995   .0103624    -0.86   0.387     .9708964    1.011519
       _cons |   .3577717   .0107766   -34.12   0.000     .3372612    .3795294
------------------------------------------------------------------------------
predict s,s
```

(c) Compare the model based and Kaplan-Meier survival curves. Comment on the agreement between the two.

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



(d) In a previous analysis of this data, it was proposed to incorporate the effect of the number of positive lymph nodes using the following transformation[**?**].

```
enodes = exp(-0.12*nodes)
```

Add `enodes` in the model.

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

Iteration 0:   log likelihood = -3436.6645
Iteration 1:   log likelihood = -3436.5582
Iteration 2:   log likelihood = -3436.5582

Log likelihood = -3436.5582                     Number of obs   =       2982

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      hormon |    .758139   .0621847    -3.38   0.001     .6455515    .8903624
      enodes |   .1204451   .0108253   -23.55   0.000     .1009916    .1436458
       _rcs1 |    2.81705   .0714888    40.81   0.000     2.680362    2.960708
       _rcs2 |   1.301295   .0286202    11.97   0.000     1.246393    1.358616
       _rcs3 |   .9941661   .0109571    -0.53   0.596     .9729209    1.015875
       _cons |     1.9542   .1396134     9.38   0.000     1.698857    2.247922
------------------------------------------------------------------------------
```

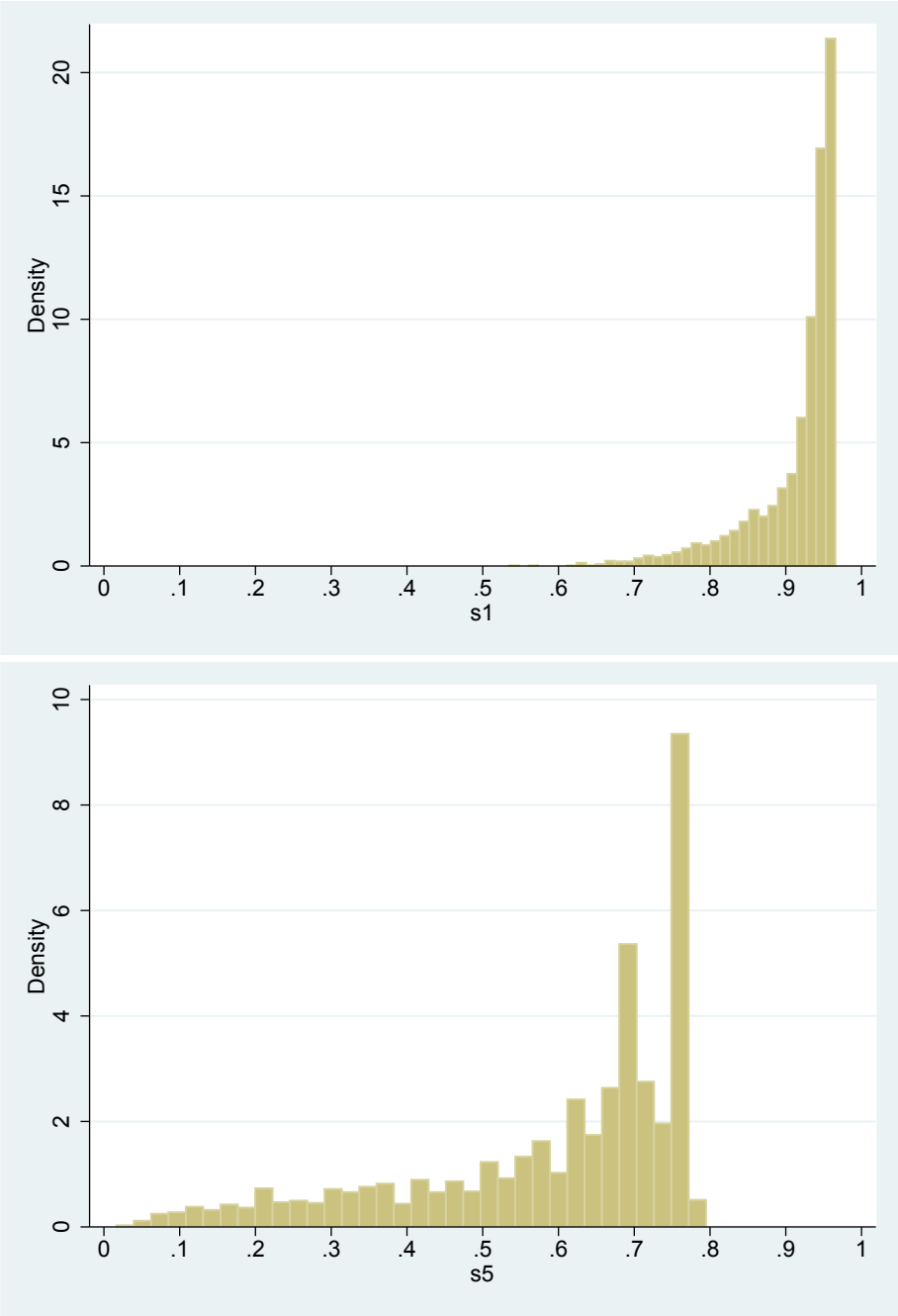
The hazard ratio is now ¡1 and significant indicating strong confounding by the number of positive lymph nodes

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

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

Iteration 0:   log likelihood = -3406.0933
Iteration 1:   log likelihood = -3405.9871
Iteration 2:   log likelihood = -3405.9871

Log likelihood = -3405.9871                     Number of obs   =       2982

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      hormon |    .801801   .0677085    -2.62   0.009     .6794953    .9461211
             |
        size |
  >20-50mmm  |   1.377321   .0818621     5.39   0.000     1.225868    1.547487
     >50 mm  |   1.650905   .1493306     5.54   0.000     1.382699    1.971135
             |
      enodes |   .1527453   .0149837   -19.15   0.000     .1260284    .1851258
     agercs1 |   .9134167   .0245823    -3.37   0.001      .866485    .9628904
     agercs2 |   .9498021   .0249017    -1.96   0.049     .9022285    .9998842
     agercs3 |   1.044863   .0275222     1.67   0.096     .9922897    1.100222
       _rcs1 |   2.835224   .0721517    40.95   0.000     2.697278    2.980225
       _rcs2 |    1.29989   .0286378    11.91   0.000     1.244956    1.357249
       _rcs3 |   .9947648   .0110285    -0.47   0.636     .9733825    1.016617
       _cons |   1.316125   .1253388     2.88   0.004      1.09203    1.586207
------------------------------------------------------------------------------
```

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

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



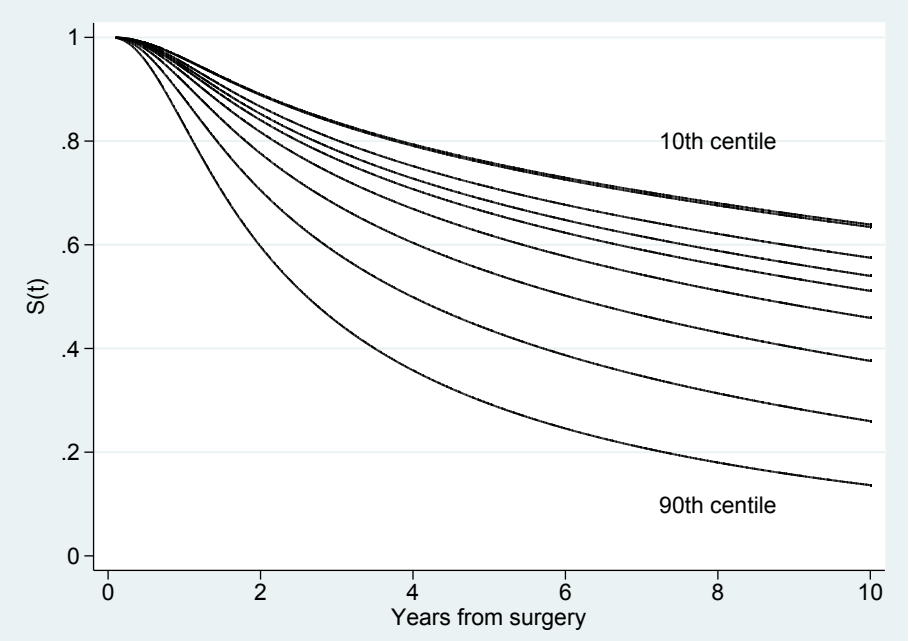
(g) Predict a prognostic index. This is the predicted values of the linear predictor without the spline terms. This can be used to classify into risk groups. We will plot from the 10th to the 90th centile of the prognostic index to show the range in predicted survival probability in the stiud population.

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

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

Iteration 0:   log likelihood = -3406.0875
Iteration 1:   log likelihood = -3405.9871
Iteration 2:   log likelihood = -3405.9871

Log likelihood = -3405.9871                     Number of obs   =       2982

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
          xb |          1   .0402395    24.85   0.000      .921132    1.078868
       _rcs1 |   1.042121   .0254314    40.98   0.000     .9922764    1.091965
       _rcs2 |   .2622797   .0220263    11.91   0.000     .2191089    .3054505
       _rcs3 |   -.005249   .0110804    -0.47   0.636    -.0269661    .0164682
       _cons |    .274692   .0518997     5.29   0.000     .1729706    .3764135
------------------------------------------------------------------------------
```

The likelhoods are the same as we are just including the same component of the linear predictor in the model.
Now obtain predictions from the 10th to the 90th centile and plot the resulting functions.

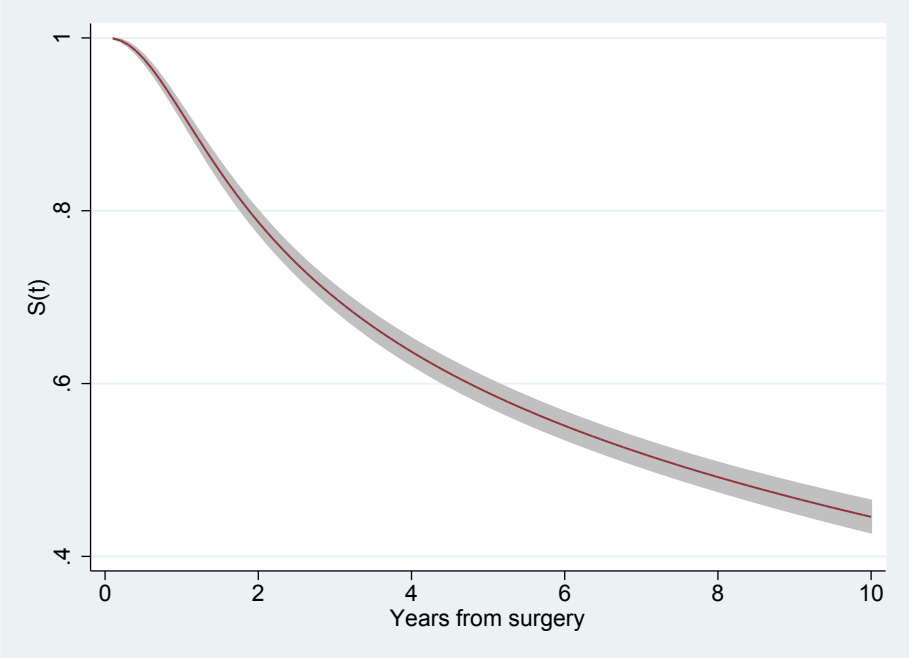
```
forvalues i = 10(10)90 {
  centile xb, centile(‘i’)
  predict s_xb‘i’, surv at(xb ‘r(c_1)’)
}
twoway (line s_xb?? _t, sort lcolor(black ..)) ///
  , legend(off) ///
  ylabel(0(0.2)1, angle(h)) ///
  xtitle("Years from surgery") ///
  ytitle("S(t)") ///
  text(0.8 8 "10th centile") ///
  text(0.1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



(h) Refit the original model and and obtain the average survival curve for the study population as a whole.

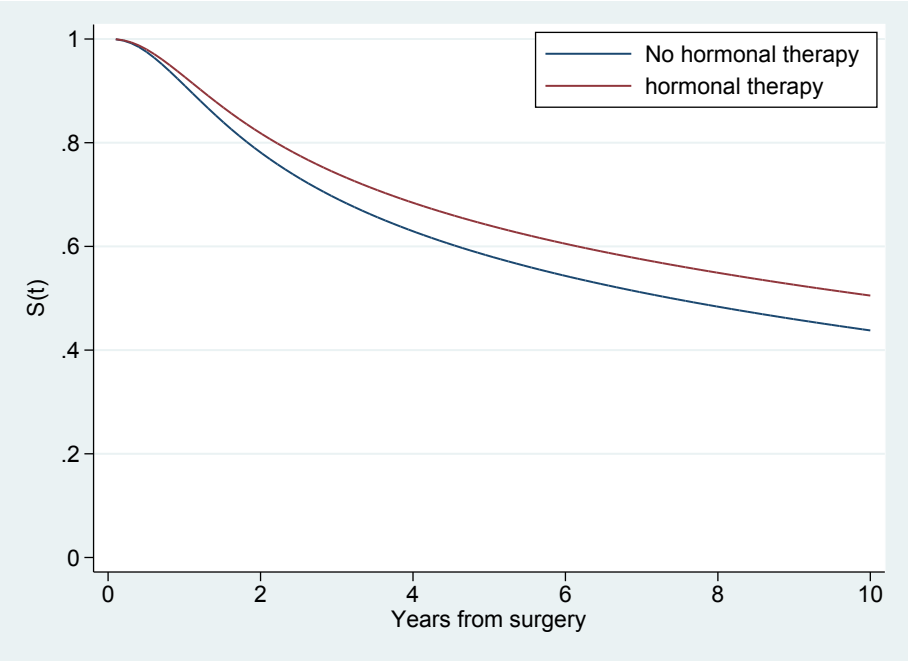
```
stpm2 hormon i.size enodes agercs*, scale(hazard) df(3) eform
```

```
range timevar 0 10 100
predict s_mean, meansurv timevar(timevar) ci
twoway  (rarea s_mean_lci s_mean_uci timevar, sort pstyle(ci)) ///
(line s_mean timevar, sort) ///
, xtitle("Years from surgery") ///
ytitle("S(t)") ///
legend(off)
```



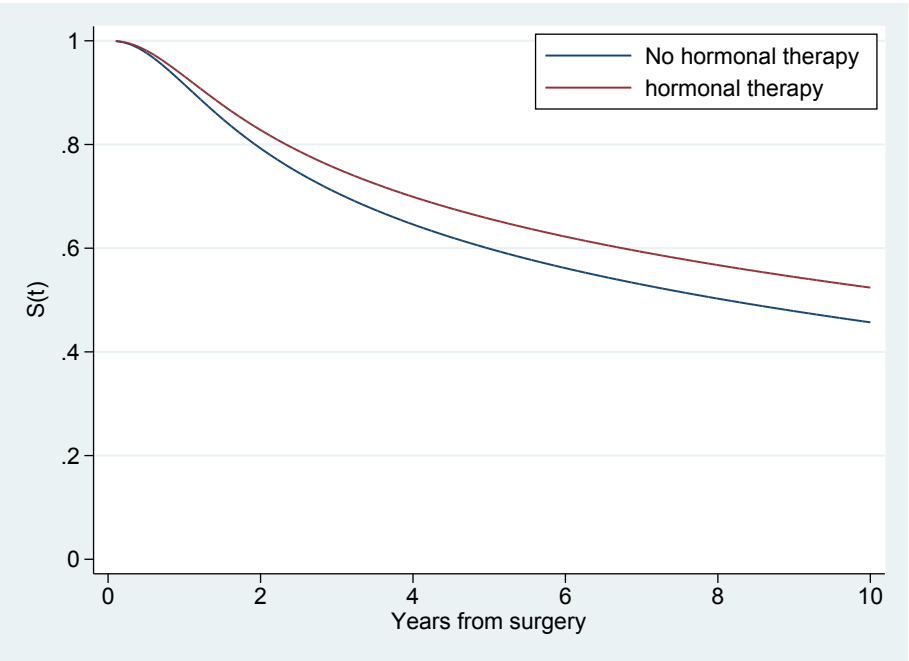
(i) Obtain the adjusted survival curves by hormonal therapy status standardising over the covariate pattern of the whole study population. Use the `meansurv` option combined with the `at()` option.

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



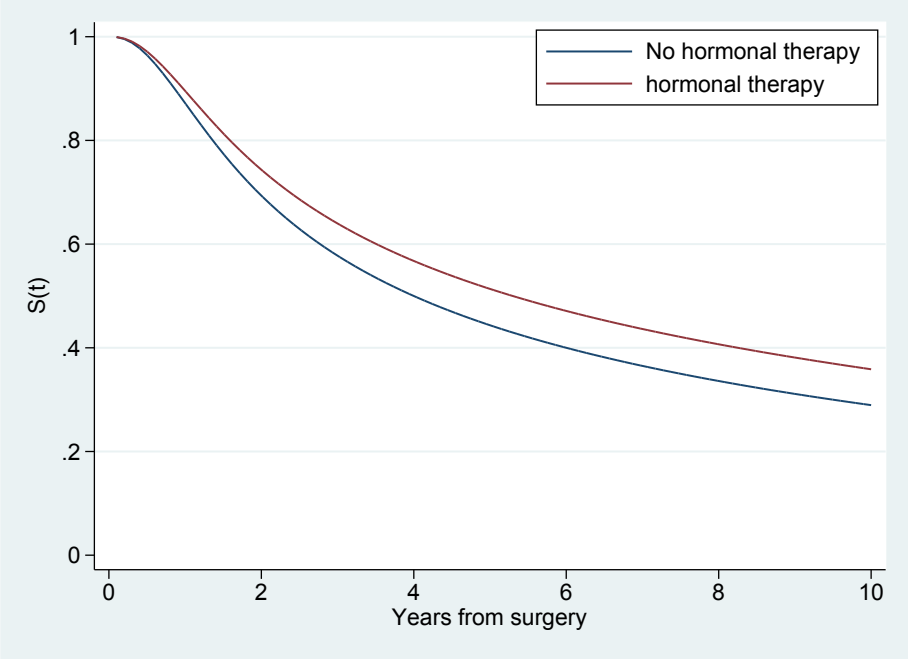
(j) Obtain the adjusted survival curves by hormonal therapy status standardising over the covariate pattern of those not on hormonal therapy.

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



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

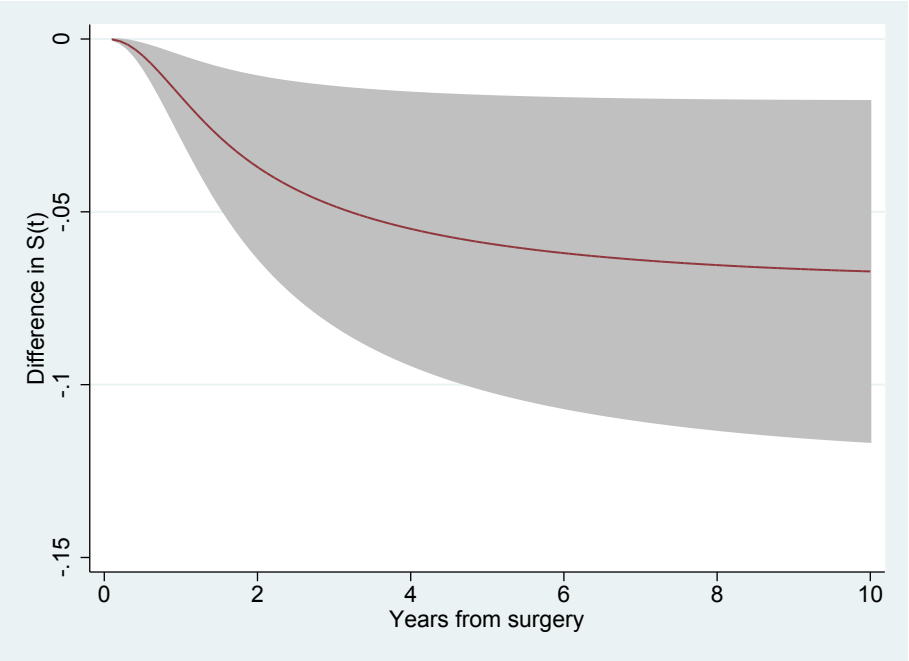
```
predict s_h0c if hormon==1, meansurv at(hormon 0) timevar(timevar) ci
predict s_h1c if hormon==1, meansurv at(hormon 1) timevar(timevar) ci
twoway  (line s_h0c timevar, sort) ///
  (line s_h1c timevar, sort) ///
  , xtitle("Years from surgery") ///
  ytitle("S(t)") ///
  ylabel(0(.2)1,angle(h)) ///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



Those on hormonal therapy tend to have more severe disease and so the survival curve is higher.

(l) Now calculate and plot the difference in adjusted survival curves.

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

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


Difference in S(t)
0
-.05
-.1
-.15
0
2
4
6
8
10
Years from surgery

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

(a)
```
. * stset the data
. stset exit, fail(status==1) enter(dx) origin(dx) scale(365.24) id(id)

                id:  id
     failure event:  status == 1
obs. time interval:  (exit[_n-1], exit]
 enter on or after:  time dx
 exit on or before:  failure
    t for analysis:  (time-origin)/365.24
            origin:  time dx

------------------------------------------------------------------------------
       7775  total observations
          0  exclusions
------------------------------------------------------------------------------
       7775  observations remaining, representing
       7775  subjects
       1913  failures in single-failure-per-subject data
  51276.908  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  20.96156
```

There are 1913 deaths (events) among 7775 patients.

(b) The estimated HR changes from 0.627167 to 0.700238 on adjusting for age, period, and stage (and to 0.749139 if we adjust for subsite). Some, but not a lot of, confounding.

(c) We would expect similar estimates (and standard errors) from the three models since we are fitting what is conceptually the same model 3 times just with a different approach to modelling the baseline hazard. We would expect the results from Poisson regression to be more different to the other two since it is modelling the baseline hazard crudely (a step function assuming the hazard is constant within 5-year intervals). We see, however, that the estimated HRs are quite robust to this.

```
. estimates table cox fpm pois, eform b(%7.3f) se(%7.3f) eq(1)

-------------------------------------------
    Variable |   cox       fpm      pois
-------------+-----------------------------
#1           |
         sex |
        Male |  (base)    (base)    (base)
             |
      Female |   0.700     0.699     0.697
             |   0.033     0.033     0.033
             |
      agegrp |
       0-44  |  (base)    (base)    (base)
             |
       45-59 |   1.286     1.288     1.294
             |   0.087     0.087     0.087
       60-74 |   1.712     1.717     1.733
             |   0.111     0.111     0.112
         75+ |   2.678     2.697     2.728
             |   0.200     0.202     0.204
             |
    year8594 |
Diagnosed..  |  (base)    (base)    (base)
             |
```

```
Diagnosed..  |   0.799     0.801     0.817
             |   0.038     0.038     0.039
             |
       stage |
     Unknown |  (base)    (base)    (base)
             |
   Localised |   1.039     1.038     1.040
             |   0.071     0.071     0.071
    Regional |   4.825     4.842     4.855
             |   0.441     0.443     0.443
      Distant |  13.618    13.839    13.362
             |   1.088     1.105     1.056
```

(d) There were 1913 events so with 1:1 matching we would expect an absolute maximum of double this (3826) unique individuals in the NCC. However, since individuals can be both cases and controls, or be controls for multiple cases we will see fewer unique individuals.

(e) i. `_time` is the underlying time scale upon which we have matched controls to cases. In this example it is time since diagnosis.

ii. There are an equal number of cases and controls, also within each age stratum. This is not always the case, since it is possible that no eligible controls exist for some cases.

```
. tab agegrp _case, missing

           | 0 for controls; 1 for
  Age in 4 |         cases
categories |         0          1 |     Total
-----------+----------------------+----------
      0-44 |       386        386 |       772
     45-59 |       522        522 |     1,044
     60-74 |       640        640 |     1,280
       75+ |       365        365 |       730
-----------+----------------------+----------
     Total |     1,913      1,913 |     3,826
```

iii. There are 3,247 unique individuals among the 3,826 cases and controls.

```
. codebook id
-----------------------------------------
id                       Unique patient ID
-----------------------------------------

 type:  numeric (int)

 range:  [4,7773]         units:  1
 unique values:  3,247    missing .:  0/3,826
```

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

Conditional (fixed-effects) logistic regression

                                  Number of obs     =      3,826
                                  LR chi2(5)        =     530.95
                                  Prob > chi2       =     0.0000
Log likelihood = -1060.5158       Pseudo R2         =     0.2002

------------------------------------------------------------------------------
     _case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
       sex |
      Male |          1  (base)
    Female |   .7263021   .0541607    -4.29   0.000     .6275421    .8406047
           |
  year8594 |
```

```
    75-84  |          1  (base)
    85-94  |   .7069653   .0568284   -4.31   0.000   .6039145    .8276006
           |
     stage |
  Unknown  |          1  (base)
Localised  |   .9390677   .0912807   -0.65   0.518   .7761705    1.136153
 Regional  |   4.467645   .8035128    8.32   0.000   3.140427    6.355776
  Distant  |   16.67736   3.559866   13.18   0.000   10.97575    25.34082
------------------------------------------------------------------------------
```

i. Rate ratio (or hazard ratio).

ii. Yes it is similar. We expect it to be similar, since we are estimating the same underlying quantity. We would not expect it to be identical to the full cohort estimate due to sampling variation.

iii. Yes, but the standard errors are larger and the confidence intervals wider.

(g)

	Outside subcohort	Inside subcohort	Total
Non-cases	4,392	1,470	5,862
Cases	1,440	473	1,913
Total	5,832	1,943	7,775

(h) The exact sampling fraction of the subcohort is $1943/7775 = 0.2499$. The exact sampling fraction of non-cases is $1470/5862 = 0.2508$.

(i) Hopefully the weights are as you expected. Ask if you don't follow. All cases have weight 1 since we included all cases. The controls have weight of approximately 4; we took a 25% sample so each sampled control represents 4 individuals. Non-cases outside the subcohort do not contribute to the analysis and have a missing weight.

```
. tab wt, missing

         wt |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      1,913       24.60       24.60
   3.987755 |      1,470       18.91       43.51
          . |      4,392       56.49      100.00
------------+-----------------------------------
      Total |      7,775      100.00
```

(j) Note that Stata reports `4392 weights invalid PROBABLE ERROR`.

(k) The first column is the analysis of the full cohort. The three approaches to analysing the case-cohort study give similar estimates to each other. Estimates are also similar to the full cohort, except with larger standard errors.

```
. estimates table cox cox_cc fpm_cc pois_cc, eform b(%7.3f) se(%7.3f) eq(1)
-------------------------------------------------------
    Variable |   cox     cox_cc    fpm_cc    pois_cc
-------------+-----------------------------------------
#1           |
         sex |
        Male |  (base)    (base)    (base)    (base)
             |
      Female |   0.700     0.684     0.683     0.680
             |   0.033     0.051     0.051     0.050
             |
      agegrp |
```

```
       0-44  |  (base)    (base)    (base)    (base)
             |
      45-59  |   1.286     1.284     1.288     1.293
             |   0.087     0.130     0.131     0.130
      60-74  |   1.712     1.613     1.618     1.632
             |   0.111     0.164     0.166     0.166
        75+  |   2.678     2.519     2.538     2.558
             |   0.200     0.331     0.337     0.331
             |
    year8594 |
Diagnosed..  |  (base)    (base)    (base)    (base)
             |
Diagnosed..  |   0.799     0.822     0.824     0.843
             |   0.038     0.061     0.062     0.062
             |
       stage |
    Unknown  |  (base)    (base)    (base)    (base)
             |
  Localised  |   1.039     1.027     1.027     1.030
             |   0.071     0.090     0.090     0.091
   Regional  |   4.825     5.172     5.196     5.204
             |   0.441     0.748     0.756     0.757
    Distant  |  13.618    13.666    13.894    13.551
             |   1.088     2.006     2.062     1.903
```

(l) Following is our output when we generated and analysed a nested case-control study 5 times. 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
-------------------------------------------------------------------------------
```

(m) With 5 controls per case we will come very close to analysing the full cohort (i.e., nothing to gain by doing a nested case-control study). However, in a more realistic scenario (where the outcome is rare) it would be reasonable to select 5 controls per case.

(n)

(o)

181. **Calculating SMRs/SIRs**

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

. gen bdate = dx-(age*365.25)
. 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 observations
        0  exclusions
------------------------------------------------------------------------------
     7775  observations remaining, representing
     7775  subjects
     3047  failures in single-failure-per-subject data
  51275.5  total analysis time at risk and under observation
                                             at risk from t =         0
                                  earliest observed entry t =         0
                                       last observed exit t =  101.4586

. stsplit _age, at(0(1)110) trim
(no obs. trimmed because none out of range)
(47427 observations (episodes) created)
```

(b)
```
. stsplit _year, after(time=d(1/1/1975)) at(0(1)22) trim
(no obs. trimmed because none out of range)
(48864 observations (episodes) created)

. tab _year

      _year |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        244        0.23        0.23
          1 |        675        0.65        0.88
          2 |      1,045        1.00        1.89
          3 |      1,428        1.37        3.26
          :
        output omitted
          :
         18 |      9,302        8.94       81.09
         19 |      9,824        9.44       90.53
         20 |      9,857        9.47      100.00
------------+-----------------------------------
      Total |    104,066      100.00
```

To make results easier to interpret, we replace `_year` with `_year`1975+.

```
. replace _year=1975+_year
_year was byte now int
(104066 real changes made)
      _year |      Freq.     Percent        Cum.
------------+-----------------------------------
       1975 |        244        0.23        0.23
       1976 |        675        0.65        0.88
       1977 |      1,045        1.00        1.89
          :
        output omitted
          :
       1992 |      8,784        8.44       72.15
       1993 |      9,302        8.94       81.09
       1994 |      9,824        9.44       90.53
       1995 |      9,857        9.47      100.00
------------+-----------------------------------
      Total |    104,066      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     1     0
       10 |    2     1     0     0     0
       20 |    8    10    10     9     1
       30 |   19    44    49    28     6
       40 |   40    62    75    99    33
       50 |   43    98   103   135    38
       60 |   80   121   177   181    54
       70 |   51   153   224   270    67
       80 |   30    82   153   285    79
       90 |    1    12    34    61    14
      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 |    0.797    17.641    13.568     0.870
       10 |   25.726    36.717    66.935    82.860    11.577
       20 |  152.056   356.272   580.056   725.567   124.215
       30 |  315.055  1053.143  1645.727  1915.429   392.845
       40 |  462.774  1368.987  2696.640  4070.498   853.771
       50 |  564.616  1677.997  2998.889  4476.847  1030.195
       60 |  562.485  1553.928  3024.645  4662.907  1065.254
       70 |  375.063  1298.308  2410.884  3710.084   870.622
       80 |   95.522   376.986   956.702  1795.746   439.716
       90 |    9.040    30.828    87.083   183.300    44.799
      100 |               0.626               2.710
------------------------------------------------------------
```

(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  0.074  0.000
       10 | 0.078  0.027  0.000  0.000  0.000
       20 | 0.053  0.028  0.017  0.012  0.008
       30 | 0.060  0.042  0.030  0.015  0.015
       40 | 0.086  0.045  0.028  0.024  0.039
       50 | 0.076  0.058  0.034  0.030  0.037
       60 | 0.142  0.078  0.059  0.039  0.051
       70 | 0.136  0.118  0.093  0.073  0.077
       80 | 0.314  0.218  0.160  0.159  0.180
       90 | 0.111  0.389  0.390  0.333  0.313
      100 |        1.597         1.107
-------------------------------------------
```

(e)
```
. sort _year sex _age

. merge m:1 _year sex _age using popmort
. tab _merge

                _merge |      Freq.     Percent        Cum.
-----------------------+-----------------------------------
 using only (2) |      7,220        6.49        6.49
    matched (3) |    104,066       93.51      100.00
-----------------------+-----------------------------------
                 Total |    111,286      100.00

. drop if _merge==2
(7220 observations deleted)
. drop _merge
```

(f)
```
. gen mortrate=(-ln(prob))
. gen e=_y*mortrate
. list id e _d mortrate in 1/20
```

	id	e	_d	mortrate
1.	1730	.0004422	0	.0010205
2.	1703	.0004439	0	.0016013
3.	1692	.001161	0	.0018417
4.	1608	.0016129	0	.0017014
5.	1585	.0007388	0	.0019519
6.	1539	.0018245	0	.0018918
7.	1522	.0015179	1	.0019118
8.	1504	.0002408	1	.0019118
9.	1479	.0002808	0	.0020822
10.	1480	.000988	0	.002002
11.	1467	.0003686	0	.002002
12.	1457	.0008306	0	.0024029
13.	1423	.0023079	0	.002463
14.	1420	.0011211	0	.002463
15.	1384	.0022039	0	.0027638
16.	1322	.0024838	0	.0031148
17.	1326	.0013751	0	.0031148
18.	1338	.0001364	1	.0031148
19.	1309	.0016526	0	.0035664
20.	1295	.0034394	0	.0035664

(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)=2.417(2.332:2.504)

. strate, smr(mortrate)

Estimated SMRs and lower/upper bounds of 95% confidence intervals
(104066 records included in the analysis)
```

D	E	SMR	Lower	Upper
3047	1260.74	2.417	2.333	2.504

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

Estimated SMRs and lower/upper bounds of 95% confidence intervals
(104066 records included in the analysis)

+-----------------------------------------------------+
|     stage      D        E      SMR    Lower    Upper |
|-----------------------------------------------------|
|   Unknown    557   283.83    1.962    1.806    2.132 |
| Localised   1795   913.15    1.966    1.877    2.059 |
|  Regional    260    37.47    6.938    6.144    7.835 |
|   Distant    435    26.29   16.546   15.062   18.176 |
+-----------------------------------------------------+
```

182. Using strs for calculating SMRs

```
. use melanoma, clear
. stset exit, fail(status == 1 2) origin(dx) entry(dx) scale(365.25) id(id)
. strs using popmort, br(0(1)21) mergeby(_year sex _age) notables save(replace)

. use grouped.dta, clear
(Collapsed (or grouped) survival data)

. list start n d w p cp d_star, sum(d d_star)
```

	start	n	d	w	p	cp	d_star
1.	0	7775	571	2	0.9266	0.9266	189.4
2.	1	7202	652	450	0.9066	0.8400	164.4
3.	2	6100	446	401	0.9244	0.7765	135.2
4.	3	5253	310	366	0.9389	0.7290	115.4
5.	4	4577	227	339	0.9485	0.6914	99.5
6.	5	4011	182	331	0.9527	0.6587	86.6
7.	6	3498	132	314	0.9605	0.6327	76.8
8.	7	3052	97	330	0.9664	0.6114	68.8
9.	8	2625	90	321	0.9635	0.5891	61.0
10.	9	2214	66	281	0.9682	0.5704	51.5
11.	10	1867	71	213	0.9597	0.5474	44.2
12.	11	1583	60	210	0.9594	0.5251	36.9
13.	12	1313	26	183	0.9787	0.5140	32.1
14.	13	1104	30	199	0.9701	0.4986	27.3
15.	14	875	23	163	0.9710	0.4842	21.9
16.	15	689	19	127	0.9696	0.4694	17.6
17.	16	543	15	130	0.9686	0.4547	14.2
18.	17	398	12	113	0.9649	0.4387	10.9
19.	18	273	7	96	0.9689	0.4251	8.6
20.	19	170	8	82	0.9380	0.3987	4.8
21.	20	80	3	77	0.9277	0.3699	1.2
Sum		3047					1267.8

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

	obs	exp	smr	LL	UL
1.	3047	1267.8	2.403313	2.318728	2.490194

200. **Calculating expected survival by hand**

(a) The first two probabilities can be seen below:

```
. use popmort
. list if sex==1 & _age==72 & _year==1989
      +--------------------------+
      | sex   _year   _age   prob |
      |--------------------------|
 8129. |   1    1989     72   .949 |
      +--------------------------+

. list if sex==1 & _age==73 & _year==1990
      +----------------------------+
      | sex   _year   _age     prob |
      |----------------------------|
 8342. |   1    1990     73   .94338 |
      +----------------------------+
```

(b) The probabilities are `0.97567  0.97354  0.97066  0.97357  0.96979`.

(c) The estimated 5-year expected survival is 0.81592 using the Ederer I method and 0.81355 using the Ederer II method. The results are contained in the Excel file `\solutions\exercise200.xls`.

(d) The output from `strs` is shown below.

`cp_e1` Ederer I estimate of the expected survival rate

`cp_e2` Ederer II estimate of the expected survival rate

```
. strs using popmort, br(0(1)5) mergeby(_year sex _age) ///
>            ederer1 list(n d w cp_e1 cp_e2)

  +-----------------------------------------------+
  | start   end    n   d   w    cp_e1    cp_e2 |
  |-----------------------------------------------|
  |     0     1   35   8   0   0.9640   0.9640 |
  |     1     2   27   2   2   0.9272   0.9268 |
  |     2     3   23   5   4   0.8900   0.8884 |
  |     3     4   14   2   1   0.8529   0.8488 |
  |     4     5   11   0   1   0.8159   0.8135 |
  +-----------------------------------------------+
```

The estimated 5-year expected survival is 0.81592 using the Ederer I method, 0.81355 using the Ederer II method, and 0.83080 using the Hakulinen method (not shown in the table). The estimate are based on only 35 patients so you should not read too much into the differences between the different methods.

201. **Life-table estimates of relative survival using `strs`**

(a) I will only show the estimates for the most recent period.

year8594 = Diagnosed 85-94

end	n	d	w	p	p_star	r	cp	cp_e2	cr_e2
1.00	3173	88	0	0.9723	0.9753	0.9969	0.9723	0.9753	0.9969
2.00	3085	180	297	0.9387	0.9748	0.9630	0.9127	0.9508	0.9599
3.00	2608	131	296	0.9467	0.9754	0.9707	0.8641	0.9273	0.9318
4.00	2181	119	271	0.9418	0.9757	0.9652	0.8138	0.9049	0.8994
5.00	1791	84	246	0.9496	0.9767	0.9723	0.7728	0.8837	0.8745
6.00	1461	60	239	0.9553	0.9766	0.9781	0.7383	0.8631	0.8554
7.00	1162	38	217	0.9639	0.9769	0.9868	0.7116	0.8431	0.8440
8.00	907	23	253	0.9705	0.9754	0.9950	0.6907	0.8224	0.8398
9.00	631	14	241	0.9726	0.9738	0.9987	0.6717	0.8009	0.8387
10.00	376	6	208	0.9779	0.9740	1.0041	0.6569	0.7801	0.8421

Here we have used annual intervals. The 5-year relative survival ratio is 0.8745.

i. The excess mortality is highest in the second interval. We can tell this as the interval specific relative survival is lowest in this interval.

ii. Remember that these patients are diagnosed with localised melonoma. It seems reasonable that they may not experience high excess mortality immediately after diagnosis, but there may be higher excess mortality later in follow-up due to progression of the disease.

iii. If a cure point was reached, the interval specific relative survival would be 1 (that is, the survival in the interval was the same as the general population). We can see that the interval specific relative survival does appear to be reach, and level out, at 1 over the follow-up.

(b)
```
-> year8594 = Diagnosed 85-94
```

end	n	d	w	p	p_star	r	cp	cr	lo_cr	hi_cr
0.50	3173	40	0	0.9874	0.9874	1.0000	0.9874	1.0000	0.9954	1.0034
1.00	3133	48	0	0.9847	0.9878	0.9968	0.9723	0.9968	0.9903	1.0021
1.50	3085	88	140	0.9708	0.9871	0.9835	0.9439	0.9804	0.9714	0.9882
2.00	2857	92	157	0.9669	0.9879	0.9788	0.9126	0.9596	0.9485	0.9695
2.50	2608	70	147	0.9724	0.9874	0.9848	0.8874	0.9450	0.9323	0.9565
3.00	2391	61	149	0.9737	0.9881	0.9854	0.8641	0.9312	0.9172	0.9441
3.50	2181	66	131	0.9688	0.9876	0.9810	0.8371	0.9135	0.8980	0.9278
4.00	1984	53	140	0.9723	0.9885	0.9836	0.8139	0.8985	0.8818	0.9141
4.50	1791	52	117	0.9700	0.9881	0.9817	0.7895	0.8821	0.8640	0.8990
5.00	1622	32	129	0.9795	0.9887	0.9907	0.7733	0.8738	0.8549	0.8917
5.50	1461	39	116	0.9722	0.9880	0.9840	0.7518	0.8598	0.8396	0.8789
6.00	1306	21	123	0.9831	0.9888	0.9943	0.7391	0.8549	0.8338	0.8748
6.50	1162	24	103	0.9784	0.9882	0.9901	0.7231	0.8464	0.8243	0.8675
7.00	1035	14	114	0.9857	0.9886	0.9971	0.7128	0.8440	0.8209	0.8659
7.50	907	15	132	0.9822	0.9874	0.9947	0.7001	0.8395	0.8153	0.8625
8.00	760	8	121	0.9886	0.9879	1.0007	0.6920	0.8401	0.8149	0.8640
8.50	631	9	105	0.9844	0.9866	0.9978	0.6813	0.8382	0.8116	0.8635
9.00	517	5	136	0.9889	0.9871	1.0018	0.6737	0.8397	0.8117	0.8663
9.50	376	3	119	0.9905	0.9867	1.0039	0.6673	0.8430	0.8134	0.8711
10.00	254	3	89	0.9857	0.9871	0.9986	0.6578	0.8417	0.8090	0.8728

The estimates at 10 years are quite similar 0.8417 with the 6-monthly splits compared to 0.8421 with the yearly splits.

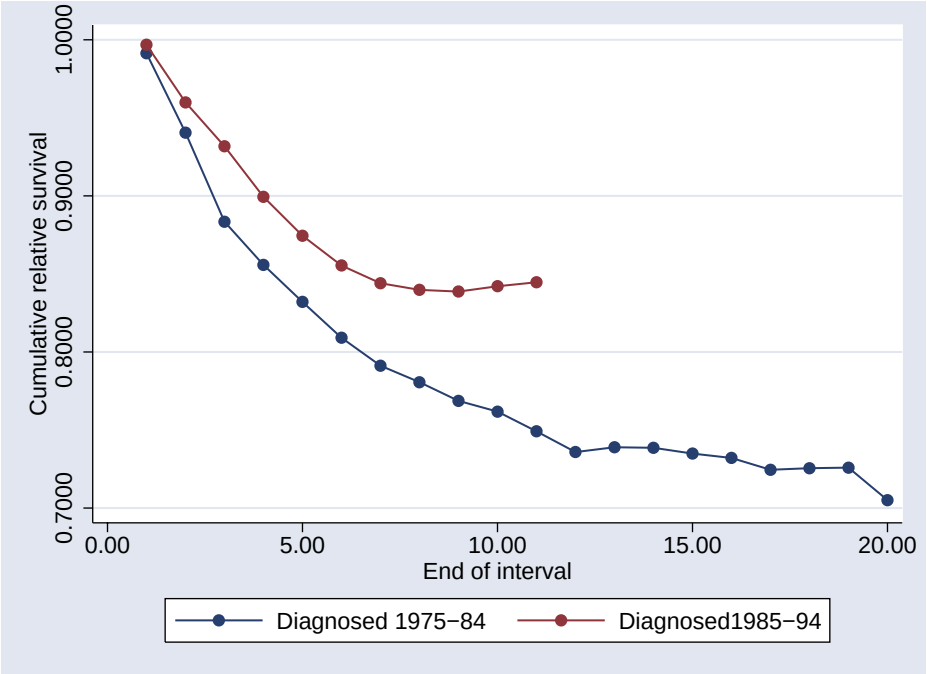
(c)
```
-> year8594 = Diagnosed 85-94
```

end	n	d	w	p	p_star	r	cp	cp_e2	cr_e2
.25	3173	13	0	0.9959	0.9936	1.0023	0.9959	0.9936	1.0023
.5	3160	27	0	0.9915	0.9937	0.9977	0.9874	0.9874	1.0000
.75	3133	19	0	0.9939	0.9938	1.0001	0.9814	0.9813	1.0001
1	3114	29	0	0.9907	0.9939	0.9968	0.9723	0.9753	0.9969
2	3085	180	297	0.9387	0.9748	0.9630	0.9127	0.9507	0.9599
3	2608	131	296	0.9467	0.9754	0.9707	0.8641	0.9273	0.9318
4	2181	119	271	0.9418	0.9757	0.9652	0.8138	0.9048	0.8994
5	1791	84	246	0.9496	0.9767	0.9723	0.7728	0.8837	0.8745
6	1461	60	239	0.9553	0.9766	0.9781	0.7383	0.8631	0.8554
7	1162	38	217	0.9639	0.9769	0.9868	0.7116	0.8431	0.8441
8	907	23	253	0.9705	0.9754	0.9950	0.6907	0.8224	0.8398
9	631	14	241	0.9726	0.9738	0.9987	0.6717	0.8009	0.8387
10	376	6	208	0.9779	0.9740	1.0041	0.6569	0.7800	0.8421

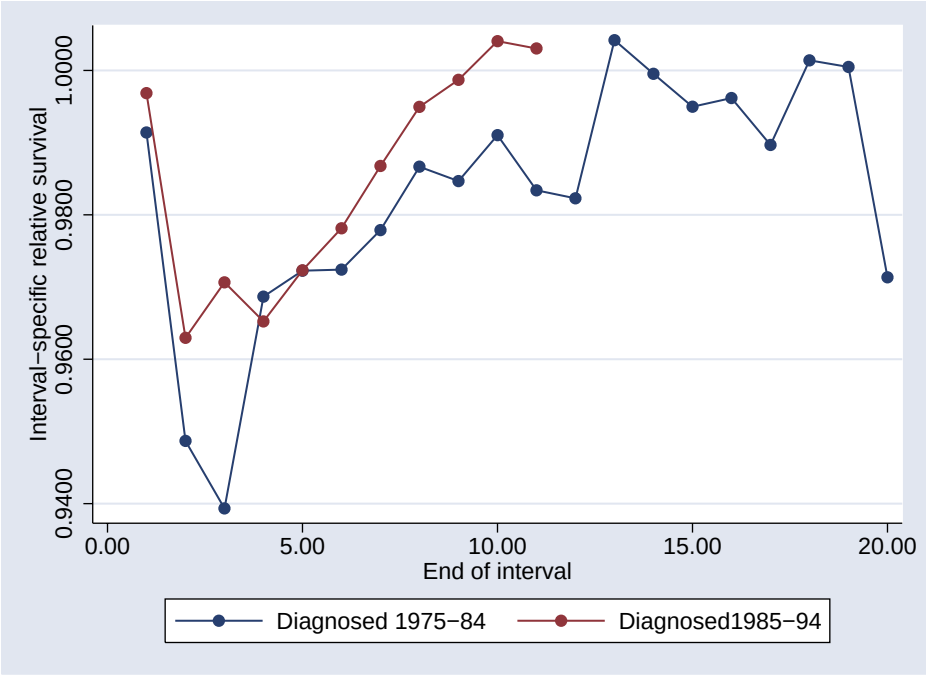
The 5 and 10 year estimates are very similar.

(d) Only the patients diagnosed in the early period have a potential follow-up of 20 years.

(e)


Cumulative relative survival
0.7000
0.8000
0.9000
1.0000
0.00
5.00
10.00
15.00
20.00
End of interval
Diagnosed 1975−84
Diagnosed1985−94

(f)


Interval−specific relative survival
0.9400
0.9600
0.9800
1.0000
0.00
5.00
10.00
15.00
20.00
End of interval
Diagnosed 1975−84
Diagnosed1985−94

(g)
```
. strs using popmort if stage==1, br(0(1)20) mergeby(_year sex _age) ///
>  by(year8594) list(start n d w cr_e1 cr_e2 cr_hak) ederer1 potfu(potfu)

-> year8594 = Diagnosed 85-94
```

start	end	n	d	w	cr_e1	cr_e2	cr_hak
0	1	3173	88	0	0.9969	0.9969	0.9969
1	2	3085	180	297	0.9599	0.9599	0.9598
2	3	2608	131	296	0.9325	0.9318	0.9324
3	4	2181	119	271	0.9014	0.8994	0.9011
4	5	1791	84	246	0.8789	0.8745	0.8780
5	6	1461	60	239	0.8623	0.8554	0.8606
6	7	1162	38	217	0.8539	0.8440	0.8513
7	8	907	23	253	0.8519	0.8398	0.8486
8	9	631	14	241	0.8521	0.8387	0.8483
9	10	376	6	208	0.8574	0.8421	0.8530
10	11	162	2	160	0.8612	0.8447	0.8564

The estimates are quite similar, although there are some differences for the long-term estimates.

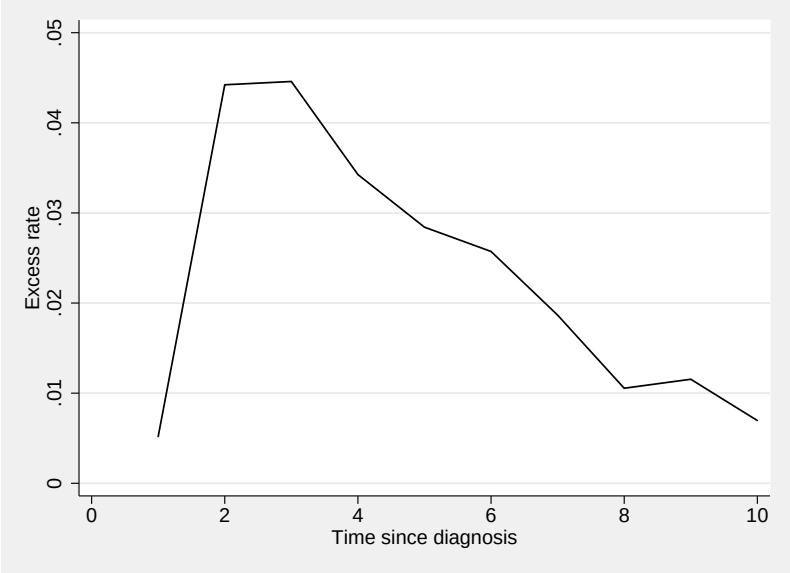
(h)
```
. strs using popmort, br(0(‘=1/12’)20) mergeby(_year sex _age) ///
>   by(year8594) pohar list(start n d w cr_e2 cns_pp) save(replace)

. use grouped, clear
. list start end cr_e2 cns_pp if mod(end,1)==0 & year8594, noobs
```

start	end	cr_e2	cns_pp
.9167	1	0.9969	0.9970
1.917	2	0.9594	0.9583
2.917	3	0.9306	0.9271
3.917	4	0.8977	0.8917
4.917	5	0.8730	0.8667
5.917	6	0.8540	0.8447
6.917	7	0.8435	0.8379
7.917	8	0.8396	0.8376
8.917	9	0.8386	0.8176
9.917	10	0.8407	0.8276
10.92	11	0.7665	0.7205

(i) i.
```
. use grouped, clear
. gen obs_rate = 1000*d/y
. gen exp_rate = 1000*d_star/y
. gen excess_rate = 1000*(d-d_star)/y
. list start end d d_star y obs_rate exp_rate excess_rate
```

	start	end	d	d_star	y	obs_rate	exp_rate	excess~e
1.	0	1	151	123.7	5257.0	28.72	23.54	5.19
2.	1	2	329	114.4	4864.8	67.63	23.51	44.12
3.	2	3	287	98.3	4242.2	67.65	23.18	44.48
4.	3	4	211	84.0	3717.6	56.76	22.60	34.16
5.	4	5	166	73.3	3271.5	50.74	22.42	28.32
6.	5	6	138	64.4	2870.3	48.08	22.45	25.63
7.	6	7	105	58.0	2529.7	41.51	22.92	18.58
8.	7	8	75	52.0	2196.1	34.15	23.70	10.46
9.	8	9	68	46.4	1892.3	35.94	24.51	11.42
10.	9	10	50	39.1	1588.0	31.49	24.63	6.86

ii. The excess mortality rate is similar to the cause-specific mortality rate shown in question 111.



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

	end	age	_age
1.	1	55.52238	55.52238
2.	2	55.06794	56.06775
3.	3	54.11723	56.11679
4.	4	53.29431	56.29431
5.	5	52.42989	56.42989
6.	6	51.86148	56.86148
7.	7	51.29445	57.29445
8.	8	50.88644	57.88644
9.	9	50.52067	58.52067
10.	10	49.91178	58.91178

202. **Life-table estimates of cause-specific survival using `ltable` and `strs`**

(a)
```
. use melanoma if stage==1, clear
. // Estimate cause-specific survival using -strs-
. stset surv_mm, fail(status==1) id(id) scale(12)
. strs using popmort, br(0(1)20) mergeby(_year sex _age) list(n d w p cp)
```

start	end	n	d	w	p	cp
0	1	5318	71	81	0.9865	0.9865
1	2	5166	228	400	0.9541	0.9413
2	3	4538	202	381	0.9535	0.8975
3	4	3955	138	344	0.9635	0.8648
4	5	3473	100	312	0.9699	0.8387
5	6	3061	80	298	0.9725	0.8157
6	7	2683	56	267	0.9780	0.7977
7	8	2360	35	293	0.9842	0.7851
8	9	2032	34	275	0.9821	0.7710
9	10	1723	16	243	0.9900	0.7633

```
[output omitted]
```

(b)
```
. // Estimate cause-specific survival using -ltable-
. generate csr_fail=0
. replace csr_fail=1 if status==1
. ltable surv_mm csr_fail, interval(12)
```

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
0	12	5318	71	81	0.9865	0.0016	0.9831	0.9893
12	24	5166	228	400	0.9413	0.0033	0.9344	0.9474
24	36	4538	202	381	0.8975	0.0043	0.8887	0.9057
36	48	3955	138	344	0.8648	0.0050	0.8546	0.8743
48	60	3473	100	312	0.8387	0.0055	0.8276	0.8491
60	72	3061	80	298	0.8157	0.0059	0.8037	0.8269
72	84	2683	56	267	0.7977	0.0062	0.7852	0.8097
84	96	2360	35	293	0.7851	0.0065	0.7721	0.7976
96	108	2032	34	275	0.7710	0.0068	0.7573	0.7841
108	120	1723	16	243	0.7633	0.0070	0.7492	0.7768

```
[output omitted]
```

As expected, both commands give identical estimates of cause-specific survival.

(c) Both cause-specific survival and relative survival estimate the same underlying theoretical quantity (net survival) and should therefore be similar, which they are.

	start	end	CSR	RSR
1.	0	1	0.9865	0.9947
2.	1	2	0.9413	0.9519
3.	2	3	0.8975	0.9109
4.	3	4	0.8648	0.8808
5.	4	5	0.8387	0.8564
6.	5	6	0.8157	0.8350
7.	6	7	0.7977	0.8196
8.	7	8	0.7851	0.8111
9.	8	9	0.7710	0.8018
10.	9	10	0.7633	0.7964
11.	10	11	0.7533	0.7843
12.	11	12	0.7422	0.7704
13.	12	13	0.7406	0.7736
14.	13	14	0.7369	0.7732
15.	14	15	0.7333	0.7694
16.	15	16	0.7302	0.7664
17.	16	17	0.7203	0.7585
18.	17	18	0.7175	0.7596
19.	18	19	0.7132	0.7599
20.	19	20	0.7132	0.7382

The following Stata commands were used.

```
use melanoma if stage==1, clear

// Estimate cause-specific survival using -strs-
stset surv_mm, fail(status==1) id(id) scale(12)
strs using popmort, br(0(1)20) mergeby(_year sex _age) list(n d w p cp) savgroup(csr,replace)

// Estimate relative survival using -strs-
stset surv_mm, fail(status==1 2) id(id) scale(12)
strs using popmort, br(0(1)20) mergeby(_year sex _age) list(n d w cr) savgroup(rsr,replace)

use rsr, clear
gen SE_RSR=se_cp/cp_e2
rename cr RSR
keep start RSR SE_RSR
save rsr, replace

use csr, clear
rename cp CSR
rename se_cp SE_CSR
keep start end CSR SE_CSR
save csr, replace

merge 1:1 start using rsr

format  CSR SE_CSR  RSR SE_RSR %6.4f
list start end CSR RSR
```

203. **Period estimation of relative survival**

First produce period estimates of relative survival by sex.

```
. use melanoma, clear
. keep if stage==1

. /* stset the data with time since diagnosis as the timescale */
. /* restrict person-time at risk to that within the period window (01jan1994-31dec1995) */
. stset exit, enter(time mdy(1,1,1994)) exit(time mdy(12,31,1995)) ///
>   origin(dx) f(status==1 2) id(id) scale(365.24)

. strs using popmort, br(0(1)10) mergeby(_year sex _age) ///
>   by(sex) list(n d p r cr_e2 se_cp)

-> sex = Male
+-------------------------------------------------------------+
  | start   end     n    d        p        r    cr_e2    se_cp |
  |-------------------------------------------------------------|
  |     0     1   307    9   0.9618   0.9900   0.9900   0.0125 |
  |     1     2   445   22   0.9260   0.9557   0.9462   0.0186 |
  |     2     3   407   18   0.9342   0.9634   0.9115   0.0219 |
  |     3     4   377   18   0.9285   0.9582   0.8734   0.0244 |
  |     4     5   340   13   0.9440   0.9705   0.8476   0.0258 |
  |-------------------------------------------------------------|
  |     5     6   340   15   0.9328   0.9586   0.8125   0.0270 |
  |     6     7   320    7   0.9679   0.9939   0.8076   0.0274 |
  |     7     8   321    9   0.9589   0.9865   0.7967   0.0277 |
  |     8     9   273    7   0.9620   0.9895   0.7883   0.0281 |
  |     9    10   234    8   0.9468   0.9737   0.7676   0.0288 |
  +-------------------------------------------------------------+

-> sex = Female
  +-------------------------------------------------------------+
  | start   end     n    d        p        r    cr_e2    se_cp |
  |-------------------------------------------------------------|
  |     0     1   338    8   0.9679   0.9883   0.9883   0.0111 |
  |     1     2   491   16   0.9536   0.9756   0.9642   0.0153 |
  |     2     3   482   14   0.9556   0.9784   0.9434   0.0181 |
  |     3     4   449   23   0.9228   0.9438   0.8905   0.0216 |
  |     4     5   414   14   0.9493   0.9679   0.8619   0.0231 |
  |-------------------------------------------------------------|
  |     5     6   410    8   0.9708   0.9890   0.8524   0.0238 |
  |     6     7   421   11   0.9613   0.9810   0.8362   0.0244 |
  |     7     8   404    2   0.9929   1.0146   0.8484   0.0245 |
  |     8     9   353    2   0.9916   1.0151   0.8612   0.0247 |
  |     9    10   312    3   0.9846   1.0051   0.8655   0.0251 |
  +-------------------------------------------------------------+
```

Now, re-stset the data and estimate relative survival for the complete cohort.

```
stset exit, enter(time dx) origin(dx) failure(status==1 2) id(id) scale(365.24)
strs using popmort, br(0(1)10) mergeby(_year sex _age) ///
by(sex) list(n d w p r cr_e2 se_cp)
```

The 10-year cumulative relative survival for males is now 0.7616 and 0.8239 for females.

204. **Period estimation of relative survival**

(a) i. The period estimate should be higher (the cohort estimate will be weighted down by patients diagnosed in the past).

ii. The period estimate should be a better predictor of the survival of newly diagnosed patients.

(b)
```
. use melanoma if stage==1 & yydx<=1983, clear
. stset exit, origin(dx) entry(dx) fail(status==1 2) id(id) ///
  exit(time mdy(12,31,1983)) scale(365.24)

. strs using popmort if (yydx <=1983), br(0(1)15) mergeby(_year sex _age)

        failure _d:  status == 1 2
  analysis time _t:  (exit-origin)/365.24
            origin:  time dx
  enter on or after:  time dx
  exit on or before:  time mdy(12,31,1983)
                id:  id

No late entry detected - p is estimated using the actuarial method
```

start	end	n	d	w	p	p_star	r	cr_e2	lo_cr_e2	hi_cr_e2
0	1	1890	51	250	0.9711	0.9789	0.9921	0.9921	0.9829	0.9991
1	2	1589	110	294	0.9237	0.9783	0.9442	0.9367	0.9198	0.9515
2	3	1185	105	217	0.9025	0.9786	0.9222	0.8638	0.8404	0.8851
3	4	863	46	158	0.9413	0.9789	0.9616	0.8307	0.8036	0.8555
4	5	659	30	148	0.9487	0.9784	0.9697	0.8055	0.7750	0.8337

The estimated 5-year RSR is 0.8055.

(c) We expect this estimate to be higher because we are excluding two years where survival is lower.

```
. strs using popmort if (1977 <= yydx) & (yydx <=1983), br(0(1)15) mergeby(_year sex _age)
```

start	end	n	d	w	p	p_star	r	cr_e2	lo_cr_e2	hi_cr_e2
0	1	1579	39	249	0.9732	0.9787	0.9944	0.9944	0.9845	1.0017
1	2	1291	79	292	0.9310	0.9778	0.9521	0.9468	0.9283	0.9624
2	3	920	76	217	0.9063	0.9785	0.9263	0.8770	0.8505	0.9005
3	4	627	27	158	0.9507	0.9787	0.9714	0.8519	0.8212	0.8796
4	5	442	16	148	0.9565	0.9780	0.9781	0.8332	0.7977	0.8654

The estimated 5-year RSR is now 0.8332.

(d)
```
. use melanoma if stage==1, clear
. stset exit, origin(dx) enter(time mdy(1,1,1983)) exit(time mdy(12,31,1983))
        f(status==1 2) id(id) scale(365.24)
. strs using popmort, br(0(1)15) mergeby(_year sex _age)

        failure _d:  status == 1 2
  analysis time _t:  (exit-origin)/365.24
            origin:  time dx
 enter on or after:  time mdy(1,1,1983)
 exit on or before:  time mdy(12,31,1983)
                id:  id
```

start	end	n	d	y	p	p_star	r	cr_e2	lo_cr_e2	hi_cr_e2
0	1	557	9	272.4	0.9675	0.9779	0.9894	0.9894	0.9597	1.0052
1	2	533	14	272.9	0.9500	0.9786	0.9708	0.9605	0.9204	0.9880
2	3	402	21	189.0	0.8948	0.9807	0.9125	0.8764	0.8184	0.9220
3	4	321	11	148.3	0.9285	0.9774	0.9500	0.8326	0.7654	0.8877
4	5	309	9	156.3	0.9441	0.9775	0.9658	0.8041	0.7322	0.8648

The period estimate of the 5-year relative survival is 0.8041.

(e)
```
. use melanoma if stage==1, clear
. stset exit, origin(dx) enter(time mdy(1,1,1982)) exit(time mdy(12,31,1983))
        f(status==1 2) id(id) scale(365.24)
. strs using popmort, br(0(1)15) mergeby(_year sex _age)
```

start	end	n	d	y	p	p_star	r	cr_e2	lo_cr_e2	hi_cr_e2
0	1	814	20	563.7	0.9651	0.9790	0.9858	0.9858	0.9668	0.9983
1	2	739	35	480.6	0.9298	0.9788	0.9499	0.9365	0.9062	0.9604
2	3	582	39	351.9	0.8951	0.9791	0.9142	0.8561	0.8143	0.8918
3	4	488	18	312.6	0.9440	0.9781	0.9651	0.8263	0.7800	0.8667
4	5	440	14	294.9	0.9536	0.9779	0.9752	0.8058	0.7563	0.8497

The period estimate of the 5-year relative survival corresponding to the new analysis window is 0.8058.

(f)
```
. use melanoma if stage==1, clear
. stset exit, origin(dx) entry(dx) fail(status==1 2) id(id) scale(365.24)
. strs using popmort if(yydx==1983), br(0(1)15) mergeby(_year sex _age)
```

start	end	n	d	w	p	p_star	r	cr_e2	lo_cr_e2	hi_cr_e2
0	1	254	10	0	0.9606	0.9782	0.9821	0.9821	0.9488	1.0005
1	2	244	13	0	0.9467	0.9786	0.9675	0.9501	0.9056	0.9809
2	3	231	11	0	0.9524	0.9800	0.9718	0.9233	0.8718	0.9620
3	4	220	10	0	0.9545	0.9799	0.9741	0.8994	0.8424	0.9443
4	5	210	11	0	0.9476	0.9780	0.9690	0.8715	0.8093	0.9224

The actual 5-year relative survival for patients diagnosed in 1983 is 0.8715.

(g)
```
. strs using popmort if(yydx==1984), br(0(1)15) mergeby(_year sex _age)
```

start	end	n	d	w	p	p_star	r	cr_e2	lo_cr_e2	hi_cr_e2
0	1	255	7	0	0.9725	0.9805	0.9919	0.9919	0.9621	1.0065
1	2	248	17	0	0.9315	0.9802	0.9503	0.9426	0.8978	0.9738
2	3	231	13	0	0.9437	0.9799	0.9631	0.9078	0.8552	0.9479
3	4	218	7	0	0.9679	0.9788	0.9889	0.8977	0.8410	0.9423
4	5	211	12	0	0.9431	0.9793	0.9631	0.8646	0.8025	0.9155

The actual 5-year relative survival for patients diagnosed in 1984 is 0.8646.

(h) The estimates of the 5-year relative survival and confidence intervals are summarized in the table.

Method	Estimate	95% C.I
Cohort (Ederer II, 1975-1983)	0.8055	(0.7750, 0.8337)
Cohort (Ederer II, 1977-1983)	0.8332	(0.7977, 0.8654)
Period (Ederer II, Jan83 - Dec83)	0.8041	(0.7322, 0.8648)
Period (Ederer II, Jan82 - Dec83)	0.8058	(0.7563, 0.8497)
Actual (Diagnosed in 1983)	0.8715	(0.8093, 0.9224)
Actual (Diagnosed in 1984)	0.8646	(0.8025, 0.9155)

Table 1: Comparison of the 5-year relative survival estimates

Yes, period analysis provide a more accurate prediction of the future prognosis of recently diagnosed patients (i.e., the period estimates are more similar to the actual survival estimates than the cohort estimates). However, the confidence intervals for the period estimates are wider than the confidence intervals for the cohort estimates since we have imposed a restriction to what information is included in the calculations (i.e, fewer events are included in the analysis).

(i) The period estimate of relative survival will be equal to the cohort estimate.

(j) See figure below.

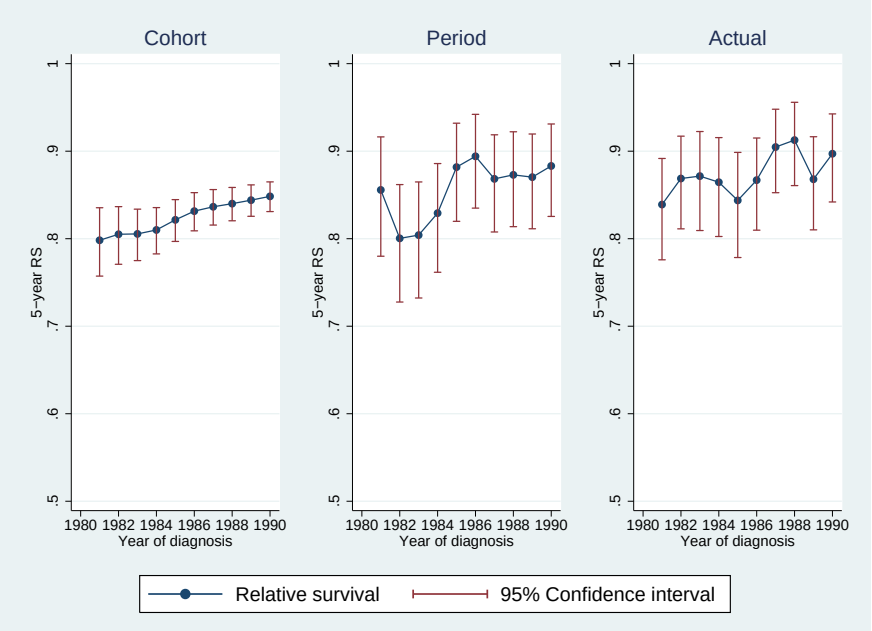


Figure 31: Comparison of estimates of 5-year cohort, period and actual relative survival for different years of diagnosis.

210. **Modelling relative survival**

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

Generalized linear models                         No. of obs      =        80
Optimization     : ML                             Residual df     =        70
                                                  Scale parameter =         1
Deviance         =   76.0143154                   (1/df) Deviance =  1.085919
Pearson          =  75.40696725                   (1/df) Pearson  =  1.077242

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = log(u-d*)               [Relative survival]

                                                  AIC             =  5.460814
Log likelihood   = -208.4325474                   BIC             = -230.7275

------------------------------------------------------------------------------
             |                 OIM
           d |      exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         end |
           2 |   6.764551   2.033588     6.36   0.000     3.752728    12.19357
           3 |   7.239822   2.180328     6.57   0.000     4.012195    13.06393
           4 |   5.423029   1.677824     5.46   0.000     2.957262    9.944753
           5 |   4.660075    1.47575     4.86   0.000     2.505156     8.66864
             |
         sex |
      Female |   .5644476   .0547487    -5.90   0.000     .4667251    .6826312
             |
    year8594 |
Diagnosed 85-94 |     .62682   .0611382    -4.79   0.000     .5177488    .7588685
             |
      agegrp |
       45-59 |   1.378033   .1724529     2.56   0.010     1.078293    1.761094
       60-74 |    1.89259   .2426843     4.98   0.000     1.472001    2.433353
         75+ |   3.239937   .5557873     6.85   0.000     2.314831    4.534756
             |
       _cons |   .0066668   .0020381   -16.39   0.000     .0036619    .0121376
        ln(y) |          1  (exposure)
------------------------------------------------------------------------------
```

(a) Excess mortality was much lower during the first year following diagnosis. This is not the usual pattern. For most cancer sites, excess mortality is highest during the first year. Localised skin melanoma, however, is not immediately fatal. A possible explanation for the observed pattern is that these patients were diagnosed with what was classified as localised skin melanoma, although if the primary tumour was excised and patient died due to the melanoma then it is highly probable that micrometastases were present at the time of diagnosis. These micrometastases were, however, undetectable at diagnosis and it took approximately one year for tumours to form in other organs leading to the death of the patient.

(b) A summary of estimated hazard ratios and standard errors is shown in the table below. Note that the models we fitted in exercise 120 for cause-specific mortality were for the first 10 years of follow-up whereas the model we fitted in the previous part was for 5 years. I have also included the results for the excess mortality model for 10 years in the table below.

Variable	Cox	Poisson	Excess5	Excess10
sex	0.588814	0.587547	0.564448	0.605145
	0.038538	0.038456	0.054749	0.052059
year8594	0.716884	0.722411	0.626820	0.636971
	0.047445	0.047813	0.061138	0.056469
agegrp				
1	1.326397	1.327795	1.378033	1.226416
	0.124911	0.125042	0.172453	0.130557
2	1.857323	1.862376	1.892590	1.576938
	0.168787	0.169244	0.242684	0.179360
3	3.372652	3.400287	3.239937	2.874281
	0.352227	0.355140	0.555787	0.453919

Cox: Cox model for cause-specific mortality, 10 year follow-up
Poisson: Poisson model for cause-specific mortality, 10 years
Excess5: Poisson model for excess mortality, 5 year follow-up
Excess10: Poisson model for excess mortality, 10 year follow-up

The hazard ratios from each model represent the same underlying concept, a ratio of net mortality rates. All models assume proportional hazards. We would expect the hazard ratios to be similar and they are. There will be differences between the cause-specific mortality models and the excess mortality models due to the appropriateness of the underlying assumptions (i.e., accuracy of coding cancer as the cause of death and our ability to estimate expected mortality).

(c) `. glm`

```
Generalized linear models                         No. of obs      =        80
Optimization     : ML                             Residual df     =        70
                                                  Scale parameter =         1
Deviance         =  76.01431531                   (1/df) Deviance =  1.085919
Pearson          =   75.4069672                   (1/df) Pearson  =  1.077242

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = log(u-d*)               [Relative survival]

                                                  AIC             =  5.460814
Log likelihood   = -208.4325474                   BIC             = -230.7275
```

d	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
end						
2	1.911696	.3006243	6.36	0.000	1.322483	2.500909
3	1.979597	.3011577	6.57	0.000	1.389338	2.569855
4	1.690654	.3093887	5.46	0.000	1.084264	2.297045
5	1.539031	.3166795	4.86	0.000	.918351	2.159712
sex						
Female	-.5719077	.0969952	-5.90	0.000	-.7620148	-.3818005
year8594						
Diagnosed 85-94	-.4670959	.0975371	-4.79	0.000	-.6582651	-.2759268
agegrp						
45-59	.3206573	.1251442	2.56	0.010	.075379	.5659355
60-74	.6379465	.1282286	4.98	0.000	.386623	.88927
75+	1.175554	.1715426	6.85	0.000	.8393367	1.511771
_cons	-5.010609	.3057016	-16.39	0.000	-5.609774	-4.411445
ln(y)	1	(exposure)				

This is the exact same model, except the β (log RER) estimates are now presented rather than $\exp(\beta)$ (RER). The standard errors and confidence intervals will be different but the test statistic (z) and p-values are the same. Note that if you exponentiate the confidence limits you will get the limits for the excess hazard ratio as shown in part (a).

(d) In order to model non-proportional excess hazards by age we include an age*follow-up interaction term in the model.

```
. glm d i.sex i.year8594 i.end##i.agegrp, ///
fam(pois) link(rs d_star) lnoff(y) eform

Generalized linear models                         No. of obs      =        80
Optimization     : ML                             Residual df     =        58
                                                  Scale parameter =         1
Deviance         =  70.61626656                   (1/df) Deviance =  1.217522
Pearson          =  69.92575924                   (1/df) Pearson  =  1.205617

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = log(u-d*)               [Relative survival]

                                                  AIC             =  5.693338
Log likelihood   =  -205.733523                   BIC             = -183.5413

------------------------------------------------------------------------------
             |                 OIM
           d |      exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
      Female |   .5672839   .0551669    -5.83   0.000     .4688386    .6864004
             |
    year8594 |
Diagnosed 85-94 |   .6213308   .0608958    -4.86   0.000     .5127406    .7529185
             |
         end |
           2 |    5.53995   2.910466     3.26   0.001     1.978422    15.51289
           3 |   6.608943   3.450353     3.62   0.000      2.37543    18.38746
           4 |   5.398605   2.872583     3.17   0.002     1.902653    15.31805
           5 |   5.536886    2.95309     3.21   0.001     1.946608    15.74899
             |
      agegrp |
       45-59 |   1.643743   1.058438     0.77   0.440     .4652954     5.80683
       60-74 |   1.310152   1.248959     0.28   0.777     .2022448    8.487237
         75+ |   1.175077   3.191062     0.06   0.953     .0057349    240.7717
             |
  end#agegrp |
     2#45-59 |   .9584949   .6565512    -0.06   0.951     .2503413     3.66984
     2#60-74 |   1.799522   1.764561     0.60   0.549     .2633256    12.29763
       2#75+ |   3.535409   9.651144     0.46   0.644     .0167797    744.8931
     3#45-59 |   .9032407   .6163441    -0.15   0.881     .2371206    3.440627
     3#60-74 |   1.556089   1.525173     0.45   0.652     .2278998    10.62489
       3#75+ |   2.555255   6.986935     0.34   0.732     .0120216    543.1354
     4#45-59 |   .7660645   .5370607    -0.38   0.704     .1938733    3.027001
     4#60-74 |   1.412114   1.405607     0.35   0.729     .2007198    9.934574
       4#75+ |   2.415016   6.646823     0.32   0.749      .010969    531.7075
     5#45-59 |    .642165   .4576746    -0.62   0.534     .1588512    2.595988
     5#60-74 |   .7916966    .819203    -0.23   0.821     .1041798    6.016365
       5#75+ |   2.623852   7.226541     0.35   0.726     .0118736    579.8232
             |
       _cons |   .0070366   .0035006    -9.96   0.000      .002654    .0186563
        ln(y) |          1  (exposure)
------------------------------------------------------------------------------

. lrtest Grouped

Likelihood-ratio test                             LR chi2(12) =      5.40
(Assumption: Grouped nested in .)                 Prob > chi2 =    0.9433
```

Age has 4 levels and follow-up 5 levels so this model uses $(4 \times 5) - 1 = 19$ parameters to model the joint effect of age and follow-up. The previous (main effects) model used only $(4-1)+(5-1) = 7$ parameters to model the joint effect of age and follow-up. The interaction model therefore involves estimating an additional 12 parameters. We can use the likelihood ratio test to determine whether these 12 parameters are statistically significant. If they are, then we conclude that the excess hazards are not proportional across age groups.

The change in deviance (i.e. change in $-2\times$ log likelihood) is $76.01 - 70.62 = 5.39$. The change in the number of residual degrees of freedom (equivalent to the number of parameters excluded from the model) is $70 - 58 = 12$. Under the null hypothesis that both models describe the data equally well, the test statistic (change in deviance) will follow a χ^2 distribution with 12 degrees of freedom. The critical value at the $\alpha = 0.10$ level for a χ^2_{12} variate is 18.5. Since the test statistic is considerably less than 18.5 we conclude that there is no evidence of non-proportional hazards across age groups.

(e)

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

est table Grouped Individual
```

```
----------------------------------------
    Variable |  Grouped     Individual
-------------+--------------------------
         end |
           2 |  1.9116958    1.9149755
           3 |  1.9795967    1.9637888
           4 |  1.6906545    1.6786063
           5 |  1.5390315    1.5539051
             |
         sex |
           2 | -.57190767   -.59569368
             |
    year8594 |
           1 | -.46709592   -.46506336
             |
      agegrp |
           1 |  .32065726    .32554278
           2 |  .63794651    .65400744
           3 |   1.175554    1.1427964
             |
       _cons | -5.0106094   -5.0046128
----------------------------------------
```

The estimates change slightly. Estimating a standard Poisson regression model (with logarithmic link and offset $\ln(y_j)$) gives identical estimates for both individual and collapsed data. Modelling excess mortality based on collapsed data, however, leads to slightly different estimates to those obtained from subject-band observations since the expected number of deaths d^* varies within each covariate pattern because we are grouping across ages.

(f)
```
. ml model lf esteve (d=i.end sex year8594 i.agegrp)
. ml maximize, eform("RER")

                                                  Number of obs   =      22450
                                                  Wald chi2(9)    =     163.72
Log likelihood = -3570.6796                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
           d |        RER   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         end |
           2 |   6.786736   2.018467     6.44   0.000     3.788803    12.15681
           3 |   7.126238    2.12553     6.58   0.000     3.971678    12.78635
           4 |   5.358053   1.642111     5.48   0.000     2.938562    9.769655
           5 |   4.729879   1.478354     4.97   0.000     2.563312    8.727674
             |
         sex |
      Female |     .55118   .0533262    -6.16   0.000     .4559744    .6662641
             |
    year8594 |
Diagnosed 85-94 |   .6280952   .0608957    -4.80   0.000     .5193964    .7595424
             |
      agegrp |
       45-59 |   1.384782   .1737737     2.59   0.009     1.082844    1.770913
       60-74 |   1.923233   .2449984     5.13   0.000     1.498298    2.468684
         75+ |   3.135519   .5420519     6.61   0.000     2.234386    4.400081
             |
       _cons |    .006707   .0020304   -16.53   0.000     .0037055    .0121396
------------------------------------------------------------------------------
```

(g) `. glm ns i.end i.sex i.year8594 i.agegrp, fam(bin n_prime) link(ht p_star) eform`

```
Generalized linear models                         No. of obs      =        80
Optimization     : ML                             Residual df     =        70
                                                  Scale parameter =         1
Deviance         =  75.70484452                   (1/df) Deviance =  1.081498
Pearson          =  74.95252964                   (1/df) Pearson  =   1.07075

Variance function: V(u) = u*(1-u/n_prime)         [Binomial]
Link function    : g(u) = log(-log(u/ps))         [Hakulinen-Tenkanen]

                                                  AIC             =  5.378895
Log likelihood   = -205.1558173                   BIC             =  -231.037

------------------------------------------------------------------------------
             |                 OIM
          ns |      exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         end |
           2 |   6.687235   1.996094     6.37   0.000     3.725364    12.00396
           3 |   7.106638   2.125751     6.56   0.000     3.954132    12.77254
           4 |   5.328617   1.637342     5.44   0.000      2.91785    9.731193
           5 |   4.587648   1.443228     4.84   0.000     2.476342    8.499034
             |
         sex |
      Female |   .5648933   .0548761    -5.88   0.000     .4669573    .6833696
             |
    year8594 |
Diagnosed 85-94 |   .6287558   .0614404    -4.75   0.000     .5191643    .7614813
             |
      agegrp |
       45-59 |    1.38386   .1732513     2.59   0.009     1.082746    1.768714
       60-74 |   1.894699   .2430729     4.98   0.000     1.473461    2.436361
         75+ |   3.193153    .550401     6.74   0.000     2.277712    4.476521
             |
       _cons |   .0067489     .00205   -16.46   0.000     .0037212    .0122402
------------------------------------------------------------------------------
```

```
. est table Grouped Individual Esteve Hakulinen, eform equations(1) ///
> b(%9.6f) modelwidth(10) title("Excess hazard ratios for various models")
Excess hazard ratios for various models
-------------------------------------------------------------
    Variable |  Grouped     Individual     Esteve     Hakulinen
-------------+-----------------------------------------------
         end |
           2 |   6.764551     6.786773     6.786735     6.687229
           3 |   7.239822     7.126276     7.126238     7.106632
           4 |   5.423029     5.358083     5.358053     5.328613
           5 |   4.660075     4.729905     4.729879     4.587643
             |
         sex |
           2 |   0.564448     0.551180     0.551180     0.564893
             |
    year8594 |
           1 |   0.626820     0.628095     0.628095     0.628756
             |
      agegrp |
           1 |   1.378033     1.384782     1.384782     1.383860
           2 |   1.892590     1.923233     1.923233     1.894699
           3 |   3.239937     3.135524     3.135519     3.193154
             |
       _cons |   0.006667     0.006707     0.006707     0.006749
-------------------------------------------------------------
```

(h)
```
use melanoma, clear
stset surv_mm, fail(status==1 2) id(id) scale(12)
strs using popmort, br(0(1)10) mergeby(_year sex _age) by(sex year8594 agegrp stage) save(replace) no
use grouped if end < 6, clear
glm d i.end i.stage i.sex i.year8594 i.agegrp, fam(pois) link(rs d_star) /// lnoff(y) eform
------------------------------------------------------------------------------
             |                 OIM
           d |      exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         end |
           2 |   1.618791   .1227919     6.35   0.000     1.395159    1.878269
           3 |   1.374816   .1206067     3.63   0.000     1.157637     1.63274
           4 |   1.016548   .1088242     0.15   0.878     .8241467    1.253867
           5 |    .822694   .1050734    -1.53   0.126     .6405072    1.056702
             |
       stage |
   Localised |   .7963889   .0777853    -2.33   0.020      .657637    .9644155
    Regional |   5.123679   .5804108    14.42   0.000     4.103532    6.397439
     Distant |   14.38884   1.464181    26.20   0.000     11.78716    17.56477
             |
         sex |
      Female |   .7430209   .0464227    -4.75   0.000     .6573844    .8398131
             |
    year8594 |
Diagnosed 85-94 |   .8016653   .0487215    -3.64   0.000     .7116411    .9030778
             |
      agegrp |
       45-59 |   1.303072   .1066735     3.23   0.001     1.109907    1.529856
       60-74 |   1.658162   .1365256     6.14   0.000     1.411051    1.948548
         75+ |   2.209734   .2392121     7.32   0.000     1.787286    2.732032
             |
       _cons |   .0292023   .0035685   -28.92   0.000     .0229827    .0371051
        ln(y) |          1  (exposure)
------------------------------------------------------------------------------
```

There is strong evidence that the effect of stage is non-proportional (p less than 0.0001).

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

(a) The number of observations in each data set is shown below

```
. use vnarrowint_ind, clear
(Survival data containing individual subject-band observations)
. display "There are " _N " observations in the individual level data"
There are 369512 observations in the individual level data

. use vnarrowint_grp, clear
(Collapsed (or grouped) survival data)
. display "There are " _N " observations in the grouped level data"
There are 1072 observations in the grouped level data
```

(b) The proportional excess hazards model using restricted cubic splines gives

```
. glm, eform

Generalized linear models                         No. of obs      =      1072
Optimization     : ML                             Residual df     =      1061
                                                  Scale parameter =         1
Deviance         =  1225.130012                   (1/df) Deviance =  1.154694
Pearson          =  1159.737352                   (1/df) Pearson  =  1.093061

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = log(u-d*)               [Relative survival]

                                                  AIC             =   3.33551
Log likelihood   =  -1776.83319                   BIC             = -6177.765

------------------------------------------------------------------------------
             |                 OIM
           d |      exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rcs1 |   10.99268   12.02477     2.19   0.028     1.288201    93.80449
        rcs2 |   623.3345   3501.583     1.15   0.252     .0103061    3.77e+07
        rcs3 |     .16017   .3586946    -0.82   0.413     .0019877    12.90667
        rcs4 |    1.24268   .6350355     0.43   0.671     .4564335    3.383305
        rcs5 |    .856384   .1130593    -1.17   0.240       .66114    1.109286
     agegrp2 |   1.346867   .1067849     3.76   0.000     1.153023    1.573299
     agegrp3 |   1.872594   .1483013     7.92   0.000     1.603364    2.187032
     agegrp4 |   2.899927   .2957647    10.44   0.000     2.374503    3.541616
      female |   .5665107   .0339221    -9.49   0.000     .5037777    .6370555
    year8594 |   .6733995   .0397644    -6.70   0.000     .5998037    .7560256
       _cons |   .0242622   .0065903   -13.69   0.000     .0142469     .041318
      ln(y) |          1  (exposure)
------------------------------------------------------------------------------
```

The estimated excess hazard ratios are similar to those obtained from the piecewise model. Even if we have now more accurately modelled the baseline hazard we don't see a great effect on the hazard ratios compared to the model where we used a step function (annual intervals) for the baseline. This is generally true – assuming a step function for the baseline usually gives reasonable estimates for hazard **ratios** even though we do not have a great model for the hazard **rates**.

(c) The graph is shown below

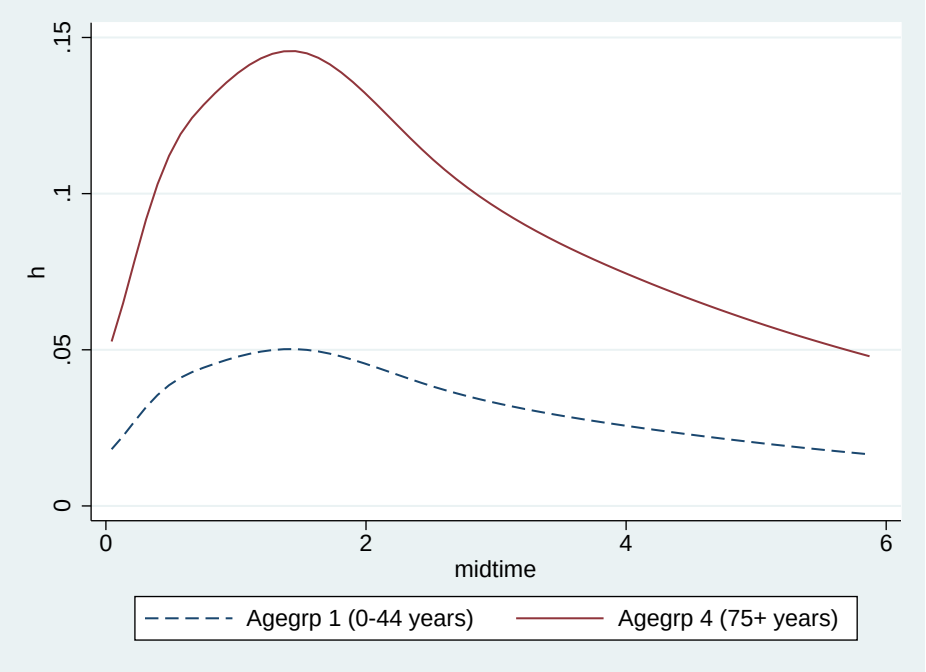


Figure 32: Predicted excess hazards for 2 age groups.

and on the log scale

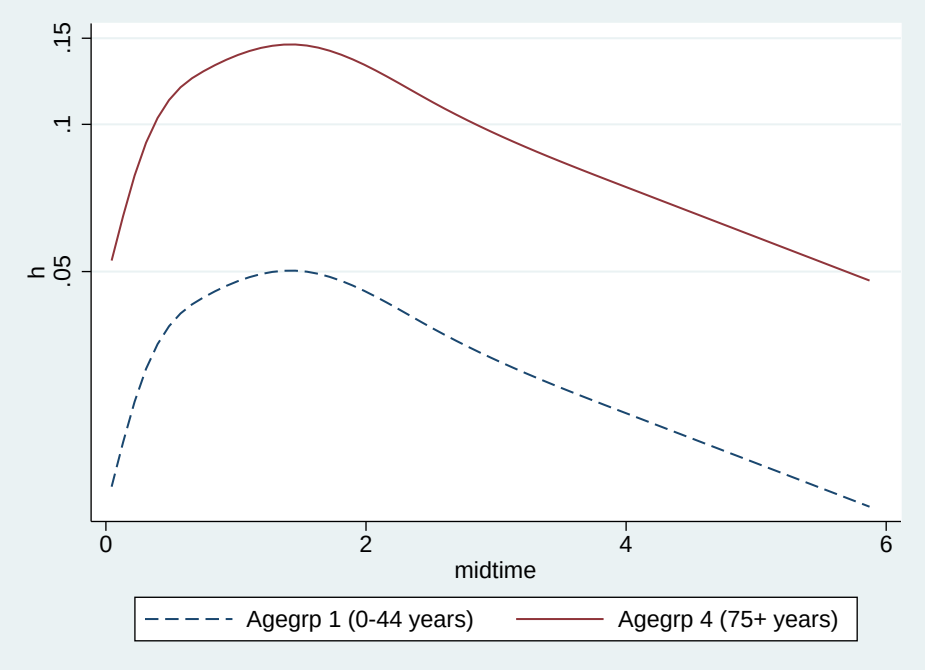


Figure 33: Predicted excess hazards for 2 age groups (log scale).

The lines are parallel as this is a proportional excess hazards model.

(d) The likelihood ratio test gives

```
. lrtest M_sp_peh

Likelihood-ratio test                                 LR chi2(15) =       8.46
(Assumption: M_sp_peh nested in .)                    Prob > chi2 =     0.9042
```

Little evidence of a time dependent effect (P=0.9042).

(e) The time-dependent excess hazard ratios are shown below.

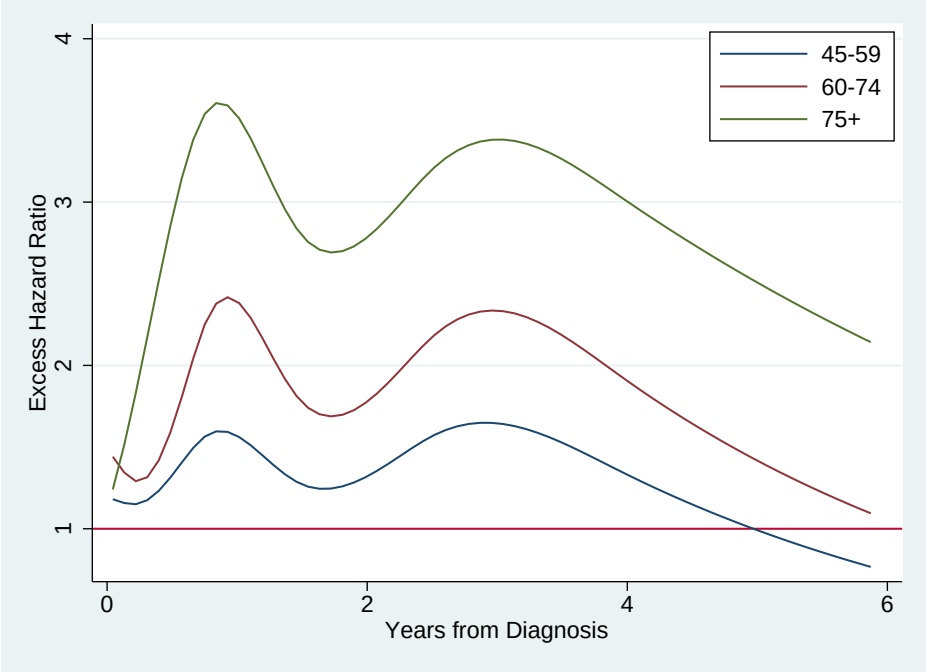


Figure 34: Time-dependent excess hazard ratios

The splines lead to a wavy appearance. Remember this is model is more complex than necessary as there is not evidence of time-dependent hazard ratios.

(f) The graph is shown below

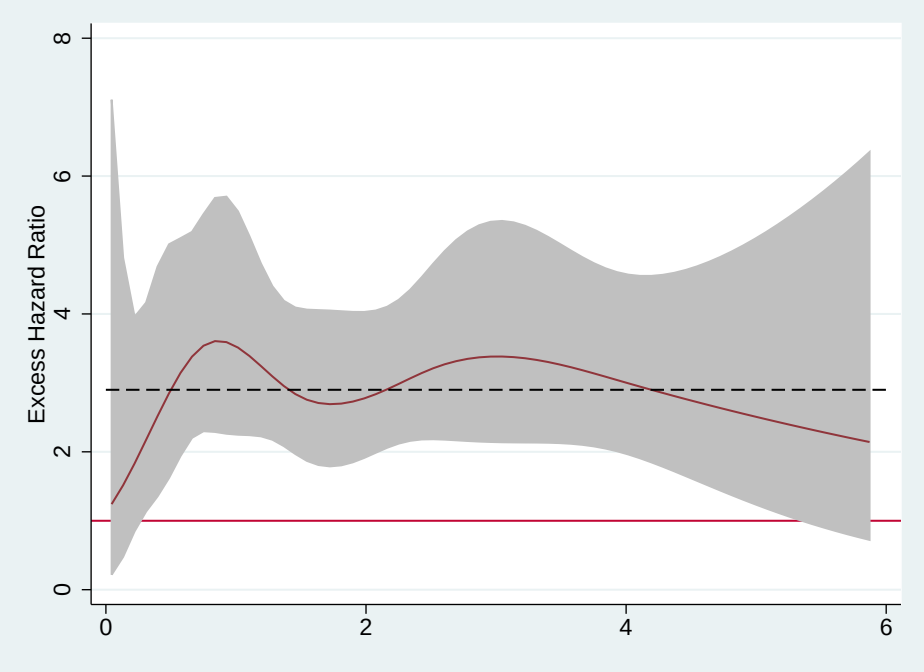


Figure 35: Time-dependent excess hazard ratios (Age Group 4)

With the confidence intervals it appears that proportionality is a reasonable assumption. A reference line at the estimated excess hazard ratio for the proportional excess hazards model has been added.

(g) The predicted survival function is shown below.

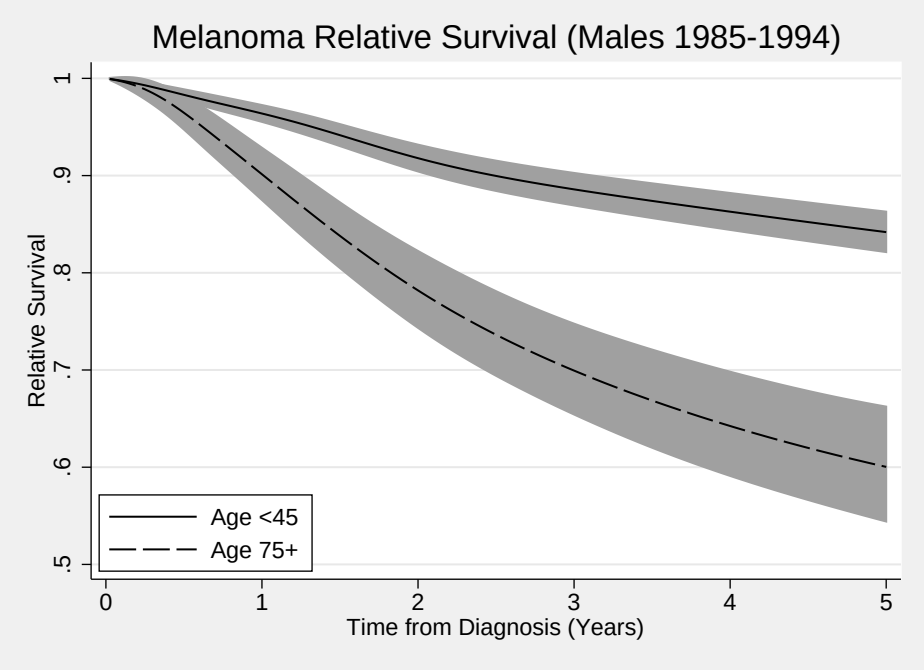


Figure 36: Predicted relative survival curves from Poisson model

(h) Proportional excess hazards model using fractional polynomials is shown below.

```
------------------------------------------------------------------------------
             |                 OIM
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Imidt__1 |   2.301804   .3166709     7.27   0.000      1.68114    2.922468
    Imidt__2 |  -1.038997   .1242679    -8.36   0.000    -1.282557    -.795436
     agegrp2 |   .2976895   .0793019     3.75   0.000     .1422607    .4531184
     agegrp3 |    .626828   .0792083     7.91   0.000     .4715826    .7820733
     agegrp4 |   1.066083   .1018715    10.46   0.000     .8664183    1.265747
      female |  -.5679877    .059873    -9.49   0.000    -.6853366   -.4506387
    year8594 |  -.3942774   .0590268    -6.68   0.000    -.5099677    -.278587
       _cons |  -2.938792   .0736832   -39.88   0.000    -3.083209   -2.794376
           y | (exposure)
------------------------------------------------------------------------------
```

The estimated excess hazard ratios from the fractional polynomial and spline models are compared below

```
. estimates       table M_sp_peh M_mfp_peh, eform ///
>                                 keep(agegrp2 agegrp3 agegrp4 female year8594)

----------------------------------------
    Variable |  M_sp_peh    M_mfp_peh
-------------+--------------------------
     agegrp2 |  1.3468667    1.3467436
     agegrp3 |  1.8725939    1.8716642
     agegrp4 |  2.8999271    2.9039817
      female |  .56651073     .5666646
    year8594 |  .67339957    .67416705
----------------------------------------
```

(i) The fractional polynomial model incorporating time-dependent effects also shows little evidence of non-proportionality.

(j) A comparison of the excess hazard ratios from the spline models using individual level and grouped data is shown below.

```
. estimates       table M_sp_peh M_sp_ind_peh, eform ///
>                                 keep(agegrp2 agegrp3 agegrp4 female year8594)

----------------------------------------
    Variable |  M_sp_peh    M_sp_ind~h
-------------+--------------------------
     agegrp2 |  1.3468667    1.3499776
     agegrp3 |  1.8725939    1.8621516
     agegrp4 |  2.8999271    2.7994952
      female |  .56651073    .55892647
    year8594 |  .67339957    .67556416
----------------------------------------
```

230. **Flexible Parametric Relative Survival Models**

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

```
. stpm2, df(3) scale(hazard) bhazard(rate)

Log likelihood = -8590.0249                     Number of obs   =       7775

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
       _rcs1 |   .8252308   .0249859    33.03   0.000     .7762595    .8742022
       _rcs2 |   .2110309   .0235833     8.95   0.000     .1648085    .2572534
       _rcs3 |   .0631928   .0109672     5.76   0.000     .0416974    .0846882
       _cons |  -1.813097   .0314253   -57.70   0.000     -1.87469   -1.751505
------------------------------------------------------------------------------
```

There are 3 spline variables calculated due to the `df(3)` option.

(b) The predicted relative survival and excess mortality rate functions are shown in Figures 37 and 38.

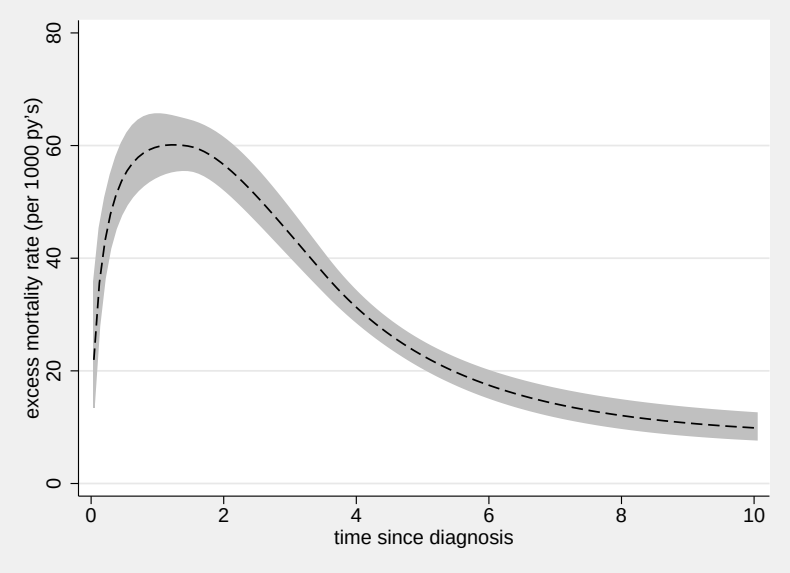


Figure 37: Localised skin melanoma. Predicted relative survival from a flexible parametric model.

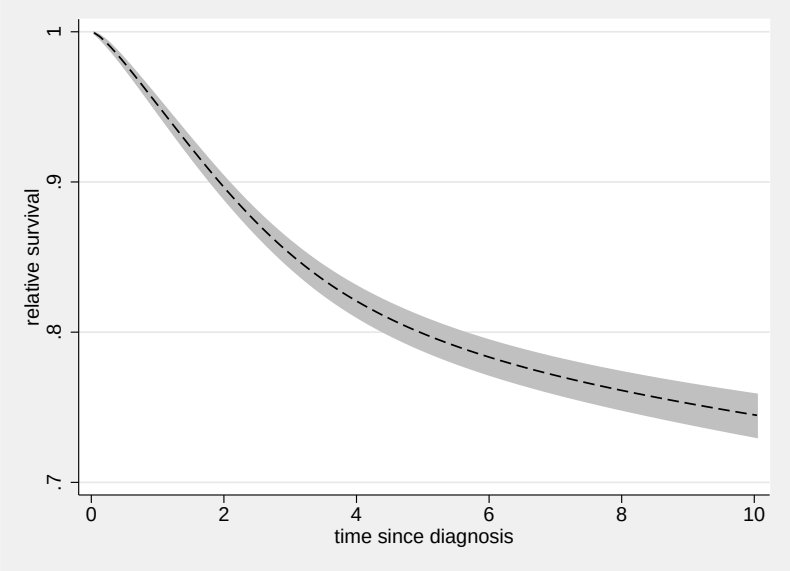


Figure 38: Localised skin melanoma. Predicted excess mortality rate from a flexible parametric model.

(c) The predicted excess hazard rates are shown in Figure 39 and the predicted relative survival functions are shown in Figure 40.

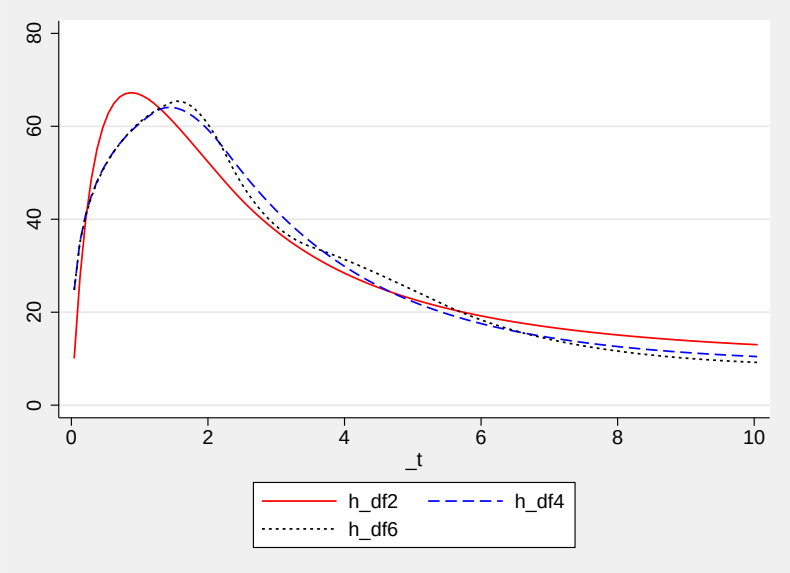


Figure 39: Localised skin melanoma. Predicted hazard functions for 2, 4 and 6 df for baseline.

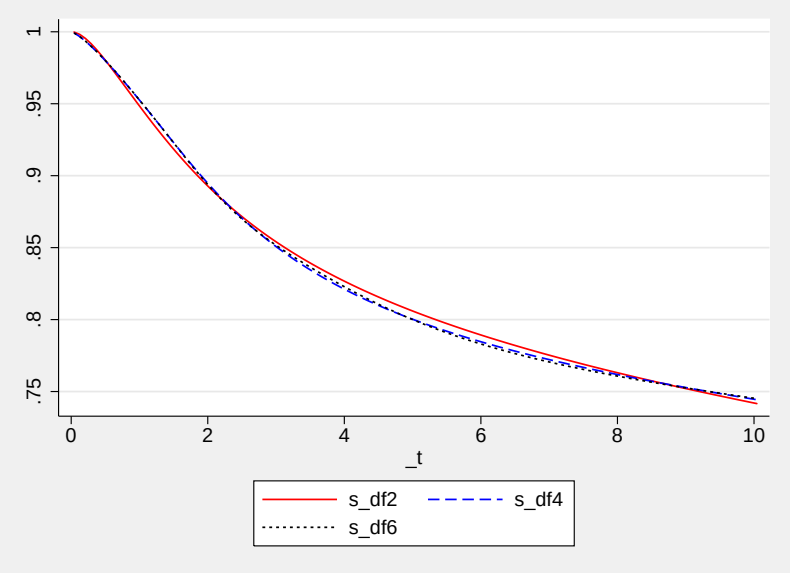


Figure 40: Localised skin melanoma. Predicted relative survival functions for 2, 4 and 6 df for baseline.

The AIC and BIC for each model are shown below

```
. estimates stats df2 df4 df6, n(2773)

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
         df2 |   2773           .   -8598.883      3     17203.77    17221.55
         df4 |   2773           .   -8588.117      5     17186.23    17215.87
         df6 |   2773           .   -8587.141      7     17188.28    17229.78
-----------------------------------------------------------------------------
               Note:  N=2773 used in calculating BIC
```

4 df is selected using both AIC and BIC.

(d) The results of fitting the proportional excess hazards model is shown below.

```
. stpm2 agegrp2 agegrp3 agegrp4 female year8594, bhazard(rate) ///
>    df(3) scale(hazard) eform

Log likelihood = -8485.5808                     Number of obs   =       7775

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
     agegrp2 |   1.285618   .0963736     3.35   0.001      1.10995    1.489089
     agegrp3 |   1.730903   .1312127     7.24   0.000     1.491924    2.008163
     agegrp4 |   2.617489    .262472     9.60   0.000     2.150451    3.185959
      female |   .5817067   .0335759    -9.39   0.000      .519485    .6513811
    year8594 |   .6791693   .0390472    -6.73   0.000     .6067925     .760179
       _rcs1 |   2.315801   .0553603    35.13   0.000       2.2098    2.426887
       _rcs2 |   1.228525   .0273486     9.25   0.000     1.176075    1.283314
       _rcs3 |   1.069712   .0112641     6.40   0.000     1.047861    1.092018
       _cons |   .1946417   .0131462   -24.23   0.000     .1705083    .2221909
------------------------------------------------------------------------------
```

The estimates are broadly similar to the other models.

(e) The excess mortality rates are shown in Figure 41, and Figure 42 shows these on the log scale.

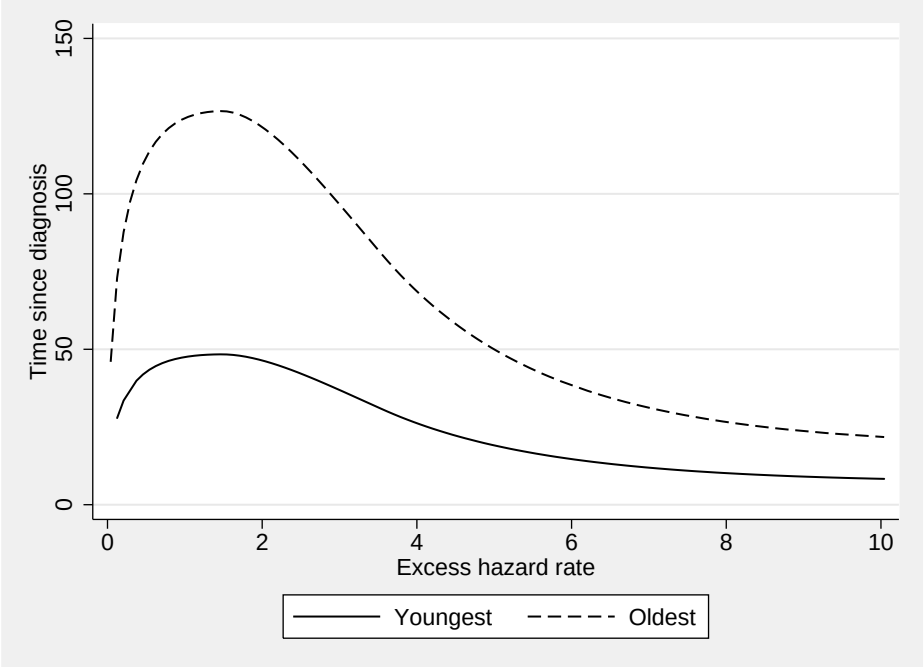


Figure 41: Excess Mortality Rates

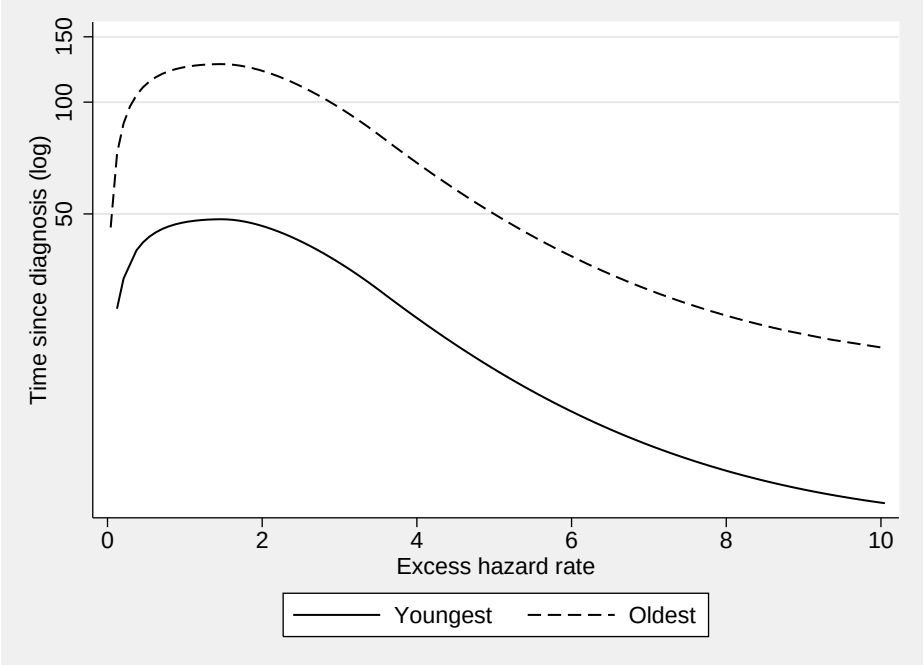


Figure 42: Excess Mortality Rates (log scale)

There is a constant difference between the predicted hazard rates on the log scale as this is a proportional hazards model.

(f) The model with time-dependent effects for age is shown below

```
. stpm2 agegrp2 agegrp3 agegrp4 female year8594, bhazard(rate) df(3) scale(hazard) ///
>            tvc(agegrp2 agegrp3 agegrp4) dftvc(2)

Log likelihood = -8479.6437                     Number of obs   =       7775

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
     agegrp2 |   .2741607   .0798607     3.43   0.001     .1176365    .4306849
     agegrp3 |    .555553   .0812003     6.84   0.000     .3964034    .7147026
     agegrp4 |    .934842    .110683     8.45   0.000     .7179073    1.151777
      female |  -.5457334   .0579363    -9.42   0.000    -.6592864   -.4321804
    year8594 |  -.3873942   .0576354    -6.72   0.000    -.5003576   -.2744309
       _rcs1 |    .851634   .0459294    18.54   0.000      .761614     .941654
       _rcs2 |   .1365924   .0357271     3.82   0.000     .0665685    .2066162
       _rcs3 |   .0697446   .0112343     6.21   0.000     .0477257    .0917635
_rcs_ageg~21 |  -.0210178   .0626366    -0.34   0.737    -.1437832    .1017477
_rcs_ageg~22 |   .0706612   .0480804     1.47   0.142    -.0235747     .164897
_rcs_ageg~31 |  -.0256869   .0665868    -0.39   0.700    -.1561946    .1048208
_rcs_ageg~32 |   .1174402   .0534022     2.20   0.028     .0127739    .2221065
_rcs_ageg~41 |   -.037214   .0856722    -0.43   0.664    -.2051286    .1307005
_rcs_ageg~42 |   .1407585   .0726802     1.94   0.053    -.0016921    .2832092
       _cons |  -1.655974   .0703345   -23.54   0.000    -1.793827   -1.518121
------------------------------------------------------------------------------
```

The predicted excess hazard rates are shown in Figure 43. This is shown on the log scale. Note that as we have introduced time-dependent effects there is no longer a constant difference between the lines.

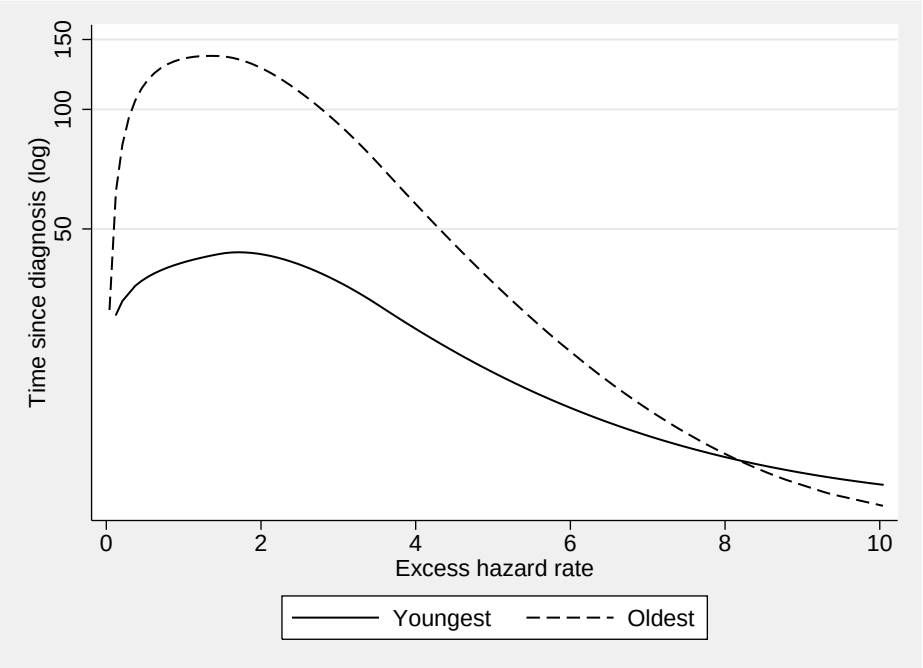


Figure 43: Excess Mortality Rates

(g) The excess mortality rate ratio for age group as a function of time is shown in Figure 44

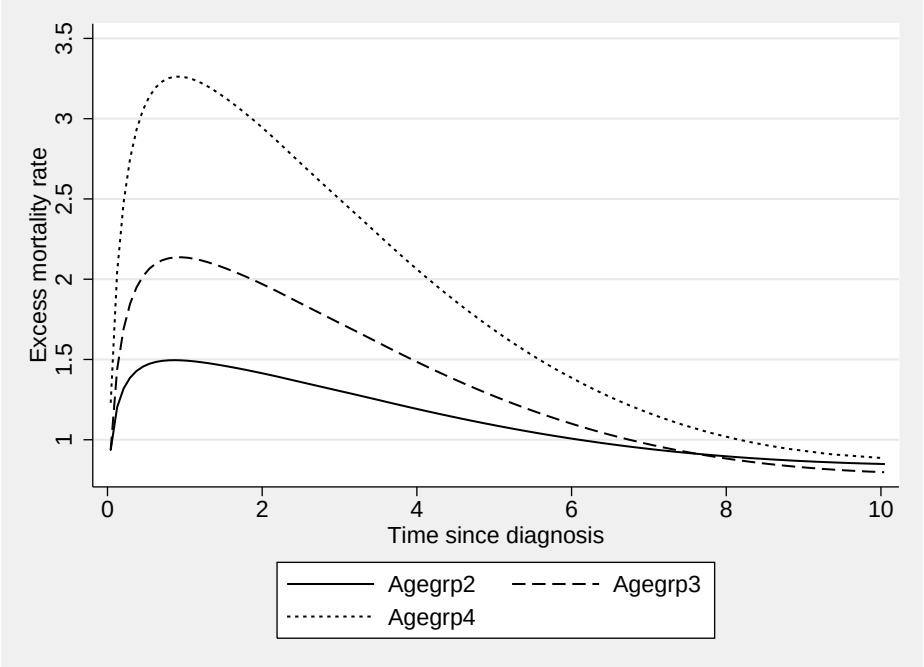


Figure 44: Excess Mortality Rate Ratio

The excess mortality rate ratio for the oldest age group is shown with 95% CI in Figure 45

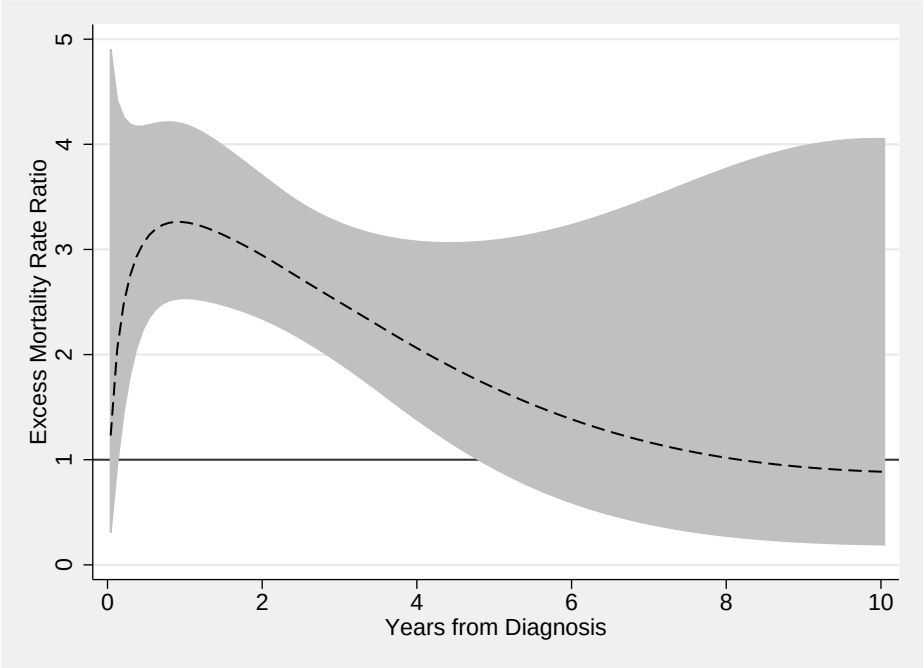


Figure 45: Excess Mortality Rate Ratio

(h) The difference in relative suvival functions is shown in Figure 46. Note that we have had to select the curves for males in 1985-1994 as there are differences in predicted relative survival curves at other levels of the covariates.

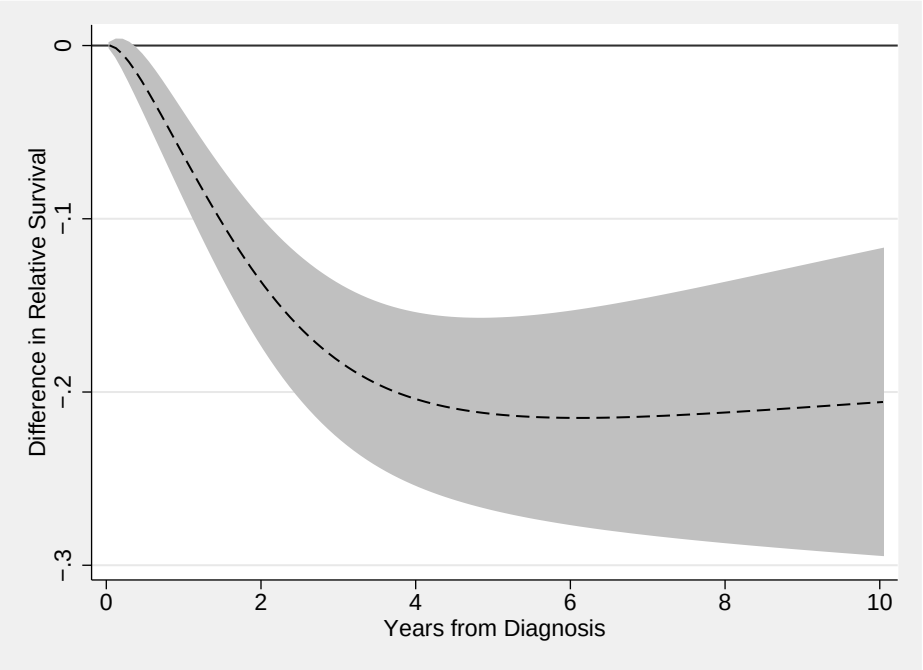


Figure 46: Difference in Relative Survival (oldest - youngest group).

(i) The difference in excess mortality rates is shown in Figure 47.

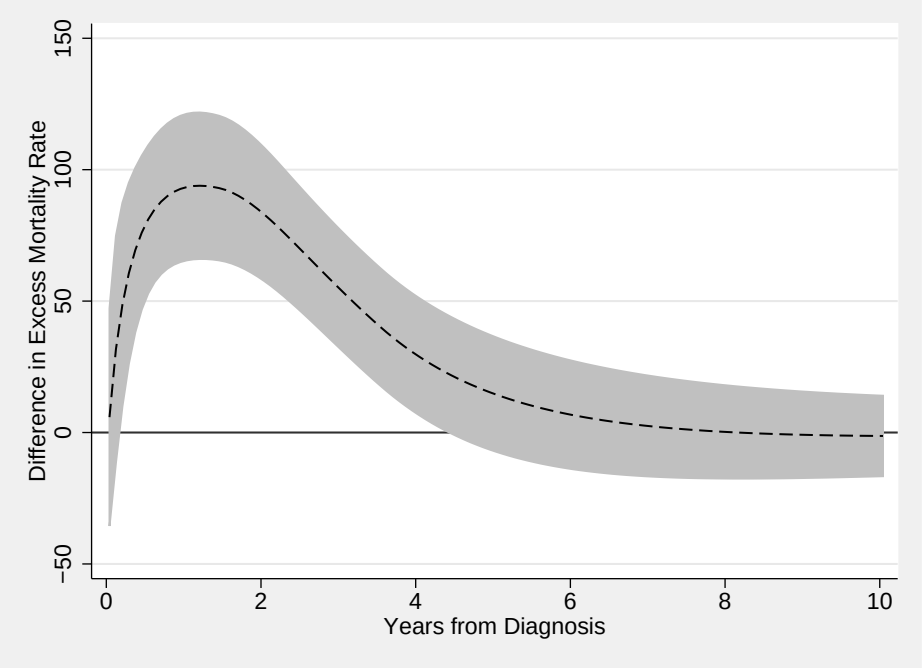


Figure 47: Difference in Excess Mortality Rates (oldest - youngest group).

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

(a)
```
. use colon, clear
(Colon carcinoma, diagnosed 1975-94, follow-up to 1995)

. stset surv_mm, failure(status=1,2) scale(12) id(id) exit(time 60.5)

                id:  id
     failure event:  status == 1 2
obs. time interval:  (surv_mm[_n-1], surv_mm]
 exit on or before:  time 60.5
    t for analysis:  time/12

--------------------------------------------------------------------------------
    15564  total observations
        0  exclusions
--------------------------------------------------------------------------------
    15564  observations remaining, representing
    15564  subjects
     9384  failures in single-failure-per-subject data
 37866.33  total analysis time at risk and under observation
                                             at risk from t =         0
                                  earliest observed entry t =         0
                                       last observed exit t =  5.041667

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

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            15,564  (_merge==3)
    -----------------------------------------

. keep if age<=90
(186 observations deleted)
```

(b) .

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

Iteration 0:   log likelihood = -18636.652
Iteration 1:   log likelihood =  -18238.12
Iteration 2:   log likelihood = -18215.577
Iteration 3:   log likelihood = -18215.412
Iteration 4:   log likelihood = -18215.412

Log likelihood = -18215.412                     Number of obs   =      15378
------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
       _rcs1 |   2.572336   .0286756    84.75   0.000     2.516742    2.629157
       _rcs2 |   1.280017   .0108216    29.20   0.000     1.258982    1.301403
       _rcs3 |   .9696206   .0047501    -6.30   0.000     .9603551    .9789755
       _rcs4 |   1.020331   .0028292     7.26   0.000     1.014801    1.025891
       _rcs5 |   1.005988   .0019324     3.11   0.002     1.002208    1.009782
       _cons |   .3602999     .00487   -75.52   0.000     .3508802    .3699726
------------------------------------------------------------------------------

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

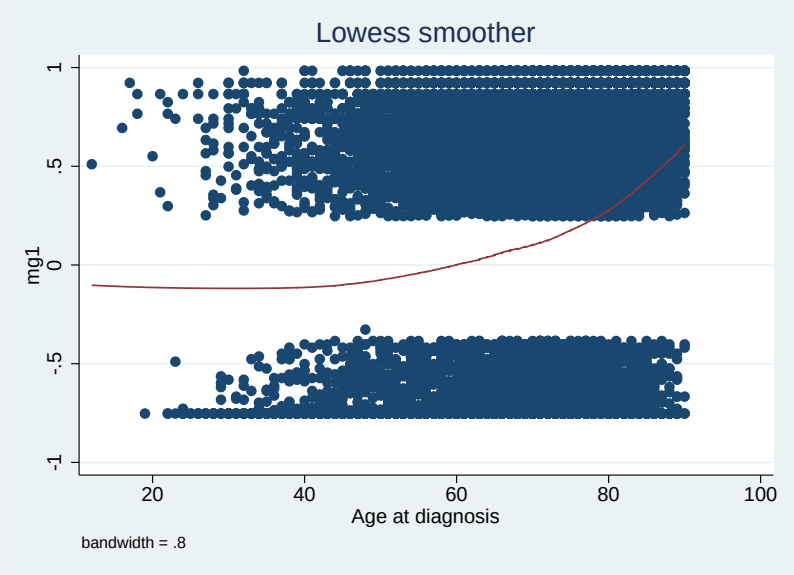


Figure 48: Colon Cancer. Martingale-like residuals from null model.

The Martingale-like residuals give us an indication of the functional form needed to model the effect of age. The effect looks non-linear with a steeper gradient for older ages.

(c)
```
Iteration 0:   log likelihood = -18536.121
Iteration 1:   log likelihood = -18131.804
Iteration 2:   log likelihood = -18110.238
Iteration 3:   log likelihood = -18110.113
Iteration 4:   log likelihood = -18110.113

Log likelihood = -18110.113                     Number of obs   =      15378
------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
         age |   1.015722    .001112    14.25   0.000     1.013545    1.017904
       _rcs1 |   2.598095   .0286852    86.48   0.000     2.542477     2.65493
       _rcs2 |   1.276285   .0106858    29.14   0.000     1.255512    1.297401
       _rcs3 |   .9688373    .004718    -6.50   0.000     .9596342    .9781286
       _rcs4 |   1.019811   .0028238     7.08   0.000     1.014292    1.025361
       _rcs5 |   1.005751   .0019385     2.98   0.003     1.001958    1.009557
       _cons |   .1255356    .009613   -27.10   0.000     .1080401    .1458642
------------------------------------------------------------------------------
```

For every yearly increase in age there is a 1.57% increase in the excess mortality rate.

(d)
```
. partpred hr_age_lin, for(age) ref(age 50) ci(hr_age_lin_lci hr_age_lin_uci) eform
note: confidence intervals calculated using Z critical values

. twoway  (rarea hr_age_lin_lci hr_age_lin_uci age, sort) ///
>                 (line hr_age_lin age, sort lpattern(solid)) ///
>                 , legend(off) ytitle("Hazard Ratio") scheme(sj) ///
>                 ylabel(0.5 1 2 4 8,angle(h) format(%3.1f)) name(hr_age_lin,replace) ///
>                 yscale(log) yline(1)
```

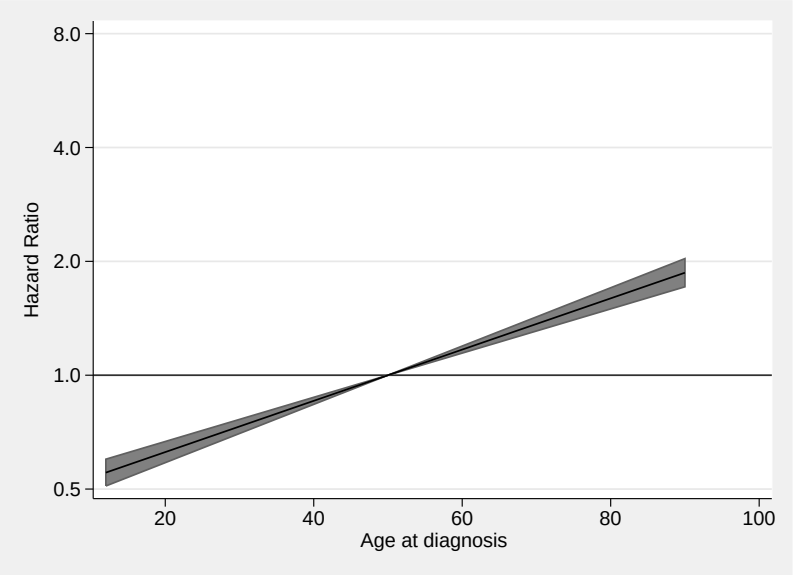


Figure 49: Colon Cancer. Excess mortality rate ratios for age at diagnosis with age 50 as the reference from model with linear effect of age.

If the assumption of linearity is reasonable (which it is not) then the excess mortality rate is about twice as high for an individual diagnosed at age 80 years compared to an individual aged 50. There is about a 20% reduction in the excess mortality rate for a woman diagnosed at age 30 compared to a woman aged 50.

(e)
```
. predict mg2, martingale
. lowess mg2 age, name(mg2, replace
```

Figure 50: Colon Cancer. Martingale-like residuals from model with linear effect of age.

This is slightly better than the initial Martingale plot, but there is evidence that the linear function of age is not capturing the observed relationship. We should therefore consider modelling the association with age as a non-linear function.

(f)
```
. rcsgen age, gen(rcsage) df(4) orthog
Variables rcsage1 to rcsage4 were created

. matrix Rage = r(R)
. global knotsage ‘r(knots)’
```

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

Iteration 0:   log likelihood = -18471.769
Iteration 1:   log likelihood = -18074.917
Iteration 2:   log likelihood = -18053.403
Iteration 3:   log likelihood = -18053.278
Iteration 4:   log likelihood = -18053.278

Log likelihood = -18053.278                     Number of obs   =      15378
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
     rcsage1 |   .1919882   .0122561    15.66   0.000     .1679667    .2160097
     rcsage2 |  -.1094588   .0120182    -9.11   0.000     -.133014   -.0859036
     rcsage3 |  -.0541316   .0121954    -4.44   0.000    -.0780342    -.030229
     rcsage4 |  -.0564078   .0122476    -4.61   0.000    -.0804126    -.032403
       _rcs1 |   .9576337       .011    87.06   0.000     .9360741    .9791932
       _rcs2 |    .244101   .0083404    29.27   0.000     .2277541    .2604479
       _rcs3 |   -.031556   .0048635    -6.49   0.000    -.0410883   -.0220237
       _rcs4 |   .0197156   .0027755     7.10   0.000     .0142757    .0251555
       _rcs5 |   .0059428   .0019377     3.07   0.002      .002145    .0097407
       _cons |  -.9981278   .0132523   -75.32   0.000    -1.024102   -.9721538
------------------------------------------------------------------------------

. predict mg3, martingale
. lowess mg3 age, name(mg3, replace)
```

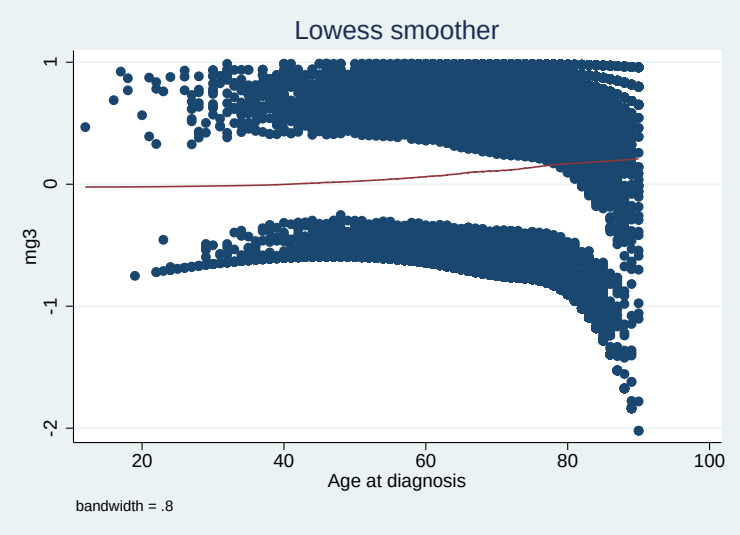


Figure 51: Colon Cancer. Martingale-like residuals from model with restricted cubic splines used to model the effect of age.

There is now very little association with age at diagnosis indicating that our splines have captured the non-linear effect of age at diagnosis.

(g)
```
. range temptime 0 5 200
(15178 missing values generated)

. foreach age in 40 60 80 {
  2.         rcsgen , scalar(‘age’) rmatrix(Rage) gen(c) knots($knotsage)
  3.         predict h‘age’, hazard at(rcsage1 ‘=c1’ rcsage2 ‘=c2’ rcsage3 ‘=c3’ rcsage4 ‘=c4’) ///
>                 timevar(temptime) per(1000)
  4.         predict s‘age’, survival at(rcsage1 ‘=c1’ rcsage2 ‘=c2’ rcsage3 ‘=c3’ rcsage4 ‘=c4’) ///
>                 timevar(temptime)
  5. }
Scalars c1 to c4 were created
Scalars c1 to c4 were created
Scalars c1 to c4 were created

. twoway (line h40 h60 h80 temptime), ///
>                 yscale(log) ytitle("Excess Mortality Rate (1000 py’s)") ///
>                 xtitle("Years from Diagnosis") ///
>                 legend(order(1 "40 yrs" 2 "60 yrs" 3 "80 yrs") cols(1) ring(0) pos(1)) ///
>                 ylabel(50 100 200 400 600 800 1000,angle(h)) ///
>                 name(hazard, replace) scheme(sj)

. twoway (line s40 s60 s80 temptime), ///
>                 ytitle("Relative Survival") ///
>                 xtitle("Years from Diagnosis") ///
>                 legend(order(1 "40 yrs" 2 "60 yrs" 3 "80 yrs") cols(1) ring(0) pos(1)) ///
>                 ylabel(0(0.2)1,angle(h) format(%3.1f)) ///
>                 name(survival, replace) scheme(sj)
```

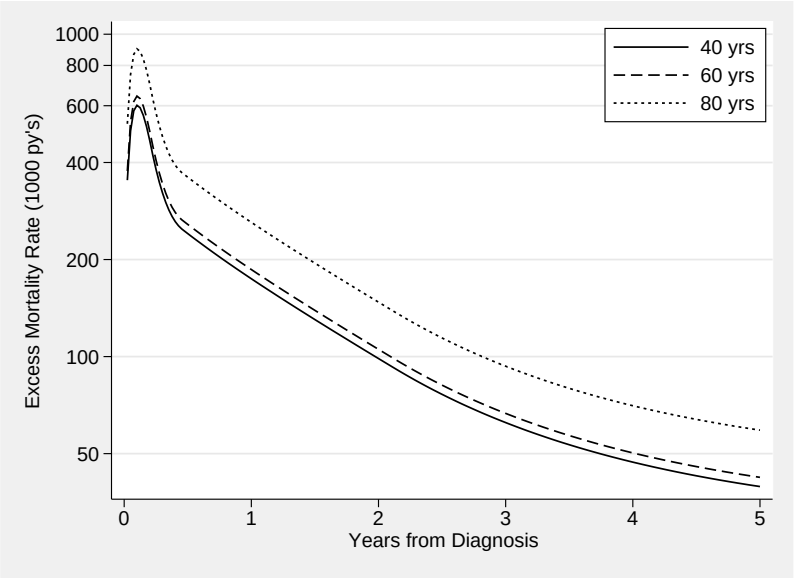


Figure 52: Colon Cancer. Predicted excess mortality rates functions for selected ages.

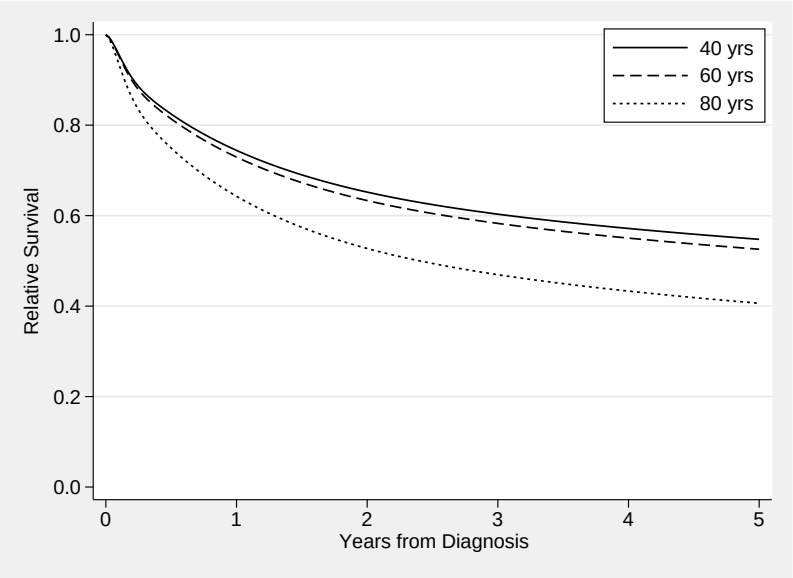


Figure 53: Colon Cancer. Predicted relative survival for selected ages.

The excess mortality rates and relative survival functions are fairly similar for 40- and 60-year-olds. There is a noticeable difference for those aged 80 at diagnosis. Note that in a proportional excess hazards model with a *linear* effect for age there would be an equal distance between the lines in the equivalent of Figure 52 (on the log scale), i.e., the distance between age 40 and age 60 would be the same as the distance between age 60 and age 80.

(h)
```
. gen t1 = 1

. predict s1, survival timevar(t1) ci

. twoway  (rarea s1_lci s1_uci age, sort) ///
>                 (line s1 age, sort lpattern(solid)) ///
>                 , legend(off) ytitle("1 year relative survival") scheme(sj) ///
>                 ylabel(0(0.2)1,angle(h) format(%3.1f)) name(s1,replace)
```

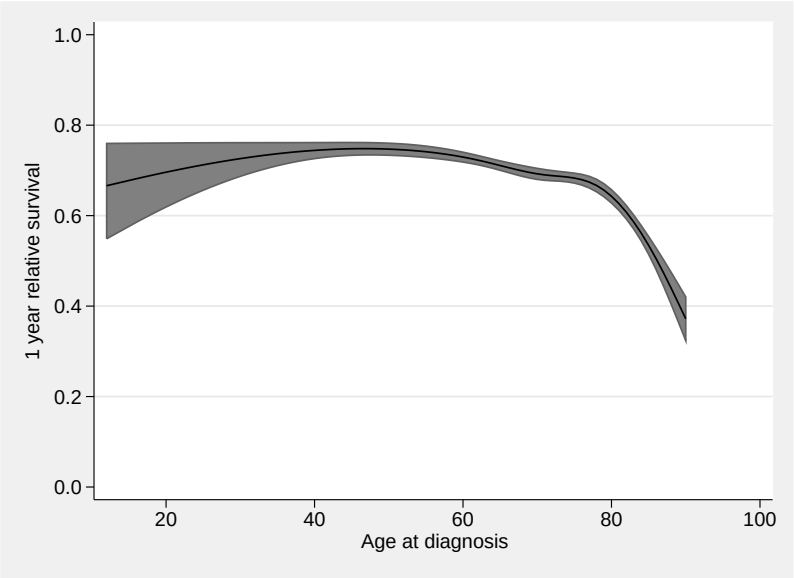


Figure 54: Colon Cancer. Predicted 1-year relative survival as a function of age.

There is fairly similar relative survival up to the age of 80 year at diagnosis. There is then a large drop in the predicted 1-year relative survival for those aged 99 at diagnosis.

(i)
```
. gen t5 = 5
. predict s5, survival timevar(t5) ci
. twoway  (rarea s5_lci s5_uci age, sort) ///
>                 (line s5 age, sort lpattern(solid)) ///
>                 , legend(off) ytitle("5 year relative survival") scheme(sj) ///
>                 ylabel(0(0.2)1,angle(h) format(%3.1f)) name(s5,replace)
```

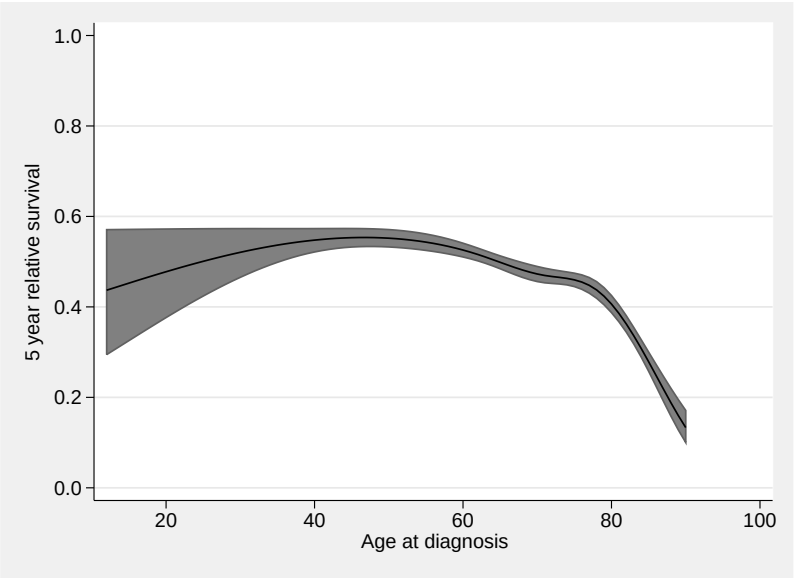


Figure 55: Colon Cancer. Predicted 5-year relative survival as a function of age.

A broadly similar pattern to the 1-year relative survival, but lower.

(j) .

```
. gen condsurv = s5/s1
. twoway  (line condsurv age, sort lpattern(solid)) ///
>                 , legend(off) ytitle("5 year conditional relative survival") scheme(sj) ///
>                 ylabel(0(0.2)1,angle(h) format(%3.1f))  name(condsurv,replace)
```

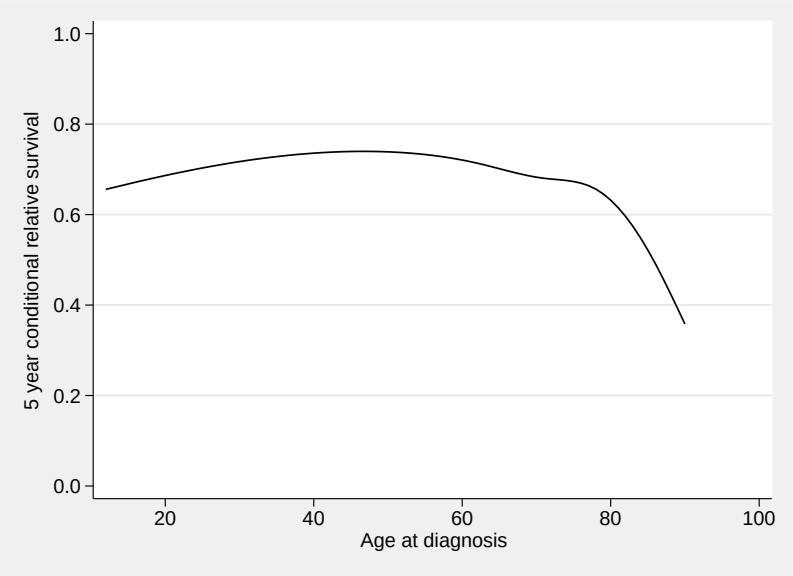


Figure 56: Colon Cancer. Predicted 5-year relative survival conditional on survival to 2 years as a function of age.

The shape of the curves is broadly similar. This is, at least in part, due to the proportional excess hazards assumption.

(k)

```
. rcsgen , scalar(50) rmatrix(Rage) gen(c) knots($knotsage)
Scalars c1 to c4 were created

. partpred hr_age_rcs, for(rcsage1-rcsage4) ///
>          ref(rcsage1 '=c1' rcsage2 '=c2' rcsage3 '=c3' rcsage4 '=c4') ///
>          eform ci(hr_age_rcs_lci hr_age_rcs_uci)
note: confidence intervals calculated using Z critical values

. twoway  (rarea hr_age_rcs_lci hr_age_rcs_uci age, sort) ///
>                 (line hr_age_rcs age, sort lpattern(solid)) ///
>                 , legend(off) ytitle("Hazard Ratio") scheme(sj) ///
>                 ylabel(0.5 1 2 4 8,angle(h) format(%3.1f)) name(hr_age,replace) ///
>                 yscale(range(0.5 1) log) yline(1)
```

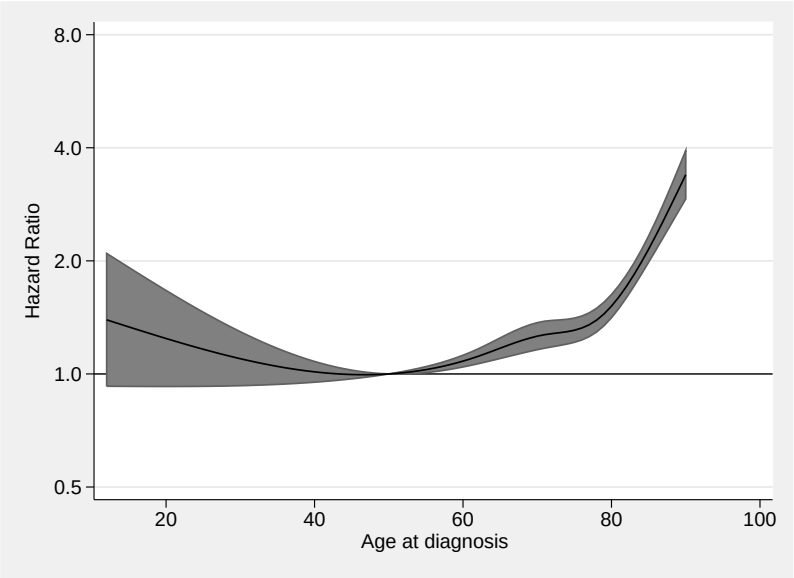


Figure 57: Colon Cancer. Hazard ratio for age with age 50 as the baseline from a model using restricted cubic splines to model the non-linear effect of age.

The reference age is 50 and so there is not a confidence interval at this point.The hazard ratio is close to 1 for those aged under 50 at diagnosis. There is a slight increase from ages 60-80. Then there is a a much steeper rise. For example, those aged 90 at diagnosis have an excess mortality rate about 3 times higher than that of a 50 year old.

(l)
```
. forvalues i = 3/5 {
  2.         capture drop rcsage*
  3.         rcsgen age, gen(rcsage) df(‘i’) orthog
  4.         matrix Rage = r(R)
  5.         global knotsage ‘r(knots)’
  6.         stpm2 rcsage*, scale(hazard) df(5) bhazard(rate) eform
  7.         estimates store m‘i’
  8.         rcsgen , scalar(50) rmatrix(Rage) gen(c) knots($knotsage)
  9.         local reflist
 10.         forvalues j = 1/‘i’ {
 11.                 local reflist ‘reflist’ rcsage‘j’ ‘=c‘j’’
 12.         }
 13.         di "‘reflist"’
 14.         partpred hr_age_rcs_df‘i’, for(rcsage*) ref(‘reflist’) ///
>                                 eform ci(hr_age_rcs_df‘i’_lci hr_age_rcs_df‘i’_uci)
 15. }
```

```
Variables rcsage1 to rcsage3 were created

Iteration 0:   log likelihood =  -18476.97
Iteration 1:   log likelihood = -18081.153
Iteration 2:   log likelihood = -18059.574
Iteration 3:   log likelihood = -18059.446
Iteration 4:   log likelihood = -18059.446

Log likelihood = -18059.446                     Number of obs   =      15378
------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
     rcsage1 |   1.212247   .0148471    15.72   0.000     1.183494    1.241699
     rcsage2 |   .8938071   .0110246    -9.10   0.000     .8724585    .9156782
     rcsage3 |   .9392246   .0116539    -5.05   0.000     .9166588    .9623459
       _rcs1 |    2.60433   .0286416    87.03   0.000     2.548794    2.661076
       _rcs2 |   1.276556   .0106458    29.28   0.000      1.25586    1.297593
       _rcs3 |   .9689861   .0047114    -6.48   0.000     .9597957    .9782645
       _rcs4 |   1.019918   .0028296     7.11   0.000     1.014387    1.025479
       _rcs5 |   1.005924   .0019481     3.05   0.002     1.002113    1.009749
       _cons |   .3687007   .0048838   -75.33   0.000     .3592518    .3783982
------------------------------------------------------------------------------
Scalars c1 to c3 were created
rcsage1 -1.542618497661927 rcsage2 -.1386906436710333 rcsage3 -.8850646862366336
note: confidence intervals calculated using Z critical values
Variables rcsage1 to rcsage4 were created

Iteration 0:   log likelihood = -18471.769
Iteration 1:   log likelihood = -18074.917
Iteration 2:   log likelihood = -18053.403
Iteration 3:   log likelihood = -18053.278
Iteration 4:   log likelihood = -18053.278

Log likelihood = -18053.278                     Number of obs   =      15378
------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
     rcsage1 |   1.211656   .0148502    15.66   0.000     1.182897    1.241114
     rcsage2 |   .8963191   .0107721    -9.11   0.000     .8754528    .9176827
     rcsage3 |   .9473075   .0115528    -4.44   0.000     .9249328    .9702234
     rcsage4 |   .9451536   .0115758    -4.61   0.000     .9227356    .9681163
       _rcs1 |   2.605524   .0286607    87.06   0.000     2.549951    2.662308
       _rcs2 |   1.276473   .0106463    29.27   0.000     1.255776    1.297511
       _rcs3 |   .9689367   .0047125    -6.49   0.000     .9597443    .9782171
       _rcs4 |   1.019911   .0028308     7.10   0.000     1.014378    1.025475
       _rcs5 |   1.005961   .0019493     3.07   0.002     1.002147    1.009788
       _cons |   .3685688   .0048844   -75.32   0.000     .3591189    .3782674
------------------------------------------------------------------------------
Scalars c1 to c4 were created
rcsage1 -1.542618497661927 rcsage2 -.0812679876188867 rcsage3 -1.211410261728227
> rcsage4 .86754585983042
note: confidence intervals calculated using Z critical values
```

```
Variables rcsage1 to rcsage5 were created

Iteration 0:   log likelihood = -18470.743
Iteration 1:   log likelihood = -18073.802
Iteration 2:   log likelihood = -18052.298
Iteration 3:   log likelihood = -18052.173
Iteration 4:   log likelihood = -18052.173

Log likelihood = -18052.173                     Number of obs   =      15378
------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
     rcsage1 |    1.21137   .0148536    15.64   0.000     1.182605    1.240835
     rcsage2 |   .8976739   .0108377    -8.94   0.000     .8766818    .9191687
     rcsage3 |   .9519768   .0115978    -4.04   0.000     .9295148    .9749816
     rcsage4 |   .9475846   .0116309    -4.39   0.000     .9250605    .9706572
     rcsage5 |   .9650437    .011842    -2.90   0.004     .9421106     .988535
       _rcs1 |   2.605858   .0286669    87.06   0.000     2.550273    2.662654
       _rcs2 |   1.276463   .0106467    29.26   0.000     1.255766    1.297502
       _rcs3 |   .9689288   .0047129    -6.49   0.000     .9597356      .97821
       _rcs4 |   1.019908   .0028312     7.10   0.000     1.014374    1.025472
       _rcs5 |   1.005957   .0019495     3.06   0.002     1.002143    1.009785
       _cons |   .3685638   .0048845   -75.32   0.000     .3591136    .3782627
------------------------------------------------------------------------------
Scalars c1 to c5 were created
rcsage1 -1.542618497661927 rcsage2 -.029822586317783 rcsage3 -1.462093415374454
> rcsage4 1.072661245645584 rcsage5 -.499434087234472
note: confidence intervals calculated using Z critical values

. twoway  (line hr_age_rcs_df3* age, sort lwidth(medthick thin thin) lcolor(red..) ///
>lpattern(solid dash..)) ///
>         (line hr_age_rcs_df4* age, sort lwidth(medthick thin thin) lcolor(blue..) ///
>lpattern(solid dash..)) ///
>         (line hr_age_rcs_df5* age, sort lwidth(medthick thin thin) lcolor(midgreen..) ///
>lpattern(solid dash..)) ///
>         , legend(order(1 "df 3" 4 "df 4" 7 "df 5") ring(0) pos(11) cols(1)) ///
>         yscale(range(0.5 8) log) yline(1) ylabel(0.5 1 2 4 8) ///
>         name(df_compare,replace)

. count if _d==1
 9215

. estimates stats m3 m4 m5, n(`r(N)')

Akaike's information criterion and Bayesian information criterion
------------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+----------------------------------------------------------------
          m3 |   9215           .   -18059.45      9     36136.89    36201.05
          m4 |   9215           .   -18053.28     10     36126.56    36197.84
          m5 |   9215           .   -18052.17     11     36126.35    36204.76
------------------------------------------------------------------------------
               Note:  N=9215 used in calculating BIC
```

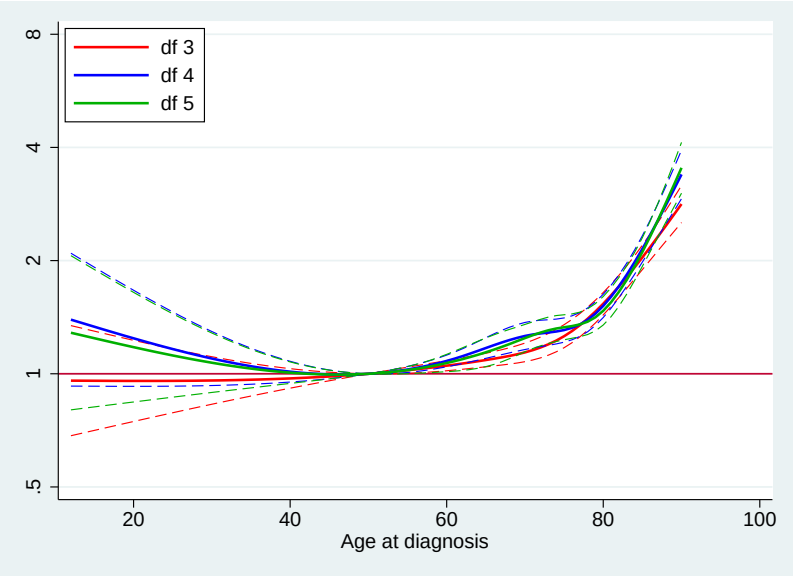


Figure 58: Colon Cancer. Comparison of non-linear hazard ratio for age for different df for the restricted cubic splines.

The graphs for 4 and 5 df are fairly similar, but there are some small differences with 3 df. The lowest AIC and BIC are for the model with 4 df.

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

(a) Load in the colon cancer data and merge in the background mortality rates as in question 230. Drop those aged over 90 years.

(b)

```
. rcsgen age, gen(rcsage) df(4) orthog
Variables rcsage1 to rcsage4 were created

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

Iteration 0:   log likelihood = -18471.769
Iteration 1:   log likelihood = -18074.917
Iteration 2:   log likelihood = -18053.403
Iteration 3:   log likelihood = -18053.278
Iteration 4:   log likelihood = -18053.278

Log likelihood = -18053.278                     Number of obs   =      15378
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
     rcsage1 |   .1919882   .0122561    15.66   0.000     .1679667    .2160097
     rcsage2 |  -.1094588   .0120182    -9.11   0.000      -.133014   -.0859036
     rcsage3 |  -.0541316   .0121954    -4.44   0.000    -.0780342    -.030229
     rcsage4 |  -.0564078   .0122476    -4.61   0.000    -.0804126    -.032403
       _rcs1 |   .9576337       .011    87.06   0.000     .9360741    .9791932
       _rcs2 |    .244101   .0083404    29.27   0.000     .2277541    .2604479
       _rcs3 |   -.031556   .0048635    -6.49   0.000    -.0410883   -.0220237
       _rcs4 |   .0197156   .0027755     7.10   0.000     .0142757    .0251555
       _rcs5 |   .0059428   .0019377     3.07   0.002      .002145    .0097407
       _cons |  -.9981278   .0132523   -75.32   0.000    -1.024102   -.9721538
------------------------------------------------------------------------------

. estimates store peh
```

(c)

```
. stpm2 rcsage1-rcsage4, scale(hazard) df(5) bhazard(rate) ///
>         tvc(rcsage1-rcsage4) dftvc(2)

Iteration 0:   log likelihood = -18452.143
Iteration 1:   log likelihood = -17933.564
Iteration 2:   log likelihood = -17902.413
Iteration 3:   log likelihood = -17902.311
Iteration 4:   log likelihood = -17902.311
```

```
Log likelihood = -17902.311                     Number of obs   =      15378
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
      rcsage1 |    .242729   .0143139    16.96   0.000     .2146743    .2707838
      rcsage2 |   -.095568   .0140304    -6.81   0.000    -.1230671    -.068069
      rcsage3 |  -.0228265   .0138512    -1.65   0.099    -.0499744    .0043214
      rcsage4 |  -.0415266     .01318    -3.15   0.002    -.0673589   -.0156943
        _rcs1 |    .988337   .0124816    79.18   0.000     .9638734     1.0128
        _rcs2 |   .2778258   .0090949    30.55   0.000     .2600001    .2956515
        _rcs3 |  -.0276664   .0049551    -5.58   0.000    -.0373783   -.0179545
        _rcs4 |   .0227319   .0028494     7.98   0.000     .0171472    .0283166
        _rcs5 |   .0074573   .0019991     3.73   0.000     .0035391    .0113755
_rcs_rcsage11 |  -.1732978   .0134101   -12.92   0.000    -.1995811   -.1470144
_rcs_rcsage12 |  -.0203696   .0083136    -2.45   0.014    -.0366641   -.0040752
_rcs_rcsage21 |   .0490377   .0130571     3.76   0.000     .0234463    .0746291
_rcs_rcsage22 |  -.0221844   .0078354    -2.83   0.005    -.0375416   -.0068272
_rcs_rcsage31 |   .0068485   .0118303     0.58   0.563    -.0163385    .0300354
_rcs_rcsage32 |   -.009837   .0072894    -1.35   0.177    -.0241238    .0044499
_rcs_rcsage41 |  -.0014651   .0098459    -0.15   0.882    -.0207627    .0178325
_rcs_rcsage42 |   .0022813   .0064509     0.35   0.724    -.0103622    .0149248
        _cons |  -1.048383   .0138684   -75.60   0.000    -1.075565   -1.021202
------------------------------------------------------------------------------

. estimates store timedep

. lrtest peh timedep

Likelihood-ratio test                                 LR chi2(8)  =    301.93
(Assumption: peh nested in timedep)                   Prob > chi2 =    0.0000
```

There is very strong evidence that the effect of age is non-proportional, i.e. proportional excess hazards is not a reasonable assumption.

(d)
```
. range temptime 0 5 200
(15178 missing values generated)

. foreach age in 40 60 80 {
  2.         rcsgen , scalar(‘age’) rmatrix(Rage) gen(c) knots($knotsage)
  3.         predict h‘age’, hazard at(rcsage1 ‘=c1’ rcsage2 ‘=c2’ rcsage3 ‘=c3’ ///
>  rcsage4 ‘=c4’) timevar(temptime) per(1000)
  4.         predict s‘age’, survival at(rcsage1 ‘=c1’ rcsage2 ‘=c2’ rcsage3 ‘=c3’ ///
>  rcsage4 ‘=c4’) timevar(temptime)
  5. }
Scalars c1 to c4 were created
Scalars c1 to c4 were created
Scalars c1 to c4 were created

. twoway (line h40 h60 h80 temptime), ///
>                 yscale(log) ytitle("Excess Mortality Rate (1000 py’s)") ///
>                 xtitle("Years from Diagnosis") ///
>                 legend(order(1 "40 yrs" 2 "60 yrs" 3 "80 yrs") ///
>cols(1) ring(0) pos(1)) ///
>                 ylabel(50 100 200 400 600 800 1000,angle(h)) ///
>                 name(hazard, replace) scheme(sj)
```

```
. twoway (line s40 s60 s80 temptime), ///
>                 ytitle("Relative Survival") ///
>                 xtitle("Years from Diagnosis") ///
>                 legend(order(1 "40 yrs" 2 "60 yrs" 3 "80 yrs") ///
>cols(1) ring(0) pos(1)) ///
>                 ylabel(0(0.2)1,angle(h) format(%3.1f)) ///
>                 name(survival, replace) scheme(sj)
```

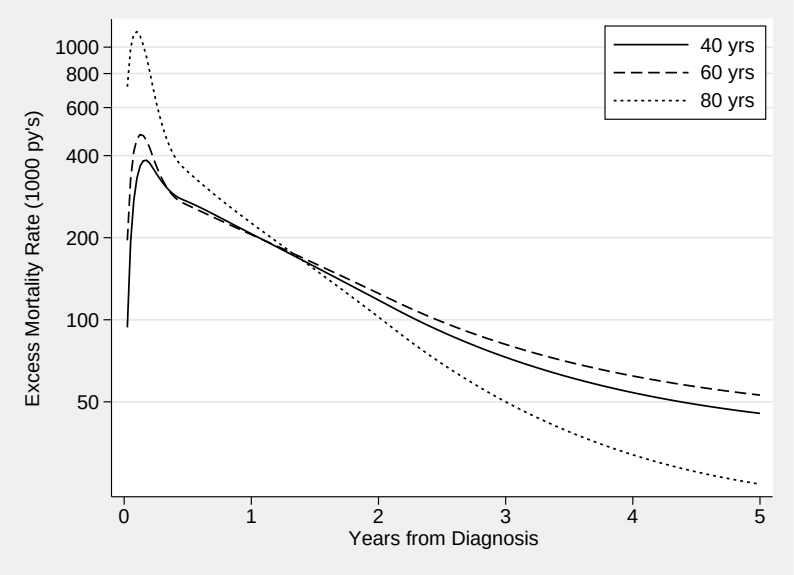


Figure 59: Colon Cancer. Excess mortality rates for selected ages at diagnosis. Age has a non-linear, time-dependent effect.

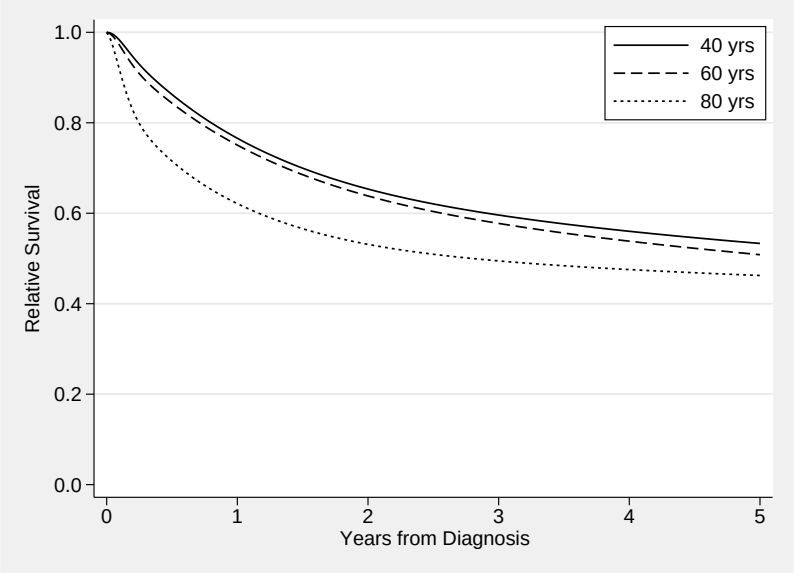


Figure 60: Colon Cancer. Relative survival for selected ages at diagnosis. Age has a non-linear, time-dependent effect.

The excess mortality rates no longer have a constant difference between them, as was the case in question q31. The most noticeable difference is for subjects aged 80 where the shape is very

different to those aged 40 and 60. For the relative survival curves, there is greater separation between those aged 80 at diagnosis and the other two curves early on in the time scale than when proportional excess hazards is assumed.

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

. twoway  (rarea s1_lci s1_uci age, sort) ///
>                 (line s1 age, sort lpattern(solid)) ///
>                 , legend(off) ytitle("1 year relative survival") scheme(sj) ///
>                 ylabel(0(0.2)1,angle(h) format(%3.1f)) name(s1,replace)
```

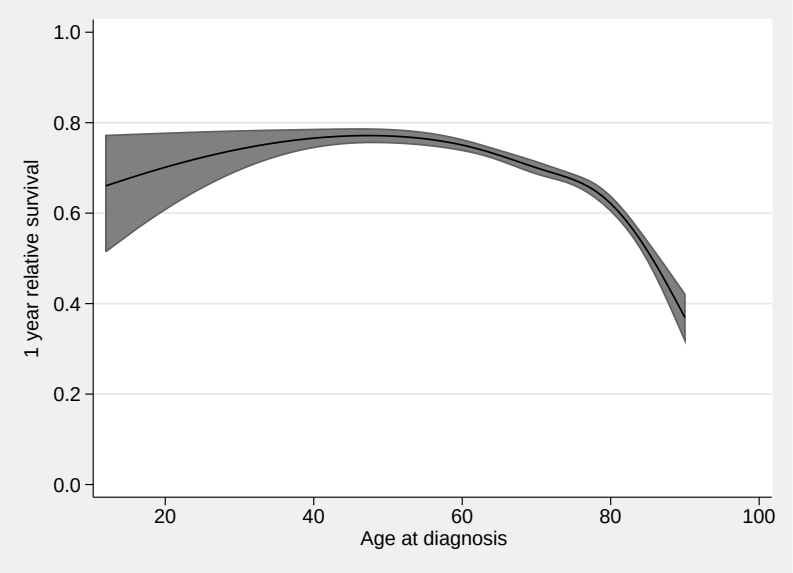


Figure 61: Colon Cancer. One year relative survival as a function of age.

(f)
```
. gen t5 = 5
. predict s5, survival timevar(t5) ci

. twoway  (rarea s5_lci s5_uci age, sort) ///
>                 (line s5 age, sort lpattern(solid)) ///
>                 , legend(off) ytitle("5 year relative survival") scheme(sj) ///
>                 ylabel(0(0.2)1,angle(h) format(%3.1f)) name(s5,replace)
```

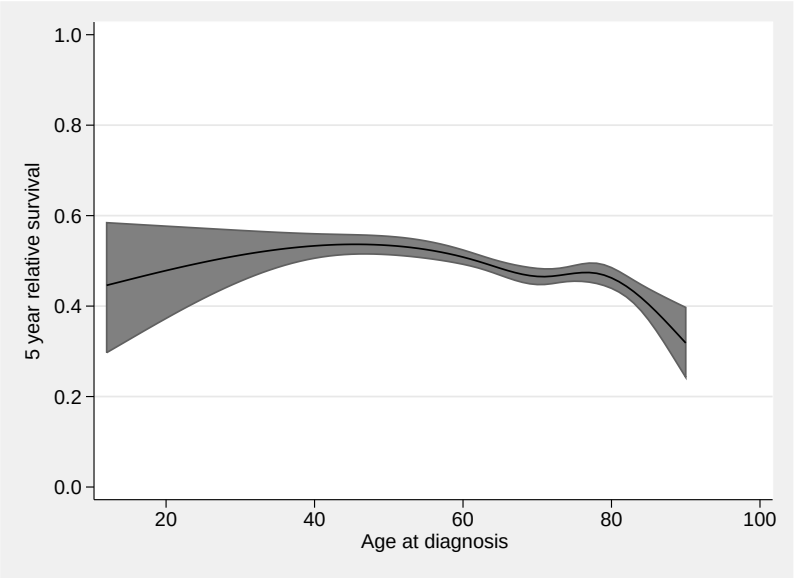


Figure 62: Colon Cancer. Five year relative survival as a function of age.

Relative survival is higher for those aged 80-90 than in the proportional excess hazards model.

(g)
```
. gen condsurv = s5/s1

. twoway  (line condsurv age, sort lpattern(solid)) ///
>                 , legend(off) ytitle("5 year conditional relative survival") scheme(sj) ///
>                 ylabel(0(0.2)1,angle(h) format(%3.1f))  name(condsurv,replace)

. predictnl condsurv2 = predict(survival timevar(t5))/predict(survival timevar(t1)) ///
>                 ,ci(condsurv2_lci condsurv2_uci)
note: confidence intervals calculated using Z critical values

. twoway  (rarea condsurv2_lci condsurv2_uci age, sort) ///
>                 (line condsurv2 age, sort lpattern(solid)) ///
>                 , legend(off) ytitle("5 year conditional relative survival") scheme(sj) ///
>                 ylabel(0(0.2)1,angle(h) format(%3.1f))  name(condsurv2,replace)
```

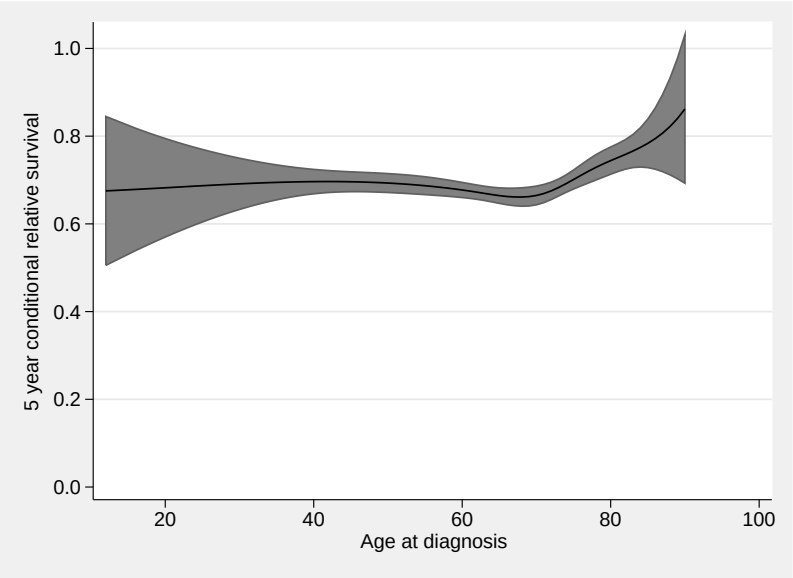


Figure 63: Colon Cancer. Five year relative survival conditional on survival to one year as a function of age.

This curve is much flatter than from the proportional excess hazards model. This illustrates that much of the difference in relative survival is due to difference in the first year after diagnosis. The proportional excess hazards model forces the same relative increase in the excess mortality rate over follow-up time. This is clearly inappropriate for the oldest age group.

(h)
```
. rcsgen , scalar(50) rmatrix(Rage) gen(ref) knots($knotsage)
Scalars ref1 to ref4 were created
. foreach age in 40 60 70 80 {
  2.         rcsgen , scalar(‘age’) rmatrix(Rage) gen(c‘age’_) knots($knotsage)
  3.         predict hr‘age’, ///
>                hrnum(rcsage1 ‘=c‘age’_1’ rcsage2 ‘=c‘age’_2’ rcsage3 ‘=c‘age’_3’ rcsage4 ‘=c‘age’_4’)
>                hrdenom(rcsage1 ‘=ref1’ rcsage2 ‘=ref2’ rcsage3 ‘=ref3’ rcsage4 ‘=ref4’) ///
>                timevar(temptime) ci
  4. }
Scalars c40_1 to c40_4 were created
Scalars c60_1 to c60_4 were created
Scalars c70_1 to c70_4 were created
Scalars c80_1 to c80_4 were created

. foreach age in 40 60 70 80 {
  2.         twoway  (rarea hr‘age’_lci hr‘age’_uci temptime, sort) ///
>                (line hr‘age’ temptime, sort lpattern(solid)) ///
>                , legend(off) ytitle("EMRR") scheme(sj) ///
>                xtitle("Years from Diagnosis") ///
>                ylabel(0.5 1 2 4 8,angle(h) format(%3.1f)) ///
>                yscale(log range(0.5 8)) yline(1, lpatter(dash)) ///
>                name(hr‘age’,replace)
  3. }

. graph combine hr40 hr60 hr70 hr80, nocopies name(hr_all,replace)
```

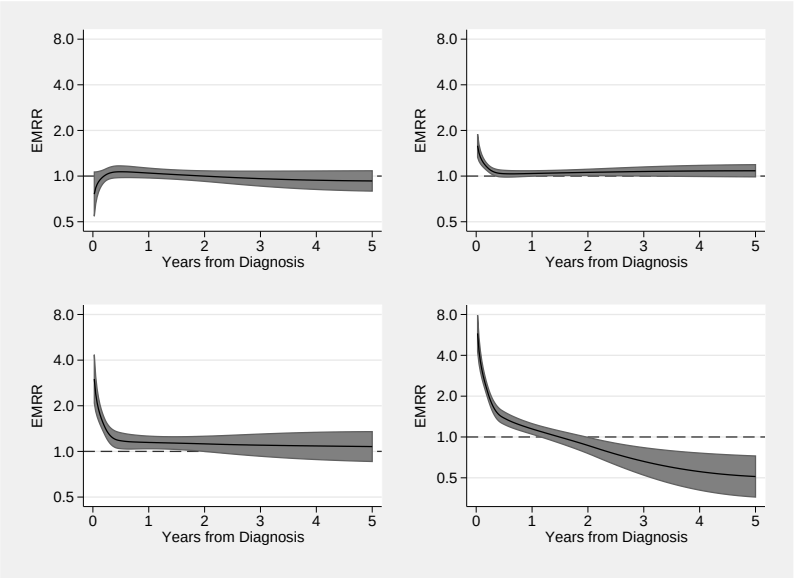


Figure 64: Colon Cancer. Time dependent excess mortality rate ratios for age. Age 50 is the reference age.

The shape of the curves are very different indicating why we had to account for non-proportional excess hazards. It appears that 40 and 50 year olds are very similar as the excess mortality rate ratio is close to 1. Both 60 and 70 year olds have an initial higher mortality rate compared to those age 50, but after about 6 months their mortality rate is similar. The shape of those aged 80 is notably different with initially a higher excess mortality rate and then a lower excess mortality rate compared to those aged 50.

(i)
```
. foreach age in 40 60 70 80 {
  2.         rcsgen , scalar(‘age’) rmatrix(Rage) gen(c‘age’_) knots($knotsage)
  3.         predict hdiff‘age’, ///
>                 hdiff1(rcsage1 ‘=c‘age’_1’ rcsage2 ‘=c‘age’_2’ rcsage3 ‘=c‘age’_3’ rcsage4 ‘=c‘age’
>                 hdiff2(rcsage1 ‘=ref1’ rcsage2 ‘=ref2’ rcsage3 ‘=ref3’ rcsage4 ‘=ref4’) ///
>                 timevar(temptime) ci per(1000)
  4.         predict sdiff‘age’, ///
>                 sdiff1(rcsage1 ‘=c‘age’_1’ rcsage2 ‘=c‘age’_2’ rcsage3 ‘=c‘age’_3’ rcsage4 ‘=c‘age’
>                 sdiff2(rcsage1 ‘=ref1’ rcsage2 ‘=ref2’ rcsage3 ‘=ref3’ rcsage4 ‘=ref4’) ///
>                 timevar(temptime) ci
  5. }
Scalars c40_1 to c40_4 were created
Scalars c60_1 to c60_4 were created
Scalars c70_1 to c70_4 were created
Scalars c80_1 to c80_4 were created

. foreach age in 40 60 70 80 {
  2.         twoway  (rarea hdiff‘age’_lci hdiff‘age’_uci temptime, sort) ///
>                 (line hdiff‘age’ temptime, sort lpattern(solid)) ///
>                 , legend(off) ytitle("") scheme(sj) ///
>                 xtitle("Years from Diagnosis") ///
>                 ylabel(-100 0  100 200 400 600 800,angle(h) format(%3.0f)) ///
>                 yscale(range(-50 900)) yline(0, lpattern(dash)) ///
>                 name(hdiff‘age’,replace)
  3. }
. graph combine hdiff40 hdiff60 hdiff70 hdiff80, nocopies ///
>         l1title("Difference in excess mortality rate (1000 py’s)") name(hdiff,replace)
```

```
. foreach age in 40 60 70 80 {
  2.         twoway  (rarea sdiff‘age’_lci sdiff‘age’_uci temptime, sort) ///
>                (line sdiff‘age’ temptime, sort lpattern(solid)) ///
>                , legend(off) ytitle("") scheme(sj) ///
>                xtitle("Years from Diagnosis") ///
>                ylabel(-0.2 -0.15 -0.1 -0.05 0,angle(h) format(%3.2f)) ///
>                yscale(range(-0.2 0.05)) yline(0, lpattern(dash)) ///
>                name(sdiff‘age’,replace)
  3. }

. graph combine sdiff40 sdiff60 sdiff70 sdiff80, nocopies ///
>         l1title("Difference in Relative Survival") name(sdiff,replace)
```

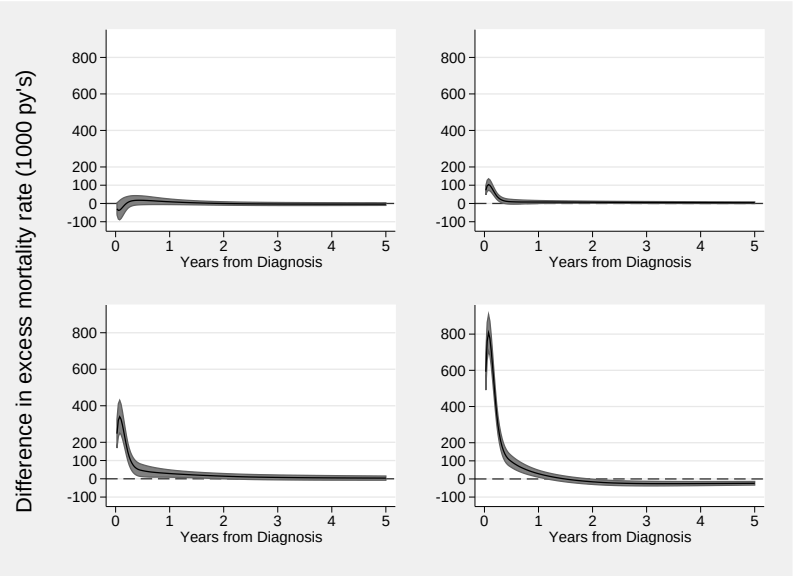


Figure 65: Colon Cancer. Differences in the excess mortality rate for selected ages. Age 50 is the reference

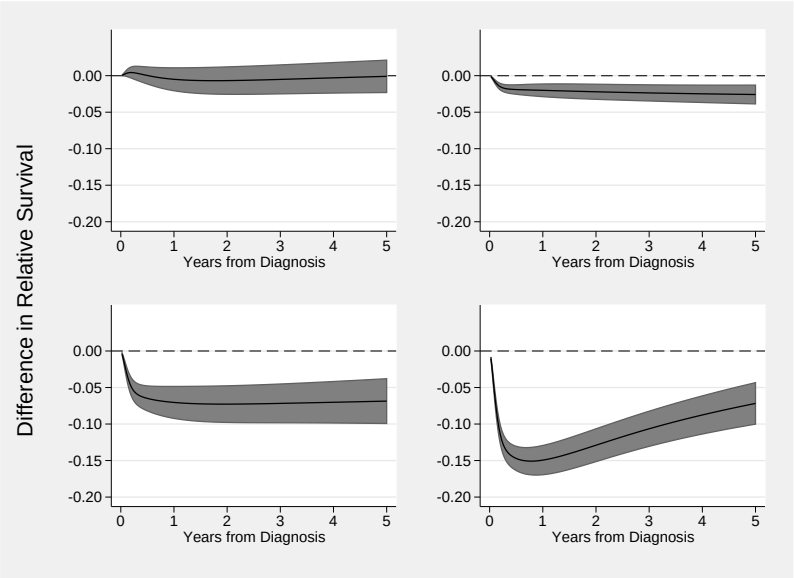


Figure 66: Colon Cancer. Differences in relative survival for selected ages. Age 50 is the reference

Note that as the excess mortality rate for colon cancers decreases as time from diagnosis increases any relative differences have less impact in absolute terms. For example, the lower excess mortality rate for 80 year olds when compared to those aged 50 after about 2 years had little impact on the absolute difference.

(j)
```
. forvalues i = 1/3 {
  2.          stpm2 rcsage*, scale(hazard) df(5) bhazard(rate) tvc(rcsage*) dftvc(‘i’)
  3.          estimates store m‘i’
  4.          predict hr_age_tvc_df‘i’, ///
>                 hrnum(rcsage1 ‘=c70_1’ rcsage2 ‘=c70_2’ rcsage3 ‘=c70_3’ rcsage4 ‘=c70_4’) ///
>                 hrdenom(rcsage1 ‘=ref1’ rcsage2 ‘=ref2’ rcsage3 ‘=ref3’ rcsage4 ‘=ref4’) ///
>                 timevar(temptime) ci
  5. }

Iteration 0:   log likelihood = -18457.159
Iteration 1:   log likelihood = -17930.981
Iteration 2:   log likelihood = -17909.858
Iteration 3:   log likelihood = -17909.781
Iteration 4:   log likelihood = -17909.781
```

```
Log likelihood = -17909.781                     Number of obs   =      15378
------------------------------------------------------------------------------
                |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
xb              |
        rcsage1 |   .2326957   .0134386    17.32   0.000     .2063566    .2590349
        rcsage2 |   -.107742   .0132165    -8.15   0.000    -.1336459   -.0818382
        rcsage3 |  -.0288056     .01316    -2.19   0.029    -.0545988   -.0030124
        rcsage4 |  -.0416388   .0127306    -3.27   0.001    -.0665903   -.0166874
          _rcs1 |   .9836702   .0113947    86.33   0.000     .9613369    1.006003
          _rcs2 |   .2752511   .0085525    32.18   0.000     .2584885    .2920137
          _rcs3 |  -.0277373    .004896    -5.67   0.000    -.0373332   -.0181413
          _rcs4 |   .0228091    .002818     8.09   0.000     .0172859    .0283324
          _rcs5 |   .0075589   .0019907     3.80   0.000     .0036572    .0114606
  _rcs_rcsage11 |  -.1465897      .0101   -14.51   0.000    -.1663854    -.126794
  _rcs_rcsage21 |   .0705756   .0096343     7.33   0.000     .0516926    .0894586
  _rcs_rcsage31 |    .010571    .009395     1.13   0.261     -.007843    .0289849
  _rcs_rcsage41 |  -.0081186   .0089113    -0.91   0.362    -.0255845    .0093473
          _cons |  -1.046086    .013823   -75.68   0.000    -1.073179   -1.018993
------------------------------------------------------------------------------

Iteration 0:   log likelihood = -18452.143
Iteration 1:   log likelihood = -17933.564
Iteration 2:   log likelihood = -17902.413
Iteration 3:   log likelihood = -17902.311
Iteration 4:   log likelihood = -17902.311

Log likelihood = -17902.311                     Number of obs   =      15378
------------------------------------------------------------------------------
                |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
xb              |
        rcsage1 |    .242729   .0143139    16.96   0.000     .2146743    .2707838
        rcsage2 |   -.095568   .0140304    -6.81   0.000    -.1230671    -.068069
        rcsage3 |  -.0228265   .0138512    -1.65   0.099    -.0499744    .0043214
        rcsage4 |  -.0415266     .01318    -3.15   0.002    -.0673589   -.0156943
          _rcs1 |    .988337   .0124816    79.18   0.000     .9638734      1.0128
          _rcs2 |   .2778258   .0090949    30.55   0.000     .2600001    .2956515
          _rcs3 |  -.0276664   .0049551    -5.58   0.000    -.0373783   -.0179545
          _rcs4 |   .0227319   .0028494     7.98   0.000     .0171472    .0283166
          _rcs5 |   .0074573   .0019991     3.73   0.000     .0035391    .0113755
  _rcs_rcsage11 |  -.1732978   .0134101   -12.92   0.000    -.1995811   -.1470144
  _rcs_rcsage12 |  -.0203696   .0083136    -2.45   0.014    -.0366641   -.0040752
  _rcs_rcsage21 |   .0490377   .0130571     3.76   0.000     .0234463    .0746291
  _rcs_rcsage22 |  -.0221844   .0078354    -2.83   0.005    -.0375416   -.0068272
  _rcs_rcsage31 |   .0068485   .0118303     0.58   0.563    -.0163385    .0300354
  _rcs_rcsage32 |   -.009837   .0072894    -1.35   0.177    -.0241238    .0044499
  _rcs_rcsage41 |  -.0014651   .0098459    -0.15   0.882    -.0207627    .0178325
  _rcs_rcsage42 |   .0022813   .0064509     0.35   0.724    -.0103622    .0149248
          _cons |  -1.048383   .0138684   -75.60   0.000    -1.075565   -1.021202
------------------------------------------------------------------------------
```

```
Iteration 0:   log likelihood = -18452.584
Iteration 1:   log likelihood = -17943.882
Iteration 2:   log likelihood = -17899.049
Iteration 3:   log likelihood = -17897.511
Iteration 4:   log likelihood = -17897.508
Iteration 5:   log likelihood = -17897.508

Log likelihood = -17897.508                     Number of obs   =      15378
------------------------------------------------------------------------------
                |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
xb              |
        rcsage1 |   .2457525   .0144575    17.00   0.000     .2174164    .2740886
        rcsage2 |  -.0944898   .0142306    -6.64   0.000    -.1223813   -.0665983
        rcsage3 |   -.026167   .0139078    -1.88   0.060    -.0534258    .0010918
        rcsage4 |  -.0427574   .0131877    -3.24   0.001    -.0686048     -.01691
          _rcs1 |   .9874801   .0125913    78.43   0.000     .9628015    1.012159
          _rcs2 |   .2721054   .0097757    27.83   0.000     .2529455    .2912654
          _rcs3 |  -.0255948   .0051167    -5.00   0.000    -.0356234   -.0155662
          _rcs4 |   .0226817   .0028583     7.94   0.000     .0170795     .028284
          _rcs5 |   .0067756    .002025     3.35   0.001     .0028067    .0107445
  _rcs_rcsage11 |  -.1723405   .0147961   -11.65   0.000    -.2013404   -.1433407
  _rcs_rcsage12 |   -.023014   .0111592    -2.06   0.039    -.0448858   -.0011423
  _rcs_rcsage13 |  -.0048067   .0049642    -0.97   0.333    -.0145364     .004923
  _rcs_rcsage21 |   .0370246    .014522     2.55   0.011     .0085621    .0654871
  _rcs_rcsage22 |  -.0346487   .0110278    -3.14   0.002    -.0562628   -.0130346
  _rcs_rcsage23 |   .0110394   .0047481     2.33   0.020     .0017333    .0203455
  _rcs_rcsage31 |   .0066997   .0125319     0.53   0.593    -.0178623    .0312618
  _rcs_rcsage32 |  -.0050684   .0091901    -0.55   0.581    -.0230806    .0129438
  _rcs_rcsage33 |   .0013141   .0044303     0.30   0.767     -.007369    .0099973
  _rcs_rcsage41 |  -.0027939   .0101801    -0.27   0.784    -.0227467    .0171588
  _rcs_rcsage42 |   .0041266    .007216     0.57   0.567    -.0100165    .0182696
  _rcs_rcsage43 |   .0001072   .0040212     0.03   0.979    -.0077743    .0079886
          _cons |  -1.046116   .0139156   -75.18   0.000    -1.073391   -1.018842
------------------------------------------------------------------------------

. twoway  (line hr_age_tvc_df1* temptime, sort lwidth(medthick thin thin) ///
>lcolor(red..) lpattern(solid dash..)) ///
>(line hr_age_tvc_df2* temptime, sort lwidth(medthick thin thin) ///
>lcolor(blue..)  lpattern(solid dash..)) ///
>(line hr_age_tvc_df3* temptime, sort lwidth(medthick thin thin) ///
>lcolor(midgreen..)  lpattern(solid dash..)) ///
>, legend(order(1 "df 1" 4 "df 2" 7 "df 3") ring(0) pos(11) cols(1)) ///
>yscale(range(0.5 8) log) yline(1) ylabel(0.5 1 2 4 8) ///
>name(df_tvc_compare,replace)

. count if _d==1
 9215

. estimates stats m1 m2 m3, n(‘r(N)’)

Akaike’s information criterion and Bayesian information criterion
------------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+----------------------------------------------------------------
          m1 |   9215           .   -17909.78     14     35847.56    35947.36
          m2 |   9215           .   -17902.31     18     35840.62    35968.94
          m3 |   9215           .   -17897.51     22     35839.02    35995.85
------------------------------------------------------------------------------
               Note:  N=9215 used in calculating BIC
```

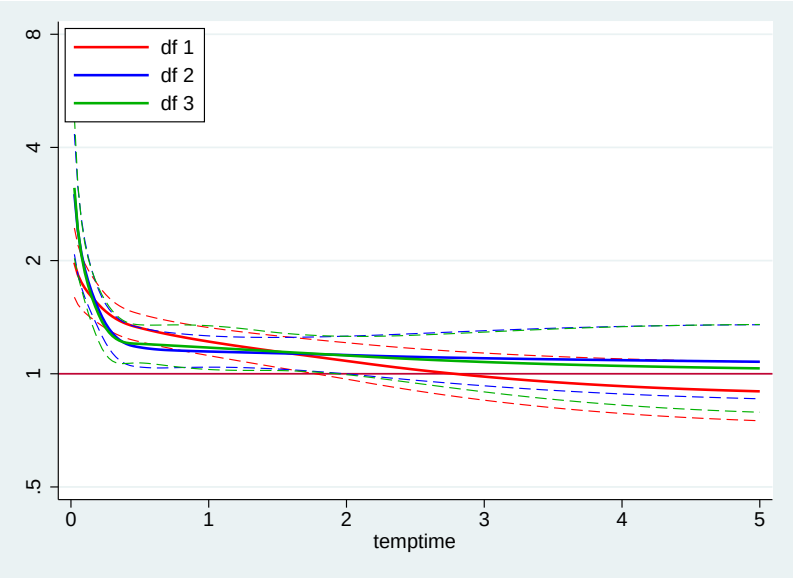


Figure 67: Colon Cancer. Sensitivity analysis for the df for the time-dependent effects.

When using 2 or 3 df the curves are similar. There is some difference when using 1 df, but conclusions would not change much. AIC gives best fitting model as 3 df and BIC gives 1 df.

240. **Age-standardised estimates of relative survival**

(a)
```
. use melanoma, clear
. keep if stage==1 /* restrict to localised */

. stset surv_mm, fail(status==1 2) id(id) scale(12)
. strs using popmort, br(0(1)15) mergeby(_year sex _age)

No late entry detected - p is estimated using the actuarial method
```

end	n	d	w	p	p_star	r	cp	cp_e2	cr_e2
1	5318	151	1	0.9716	0.9768	0.9947	0.9716	0.9768	0.9947
2	5166	329	299	0.9344	0.9763	0.9571	0.9079	0.9537	0.9519
3	4538	287	296	0.9346	0.9767	0.9569	0.8485	0.9315	0.9109
4	3955	211	271	0.9448	0.9771	0.9669	0.8017	0.9102	0.8808
5	3473	166	246	0.9504	0.9775	0.9723	0.7619	0.8897	0.8564
6	3061	138	240	0.9531	0.9775	0.9751	0.7262	0.8696	0.8350
7	2683	105	218	0.9592	0.9772	0.9815	0.6966	0.8499	0.8196
8	2360	75	253	0.9664	0.9766	0.9896	0.6732	0.8299	0.8111
9	2032	68	241	0.9644	0.9756	0.9885	0.6492	0.8097	0.8018
10	1723	50	209	0.9691	0.9756	0.9933	0.6292	0.7900	0.7964
11	1464	55	160	0.9603	0.9752	0.9847	0.6042	0.7704	0.7843
12	1249	49	157	0.9581	0.9754	0.9823	0.5789	0.7514	0.7704
13	1043	21	142	0.9784	0.9743	1.0042	0.5664	0.7321	0.7736
14	880	22	168	0.9724	0.9728	0.9995	0.5507	0.7122	0.7732
15	690	20	136	0.9678	0.9727	0.9950	0.5330	0.6928	0.7694

The crude 10-year RSR is 0.7964.

(b)
```
. strs using popmort, br(0(1)15) mergeby(_year sex _age) by(agegrp) save(replace)
. use grouped, clear
. bysort agegrp: gen n0=n[1]
. local N `r(sum)'
. gen weight=n0/`N'
. gen x=cr_e2*weight
. list agegrp n0 cr_e2 weight x if end==10, sum(n0 weight x) mean(cr_e2)
```

	agegrp	n0	cr_e2	weight	x
10.	0-44	1463	0.8317	.2751034	.2288065
25.	45-59	1575	0.8069	.296164	.2389828
40.	60-74	1536	0.7901	.2888304	.2281977
55.	75+	744	0.6838	.1399022	.0956643
Mean			0.7781		
Sum		5318		1	.7916513

The age-standardised (traditional) 10-year RSR is 0.7917.

(c)
```
. use melanoma, clear
. keep if stage==1 /* restrict to localised */
. stset surv_mm, fail(status==1 2) id(id) scale(12)
. local totalobs = _N
. bysort agegrp: gen standwei = _N/‘totalobs’

. strs using popmort [iw=standwei], br(0(1)15) mergeby(_year sex _age)
              standstrata(agegrp) notables

No late entry detected - p is estimated using the actuarial method
Adjusted survival estimates weighting stratum-specific survival in each
    group of agegrp by standwei weights.
  +-----------------------------------------+
  | start   end    cr_e2   lo_cr_e2   hi_cr_e2 |
  |-----------------------------------------|
  |     0     1   0.9947     0.9844     1.0021 |
  |     1     2   0.9506     0.9330     0.9655 |
  |     2     3   0.9083     0.8858     0.9284 |
  |     3     4   0.8765     0.8504     0.9003 |
  |     4     5   0.8504     0.8212     0.8776 |
  |-----------------------------------------|
  |     5     6   0.8280     0.7956     0.8585 |
  |     6     7   0.8126     0.7772     0.8466 |
  |     7     8   0.8047     0.7660     0.8420 |
  |     8     9   0.7932     0.7510     0.8345 |
  |     9    10   0.7917     0.7451     0.8379 |
  |-----------------------------------------|
  |    10    11   0.7739     0.7222     0.8264 |
  |    11    12   0.7529     0.6959     0.8126 |
  |    12    13   0.7598     0.6966     0.8279 |
  |    13    14   0.7578     0.6865     0.8384 |
  |    14    15   0.7590     0.6749     0.8591 |
  +-----------------------------------------+
```

Same answer as previous part (after rounding).

(d)
```
. strs using popmort [iw=standwei], br(0(1)15) mergeby(_year sex _age)
      standstrata(agegrp) brenner

No late entry detected - p is estimated using the actuarial method
Adjusted survival estimates weighting individual observations as proposed by Brenner.
  +-------------------------------------------------------------------------------+
  | end      n     d     w        p   p_star        r       cp    cp_e2    cr_e2 |
  |-------------------------------------------------------------------------------|
  |   1   5318   151     1   0.9716   0.9768   0.9947   0.9716   0.9768   0.9947 |
  |   2   5166   329   299   0.9344   0.9763   0.9571   0.9079   0.9537   0.9519 |
  |   3   4538   287   296   0.9346   0.9767   0.9569   0.8485   0.9315   0.9109 |
  |   4   3955   211   271   0.9448   0.9771   0.9669   0.8017   0.9102   0.8808 |
  |   5   3473   166   246   0.9504   0.9775   0.9723   0.7619   0.8897   0.8564 |
  |-------------------------------------------------------------------------------|
  |   6   3061   138   240   0.9531   0.9775   0.9751   0.7262   0.8696   0.8350 |
  |   7   2683   105   218   0.9592   0.9772   0.9815   0.6966   0.8499   0.8196 |
  |   8   2360    75   253   0.9664   0.9766   0.9896   0.6732   0.8299   0.8111 |
  |   9   2032    68   241   0.9644   0.9756   0.9885   0.6492   0.8097   0.8018 |
  |  10   1723    50   209   0.9691   0.9756   0.9933   0.6292   0.7900   0.7964 |
  |-------------------------------------------------------------------------------|
  |  11   1464    55   160   0.9603   0.9752   0.9847   0.6042   0.7704   0.7843 |
  |  12   1249    49   157   0.9581   0.9754   0.9823   0.5789   0.7514   0.7704 |
  |  13   1043    21   142   0.9784   0.9743   1.0042   0.5664   0.7321   0.7736 |
  |  14    880    22   168   0.9724   0.9728   0.9995   0.5507   0.7122   0.7732 |
  |  15    690    20   136   0.9678   0.9727   0.9950   0.5330   0.6928   0.7694 |
  +-------------------------------------------------------------------------------+
```

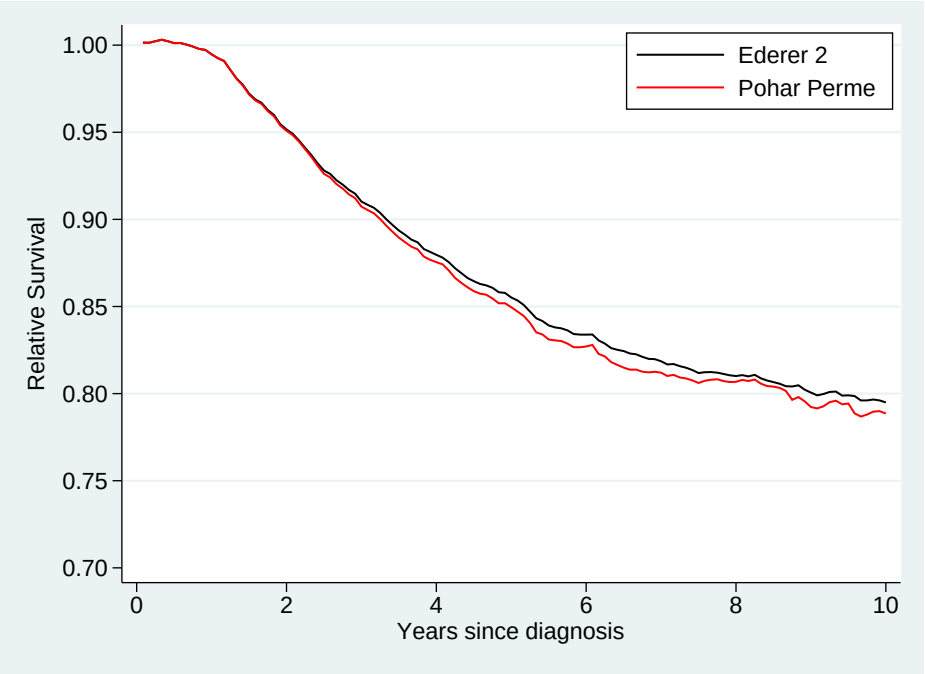
Identical to the crude life table.

(e) The differences between the methods are smaller after age standardisation.

```
. strs using popmort [iw=standwei], br(0(1)10)  mergeby(_year sex _age) ///
> list(start end n d d_star p_star w cr_e1 cr_e2 cr_hak) ///
> ederer1 potfu(potfu) pohar standstrata(agegrp)

Adjusted survival estimates weighting stratum-specific survival in each group
of agegrp by standwei weights.

  +---------------------------------------+
  | start   end    cr_e2    cr_e1   cr_hak |
  |---------------------------------------|
  |     0     1   0.9947   0.9947   0.9947 |
  |     1     2   0.9506   0.9505   0.9506 |
  |     2     3   0.9083   0.9080   0.9083 |
  |     3     4   0.8765   0.8763   0.8768 |
  |     4     5   0.8504   0.8504   0.8513 |
  |---------------------------------------|
  |     5     6   0.8280   0.8280   0.8293 |
  |     6     7   0.8126   0.8126   0.8145 |
  |     7     8   0.8047   0.8045   0.8071 |
  |     8     9   0.7932   0.7930   0.7962 |
  |     9    10   0.7917   0.7920   0.7958 |
  +---------------------------------------+
```

(f) Obtaining the Pohar Perme estimate.

```
. strs using popmort, ///
>         br(0(‘=1/12’)10) ///
>         mergeby(_year sex _age) ///
>         pohar save(replace) notables

         failure _d:  status == 1 2
   analysis time _t:  (exit-origin)/365.24
             origin:  time dx
                 id:  id

No late entry detected - p is estimated using the actuarial method

.
. use grouped, clear
(Collapsed (or grouped) survival data)

. list start end cr_e2 cns_pp if mod(end,1)==0, noobs

  +-------------------------------+
  | start   end    cr_e2   cns_pp |
  |-------------------------------|
  | .9167     1   0.9947   0.9947 |
  | 1.917     2   0.9516   0.9507 |
  | 2.917     3   0.9102   0.9071 |
  | 3.917     4   0.8797   0.8752 |
  | 4.917     5   0.8552   0.8493 |
  |-------------------------------|
  | 5.917     6   0.8338   0.8267 |
  | 6.917     7   0.8186   0.8116 |
  | 7.917     8   0.8101   0.8062 |
  | 8.917     9   0.8006   0.7918 |
  | 9.917    10   0.7950   0.7879 |
  +-------------------------------+
```


Ederer 2
Pohar Perme
Relative Survival
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
2
4
6
8
10
Years since diagnosis

241. **Age-standardised comparisons of relative survival**

(a)
```
. strs using popmort, br(0(1)10) mergeby(_year sex _age) by(year8594)
list(start end n d w cr_e2 lo_cr_e2 hi_cr_e2) save(replace )
         failure _d:  status == 1 2
   analysis time _t:  surv_mm/12
                 id:  id
No late entry detected - p is estimated using the actuarial method
-------------------------------------------------------------------------
-> year8594 = Diagnosed 75-84
  +-------------------------------------------------------------+
  | start   end      n     d   w    cr_e2   lo_cr_e2   hi_cr_e2 |
  |-------------------------------------------------------------|
  |     0     1   2145    63   1   0.9914     0.9831     0.9979 |
  |     1     2   2081   149   2   0.9405     0.9265     0.9529 |
  |     2     3   1930   156   0   0.8835     0.8657     0.8998 |
  |     3     4   1774    92   0   0.8558     0.8362     0.8741 |
  |     4     5   1682    82   0   0.8321     0.8109     0.8519 |
  |-------------------------------------------------------------|
  |     5     6   1600    78   1   0.8091     0.7867     0.8304 |
  |     6     7   1521    67   1   0.7912     0.7677     0.8137 |
  |     7     8   1453    52   0   0.7807     0.7562     0.8042 |
  |     8     9   1401    54   0   0.7687     0.7433     0.7932 |
  |     9    10   1347    44   1   0.7618     0.7355     0.7872 |
  +-------------------------------------------------------------+
-------------------------------------------------------------------------
-> year8594 = Diagnosed 85-94
  +-------------------------------------------------------------------+
  | start   end      n     d     w    cr_e2   lo_cr_e2   hi_cr_e2 |
  |-------------------------------------------------------------------|
  |     0     1   3173    88     0   0.9969     0.9904     1.0022 |
  |     1     2   3085   180   297   0.9599     0.9488     0.9699 |
  |     2     3   2608   131   296   0.9318     0.9178     0.9446 |
  |     3     4   2181   119   271   0.8994     0.8826     0.9150 |
  |     4     5   1791    84   246   0.8745     0.8554     0.8924 |
  |-------------------------------------------------------------------|
  |     5     6   1461    60   239   0.8554     0.8342     0.8754 |
  |     6     7   1162    38   217   0.8440     0.8209     0.8661 |
  |     7     8    907    23   253   0.8398     0.8145     0.8639 |
  |     8     9    631    14   241   0.8387     0.8105     0.8656 |
  |     9    10    376     6   208   0.8421     0.8098     0.8728 |
  +-------------------------------------------------------------------+
```

Based on the crude estimates the 10-year relative survival for the two periods are 0.7618 and 0.8421 respectively.

(b) Using Stata code very similar to that from question 240b we get the following estimates of the weights.

```
. use melanoma
. keep if stage==1
. stset surv_mm, fail(status== 1 2) id(id) scale(12)
. strs using popmort , br(0(1)10) mergeby(_year sex _age) by(agegrp year8594)
save(replace)
. use grouped, clear
. bysort agegrp year8594: gen n0 = n[1]
. bysort agegrp year8594: gen first = _n == 1
. bysort year8594: egen N0 = total(n0*first)
. gen weight=n0/N0
```

```
. list n0 cr_e2 weight if end==10 &  year8594==0 , sum(n0 weight ) mean(cr_e2)
     +---------------------------+
     |   n0    cr_e2      weight |
     |---------------------------|
  6. |  231   0.6422   .10769231 |
 18. |  628   0.7348   .29277389 |
 24. |  652   0.8135    .3039627 |
 26. |  634   0.7604    .2955711 |
     |---------------------------|
Mean |        0.7377             |
 Sum | 2145                    1 |
     +---------------------------+
```

Now, to estimate the age standardised 10-year relative survival for patient diagnosed in the latter period we sum the products of the weights and the corresponding age specific relative survival ratio.

```
. list n0 cr_e2 if end==10 &  year8594==1 , sum(n0) mean(cr_e2)

     +---------------+
     |   n0    cr_e2 |
     |---------------|
 45. |  908   0.8726 |
 63. |  513   0.8103 |
 70. |  941   0.8661 |
 71. |  811   0.8374 |
     |---------------|
Mean |        0.8466 |
 Sum | 3173          |
     +---------------+
.display .3039627*0.8374 + .2955711*0.8661 + .2927739*0.8726 + .1076923*0.8103
.85327007
```

The stratum specific estimates and weights for 1985-1994 are given below.

```
     +---------------------------+
     |   n0    cr_e2     weight |
     |---------------------------|
 20. |  811   0.8374   .2555941 |
 40. |  941   0.8661   .2965648 |
 60. |  908   0.8726   .2861645 |
 80. |  513   0.8103   .1616766 |
     |---------------------------|
Mean |        0.8466             |
 Sum | 3173                      |
     +---------------------------+
```

The age standardised estimate of the 10-year relative survival for patients diagnosed in 1985-1994 is 0.8533, i.e. slightly higher than the crude estimate. Standardisation did not have a large effect in this example. We would expect to see a substantial difference if the age distribution if the two groups disagreed. However, in this case the stratum specific weights are roughly the same.

(c) The estimate is the same (after rounding).

```
. gen standwei = agegrp
. recode standwei 0=0.3039627 1=0.2955711 2=0.2927739 3=0.1076923
. strs using popmort [iw=standwei], br(0(1)10) mergeby(_year sex _age)
standstrata(agegrp) by(year8594)

-> year8594 = Diagnosed 85-94
  +-----------------------------------------+
  | start   end    cr_e2   lo_cr_e2   hi_cr_e2 |
  |-----------------------------------------|
  |     0     1   0.9970     0.9845     1.0046 |
  |     1     2   0.9618     0.9410     0.9782 |
  |     2     3   0.9343     0.9080     0.9563 |
  |     3     4   0.9015     0.8697     0.9292 |
  |     4     5   0.8760     0.8393     0.9085 |
  |-----------------------------------------|
  |     5     6   0.8562     0.8149     0.8935 |
  |     6     7   0.8451     0.7988     0.8876 |
  |     7     8   0.8425     0.7904     0.8908 |
  |     8     9   0.8423     0.7826     0.8986 |
  |     9    10   0.8533     0.7821     0.9200 |
  +-----------------------------------------+
```

(d) The estimate is now 0.8454. Again, this is very similar to the estimates from part a and b indicating that the age distributions of the groups are similar.

```
. strs using popmort [iw=standwei], br(0(1)10) mergeby(_year sex _age)
 standstrata(agegrp) by(year8594)
 list(start end n d w cr_e2 lo_cr_e2 hi_cr_e2) brenner
No late entry detected - p is estimated using the actuarial method

Adjusted survival estimates weighting individual observations as proposed
by Brenner.
-> year8594 = Diagnosed 85-94
  +-------------------------------------------------------------------+
  | start   end      n     d     w    cr_e2   lo_cr_e2   hi_cr_e2 |
  |-------------------------------------------------------------------|
  |     0     1   3173    72     0   0.9970     0.9911     1.0017 |
  |     1     2   3101   158   294   0.9631     0.9528     0.9723 |
  |     2     3   2649   117   301   0.9366     0.9236     0.9485 |
  |     3     4   2231   109   272   0.9058     0.8901     0.9203 |
  |     4     5   1849    78   253   0.8818     0.8640     0.8986 |
  |-------------------------------------------------------------------|
  |     5     6   1518    54   247   0.8643     0.8445     0.8830 |
  |     6     7   1217    37   229   0.8523     0.8306     0.8729 |
  |     7     8    951    22   267   0.8471     0.8233     0.8696 |
  |     8     9    663    13   253   0.8450     0.8186     0.8701 |
  |     9    10    396     6   219   0.8454     0.8148     0.8743 |
  +-------------------------------------------------------------------+
```

(e) The problem occurs if the age distribution between the two calendar periods differs since the estimates would be standardized to different age distributions.

```
. strs using popmort [iw=standwei], br(0(‘=1/12’)10) mergeby(_year sex _age) ///
> standstrata(agegrp) by(year8594) pohar savstand(pohar_q17,replace) notables

No late entry detected - p is estimated using the actuarial method
(2 missing values generated)
(1 missing value generated)
file pohar_q17.dta saved

. use pohar_q17, clear
(Collapsed (or grouped) survival data)

. list year8594 end cns_pp lo_cns_pp hi_cns_pp cr_e2 lo_cr_e2 hi_cr_e2 ///
>         if mod(end,1)==0,noobs
```

year8594	end	cns_pp	lo_cns~p	hi_cns~p	cr_e2	lo_cr_e2	hi_cr_e2
Diagnosed 75-84	1	0.9915	0.9797	0.9964	0.9915	0.9797	0.9964
Diagnosed 75-84	2	0.9397	0.9247	0.9518	0.9398	0.9248	0.9518
Diagnosed 75-84	3	0.8805	0.8618	0.8968	0.8814	0.8629	0.8976
Diagnosed 75-84	4	0.8526	0.8315	0.8712	0.8534	0.8328	0.8717
Diagnosed 75-84	5	0.8260	0.8028	0.8467	0.8284	0.8059	0.8485
Diagnosed 75-84	6	0.8032	0.7780	0.8258	0.8059	0.7816	0.8277
Diagnosed 75-84	7	0.7828	0.7557	0.8073	0.7872	0.7614	0.8106
Diagnosed 75-84	8	0.7750	0.7462	0.8011	0.7761	0.7487	0.8010
Diagnosed 75-84	9	0.7623	0.7313	0.7901	0.7618	0.7329	0.7881
Diagnosed 75-84	10	0.7558	0.7210	0.7869	0.7558	0.7247	0.7839
Diagnosed 85-94	1	0.9971	0.9840	0.9995	0.9970	0.9843	0.9994
Diagnosed 85-94	2	0.9618	0.9511	0.9702	0.9617	0.9510	0.9700
Diagnosed 85-94	3	0.9329	0.9191	0.9443	0.9338	0.9203	0.9450
Diagnosed 85-94	4	0.8995	0.8829	0.9139	0.9009	0.8846	0.9149
Diagnosed 85-94	5	0.8751	0.8558	0.8919	0.8760	0.8572	0.8924
Diagnosed 85-94	6	0.8554	0.8334	0.8746	0.8567	0.8354	0.8754
Diagnosed 85-94	7	0.8468	0.8215	0.8687	0.8465	0.8223	0.8676
Diagnosed 85-94	8	0.8451	0.8156	0.8703	0.8435	0.8158	0.8673
Diagnosed 85-94	9	0.8298	0.7871	0.8648	0.8430	0.8102	0.8705
Diagnosed 85-94	10	0.8335	0.7758	0.8776	0.8519	0.8104	0.8850

242. **Age standardization using** `stpm2`

(a) Fit All age model

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

Iteration 0:   log likelihood = -5177.0146
Iteration 1:   log likelihood = -5065.2884
Iteration 2:   log likelihood = -5060.5916
Iteration 3:   log likelihood = -5060.2545
Iteration 4:   log likelihood =  -5060.254

Log likelihood =  -5060.254                     Number of obs   =       5318
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
       _rcs1 |   1.269834   .1363668     9.31   0.000      1.00256    1.537108
       _rcs2 |   .7862455   .2323387     3.38   0.001       .33087    1.241621
       _rcs3 |  -.0516219   .0852329    -0.61   0.545    -.2186753    .1154316
       _rcs4 |  -.0005241   .0159238    -0.03   0.974    -.0317342    .0306859
       _rcs5 |   .0136257   .0075335     1.81   0.070    -.0011396     .028391
       _cons |  -2.327084   .0632075   -36.82   0.000    -2.450969    -2.2032
------------------------------------------------------------------------------

. range temptime 0 10 100
(5218 missing values generated)

. predict rs_noage, survival timevar(temptime) ci

. list rs_noage* if temptime == 10
     +---------------------------------+
     |  rs_noage   rs_n~lci   rs_n~uci |
     |---------------------------------|
100. | .80927354   .7923437   .8249793 |
     +---------------------------------+
```

Similar to question question 240 where 10 year relative survival estimate was 0.7964.

(b) Proportional excess hazards model

```
. tab agegrp, gen(agegrp)

   Age in 4 |
 categories |      Freq.     Percent        Cum.
------------+-----------------------------------
      0-44 |      1,463       27.51       27.51
     45-59 |      1,575       29.62       57.13
     60-74 |      1,536       28.88       86.01
       75+ |        744       13.99      100.00
------------+-----------------------------------
     Total |      5,318      100.00

. stpm2 agegrp2-agegrp4, scale(hazard) df(5) bhazard(rate) eform

Iteration 0:   log likelihood = -5118.1276
Iteration 1:   log likelihood = -5049.7159
Iteration 2:   log likelihood = -5045.0574
Iteration 3:   log likelihood = -5044.5816
Iteration 4:   log likelihood = -5044.5762
Iteration 5:   log likelihood = -5044.5762
```

```
Log likelihood = -5044.5762                     Number of obs   =       5318
------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
     agegrp2 |   1.217689   .1314222     1.82   0.068     .9855262    1.504542
     agegrp3 |   1.573098   .1787931     3.99   0.000     1.258957    1.965624
     agegrp4 |   2.523066   .4004497     5.83   0.000     1.848544    3.443718
       _rcs1 |   3.514151   .4644091     9.51   0.000     2.712257    4.553131
       _rcs2 |   2.117739   .4739337     3.35   0.001     1.365778     3.28371
       _rcs3 |   .9535632   .0777112    -0.58   0.560     .8127937    1.118713
       _rcs4 |   1.002938   .0152736     0.19   0.847     .9734445    1.033325
       _rcs5 |   1.013271   .0075069     1.78   0.075     .9986643    1.028092
       _cons |   .0791188   .0069135   -29.03   0.000     .0666654    .0938985
------------------------------------------------------------------------------
```

```
. predict rs0, survival zeros timevar(temptime)
. predict rs1, survival at(agegrp2 1) zeros timevar(temptime)
. predict rs2, survival at(agegrp3 1) zeros timevar(temptime)
. predict rs3, survival at(agegrp4 1) zeros timevar(temptime)
```

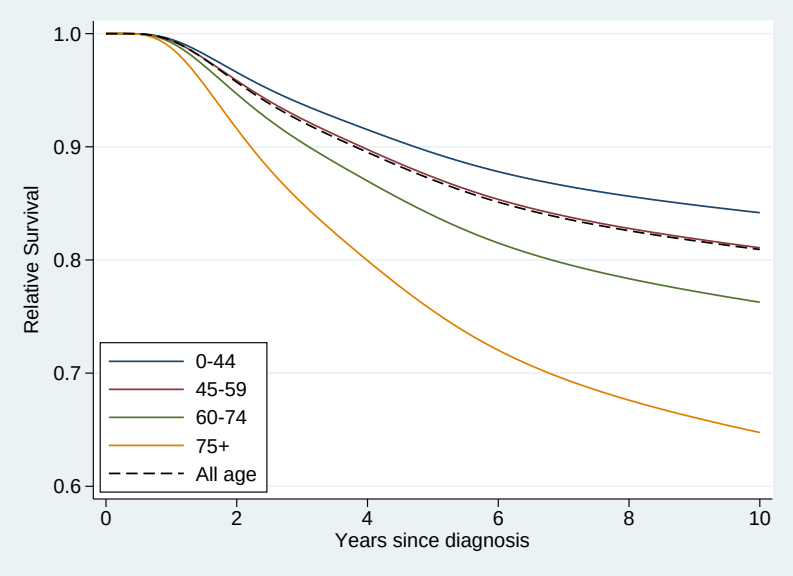


Figure 68: Melanoma Data. Relative survival by age group and all age estimate.

(c) Age standardized estimate.

```
. tab agegrp
```

Age in 4 categories	Freq.	Percent	Cum.
0-44	1,463	27.51	27.51
45-59	1,575	29.62	57.13
60-74	1,536	28.88	86.01
75+	744	13.99	100.00
Total	5,318	100.00	

```
. gen rs_stand1 = 0.2751*rs0 + 0.2962*rs1 + 0.2888*rs2 + 0.1399*rs3
(5218 missing values generated)
```

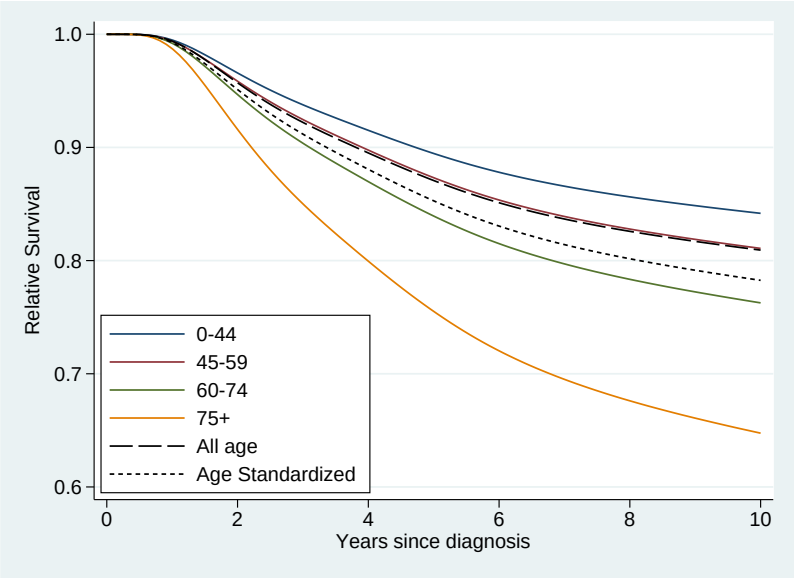


Figure 69: Melanoma Data. Relative survival by age group, all age and age standardized estimate.

Age standardized relative survival is lower than the all age estimate.

(d) Age standardized relative survival at 10 years

```
. list rs_stand1 if temptime == 10

     +----------+
     | rs_sta~1 |
     |----------|
100. | .7825714 |
     +----------+
```

Similar to question 240 where it was 0.7916. Note that we are making an assumption of proportional excess hazards in our model based estimate.

(e) The two ways of estimating age standardized relative survival give identical results.

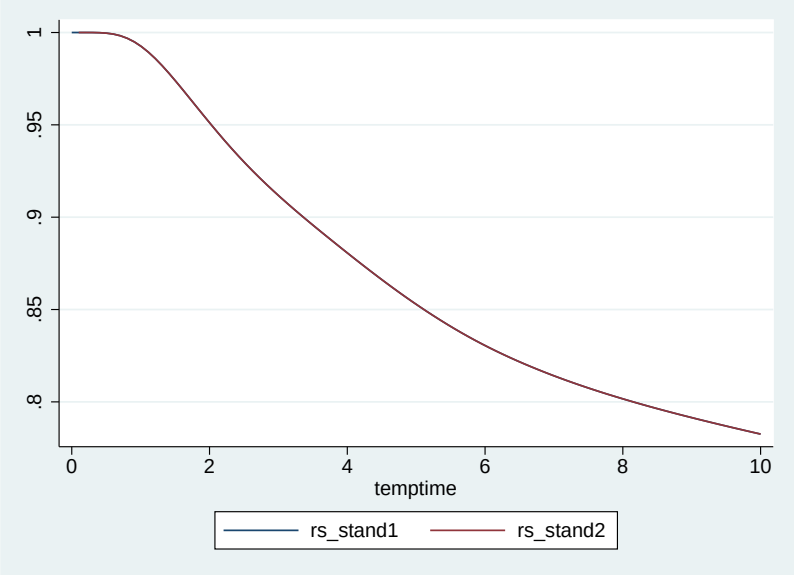


Figure 70: Melanoma Data. Age standardized survival using 2 methods of calculation.

(f) Confidence intervals for age standardized relative survival.

```
. predict rs_stand3, meansurv timevar(temptime) ci
. list rs_stand3* if temptime == 10
     +---------------------------------+
     | rs_stand3   rs_st~lci   rs_s~uci |
     |---------------------------------|
100. | .78256963   .76368048    .801926 |
     +---------------------------------+
```

The width of the confidence interval is narrower than in question 240 as we are making an additional assumption of proportional excess hazards.

(g) Proportional excess hazards model for age group and calendar period

```
. stpm2 agegrp2-agegrp4 year8594, scale(hazard) df(5) bhazard(rate)

Iteration 0:   log likelihood = -5108.8352
Iteration 1:   log likelihood = -5039.0263
Iteration 2:   log likelihood = -5033.8175
Iteration 3:   log likelihood = -5033.2664
Iteration 4:   log likelihood =  -5033.261
Iteration 5:   log likelihood =  -5033.261

Log likelihood =  -5033.261                     Number of obs   =       5318
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
     agegrp2 |   .2213339    .107728     2.05   0.040     .0101909    .4324769
     agegrp3 |   .4782906   .1131929     4.23   0.000     .2564366    .7001445
     agegrp4 |   .9539447   .1602753     5.95   0.000      .639811    1.268078
    year8594 |  -.4173968   .0888288    -4.70   0.000     -.591498   -.2432956
       _rcs1 |   1.233252    .124483     9.91   0.000       .98927    1.477234
       _rcs2 |   .7464762   .2105896     3.54   0.000     .3337282    1.159224
       _rcs3 |  -.0428173   .0772008    -0.55   0.579    -.1941281    .1084935
       _rcs4 |   .0020997   .0148513     0.14   0.888    -.0270083    .0312078
       _rcs5 |   .0132815    .007116     1.87   0.062    -.0006656    .0272286
       _cons |  -2.345981   .0933391   -25.13   0.000    -2.528923    -2.16304
------------------------------------------------------------------------------

. predict rs, survival
. table agegrp year8594, c(mean rs) format(%5.3f)

-------------------------------------------
Age in 4  |  Indicator for diagnosed during
categorie |             1985-94
s         | Diagnosed 75-84  Diagnosed 85-94
----------+--------------------------------
     0-44 |           0.834            0.915
    45-59 |           0.810            0.902
    60-74 |           0.790            0.888
      75+ |           0.763            0.875
-------------------------------------------
```

Relative survival has improved over calendar period in all age groups.

(h) Has age distribution changed?

```
. tab agegrp year8594 , col
```

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+
```

Age in 4 categories	Indicator for diagnosed during 1985-94: Diagnosed	Diagnosed	Total
0-44	652	811	1,463
	30.40	25.56	27.51
45-59	634	941	1,575
	29.56	29.66	29.62
60-74	628	908	1,536
	29.28	28.62	28.88
75+	231	513	744
	10.77	16.17	13.99
Total	2,145	3,173	5,318
	100.00	100.00	100.00

There are more subjects in the 75+ group in the latter period

(i) The age standardized relative survival in the two periods is shown below (Figure 71). The first period is the reference period.

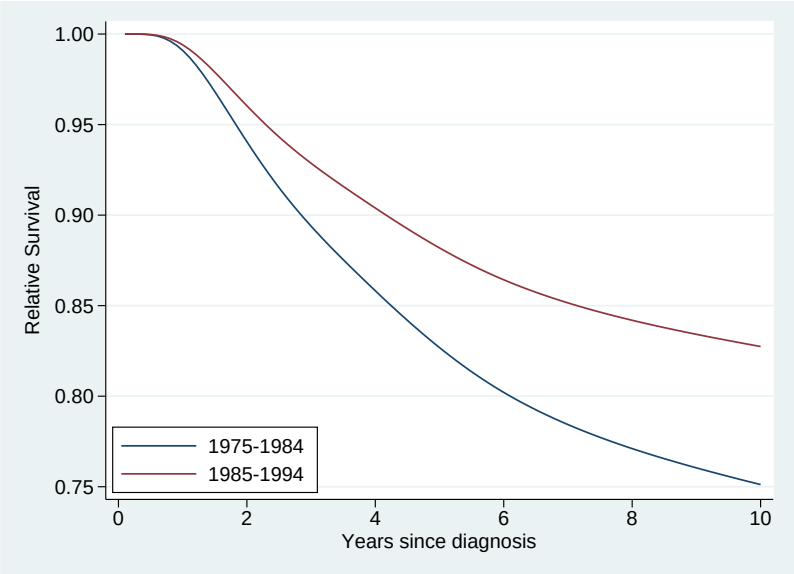


Figure 71: Melanoma Data. Age standardized survival in two calendar periods with the first period as the reference.

Clear difference between the two calendar periods.

```
. list rs_7584 rs_8594 if temptime == 10
     +-----------------------+
     |   rs_7584     rs_8594 |
     |-----------------------|
 100. | .75120467   .82744076 |
     +-----------------------+
```

There is a small difference when compared to question 241. This is likely due to the assumption of proportional excess hazards.

(j) Age-standardized estimate with 1985-1994 as the reference.

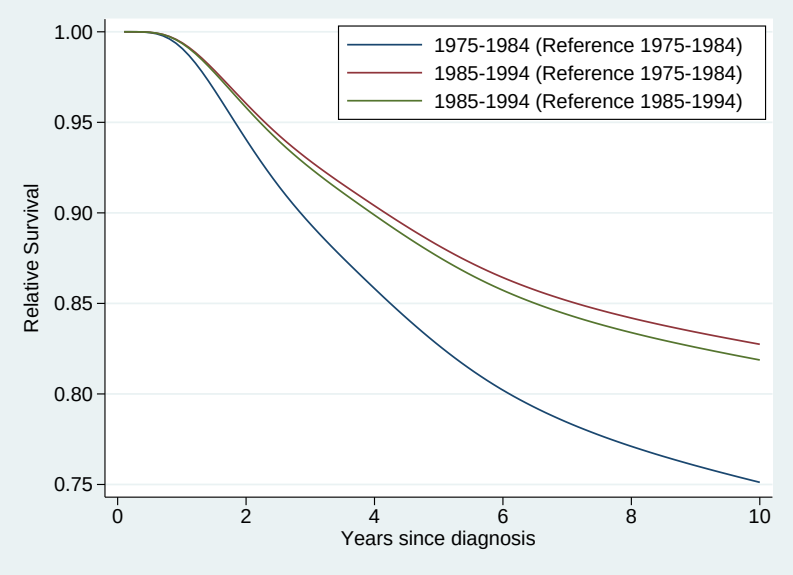


Figure 72: Melanoma Data. Age standardized survival in two calendar periods with the first period as the reference. Also shown is the age standard estimate for the second period with the second period used as the reference

The age-standardized estimate for the second period is lower when using the age distribution in the second period rather than in the first period. This is because the population in slightly older in the second calendar period and relative survival decreases with age.

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

(a) Calculate the age-standardised 5-year RSR (traditional direct standardisation - Ederer II method) using the standardisation options in `strs` for all patients diagnosed with localised melanoma 1975-1994. Use the age groups defined in the table above.

```
. use melanoma, clear
(Skin melanoma, diagnosed 1975-94, follow-up to 1995)

. keep if stage==1 /* restrict to localised */
(2457 observations deleted)

. stset surv_mm, fail(status==1 2) id(id) scale(12)

                id:  id
     failure event:  status == 1 2
obs. time interval:  (surv_mm[_n-1], surv_mm]
 exit on or before:  failure
    t for analysis:  time/12
------------------------------------------------------------------------------
     5318  total observations
        0  exclusions
------------------------------------------------------------------------------
     5318  observations remaining, representing
     5318  subjects
     1795  failures in single-failure-per-subject data
 38626.58  total analysis time at risk and under observation
                                         at risk from t =         0
                               earliest observed entry t =         0
                                    last observed exit t =  20.95833

. /* generate an age group variable for the 5 groupings */
. recode age (min/44=1) (45/54=2) (55/64=3) (65/74=4) (75/max=5), gen(agegrpICSS)
(5318 differences between age and agegrpICSS)

. label variable agegrpICSS "Age groups for ICSS"
. label define agegrpICSS 1 "0-44" 2 "45-54" 3 "55-64" 4 "65-74" 5 "75+"
. label values agegrpICSS agegrpICSS

. /*Generate the internal weights based on the age distribution of the data*/
. local totalobs = _N
. bysort agegrpICSS: gen standwei = _N/`totalobs'
. label variable standwei "Internal age group weights"

. /* Age-standardised using traditional approach implemented with iweights */
. strs using popmort [iw=standwei], br(0(1)10) mergeby(_year sex _age) ///
> list(n d w cr_e2 se_cp) standstrata(agegrpICSS) ///
> savstand(internal,replace)

        failure _d:  status == 1 2
  analysis time _t:  surv_mm/12
                id:  id

No late entry detected - p is estimated using the actuarial method
```

(b) Calculate the externally age-standardised 5-year RSR using the standardisation options in `strs` by using the ICSS 2 weights given in the table above.

```
. recode age (min/44=0.28) (45/54=0.17) (55/64=0.21) (65/74=0.20) (75/max=0.14), gen(ICSS2wei)
(5318 differences between age and ICSS2wei)
. label variable ICSS2wei "ICSS2 age group weights"
```

```
. strs using popmort [iw=ICSS2wei], br(0(1)10) mergeby(_year sex _age) ///
>  list(n d w cr_e2 se_cp) standstrata(agegrpICSS) ///
> savstand(external,replace)

        failure _d:  status == 1 2
  analysis time _t:  surv_mm/12
                id:  id

No late entry detected - p is estimated using the actuarial method
```

(c) Compare the estimates using the two different weights. Are they similar? Did you expect them to be?

```
. bys agegrpICSS: gen ind=1 if _n==1
(5313 missing values generated)

. list agegrpICSS standwei ICSS2wei if ind==1, noobs
  +--------------------------------+
  | agegrp~S   standwei   ICSS2wei |
  |--------------------------------|
  |     0-44   .2751034        .28 |
  |    45-54   .1904851        .17 |
  |    55-64   .2098533        .21 |
  |    65-74   .1846559         .2 |
  |      75+   .1399022        .14 |
  +--------------------------------+

. use internal, replace
(Age-standardized survival data)

. list end cr_e2 lo_cr_e2 hi_cr_e2 if end==5, noobs
  +-----------------------------------+
  | end    cr_e2   lo_cr_e2   hi_cr_e2 |
  |-----------------------------------|
  |   5   0.8508     0.8355     0.8648 |
  +-----------------------------------+

. use external, replace
(Age-standardized survival data)

. list end cr_e2 lo_cr_e2 hi_cr_e2 if end==5, noobs
  +-----------------------------------+
  | end    cr_e2   lo_cr_e2   hi_cr_e2 |
  |-----------------------------------|
  |   5   0.8505     0.8351     0.8647 |
  +-----------------------------------+
```

The estimates do appear to be quite similar. This is because the external weights are very similar to the internal weights for this particular dataset.

(d) Repeat part (b) using the ICSS 1 weights instead. What do you expect to happen to the standardised estimate when standardising to an older age distribution?

```
. use melanoma, clear
(Skin melanoma, diagnosed 1975-94, follow-up to 1995)

. keep if stage==1 /* restrict to localised */
(2457 observations deleted)

. stset surv_mm, fail(status==1 2) id(id) scale(12)
```

```
               id:  id
     failure event:  status == 1 2
obs. time interval:  (surv_mm[_n-1], surv_mm]
 exit on or before:  failure
    t for analysis:  time/12
--------------------------------------------------------------------------------
     5318  total observations
        0  exclusions
--------------------------------------------------------------------------------
     5318  observations remaining, representing
     5318  subjects
     1795  failures in single-failure-per-subject data
 38626.58  total analysis time at risk and under observation
                                             at risk from t =         0
                                  earliest observed entry t =         0
                                       last observed exit t =  20.95833

. recode age (min/44=1) (45/54=2) (55/64=3) (65/74=4) (75/max=5), gen(agegrpICSS)
(5318 differences between age and agegrpICSS)

. label variable agegrpICSS "Age groups for ICSS"
. label define agegrpICSS 1 "0-44" 2 "45-54" 3 "55-64" 4 "65-74" 5 "75+"
. label values agegrpICSS agegrpICSS

. recode age (min/44=0.07) (45/54=0.12) (55/64=0.23) (65/74=0.29) (75/max=0.29), gen(ICSS1wei)
(5318 differences between age and ICSS1wei)

. label variable ICSS1wei "ICSS1 age group weights"

. strs using popmort [iw=ICSS1wei], br(0(1)10) mergeby(_year sex _age) ///
>  list(n d w cr_e2 se_cp) standstrata(agegrpICSS)  ///
> savstand(externalICSS1,replace)

         failure _d:  status == 1 2
   analysis time _t:  surv_mm/12
                 id:  id

No late entry detected - p is estimated using the actuarial method

. use internal, replace
(Age-standardized survival data)

. list end cr_e2 lo_cr_e2 hi_cr_e2 if end==5, noobs
  +-----------------------------------+
  | end    cr_e2   lo_cr_e2   hi_cr_e2 |
  |-----------------------------------|
  |   5   0.8508     0.8355     0.8648 |
  +-----------------------------------+

. use external, replace
(Age-standardized survival data)

. list end cr_e2 lo_cr_e2 hi_cr_e2 if end==5, noobs
  +-----------------------------------+
  | end    cr_e2   lo_cr_e2   hi_cr_e2 |
  |-----------------------------------|
  |   5   0.8505     0.8351     0.8647 |
  +-----------------------------------+
```

```
. use externalICSS1, replace
(Age-standardized survival data)

. list end cr_e2 lo_cr_e2 hi_cr_e2 if end==5, noobs
  +-----------------------------------+
  | end    cr_e2   lo_cr_e2   hi_cr_e2 |
  |-----------------------------------|
  |   5   0.8222     0.7972     0.8443 |
  +-----------------------------------+
```

Standardising to the older age distribution results in a lower age-standardised estimate of relative survival. This is because the older patients have poorer survival.

250. **Calculating the crude probability of death from life tables.**

(a) Load the Melanoma data, drop subjects diagnosed 1975-1984 and then and use `strs` to obtain life-tables stratified by age group and sex. Use the `cuminc` option to obtain the crude probabilities of death due to cancer and due to other causes.

```
. stset surv_mm, fail(status==1 2) id(id) scale(12)

                id:  id
     failure event:  status == 1 2
obs. time interval:  (surv_mm[_n-1], surv_mm]
 exit on or before:  failure
    t for analysis:  time/12

--------------------------------------------------------------------------------
     4744  total observations
        0  exclusions
--------------------------------------------------------------------------------
     4744  observations remaining, representing
     4744  subjects
     1404  failures in single-failure-per-subject data
  22108.5  total analysis time at risk and under observation
                                          at risk from t =         0
                                earliest observed entry t =         0
                                     last observed exit t =  10.95833

. strs using popmort, br(0(1)5) mergeby(_year sex _age) by(agegrp sex) ///
>         save(replace) cuminc list(n d w cp F cp_e2 cr_e2 ci_dc ci_do) f(%7.5f)

         failure _d:  status == 1 2
   analysis time _t:  surv_mm/12
                id:  id

No late entry detected - p is estimated using the actuarial method
-------------------------------------------------------------------------------------------
-> agegrp = 0-44, sex = Male

  +-----------------------------------------------------------------------------------------+
  | start   end     n    d    w        cp         F     cp_e2     cr_e2     ci_dc     ci_do |
  |-----------------------------------------------------------------------------------------|
  |     0     1   537   25    0   0.95345   0.04655   0.99727   0.95605   0.04389   0.00267 |
  |     1     2   512   33   43   0.88930   0.11070   0.99437   0.89433   0.10535   0.00535 |
  |     2     3   436    9   43   0.86999   0.13001   0.99130   0.87762   0.12194   0.00807 |
  |     3     4   384   18   39   0.82703   0.17297   0.98810   0.83698   0.16216   0.01081 |
  |     4     5   327    6   34   0.81102   0.18898   0.98473   0.82360   0.17537   0.01361 |
  +-----------------------------------------------------------------------------------------+

-------------------------------------------------------------------------------------------
-> agegrp = 0-44, sex = Female

  +-----------------------------------------------------------------------------------------+
  | start   end     n   d    w        cp         F     cp_e2     cr_e2     ci_dc     ci_do |
  |-----------------------------------------------------------------------------------------|
  |     0     1   624   9    0   0.98558   0.01442   0.99911   0.98645   0.01354   0.00088 |
  |     1     2   615   9   52   0.97052   0.02948   0.99816   0.97231   0.02766   0.00182 |
  |     2     3   554   9   56   0.95391   0.04609   0.99712   0.95667   0.04327   0.00282 |
  |     3     4   489   8   51   0.93745   0.06255   0.99599   0.94122   0.05867   0.00389 |
  |     4     5   430   8   68   0.91851   0.08149   0.99477   0.92334   0.07647   0.00503 |
  +-----------------------------------------------------------------------------------------+
```

```
-------------------------------------------------------------------------------------------
-> agegrp = 45-59, sex = Male

  +-------------------------------------------------------------------------------------+
  | start   end     n    d    w        cp         F     cp_e2     cr_e2     ci_dc     ci_do |
  |-------------------------------------------------------------------------------------|
  |     0     1   752   51    0   0.93218   0.06782   0.99094   0.94070   0.05903   0.00879 |
  |     1     2   701   38   72   0.87891   0.12109   0.98140   0.89557   0.10353   0.01755 |
  |     2     3   591   38   64   0.81917   0.18083   0.97111   0.84354   0.15433   0.02650 |
  |     3     4   489   17   61   0.78879   0.21121   0.96025   0.82145   0.17566   0.03554 |
  |     4     5   411   16   53   0.75597   0.24403   0.94866   0.79688   0.19912   0.04491 |
  +-------------------------------------------------------------------------------------+

-------------------------------------------------------------------------------------------
-> agegrp = 45-59, sex = Female

  +-------------------------------------------------------------------------------------+
  | start   end     n    d    w        cp         F     cp_e2     cr_e2     ci_dc     ci_do |
  |-------------------------------------------------------------------------------------|
  |     0     1   612   21    0   0.96569   0.03431   0.99661   0.96897   0.03098   0.00333 |
  |     1     2   591   23   61   0.92606   0.07394   0.99298   0.93261   0.06715   0.00679 |
  |     2     3   507   16   64   0.89487   0.10513   0.98906   0.90477   0.09474   0.01039 |
  |     3     4   427   11   62   0.87001   0.12999   0.98482   0.88341   0.11581   0.01418 |
  |     4     5   354    5   49   0.85681   0.14319   0.98034   0.87399   0.12508   0.01812 |
  +-------------------------------------------------------------------------------------+

-------------------------------------------------------------------------------------------
-> agegrp = 60-74, sex = Male

  +-------------------------------------------------------------------------------------+
  | start   end     n    d    w        cp         F     cp_e2     cr_e2     ci_dc     ci_do |
  |-------------------------------------------------------------------------------------|
  |     0     1   709   61    0   0.91396   0.08604   0.96735   0.94481   0.05429   0.03175 |
  |     1     2   648   67   75   0.81366   0.18634   0.93361   0.87152   0.12395   0.06239 |
  |     2     3   506   37   63   0.75021   0.24979   0.89794   0.83548   0.15695   0.09283 |
  |     3     4   406   39   55   0.67291   0.32709   0.86090   0.78164   0.20430   0.12279 |
  |     4     5   312   27   51   0.60950   0.39050   0.82214   0.74135   0.23821   0.15230 |
  +-------------------------------------------------------------------------------------+

-------------------------------------------------------------------------------------------
-> agegrp = 60-74, sex = Female

  +-------------------------------------------------------------------------------------+
  | start   end     n    d    w        cp         F     cp_e2     cr_e2     ci_dc     ci_do |
  |-------------------------------------------------------------------------------------|
  |     0     1   661   41    0   0.93797   0.06203   0.98381   0.95340   0.04622   0.01581 |
  |     1     2   620   47   60   0.86325   0.13675   0.96623   0.89343   0.10470   0.03205 |
  |     2     3   513   31   62   0.80773   0.19227   0.94730   0.85267   0.14369   0.04857 |
  |     3     4   420   22   52   0.76263   0.23737   0.92670   0.82295   0.17154   0.06583 |
  |     4     5   346   18   48   0.72000   0.28000   0.90473   0.79582   0.19638   0.08362 |
  +-------------------------------------------------------------------------------------+
```

```
-------------------------------------------------------------------------------------------
-> agegrp = 75+, sex = Male

  +-------------------------------------------------------------------------------------+
  | start   end     n    d    w        cp         F     cp_e2     cr_e2     ci_dc     ci_do |
  |-------------------------------------------------------------------------------------|
  |     0     1   337   67    0   0.80119   0.19881   0.88853   0.90170   0.09282   0.10599 |
  |     1     2   270   61   37   0.60686   0.39314   0.78562   0.77247   0.20100   0.19214 |
  |     2     3   172   33   17   0.48438   0.51562   0.68883   0.70319   0.25207   0.26355 |
  |     3     4   122   19   19   0.40257   0.59743   0.59992   0.67104   0.27279   0.32464 |
  |     4     5    84   11   12   0.34580   0.65420   0.52181   0.66269   0.27747   0.37673 |
  +-------------------------------------------------------------------------------------+

-------------------------------------------------------------------------------------------
-> agegrp = 75+, sex = Female

  +-------------------------------------------------------------------------------------+
  | start   end     n    d    w        cp         F     cp_e2     cr_e2     ci_dc     ci_do |
  |-------------------------------------------------------------------------------------|
  |     0     1   512   68    0   0.86719   0.13281   0.91552   0.94721   0.05056   0.08225 |
  |     1     2   444   75   47   0.71252   0.28748   0.83184   0.85655   0.12977   0.15772 |
  |     2     3   322   50   32   0.59609   0.40391   0.75041   0.79436   0.17897   0.22494 |
  |     3     4   240   39   27   0.49345   0.50655   0.67530   0.73072   0.22433   0.28221 |
  |     4     5   174   23   24   0.42340   0.57660   0.60436   0.70057   0.24363   0.33298 |
  +-------------------------------------------------------------------------------------+
```

(b) How is the probability of death due to all causes, F, calculated?

This is just 1 - the survival function , i.e. `1-cp`.

(c) Why is the crude probability of death due to cancer, `ci_dc` similar to the all-cause probability of death for subjects aged 0-44?

```
. use grouped, clear
(Collapsed (or grouped) survival data)

. list agegrp start end sex F ci_dc if agegrp == 0 & sex == 1, noobs
  +-----------------------------------------------+
  | agegrp   start   end    sex         F     ci_dc |
  |-----------------------------------------------|
  |   0-44       0     1   Male   0.04655   0.04389 |
  |   0-44       1     2   Male   0.11070   0.10535 |
  |   0-44       2     3   Male   0.13001   0.12194 |
  |   0-44       3     4   Male   0.17297   0.16216 |
  |   0-44       4     5   Male   0.18898   0.17537 |
  +-----------------------------------------------+
```

They are similar as there is low probability that subjects of this age will die from other causes. Thus, if they die it is highly likely to be due to cancer.

(d) For both males and females aged 60-74 what is the probability of death due to all causes at 5 years post diagnosis? What two variables can be added together to give the probability of death due to all-causes?}

```
. list  end agegrp sex F ci_dc ci_do if  agegrp == 2 & end == 5
     +-----------------------------------------------------+
     | end   agegrp      sex         F     ci_dc     ci_do |
     |-----------------------------------------------------|
 25. |   5    60-74     Male   0.39050   0.23821   0.15230 |
 30. |   5    60-74   Female   0.28000   0.19638   0.08362 |
     +-----------------------------------------------------+

. gen F2 = ci_dc + ci_do
. list  end agegrp sex F ci_dc ci_do F2 if  agegrp == 2 & end == 5
```

```
     +-------------------------------------------------------------------+
     | end   agegrp      sex         F     ci_dc     ci_do         F2 |
     |-------------------------------------------------------------------|
 25. |   5    60-74     Male   0.39050   0.23821   0.15230   .3905036 |
 30. |   5    60-74   Female   0.28000   0.19638   0.08362   .2800009 |
     +-------------------------------------------------------------------+
```

The probability of death due to all causes is 0.39 for males and 0.28 for females. With crude mortality we partition the all-cause probability of death into that due to cancer and that due to other cause. Thus `F = ci_dc + ci_do`.

(e) What proportion of the all-cause deaths at 5 years post diagnosis are due to cancer and due to other causes for males? Compare these figures for the different age groups.

```
. gen prob_c = ci_dc / F
. gen prob_o = ci_do / F
. list  end agegrp sex F ci_dc ci_do prob_c prob_o ///
>         if  end == 5 & sex == 1, noobs
```

```
  +-----------------------------------------------------------------------------+
  | end   agegrp    sex         F     ci_dc     ci_do     prob_c     prob_o |
  |-----------------------------------------------------------------------------|
  |   5     0-44   Male   0.18898   0.17537   0.01361     .92796   .0720402 |
  |   5    45-59   Male   0.24403   0.19912   0.04491   .8159498   .1840501 |
  |   5    60-74   Male   0.39050   0.23821   0.15230   .6100003   .3899997 |
  |   5      75+   Male   0.65420   0.27747   0.37673   .4241378   .5758622 |
  +-----------------------------------------------------------------------------+
```

In the youngest age group 93% of the deaths are associated with a diagnosis of cancer at 5 years poist diagnosis. In the oldest agegroup the figure is 42%. This is due to increased probability of dying from other causes in the oldest age group.

(f) The age groups are fairly wide, explain how you would expect the crude probability of death due to cancer to differ between a 60 and 74 year old, even if the relative survival was identical.

Since the probability of death due to other cause is higher for a 74 year old than for a 60 year old then if relative survival was identical we would expect the actual probability of death due to cancer to be lower for someone aged 74 than a 60 year old.

(g) Plot the net probability of death, the crude probability of death due to cancer and the overall probability of death for males by age group. Try to understand the relationship between these various measures.

```
. gen net = 1- cr_e2

. twoway  (line F net ci_dc end if sex == 1, sort ), by(agegrp) ///
>                 legend(order(1 "Overall" 2 "Net" 3 "Crude") cols(3)) ///
>                 ylabel(0(0.1)0.6, angle(h) format(%3.1f)) ///
>                 ytitle("Probability of Death")
```

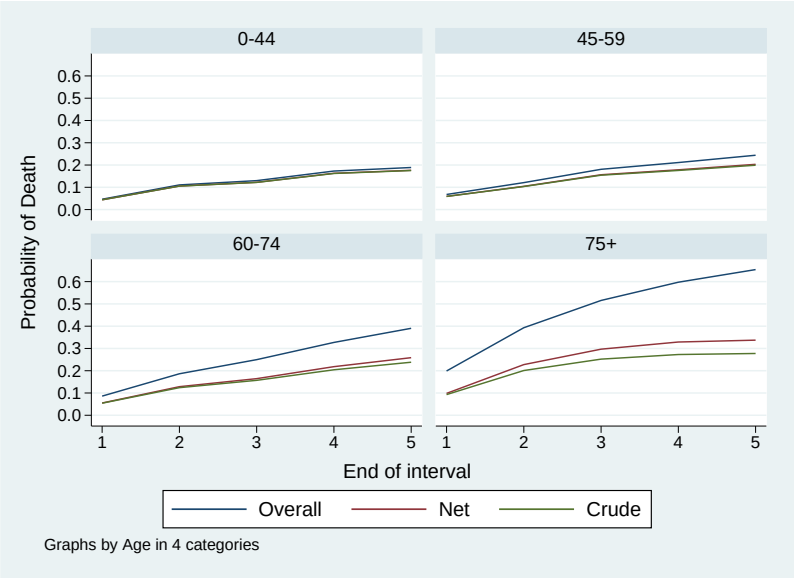


Figure 73: Melanoma Data. All cause, Net and Crude Probability of Death due to cancer.

Very little difference between the estimates in youngest age group. Increasing separation as age increases due to increased contribution of deaths due to other causes.

251. **Estimating crude mortality from flexible parametric relative survival models**

(a) Load the Melanoma data and merge in the background mortality rates. Fit a flexible parametric relative survival model including age group with time-dependent effects.
Obtain the predicted relative survival function for each age group. Calculate the estimated net mortality (1 - relative survival) and plot the four curves on a single graph. Interpret the plot.

```
. tab agegrp, gen(agegrp)

   Age in 4 |
 categories |      Freq.     Percent        Cum.
------------+-----------------------------------
       0-44 |      2,046       26.32       26.32
      45-59 |      2,238       28.78       55.10
      60-74 |      2,280       29.32       84.42
        75+ |      1,211       15.58      100.00
------------+-----------------------------------
      Total |      7,775      100.00

. stpm2 agegrp2-agegrp4, scale(hazard) bhazard(rate) df(5) ///
>         tvc(agegrp2-agegrp4) dftvc(3)

Log likelihood = -6711.5119                     Number of obs   =       7775
------------------------------------------------------------------------------
                |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
xb              |
        agegrp2 |   .3324786   .0865048     3.84   0.000     .1629322    .5020249
        agegrp3 |   .6293067   .0855523     7.36   0.000     .4616273    .7969862
        agegrp4 |   .9036727   .1108223     8.15   0.000     .6864651     1.12088
          _rcs1 |   .7609295   .0492328    15.46   0.000      .664435    .8574241
          _rcs2 |   .0904807   .0405564     2.23   0.026     .0109916    .1699698
          _rcs3 |   .0353396   .0180933     1.95   0.051    -.0001226    .0708019
          _rcs4 |   .0129533   .0085917     1.51   0.132    -.0038861    .0297927
          _rcs5 |  -.0017028   .0038639    -0.44   0.659    -.0092759    .0058702
  _rcs_agegrp21 |   .0234426   .0664294     0.35   0.724    -.1067565    .1536418
  _rcs_agegrp22 |   .0075705   .0540635     0.14   0.889    -.0983919     .113533
  _rcs_agegrp23 |   .0192523   .0242994     0.79   0.428    -.0283736    .0668782
  _rcs_agegrp31 |    .034558   .0664849     0.52   0.603    -.0957501    .1648661
  _rcs_agegrp32 |   .0029712   .0542771     0.05   0.956    -.1034101    .1093524
  _rcs_agegrp33 |   .0216319   .0241749     0.89   0.371    -.0257501    .0690139
  _rcs_agegrp41 |   .0455839   .0902801     0.50   0.614    -.1313618    .2225295
  _rcs_agegrp42 |   .0677156    .075406     0.90   0.369    -.0800775    .2155087
  _rcs_agegrp43 |   .0168436   .0324306     0.52   0.603    -.0467191    .0804063
          _cons |  -2.279718   .0653933   -34.86   0.000    -2.407886   -2.151549
------------------------------------------------------------------------------

. range temptime 0 10 1000
(6775 missing values generated)

. predict nm1, failure zeros      timevar(temptime)
. predict nm2, failure at(agegrp2 1)      zeros timevar(temptime)
. predict nm3, failure at(agegrp3 1)      zeros timevar(temptime)
. predict nm4, failure at(agegrp4 1)      zeros timevar(temptime)
```

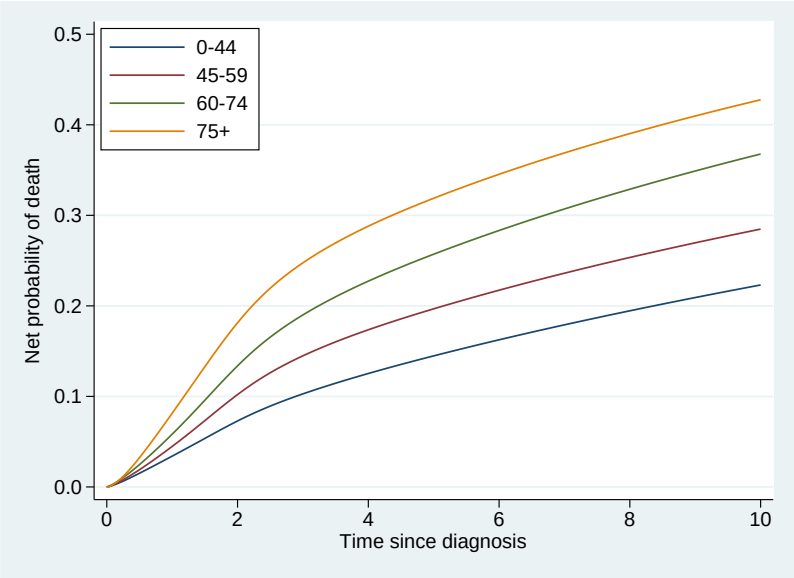


Figure 74: Melanoma Data. Net probability of death due to cancer

Figure 74 shows the estimated net probability of death due to cancers, i.e. survival in a hypothetical world where it is not possible to die of other causes.

(b) Use the `stpm2cm` command to estimate the crude probability of death. Note that `stpm2cm` will predict for individual covariate patterns and for ages at diagnosis. Perform the predictions for males aged 40, 55, 70 and 80 diagnosed in 1985.

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

. stpm2cm using popmort, at(agegrp2 1 agegrp3 0 agegrp4 0) ///
>                                                  mergeby(_year sex _age) ///
>                                                  diagage(55) diagyear(1985) ///
>                                                  sex(1) stub(cm2) nobs(1000) ///
>                                                  tgen(cm2_t)

. stpm2cm using popmort, at(agegrp2 0 agegrp3 1 agegrp4 0) ///
>                                                  mergeby(_year sex _age) ///
>                                                  diagage(70) diagyear(1985) ///
>                                                  sex(1) stub(cm3) nobs(1000) ///
>                                                  tgen(cm3_t)

. stpm2cm using popmort, at(agegrp2 0 agegrp3 0 agegrp4 1) ///
>                                                  mergeby(_year sex _age) ///
>                                                  diagage(80) diagyear(1985) ///
>                                                  sex(1) stub(cm4) nobs(1000) ///
>                                                  tgen(cm4_t)
```

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

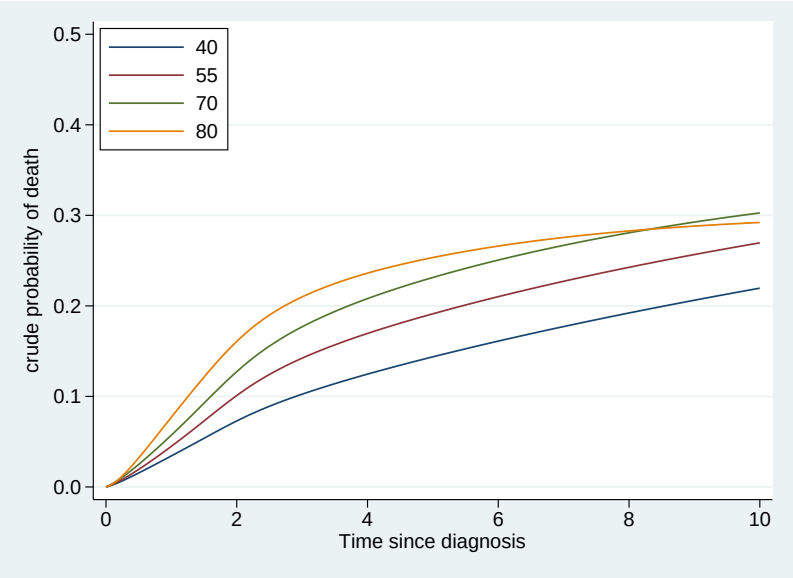


Figure 75: Melanoma Data. Crude probability of death due to cancer

Figure 75 shows the crude probability of death due to cancer. For the youngest age group there is very little difference between the net and the crude estimate since these individuals have a low risk of death due to other causes. However, there is a noticable change for the oldest group since these individuals are at increased risk of death due to other causes.

(c) Generate a similar plot but for the crude probability of death due to other causes.

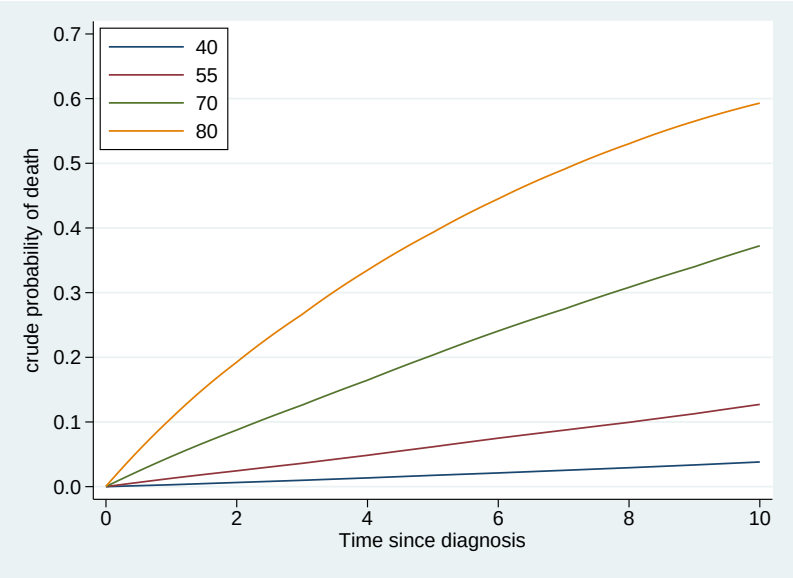


Figure 76: Melanoma Data. Crude probability of death due to other causes

Figure 76 shows that the oldest patients have the highest risk of death to other causes.

(d) A useful way of presenting crude probabilities is through stacked graphs. Generate the stacked graphs for each of the selected ages. Use the solution Do file for help.

```
. gen cm1_do = cm1_d + cm1_o
(6775 missing values generated)
```

```
. gen cm2_do = cm2_d + cm2_o
(6775 missing values generated)

. gen cm3_do = cm3_d + cm3_o
(6775 missing values generated)

. gen cm4_do = cm4_d + cm4_o
(6775 missing values generated)

. forvalues i = 1/4 {
  2.         twoway  (area cm`i'_d cm`i'_t) ///
>                         (rarea cm`i'_do cm`i'_d cm`i'_t) ///
>                         (area cm`i'_do cm`i'_t, base(1)) ///
>                         , ylabel(0(0.2)1.0, angle(h) format(%3.1f)) ///
>                         xtitle("Time since diagnosis") ytitle("crude probability of death") ///
>                         legend(order(1 "P(Dead Cancer)" 2 "P(Dead Other Causes)" 3 "P(Alive)") //
> /
>                                 cols(3)) ///
>                         name(cm_stack`i',replace)
  3. }
. grc1leg cm_stack1 cm_stack2 cm_stack3 cm_stack4, nocopies
```

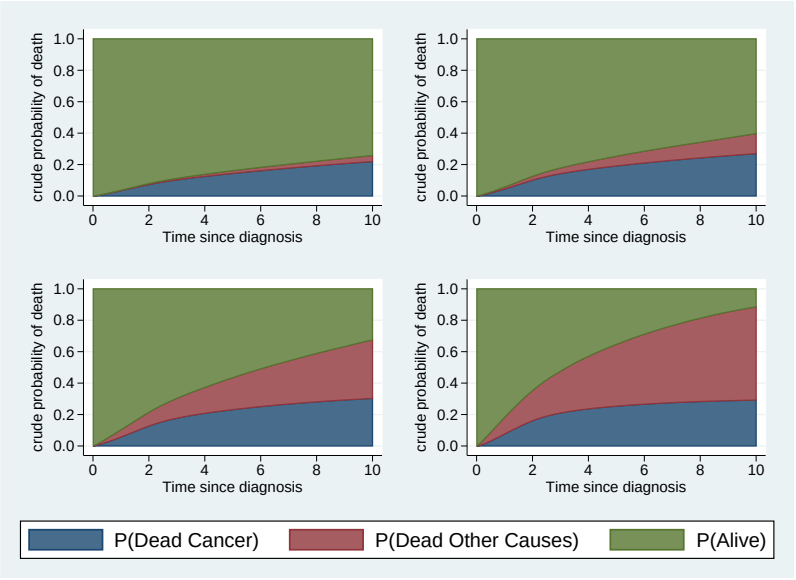


Figure 77: Melanoma Data. Crude probabilities stacked graph

(e) Advanced: Now fit a model using splines for the effect age with the spline terms allowed to be time-dependent. Calculate the crude probabilities of death and compare these to the model where age is categorized.

See the Do file for how to fit the model.

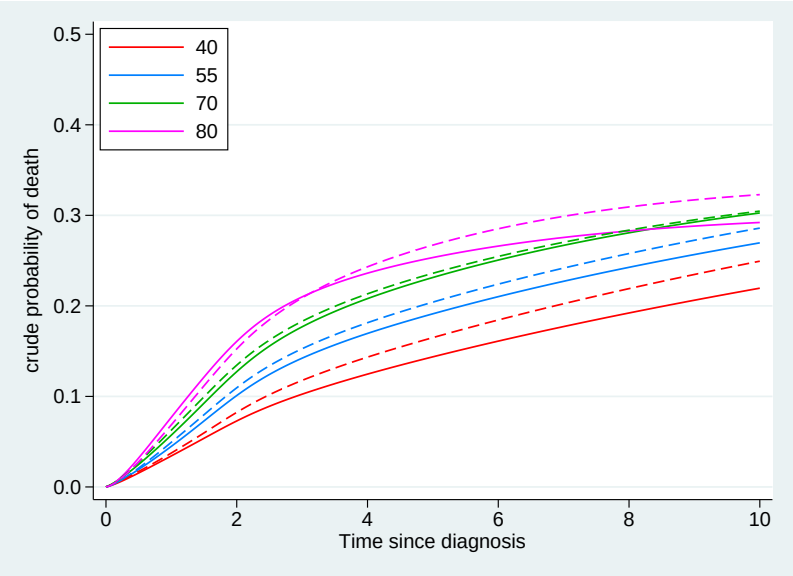


Figure 78: Melanoma Data. Comparison of crude probabilities where age is grouped and treated as a continuous variable.

260. **Estimating cure models**

(a) `_t` contains the time in years from diagnosis. The `strsmix` command requires the expected mortality rate at the event time. The first `gen` command calculates the age at the event (or censoring) time (up to a maximum age of 99). The second `gen` command calculates the calender year at the event time. The third `gen` command converts the expected survival probability into the expected mortality rate.

(b) Fitting this model gives

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

                                                Number of obs   =       6477
                                                Wald chi2(0)    =          .
Log likelihood =  -9988.719                     Prob > chi2     =          .

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pi           |
       _cons |   .4151695   .0081152    51.16   0.000      .399264     .431075
-------------+----------------------------------------------------------------
ln_lambda    |
       _cons |  -.1694096   .0257529    -6.58   0.000    -.2198843   -.1189348
-------------+----------------------------------------------------------------
ln_gamma     |
       _cons |  -.1783506   .0166044   -10.74   0.000    -.2108946   -.1458066
------------------------------------------------------------------------------
```

i. The cure fraction is 0.415 (i.e. 41.5%).

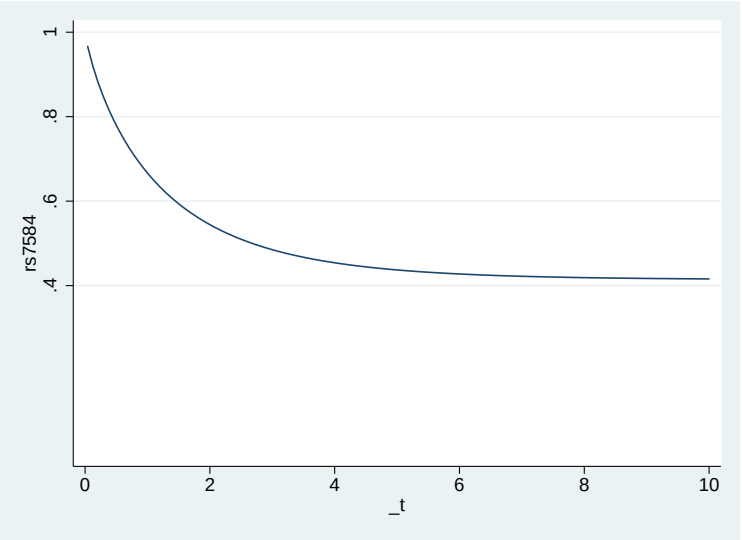


Figure 79: Relative survival in 1975-1984 for cancer of the colon

ii. Yes the relative survival curves reaches a plateau at the cure fraction. Note that if this did not appear to be the case then the cure fraction estimate would be based on extrapolation beyond the range of follow-up in the data.

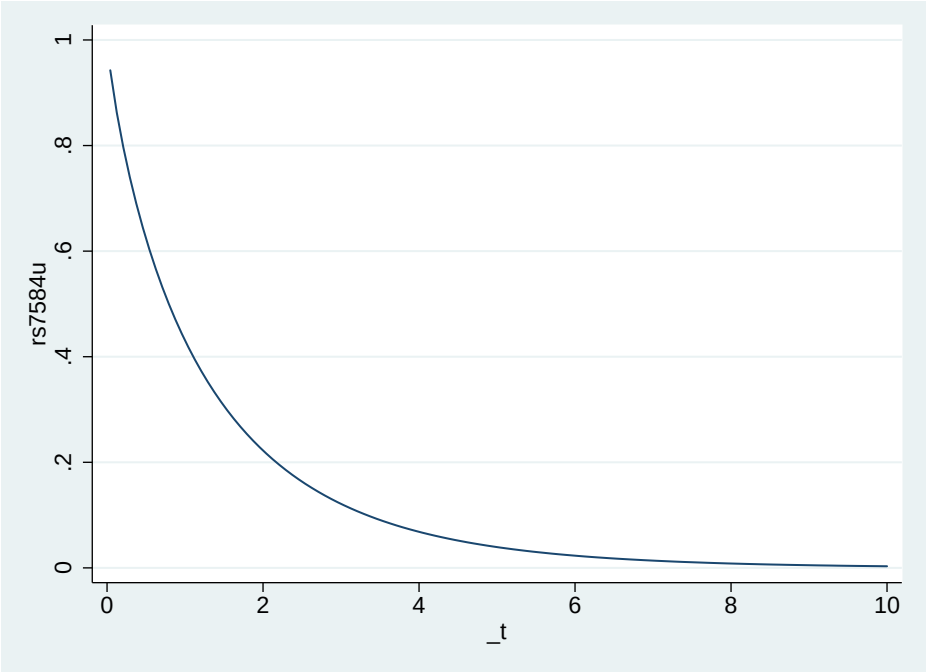


Figure 80: Relative survival for the 'uncured' in 1975-1984 for cancer of the colon

iii. Approximately 80% of the 'uncured' have died after 2 years.

iv. Median survival for the 'uncured' is approximately 0.8 years

(c) Now fitting to those diagnosed 1985-1994.

```
. strsmix if year8594==1, dist(weibull) link(identity) bhazard(rate)

                                                Number of obs   =       9087
                                                Wald chi2(0)    =          .
Log likelihood = -11339.861                     Prob > chi2     =          .

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pi           |
       _cons |     .46044   .0087593    52.57   0.000     .4432721    .4776078
-------------+----------------------------------------------------------------
ln_lambda    |
       _cons |  -.2648208   .0292473    -9.05   0.000    -.3221445   -.2074972
-------------+----------------------------------------------------------------
ln_gamma     |
       _cons |  -.2101828   .0163283   -12.87   0.000    -.2421857   -.1781799
------------------------------------------------------------------------------
```

i. The cure fraction is now 0.459 (i.e 45.9%) - a difference of 4.5%.

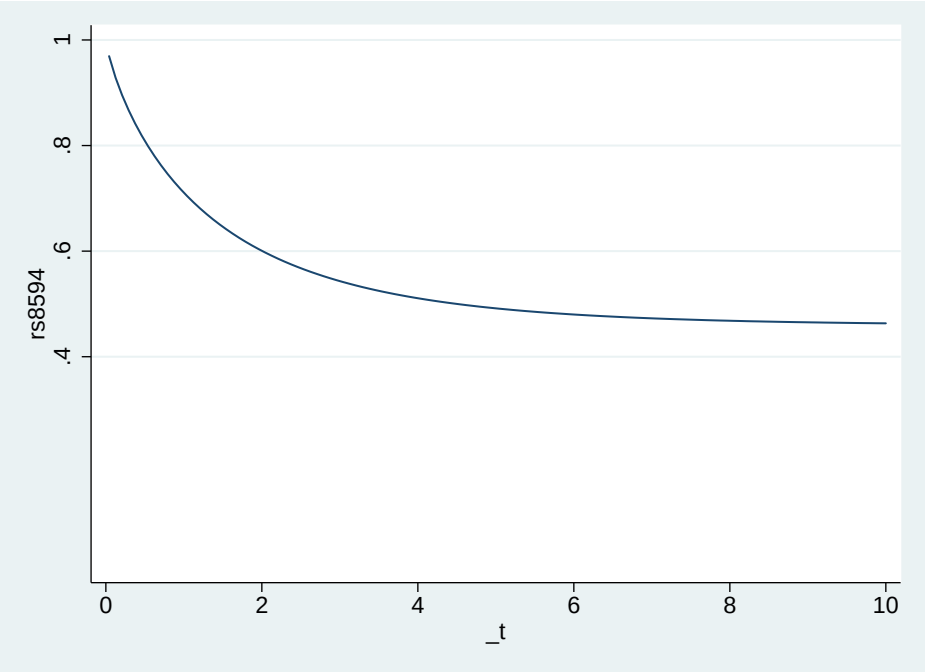


Figure 81: Relative survival in 1985-1984 for cancer of the colon

ii. Yes, the relative survival cure reaches a plateau.

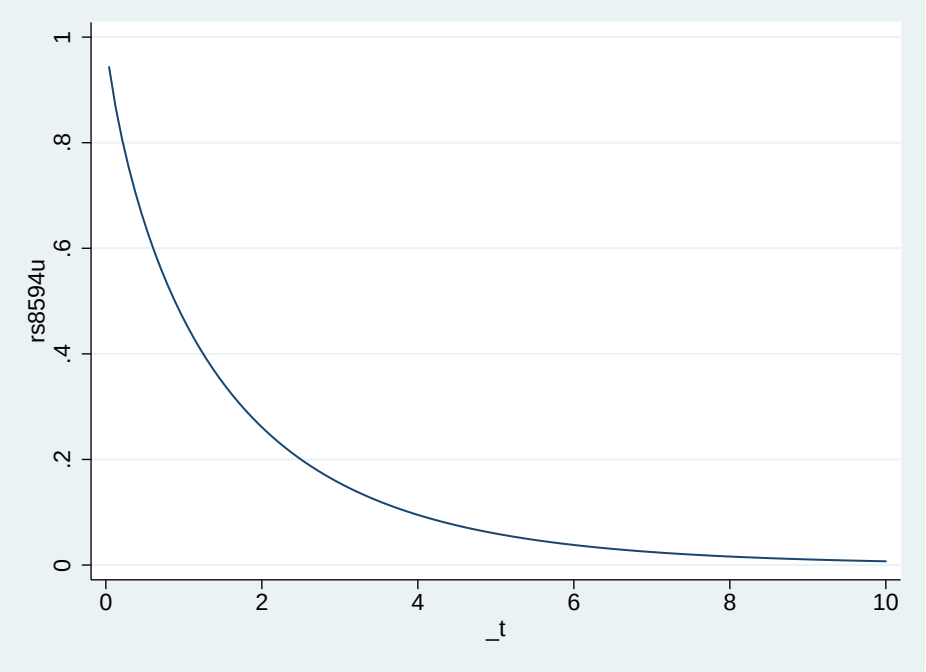


Figure 82: Relative survival for the 'uncured' in 1975-1984 for cancer of the colon

iii. At two years about 75% of the 'uncured' have died after 2 years. A reduction of about 5% in absolute terms.

iv. The median survival of the 'uncured' is about 0.9 years, a slight improvement.

(d) Including `year8594` as a covariate gives

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

```
                                                Number of obs   =      15564
                                                Wald chi2(1)    =      38.51
Log likelihood =  -21332.05                     Prob > chi2     =     0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pi           |
    year8594 |   .0618817   .0099714     6.21   0.000      .042338    .0814254
       _cons |   .4090526   .0078184    52.32   0.000     .3937288    .4243765
-------------+----------------------------------------------------------------
ln_lambda    |
       _cons |  -.2110754   .0191294   -11.03   0.000    -.2485684   -.1735825
-------------+----------------------------------------------------------------
ln_gamma     |
       _cons |  -.1925967   .0115469   -16.68   0.000    -.2152282   -.1699652
------------------------------------------------------------------------------
```

i. The estimated difference in the cure fraction is 0.062 (i.e. 6.2%). This is larger than the difference observed in b(i) and c(i).

ii. The assumption is that the survival distribution of the 'uncured' is the same in the two periods. This is because λ and γ do not vary by our covariate (`year8594`).

Allowing both λ and γ to vary by `year8594` gives

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

```
                                                Number of obs   =      15564
                                                Wald chi2(1)    =      14.37
Log likelihood =  -21328.58                     Prob > chi2     =     0.0001

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pi           |
    year8594 |   .0452705   .0119408     3.79   0.000     .0218671     .068674
       _cons |   .4151695   .0081152    51.16   0.000      .399264     .431075
-------------+----------------------------------------------------------------
ln_lambda    |
    year8594 |  -.0954111   .0389694    -2.45   0.014    -.1717897   -.0190325
       _cons |  -.1694096   .0257529    -6.58   0.000    -.2198843   -.1189348
-------------+----------------------------------------------------------------
ln_gamma     |
    year8594 |  -.0318322   .0232878    -1.37   0.172    -.0774754     .013811
       _cons |  -.1783506   .0166044   -10.74   0.000    -.2108946   -.1458066
------------------------------------------------------------------------------
```

iii. The difference in the cure fraction is 0.045 (i.e. 4.5%). This gives the same as we observed when fitting two separate models, as this is essentially what we are doing by including `year8594` for all 3 parameters. If the distribution of the 'uncured' is not modelled appropriately then biased estimates of the cure fraction may be obtained.

iv. Using a Wald test gives

```
. test [ln_lambda][year8594] [ln_gamma][year8594], mtest
```

```
 ( 1)  [ln_lambda]year8594 = 0
 ( 2)  [ln_gamma]year8594 = 0
------------------------------------
      |        chi2     df       p
------+-----------------------------
  (1) |        6.00      1     0.0143 #
  (2) |        1.83      1     0.1761 #
------+-----------------------------
  all |        6.84      2     0.0328
------------------------------------
                  # unadjusted p-values
```

There is evidence that the survival distribution of the 'uncured' differs between the two time periods.

(e) This model can be fitted using the `xi` prefix command.

```
. tab agegrp, gen(cage)
 strsmix year8594 cage1 cage2 cage3 cage4, dist(weibull) link(logit) ///
        bhazard(rate) k1(year8594 cage1 cage2 cage3 cage4) ///
        k2(year8594 cage1 cage2 cage3 cage4) eform

                                                Number of obs   =      15564
                                                 Wald chi2(4)   =      28.29
Log likelihood = -21088.807                      Prob > chi2    =     0.0000

------------------------------------------------------------------------------
          _t |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pi           |
    year8594 |   1.231615   .0573756     4.47   0.000     1.124142    1.349363
       cage2 |    .903997   .0879128    -1.04   0.299     .7471167    1.093819
       cage3 |   .7988555    .072884    -2.46   0.014     .6680492    .9552742
       cage4 |    .869293    .080983    -1.50   0.133     .7242167    1.043431
       _cons |    .891236   .0760408    -1.35   0.177     .7539937    1.053459
-------------+----------------------------------------------------------------
ln_lambda    |
    year8594 |  -.1118244   .0392174    -2.85   0.004     -.188689   -.0349597
       cage2 |   .0856077    .084418     1.01   0.311    -.0798484    .2510639
       cage3 |   .2501009   .0791222     3.16   0.002     .0950243    .4051775
       cage4 |    1.00063   .0845808    11.83   0.000     .8348543    1.166405
       _cons |  -.5465794   .0750655    -7.28   0.000    -.6937052   -.3994537
-------------+----------------------------------------------------------------
ln_gamma     |
    year8594 |  -.0241314   .0224827    -1.07   0.283    -.0681968     .019934
       cage2 |  -.0614646    .056022    -1.10   0.273    -.1712656    .0483365
       cage3 |  -.1322088   .0518933    -2.55   0.011    -.2339179   -.0304997
       cage4 |  -.1330111   .0527858    -2.52   0.012    -.2364693   -.0295528
       _cons |  -.0000647   .0498729    -0.00   0.999    -.0978138    .0976845
------------------------------------------------------------------------------
```

i. The parameter estimates for the cure fraction are now odds ratios. Thus the odds of cure are 23% higher in 1985-1994 when compared to 1975-1984. For age group 0-44 is the reference category. The odds of cure are 10% lower in the 45-59 age group, 21% lower in the 60-74 age group and 14% lower in the 75+ age group. Only the 60-84 age group is significant at the 5% level. The needs to be a degree of caution here as the Weibull cure models tends to not fit well to the oldest age group and more complex models may be necessary.

ii. The predicted median survival for the 'uncured' is obtained using

```
. predict med, centile
. bysort agegrp year8594: gen flag = (_n==1)
```

```
. list agegrp year8594 med if flag==1, noobs

  +-------------------------------------+
  | agegrp          year8594         med |
  |-------------------------------------|
  |   0-44   Diagnosed 75-84    1.197311 |
  |   0-44   Diagnosed 85-94   1.3485631 |
  |  45-59   Diagnosed 75-84    1.105672 |
  |  45-59   Diagnosed 85-94   1.2519877 |
  |  60-74   Diagnosed 75-84   .92317295 |
  |-------------------------------------|
  |  60-74   Diagnosed 85-94   1.0500786 |
  |    75+   Diagnosed 75-84   .39166079 |
  |    75+   Diagnosed 85-94   .43631407 |
  +-------------------------------------+
```

This table shows how median survival increases with time period in each age group. In addition median survival for the 'uncured' decreases with age.

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

(a)

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

Iteration 0:   log likelihood = -21851.481
Iteration 1:   log likelihood = -21147.216
Iteration 2:   log likelihood = -21095.674
Iteration 3:   log likelihood = -21095.385
Iteration 4:   log likelihood = -21095.385

Log likelihood = -21095.385                     Number of obs   =      15564

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
    year8594 |  -.1556103    .025088    -6.20   0.000    -.2047819   -.1064388
       _rcs1 |   .9889082   .0117887    83.89   0.000     .9658028    1.012014
       _rcs2 |   .0353623    .006665     5.31   0.000      .022299    .0484255
       _rcs3 |   .0684074   .0045871    14.91   0.000     .0594168     .077398
       _rcs4 |   .0530653   .0039162    13.55   0.000     .0453896     .060741
       _rcs5 |   .0410339   .0032154    12.76   0.000     .0347319    .0473359
       _rcs6 |  (omitted)
       _cons |  -.1110995   .0197347    -5.63   0.000    -.1497788   -.0724201
------------------------------------------------------------------------------
```

i. The coefficient -.1556103 is the log-hazard ratio (HR = 0.86) comparing the second period to the first.

ii. The cure proportion for the first period is $\exp(-\exp(-.1110995)) = .40866901$, and for the second period $\exp(-\exp(-.1110995 - .1556103)) = .4649175$.

iii.

```
. predict cure1, cure

. list cure1 if year8594==0, constant

  +---------+
  |   cure1 |
  |---------|
  | .408669 |
  +---------+
  (no variables vary in 6477 observations)

. list cure1 if year8594==1, constant

  +-----------+
  |     cure1 |
  |-----------|
  | .46491749 |
  +-----------+
  (no variables vary in 9087 observations)
```

iv. The estimated difference in the cure fraction is 0.056 (i.e. 5.6%) compared to 0.062 (i.e. 6.2%) in exercise 260.

v. The predicted median survival times are similar in the two groups, but not the same. The flexible parametric cure model is a special case of a non-mixture model. Non-mixture cure models use both the estimated cure proportions and the specified distribution function to estimate the survival function of uncured, which will lead to different survival even when no time-dependent effects are modelled.

```
. predict med1, centile(50) uncured

. list med1 if year8594==0, constant

  +-----------+
  |      med1 |
  |-----------|
  | .75329265 |
  +-----------+
  (no variables vary in 6477 observations)

. list med1 if year8594==1, constant

  +-----------+
  |      med1 |
  |-----------|
  | .80035703 |
  +-----------+
  (no variables vary in 9087 observations)
```

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

Iteration 0:   log likelihood = -21848.799
Iteration 1:   log likelihood = -21144.251
Iteration 2:   log likelihood = -21092.538
Iteration 3:   log likelihood = -21092.239
Iteration 4:   log likelihood = -21092.239

Log likelihood = -21092.239                     Number of obs   =      15564

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
    year8594 |  -.1492647   .0269617    -5.54   0.000    -.2021086   -.0964208
       _rcs1 |   1.006746   .0177333    56.77   0.000     .9719896    1.041503
       _rcs2 |   .0447082   .0094731     4.72   0.000     .0261413    .0632751
       _rcs3 |   .0692846   .0065112    10.64   0.000     .0565229    .0820462
       _rcs4 |   .0493157   .0057847     8.53   0.000     .0379779    .0606535
       _rcs5 |   .0384908   .0038595     9.97   0.000     .0309262    .0460553
       _rcs6 |  (omitted)
_rcs_y~85941 |  -.0329169   .0238804    -1.38   0.168    -.0797216    .0138878
_rcs_y~85942 |  -.0137549   .0135084    -1.02   0.309    -.0402309    .0127211
_rcs_y~85943 |   .0100166   .0086015     1.16   0.244    -.0068419    .0268752
_rcs_y~85944 |  (omitted)
       _cons |  -.1131936   .0202657    -5.59   0.000    -.1529136   -.0734736
------------------------------------------------------------------------------
```

i. The coefficient is no longer interpreted as the log-hazard ratio since the hazard ratio is varying over time.

ii. The cure proportion for the first period is $\exp(-\exp(-.1131936)) = 0.40943474$, and for the second period $\exp(-\exp(-.1131936 - .1492647)) = 0.46340289$.

iii.
```
. predict cure2, cure
```

```
. list cure2 if year8594==0, constant
  +-----------+
  |     cure2 |
  |-----------|
  | .40943473 |
  +-----------+
  (no variables vary in 6477 observations)

. list cure2 if year8594==1, constant
  +-----------+
  |     cure2 |
  |-----------|
  | .46340288 |
  +-----------+
  (no variables vary in 9087 observations)
```

iv. The estimated difference in the cure fraction is 0.054 (i.e. 5.4%), very similar to the result in a.

v. The difference in the predicted median survival times between the two groups is larger than in a, since we are now allowing more flexibility into the estimation.

```
. predict med2, centile(50) uncured

. list med2 if year8594==0, constant
  +----------+
  |     med2 |
  |----------|
  | .7406603 |
  +----------+
  (no variables vary in 6477 observations)

. list med2 if year8594==1, constant
  +-----------+
  |      med2 |
  |-----------|
  | .81717336 |
  +-----------+
  (no variables vary in 9087 observations)
```

(c) The flexible parametric cure model forces the cumulative excess hazard to be constant after the last knot, and therefore the relative survival is forced to reach a plateau. The assumption of cure should always be checked in a model that does not assume cure or by looking at empirical life table estimates.

```
. predict surv, survival
. predict survunc, survival uncured
. forvalues j=0/1 {
        twoway (line surv _t if year8594==`j', sort) ///
          (line survunc _t if year8594==`j', sort), ///
          legend(label(1 "Survival overall") ///
          label(2 "Survival for uncured")) name(period`j', replace)
  }
```

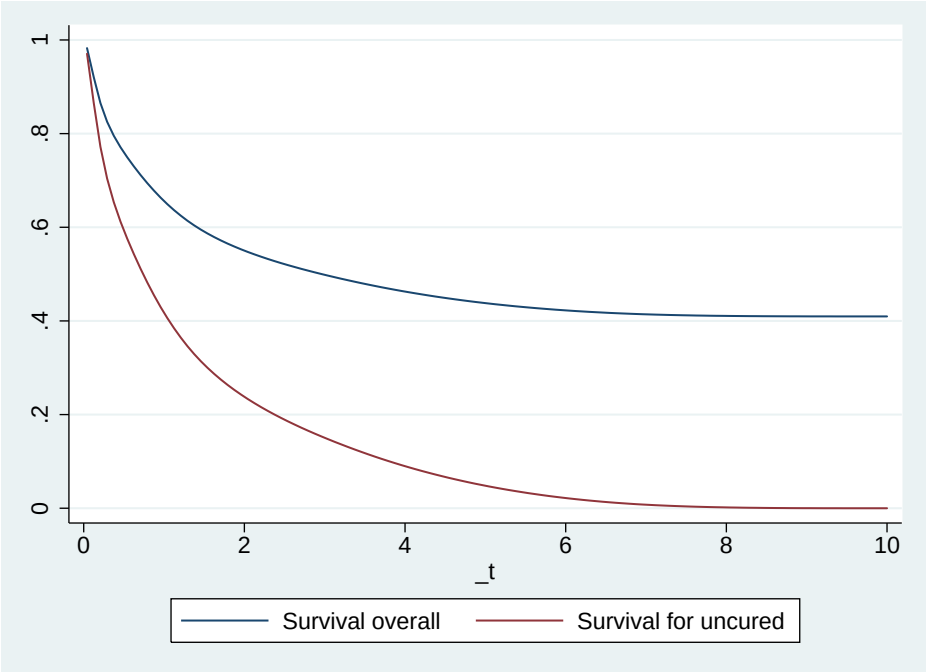


Figure 83: Relative survival overall and for the 'uncured' in 1975-1984 for cancer of the colon

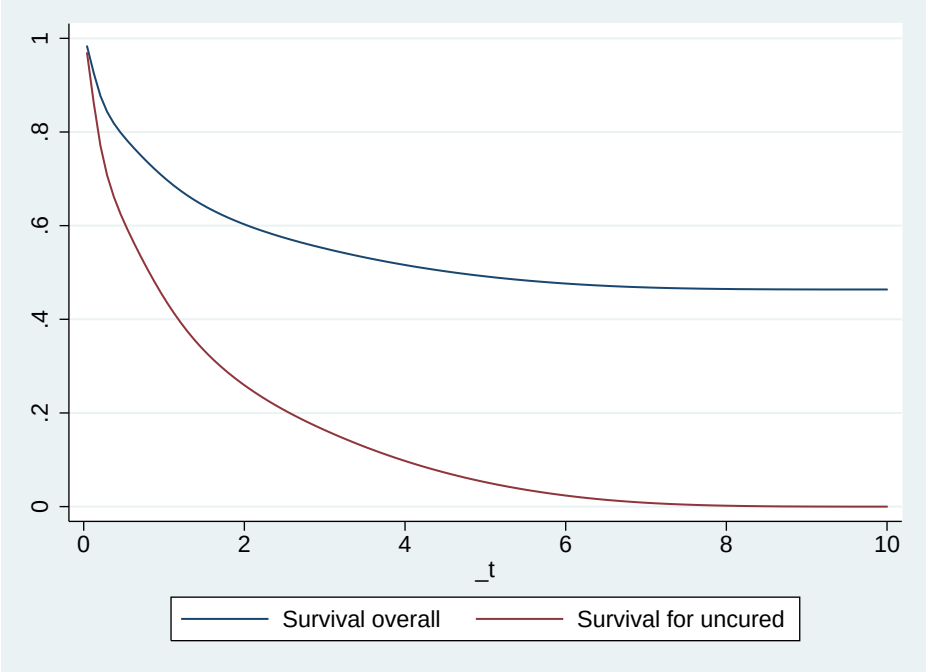


Figure 84: Relative survival overall and for the 'uncured' in 1985-1994 for cancer of the colon

280. **Creating a popmort file from the Human Mortality Database**
There are no written solutions for this exercise.

281. **Constructing a popmort file by modelling cohort data**
There are no written solutions for this exercise.

282. **Calculating excess and ï¿½avoidableï¿½ deaths from life tables.**

(a) Load the Melanoma data, drop subjects diagnosed 1975-1984.

(b) What is the difference in five-year relative survival between males and females in each age group?

```
. list agegrp sex cr_e2 if end == 5, noobs sepby(agegrp)
  +--------------------------+
  | agegrp      sex    cr_e2 |
  |--------------------------|
  |   0-44     Male   0.8236 |
  |   0-44   Female   0.9233 |
  |--------------------------|
  |  45-59     Male   0.7969 |
  |  45-59   Female   0.8740 |
  |--------------------------|
  |  60-74     Male   0.7413 |
  |  60-74   Female   0.7958 |
  |--------------------------|
  |    75+     Male   0.6627 |
  |    75+   Female   0.7006 |
  +--------------------------+
```

Five year relative survival is lower for males in all age groups.

(c) Reshape the data.

```
. bysort sex (agegrp start): gen j = _n
. gen sexlab =cond(sex==1,"_m","_f")
. drop sex
. reshape wide start end n cp cp_e2 cr_e2 agegrp, i(j) j(sexlab) string
(note: j = _f _m)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       40   ->      20
Number of variables                   9   ->      15
j variable (2 values)           sexlab   ->   (dropped)
xij variables:
                                  start   ->   start_f start_m
                                    end   ->   end_f end_m
                                      n   ->   n_f n_m
                                     cp   ->   cp_f cp_m
                                  cp_e2   ->   cp_e2_f cp_e2_m
                                  cr_e2   ->   cr_e2_f cr_e2_m
                                 agegrp   ->   agegrp_f agegrp_m
-----------------------------------------------------------------------------

. rename agegrp_m agegrp
. rename start_m start
. rename end_m end
. drop agegrp_f start_f end_f
```

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

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

. gen p_dead_m = 1 - cp_e2_m * cr_e2_m
. gen Nd_m = Nrisk_m*p_dead_m
. gen NExp_d_m = Nrisk_m*(1-cp_e2_m)
```

```
. gen ED_m = Nd_m - NExp_d_m

. format Nd_m NExp_d_m ED_m %4.1f
. list agegrp Nrisk_m p_dead_m Nd_m NExp_d_m ED_m if end==5, noobs
  +-------------------------------------------------------+
  | agegrp   Nrisk_m   p_dead_m   Nd_m   NExp_d_m   ED_m |
  |-------------------------------------------------------|
  |   0-44      53.7   .1889797   10.1        0.8    9.3 |
  |  45-59      75.2   .2440302   18.4        3.9   14.5 |
  |  60-74      70.9   .3905036   27.7       12.6   15.1 |
  |    75+      33.7   .6542017   22.0       16.1    5.9 |
  +-------------------------------------------------------+

. table agegrp if end == 5, c(sum Nd_m sum NExp_d_m sum ED_m) row format(%4.1f)
----------------------------------------------------------
_m agegrp |     sum(Nd_m)  sum(NExp_d_m)      sum(ED_m)
----------+-----------------------------------------------
     0-44 |          10.1            0.8            9.3
    45-59 |          18.4            3.9           14.5
    60-74 |          27.7           12.6           15.1
      75+ |          22.0           16.1            5.9
          |
    Total |          78.2           33.4           44.8
----------------------------------------------------------
```

i. We would expect to see 10, 18, 28 and 22 all cause deaths in the (ascending) age groups.

ii. This is given by the excess deaths, `ED_m`. In ascending age groups there are 9, 14, 15, and 6 excess deaths at 5 years post diagnosis when compared to a similar cancer free population. This is for a typical cohort diagnosed in one calendar year.

iii. There are 45 excess deaths when compared to the general population.

(e) Repeat calculations for females.

```
. bys agegrp: gen Nrisk_f = n_f[1]/10

. gen p_dead_f = 1 - cp_e2_f * cr_e2_f
. gen Nd_f = Nrisk_f*p_dead_f
. gen NExp_d_f = Nrisk_f*(1-cp_e2_f)
. gen ED_f = Nd_f - NExp_d_f

. format Nd_f NExp_d_f ED_f %4.1f
. list agegrp Nrisk_f p_dead_f Nd_f NExp_d_f ED_f if end==5, noobs
  +-------------------------------------------------------+
  | agegrp   Nrisk_f   p_dead_f   Nd_f   NExp_d_f   ED_f |
  |-------------------------------------------------------|
  |   0-44      62.4   .0814915    5.1        0.3    4.8 |
  |  45-59      61.2   .1431934    8.8        1.2    7.6 |
  |  60-74      66.1   .2800009   18.5        6.3   12.2 |
  |    75+      51.2   .5766043   29.5       20.3    9.3 |
  +-------------------------------------------------------+

. table agegrp if end == 5, c(sum Nd_f sum NExp_d_f sum ED_f) row format(%4.1f)
----------------------------------------------------------
_m agegrp |     sum(Nd_f)  sum(NExp_d_f)      sum(ED_f)
----------+-----------------------------------------------
     0-44 |           5.1            0.3            4.8
    45-59 |           8.8            1.2            7.6
    60-74 |          18.5            6.3           12.2
      75+ |          29.5           20.3            9.3
          |
    Total |          61.9           28.1           33.8
----------------------------------------------------------
```

In terms of the total number of all cause deaths, females have fewer at all ages except the 70+ group. This is because they are more females diagnosed in this group 51 vs 34, so even though females have lower relative survival they have more deaths due to a number of women in the oldest age groups being diagnosed. This leads to there being more excess deaths in this age group for women when compared to men. As a whole there are more excess deaths in men.

(f) How many deaths would be 'avoided' if males could achieve the same relative survival as females for Melanoma?

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

. format Nd_m_f AD_m %4.1f
. list agegrp Nrisk_m p_dead_m Nd_m NExp_d_m ED_m Nd_m_f AD_m if end==5, noobs
  +-------------------------------------------------------------------------+
  | agegrp   Nrisk_m   p_dead_m   Nd_m   NExp_d_m   ED_m   Nd_m_f   AD_m |
  |-------------------------------------------------------------------------|
  |   0-44      53.7   .1889797   10.1        0.8    9.3      4.9    5.3 |
  |  45-59      75.2   .2440302   18.4        3.9   14.5     12.9    5.5 |
  |  60-74      70.9   .3905036   27.7       12.6   15.1     24.5    3.2 |
  |    75+      33.7   .6542017   22.0       16.1    5.9     21.4    0.7 |
  +-------------------------------------------------------------------------+
```

There would be about 15 deaths 'avoided'. The youngest two age groups contribute most to the avoidable deaths.

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

```
. list agegrp end AD_m if agegrp==3
     +---------------------+
     | agegrp   end   AD_m |
     |---------------------|
 16. |    75+     1    1.4 |
 17. |    75+     2    2.2 |
 18. |    75+     3    2.1 |
 19. |    75+     4    1.2 |
 20. |    75+     5    0.7 |
     +---------------------+
```

This is because we can not avoid deaths for ever. Remember that we are looking at all cause deaths. If we had unlimited follow-up we would avoid no deaths at all. In the oldest age group we can actually see that we are just postponing deaths.

283. **Simulating relative survival**
There are no written solutions for this exercise.

284. **Estimating loss in expectation of life**

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

```
. use melanoma, clear
(Skin melanoma, diagnosed 1975-94, follow-up to 1995)
. gen patid = _n
. stset surv_mm, failure(status=1 2) scale(12) exit(time 120.5) id(patid)

                id:  patid
     failure event:  status == 1 2
obs. time interval:  (surv_mm[_n-1], surv_mm]
 exit on or before:  time 120.5
    t for analysis:  time/12
------------------------------------------------------------------------------
     7775  total observations
        0  exclusions
------------------------------------------------------------------------------
     7775  observations remaining, representing
     7775  subjects
     2777  failures in single-failure-per-subject data
 43384.63  total analysis time at risk and under observation
                                             at risk from t =         0
                                  earliest observed entry t =         0
                                       last observed exit t =  10.04167
```

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

```
. rcsgen age, df(4) gen(sag) orthog
Variables sag1 to sag4 were created

. rcsgen yydx, df(4) gen(syr) orthog
Variables syr1 to syr4 were created

. gen fem = sex==2
. gen _age = min(int(age + _t),99)
. gen _year = int(yydx + _t)
. sort _year sex _age
. merge m:1 _year sex _age using popmort, keep(match master) keepusing(rate)

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                             7,775  (_merge==3)
    -----------------------------------------
. drop _age _year _merge

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

```
Log likelihood = -8444.5801                     Number of obs   =       7775
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
        sag1 |   .3486966   .0355765     9.80   0.000     .2789678    .4184253
        sag2 |  -.0382469   .0368393    -1.04   0.299    -.1104506    .0339568
        sag3 |  -.0826459   .0352677    -2.34   0.019    -.1517692   -.0135225
        sag4 |  -.0171397   .0333635    -0.51   0.607     -.082531    .0482516
        syr1 |  -.0064674   .1187121    -0.05   0.957    -.2391387     .226204
        syr2 |  -.2522286   .1030806    -2.45   0.014    -.4542629   -.0501944
        syr3 |  -.1413523   .0858927    -1.65   0.100     -.309699    .0269943
        syr4 |  -.1155111   .0700542    -1.65   0.099    -.2528149    .0217927
         fem |  -.5220707   .0604833    -8.63   0.000    -.6406158   -.4035256
       _rcs1 |   .9474817   .0781558    12.12   0.000     .7942992    1.100664
       _rcs2 |   .1927113    .054332     3.55   0.000     .0862225    .2992001
       _rcs3 |   .0568751   .0304669     1.87   0.062    -.0028389    .1165892
       _rcs4 |   .0032183    .014089     0.23   0.819    -.0243957    .0308323
       _rcs5 |   .0063443   .0052562     1.21   0.227    -.0039577    .0166462
  _rcs_sag11 |   .0101007   .0305454     0.33   0.741    -.0497673    .0699687
  _rcs_sag12 |   .0327253    .026622     1.23   0.219    -.0194529    .0849034
  _rcs_sag13 |   .0204141   .0135927     1.50   0.133     -.006227    .0470553
  _rcs_sag21 |  -.0382793   .0312975    -1.22   0.221    -.0996212    .0230626
  _rcs_sag22 |  -.0024951   .0278919    -0.09   0.929    -.0571622    .0521719
  _rcs_sag23 |   .0015633   .0139492     0.11   0.911    -.0257767    .0289032
  _rcs_sag31 |  -.0148982   .0288652    -0.52   0.606     -.071473    .0416766
  _rcs_sag32 |   .0178845    .025579     0.70   0.484    -.0322494    .0680183
  _rcs_sag33 |   .0007745   .0129807     0.06   0.952    -.0246672    .0262163
  _rcs_sag41 |  -.0217533   .0278767    -0.78   0.435    -.0763907    .0328841
  _rcs_sag42 |   .0036575   .0247048     0.15   0.882    -.0447631    .0520781
  _rcs_sag43 |  -.0002257   .0126263    -0.02   0.986    -.0249727    .0245214
  _rcs_syr11 |   .1082871   .0951937     1.14   0.255    -.0782891    .2948633
  _rcs_syr12 |  -.0912392   .0569474    -1.60   0.109    -.2028541    .0203757
  _rcs_syr13 |  -.0598222   .0368902    -1.62   0.105    -.1321258    .0124813
  _rcs_syr21 |  -.1088465   .0811995    -1.34   0.180    -.2679946    .0503015
  _rcs_syr22 |   .0769735   .0481734     1.60   0.110    -.0174446    .1713916
  _rcs_syr23 |   .0206394    .030727     0.67   0.502    -.0395845    .0808632
  _rcs_syr31 |  -.1046798   .0660342    -1.59   0.113    -.2341045    .0247448
  _rcs_syr32 |   .0236841   .0431332     0.55   0.583    -.0608553    .1082236
  _rcs_syr33 |   .0266358   .0243036     1.10   0.273    -.0209984      .07427
  _rcs_syr41 |  -.0203372   .0520008    -0.39   0.696    -.1222569    .0815826
  _rcs_syr42 |   .0493604   .0349461     1.41   0.158    -.0191328    .1178536
  _rcs_syr43 |   .0196377   .0188815     1.04   0.298    -.0173694    .0566448
   _rcs_fem1 |  -.0019995   .0503392    -0.04   0.968    -.1006625    .0966635
   _rcs_fem2 |  -.0844331   .0450417    -1.87   0.061    -.1727131     .003847
   _rcs_fem3 |  -.0203553   .0212678    -0.96   0.339    -.0620393    .0213288
       _cons |  -1.378518   .0959111   -14.37   0.000      -1.5665   -1.190535
------------------------------------------------------------------------------
```

(c) We will now estimate the loss in expectation of life. To save time we don't estimate confidence intervals, although they can be obtained by removing the comments around the `ci` option.

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

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

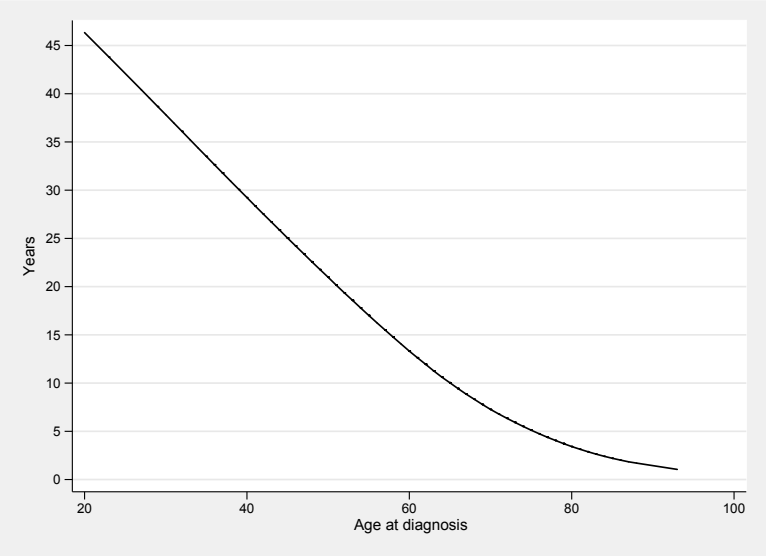


Figure 85: Melanoma Data. Loss in expectation of life

Figure 85 shows the loss in expectation of life for males diagnosed with melanoma in 1994.

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

```
. foreach age in 50 60 70 80 {
  2.         foreach sex in 1 2 {
  3.                 list age sex yydx survexp survobs ll if age==‘age’ & sex==‘sex’ & yydx==1994, consta
  4.         }
  5. }

  +-----------------------------------------------------+
  | age    sex   yydx    survexp     survobs         ll |
  |-----------------------------------------------------|
  |  50   Male   1994   26.63637   5.6614445   20.97493 |
  +-----------------------------------------------------+
  (no variables vary in 5 observations)

  +-------------------------------------------------------+
  | age      sex   yydx    survexp     survobs         ll |
  |-------------------------------------------------------|
  |  50   Female   1994   32.36633   7.2172614   25.14907 |
  +-------------------------------------------------------+
  (no variables vary in 3 observations)

  +-----------------------------------------------------+
  | age    sex   yydx    survexp     survobs         ll |
  |-----------------------------------------------------|
  |  60   Male   1994   18.49159   5.1773682   13.31423 |
  +-----------------------------------------------------+
  (no variables vary in 8 observations)
```

```
  +-----------------------------------------------------+
  | age      sex   yydx    survexp     survobs         ll |
  |-----------------------------------------------------|
  |  60   Female   1994   23.30669   6.8167728   16.48991 |
  +-----------------------------------------------------+
  (no variables vary in 8 observations)

  +-----------------------------------------------+
  | age    sex   yydx    survexp     survobs    ll |
  |-----------------------------------------------|
  |  70   Male   1994   11.53323   4.2612695   7.27196 |
  +-----------------------------------------------+
  (no variables vary in 4 observations)

  +-----------------------------------------------------+
  | age      sex   yydx   survexp     survobs         ll |
  |-----------------------------------------------------|
  |  70   Female   1994   14.8622   5.8554623   9.006738 |
  +-----------------------------------------------------+
  (no variables vary in 9 observations)

  +-----------------------------------------------+
  | age    sex   yydx    survexp     survobs    ll |
  |-----------------------------------------------|
  |  80   Male   1994   6.431057   3.0075134   3.423544 |
  +-----------------------------------------------+
  (no variables vary in 3 observations)

  +-----------------------------------------------------+
  | age      sex   yydx    survexp     survobs         ll |
  |-----------------------------------------------------|
  |  80   Female   1994   8.000338   4.1340081   3.866329 |
  +-----------------------------------------------------+
  (no variables vary in 3 observations)

. qui summ ll if yydx==1994
. display r(sum)
8767.1307
```

The total number of life years lost among patients diagnosed with melanoma in Finland in 1994 is 8767.

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

```
. replace fem=1
(3680 real changes made)

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

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

```
. gen lldiff= ll-ll_alt
. summ lldiff if yydx==1994

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      lldiff |       518    .6344759    .6386128          0   1.554199
```

```
. display r(sum)
328.6585

. foreach age in 50 60 70 80 {
  2.   list ll ll_alt lldiff age if sex==1 & age==‘age’ & yydx==1994, constant
  3. }

  +-----------------------------------+
  |       ll     ll_alt    lldiff   age |
  |-----------------------------------|
  | 20.97493   19.56192   1.41301    50 |
  +-----------------------------------+
  (no variables vary in 5 observations)

  +-----------------------------------+
  |       ll     ll_alt   lldiff   age |
  |-----------------------------------|
  | 13.31423   11.99303   1.3212    60 |
  +-----------------------------------+
  (no variables vary in 8 observations)

  +-----------------------------------+
  |      ll     ll_alt     lldiff   age |
  |-----------------------------------|
  | 7.27196   6.200533   1.071427    70 |
  +-----------------------------------+
  (no variables vary in 4 observations)

  +-----------------------------------+
  |       ll     ll_alt     lldiff   age |
  |-----------------------------------|
  | 3.423544   2.734462   .6890819    80 |
  +-----------------------------------+
```

If males diagnosed in 1994 had the same relative survival as females diagnosed in 1994, the total number of life years lost would reduce by 328 years. For a man aged 50 at diagnosis the potential gain in life expectancy is 1.4 years (1.3, 1.1 and 0.7 years for males aged 60, 70 and 80 years at diagnosis, respectively).

285. Multiple imputation for missing covariate data

(a) 15.14% of patients have missing stage

stage at diagnosis	Freq.	Percent	Cum.
Unknown	2,356	15.14	15.14
Localised	6,274	40.31	55.45
Regional	1,787	11.48	66.93
Distant	5,147	33.07	100.00
Total	15,564	100.00	

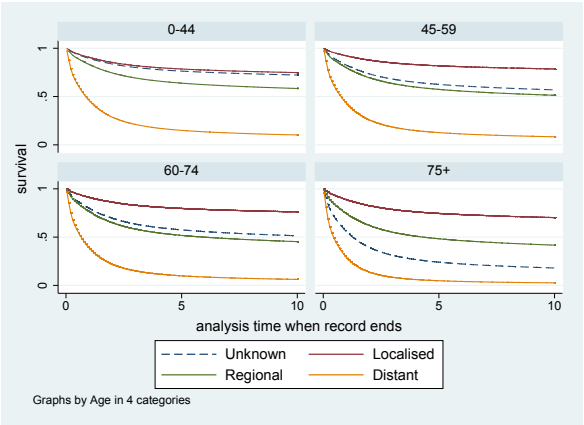
(b) Investigate the distribution of unknown stage across age group and gender. Are older patients more likely to have an unknown recorded stage?

```
. tab stage agegrp, column
```

stage at diagnosis	Age in 4 categories 0-44	45-59	60-74	75+	Total
Unknown	83	262	858	1,153	2,356
	11.29	11.06	13.01	19.65	15.14
Localised	297	993	2,716	2,268	6,274
	40.41	41.93	41.20	38.65	40.31
Regional	114	329	772	572	1,787
	15.51	13.89	11.71	9.75	11.48
Distant	241	784	2,247	1,875	5,147
	32.79	33.11	34.08	31.95	33.07
Total	735	2,368	6,593	5,868	15,564
	100.00	100.00	100.00	100.00	100.00

The oldest age-group has the largest proportion of unknown stage.

(c)



The survival of the young patients with unknown stage is relatively good (similar to those with localised) but for the oldest age group the survival for patients with unknown stage is relatively worse (closer to the survival for patients with distant metastases). This suggests that the mechanism leading to unknown stage may differ according to age.

(d)

(e) It is possible that stage is more likely to be missing for elderly patients with poor general health. It may be more likely to be missing for individuals under care in a nursing home. We do not have access to such information so a MAR assumption is unlikely to be true.

Note that the above is by no means the definitive answer. The key concept is that you consider the mechanisms that might give rise to missing data and whether or not we have data on the factors that might predict missingness.

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

Log likelihood = -18267.394                     Number of obs     =     15,564

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
       stage |
    Unknown  |   3.241262   .1571215    24.26   0.000     2.947487    3.564319
   Regional  |   2.660883   .1403817    18.55   0.000     2.399487    2.950755
    Distant  |   10.00967   .3928204    58.70   0.000     9.268619    10.80997
             |
      agegrp |
      45-59  |   1.101743   .0692448     1.54   0.123     .9740518    1.246174
      60-74  |   1.241194   .0720058     3.72   0.000     1.107793    1.390659
        75+  |   1.780897   .1042263     9.86   0.000     1.587898    1.997354
             |
       _rcs1 |   3.044807   .0362527    93.52   0.000     2.974576    3.116697
       _rcs2 |   1.320444   .0119872    30.62   0.000     1.297157    1.344149
       _rcs3 |     .99282   .0056625    -1.26   0.206     .9817836     1.00398
       _rcs4 |   1.048117   .0038603    12.76   0.000     1.040579    1.055711
       _rcs5 |   1.011472   .0029515     3.91   0.000     1.005704    1.017273
       _cons |   .0765708   .0049307   -39.90   0.000     .0674918    .0868711
------------------------------------------------------------------------------
```

(g)
```
. replace stage=. if stage==0
(2,356 real changes made, 2,356 to missing)

. stpm2 ib1.stage i.agegrp , df(5) bhaz(rate) scale(hazard) eform nolog

Log likelihood = -15353.605                     Number of obs     =     13,208

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
       stage |
   Regional  |   2.676154   .1410076    18.68   0.000     2.413576    2.967299
    Distant  |    10.3598   .4080125    59.36   0.000     9.590197    11.19117
             |
      agegrp |
      45-59  |   1.061092   .0688816     0.91   0.361     .9343219    1.205062
      60-74  |   1.204694   .0721444     3.11   0.002     1.071277    1.354727
        75+  |   1.557469   .0950547     7.26   0.000     1.381876    1.755373
             |
       _rcs1 |   3.141848   .0410415    87.64   0.000     3.062429    3.223326
       _rcs2 |   1.318276   .0133637    27.26   0.000     1.292342     1.34473
       _rcs3 |   1.001148    .006439     0.18   0.858     .9886073    1.013848
       _rcs4 |   1.050811   .0043694    11.92   0.000     1.042282     1.05941
       _rcs5 |   1.010931     .00336     3.27   0.001     1.004367    1.017538
       _cons |   .0807468   .0053034   -38.31   0.000     .0709936    .0918398
------------------------------------------------------------------------------
```

(h) We did the analysis with 100 imputations with the following imputation model.

```
. mi impute chained (mlogit) stage = i.subsite sex i.agegrp H _d, add(100)
```

The distribution of imputed values was as follows.

id	agegrp	_t	_d	localised	regional	distant
2287	45-59	0.04	1	7	11	82
3362	75+	6.21	1	76	22	2
3501	75+	10.0	0	87	12	1

Obtaining answers close to those above is not especially important. The aim of this exercise is for you to get insight into the process of multiple imputation by performing the same task we will be asking the computer to perform for us. One of the key points is that we are imputing a distribution for the missing values, not just a single best value. The second key point was to think about how the known covariates, and the value of the outcome, are associated with the distribution of the missing values.

Those of you without knowledge of cancer and cancer registration may have struggled. This was intentional. Subject matter knowledge is crucial when imputing missing values. We need knowledge of the process by which stage is assessed, classified, and registered along with knowledge of why it might be missing.

Age and survival time are considerably more important than sex and subsite in imputing missing stage. many of you will have realised that information on age in years would have been useful. Absolutely! We used age in groups hoping you would realise that it is suboptimal. Similarly, cause of death information would also have been useful. We'll explore these issues more later.

How might we assess this more formally. Let's consider patient 3362. Among patients of that age and sex, we wish to estimate the probability that stage takes a given value conditional on survival time being 6.2 years. We can apply Bayes' theorem.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{1}$$

where in our example A is stage=s and B is survival time T equal to 6.2 years. Note that we don't want the survivor function, $P(T > 6.2)$, but the probability density function. The probability that the survival time is exactly 6.2 years is zero, so we'll evaluate the probability that the survival time is within 6.2 and 6.3 years. Recall that

$$f(t) = S(t)h(t) \tag{2}$$

We will use $S(6.2) \times (H(6.3) - H(6.2))$ as an approximation to f(6.2).

We can use the following command to obtain $S(6.2)$, $H(6.3)$ and $H(6.2)$. We also need $P(B)$, the density function for all patients.

```
. sts list if agegrp==3 & sex==2, by(stage) at(6.2 6.3) cumhaz
. sts list if agegrp==3 & sex==2, at(6.2 6.3) cumhaz
. sts list if agegrp==3 & sex==2, by(stage) at(6.2 6.3)
. sts list if agegrp==3 & sex==2, at(6.2 6.3)
```

Results are summarised below.

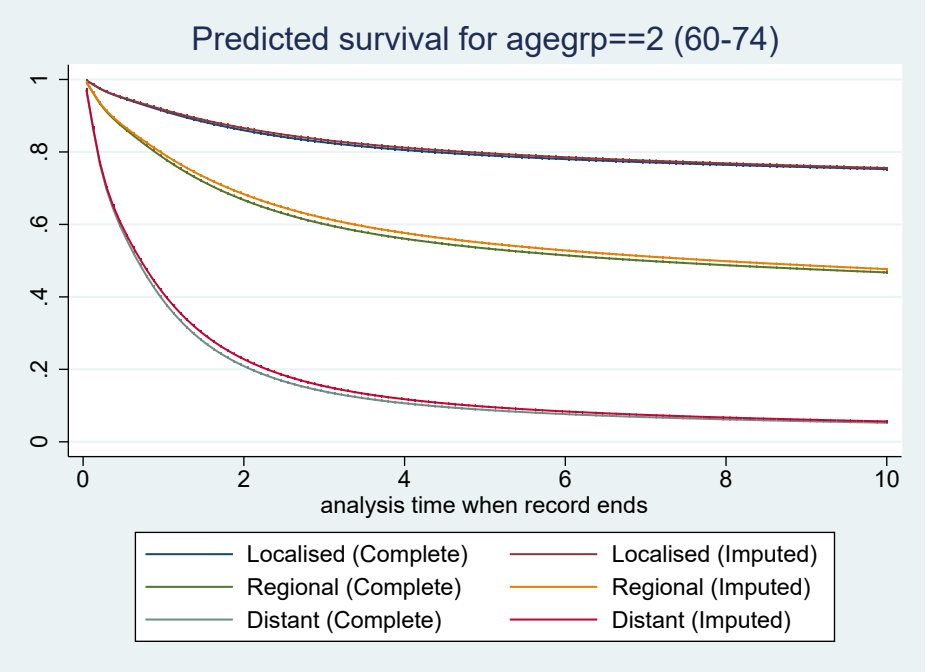
A	$Pr(A)$	$S(6.2)$	$H(6.2)$	$H(6.3)$	diff	$P(B\|A)$	$P(B)$	$P(A\|B)$
localised	0.40	0.3344	1.0828	1.0955	0.0127	0.0042	0.0022	0.7586
regional	0.12	0.1631	1.7764	1.7995	0.0231	0.0038	0.0022	0.2019
distant	0.48	0.0143	4.0415	4.0532	0.0117	0.0002	0.0022	0.0359

That became more complicated than I had anticipated, but we see that our stage distribution (76/20/4) is very close to the distribution of imputed values obtained by Stata (76/22/2).

(i) See the solution to the previous part, where we used 100 imputations.

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

Imputations (10):
  .........10 done

Multiple-imputation estimates                   Imputations       =         10
                                                Number of obs     =     15,564
                                                Average RVI       =     0.0612
                                                Largest FMI       =     0.1812
DF adjustment:   Large sample                   DF:     min       =     291.86
                                                        avg       = 278,736.11
Within VCE type:          OIM                           max       = 2036800.14

 ( 1)  [xb]_rcs1 - [dxb]_d_rcs1 = 0
 ( 2)  [xb]_rcs2 - [dxb]_d_rcs2 = 0
 ( 3)  [xb]_rcs3 - [dxb]_d_rcs3 = 0
 ( 4)  [xb]_rcs4 - [dxb]_d_rcs4 = 0
 ( 5)  [xb]_rcs5 - [dxb]_d_rcs5 = 0
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
       stage |
   Regional  |    .971939   .0539502    18.02   0.000     .8657582     1.07812
    Distant  |   2.328714   .0385613    60.39   0.000     2.253008     2.40442
             |
      agegrp |
      45-59  |   .0791556   .0635025     1.25   0.213    -.0453164    .2036275
      60-74  |   .2109143    .059261     3.56   0.000     .0947338    .3270947
        75+  |   .5465154   .0591999     9.23   0.000      .430476    .6625547
             |
       _rcs1 |   1.144751   .0121193    94.46   0.000     1.120996    1.168506
       _rcs2 |   .2693084   .0091656    29.38   0.000     .2513442    .2872726
       _rcs3 |  -.0091212   .0058455    -1.56   0.119    -.0205785    .0023361
       _rcs4 |   .0470241     .00389    12.09   0.000     .0393986    .0546496
       _rcs5 |   .0116256   .0030814     3.77   0.000     .0055861     .017665
       _cons |  -2.572192   .0645294   -39.86   0.000    -2.698717   -2.445667
-------------+----------------------------------------------------------------
dxb          |
     _d_rcs1 |   1.144751   .0121193    94.46   0.000     1.120996    1.168506
     _d_rcs2 |   .2693084   .0091656    29.38   0.000     .2513442    .2872726
     _d_rcs3 |  -.0091212   .0058455    -1.56   0.119    -.0205785    .0023361
     _d_rcs4 |   .0470241     .00389    12.09   0.000     .0393986    .0546496
     _d_rcs5 |   .0116256   .0030814     3.77   0.000     .0055861     .017665
------------------------------------------------------------------------------


.
. // predict survival using -mi predictnl-
. mi predictnl survimp2 = predict(survival at(agegrp 2)) using mi_stpm2
(2356 missing values generated)
```

(k)


Predicted survival for agegrp==2 (60-74)
0
.2
.4
.6
.8
1
0
2
4
6
8
10
analysis time when record ends
Localised (Complete)
Regional (Complete)
Distant (Complete)
Localised (Imputed)
Regional (Imputed)
Distant (Imputed)