Biostatistics III: Survival analysis for epidemiologists
Solutions to exercises

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

http://biostat3.net

Contents

1 Exercise solutions

100. **Kaplan-Meier estimates of survival**

The hand-calculated results can be found in the Excel file `solution_exercise100.xls` and in the Stata output for exercise 101.

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

First, prepare the data for survival time analysis by using stset.

```
. 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

Read the table as follows: If you want the 2-year survival proportion, read off the nearest line where time is $\leq$ 24 months. That is, on the row where time is 22 the survival proportion is 0.7617. This is the probability of surviving up until 24 months (as the function will be flat from time 22 until the next event at time 27).

To produce a graph of the Kaplan-Meier estimates:

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

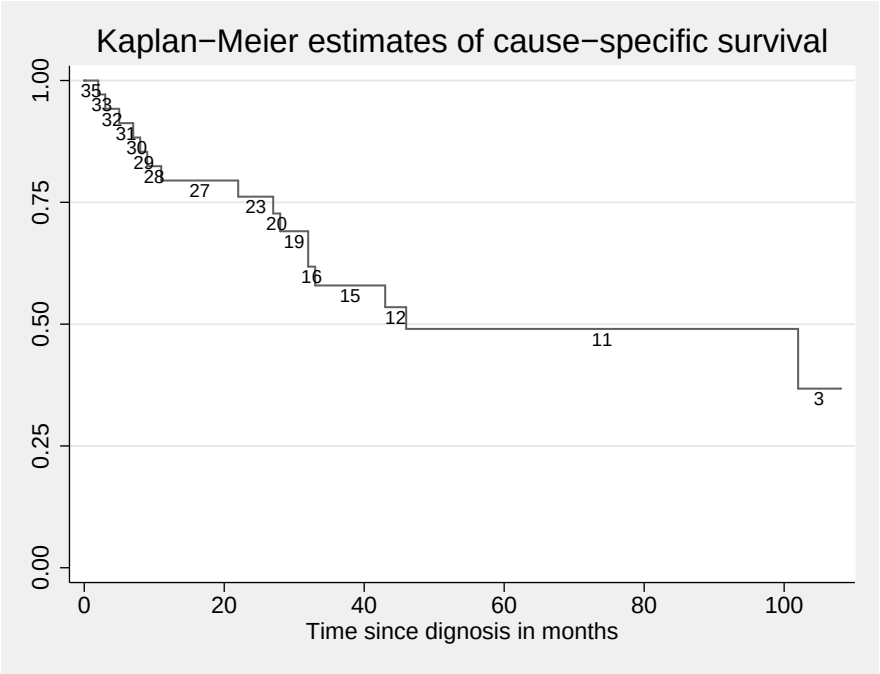


Figure 1: Kaplan-Meier plot of the cause-specific survivor function for sample of 35 patients diagnosed with colon carcinoma. The number at risk at each time point are shown on the curve.

EXTRA: Actuarial method 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

Read the table as follows: If you want the 2-year survival proportion, you must survive the second year, namely the interval 12-24. The survival proportion is 0.7658. This is the probability of surviving up until 24 months.

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

To calculate the Kaplan-Meier estimate for 10-year survival, use `stset` with the two different variables surv_yy and surv_mm. To find the 10-year survival proportion, read off the `sts list` output at the nearest line where time is ≤10 years or ≤120 months.

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

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

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

	Kaplan-Meier
Years	0.7729
Months	0.7645

(a) Using surv_mm is most appropriate, since using months as the time unit will reduce the number of ties (fewer events and censorings will occur at the same time point). If year is the time unit, then all events and censorings within 12 months of each year will be considered as happening on the same time. This will influence the precision of the method.

(b) Both estimates are biased, but using month (meaning fewer ties) will make the bias smaller. The Kaplan-Meier method assumes that all individuals at the time point where survival is estimated are at risk at the start of that time point. If there are ties, this assumption will lead to an overestimate of the number of persons at risk during the time point (time interval). This will therefore underestimate the interval-specific mortality (d/l) at that time point, and consequently the cumulative survival proportion will be overestimated. On average, we do not expect those who end their follow-up during a year to contribute risktime that full year, which is what the method assumes by allowing them to all be at risk at the start of the time point. The fewer the ties, the smaller this bias will be. Hence, we prefer to use time in months rather than years, as the bias will be smaller.

(c) **EXTRA:** The actuarial method is most appropriate because it deals with ties (events and censorings at the same time) in a more appropriate manner. The actuarial method assumes that those who are censored during the time interval only contribute with half of the risk time for that interval, i.e. they are assumed to be censored on average after half a year (or month). The fact that there are a reasonably large number of ties in these data means that there is a difference between the Kaplan-Meier and actuarial estimates, and the actuarial method will be most appropriate. However, when there are fewer ties (if time unit is months), then the two estimation methods are very similar, and the bias from the Kaplan-Meier method is negligible.

EXTRA: How to obtain the actuarial estimates:

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

. ltable surv_yy csr_fail
. ltable surv_mm csr_fail
```

	Kaplan-Meier	Actuarial
Years	0.7729	0.7633
Months	0.7645	0.7637

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)
```

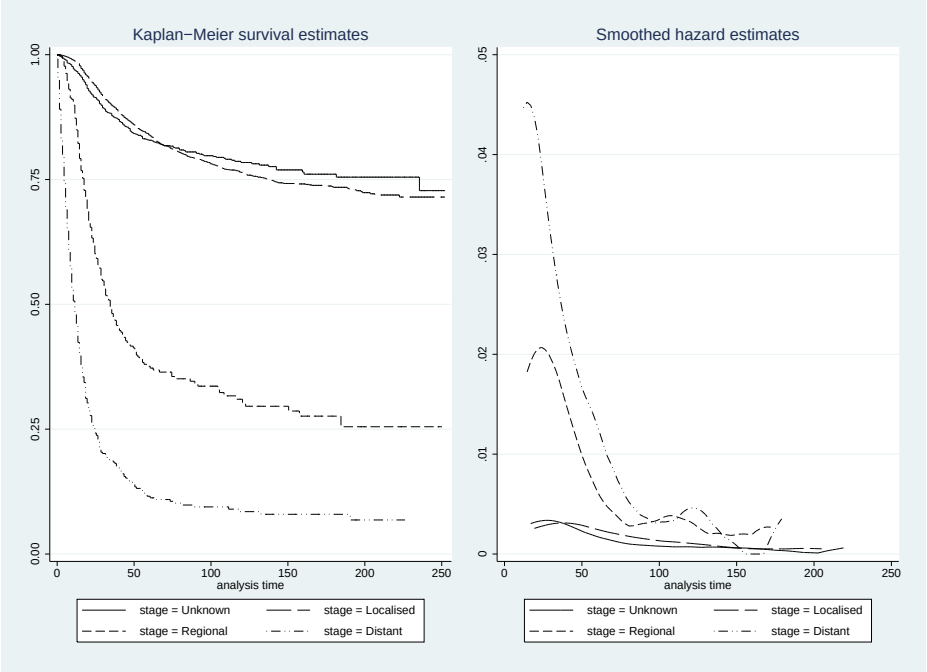


Figure 2: Skin melanoma. Kaplan-Meier estimates of cause-specific survival and mortality rate for each stage.

(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)
```

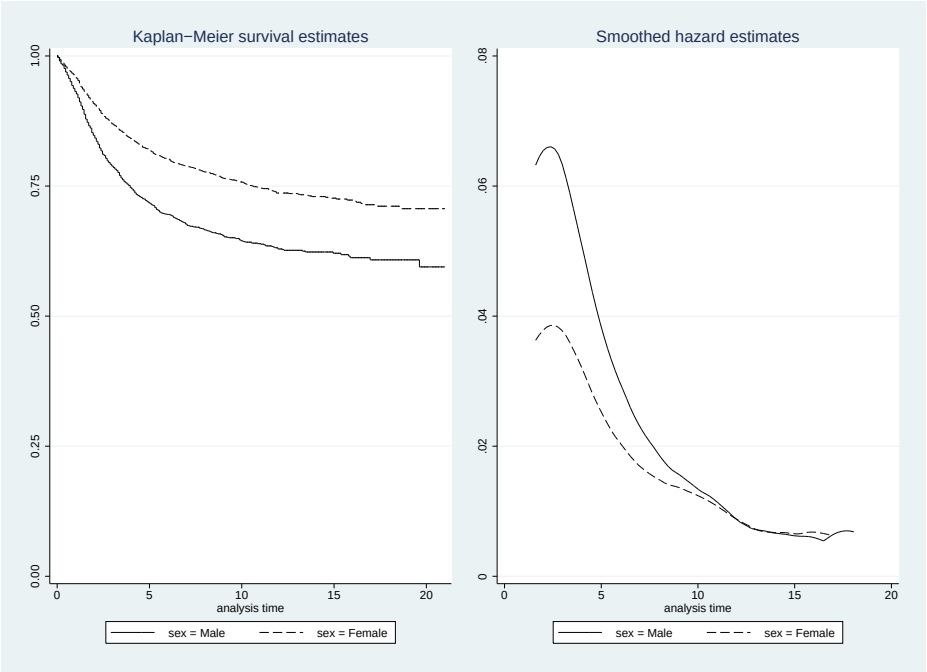


Figure 3: Skin melanoma (all stages). Kaplan-Meier estimates of cause-specific survival and mortality for each sex.

(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)
```

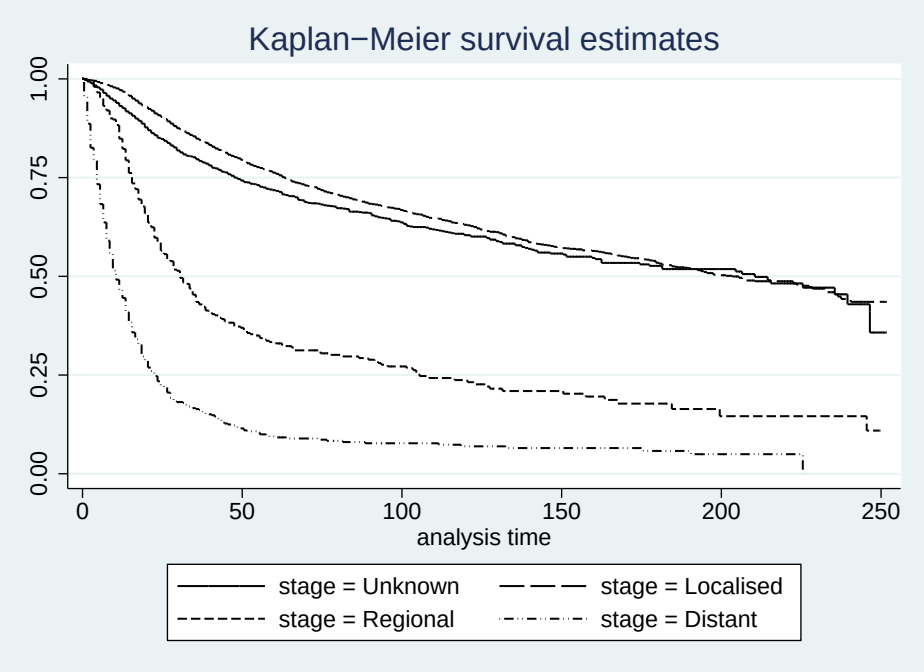


Figure 4: Skin melanoma (all stages). Kaplan-Meier estimates of all-cause survival for each stage.

(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)
```

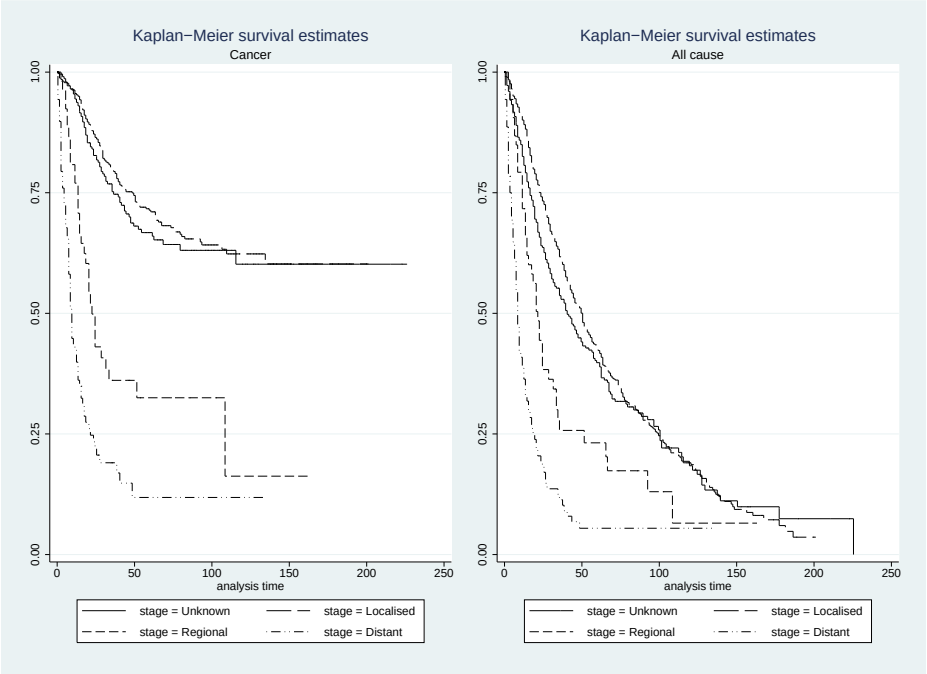


Figure 5: Skin melanoma (all stages). Kaplan-Meier estimates of all-cause survival versus cause-specific survival for each stage.

(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)
```

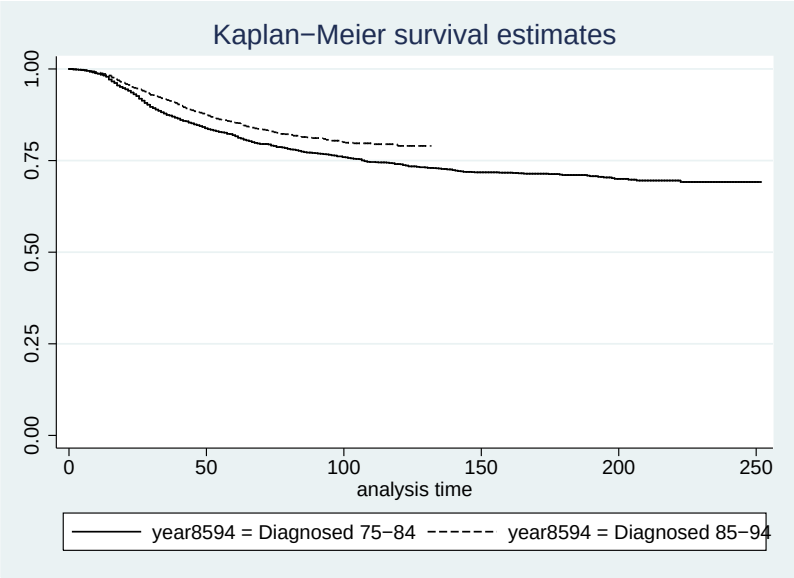


Figure 6: Skin melanoma. Kaplan-Meier plot of the cause-specific survivor function for each calendar period of diagnosis

(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)`

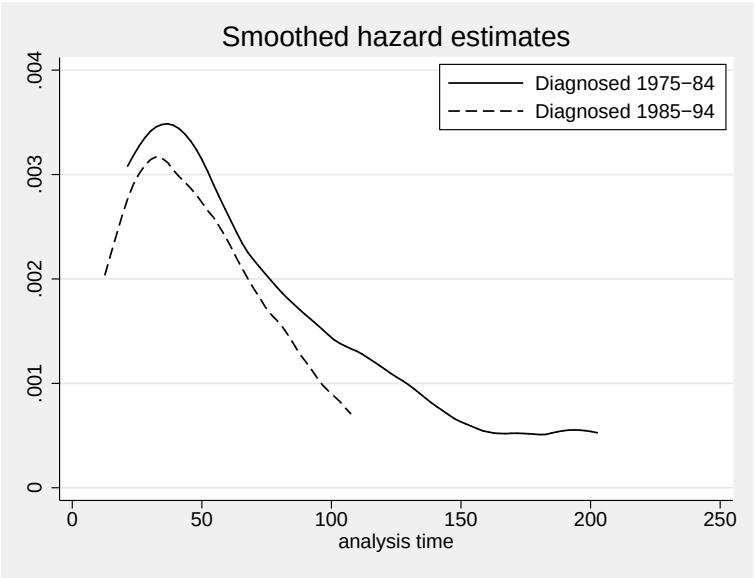


Figure 7: Skin melanoma. Plot of the cause-specific hazard for each calendar period of diagnosis

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)
```

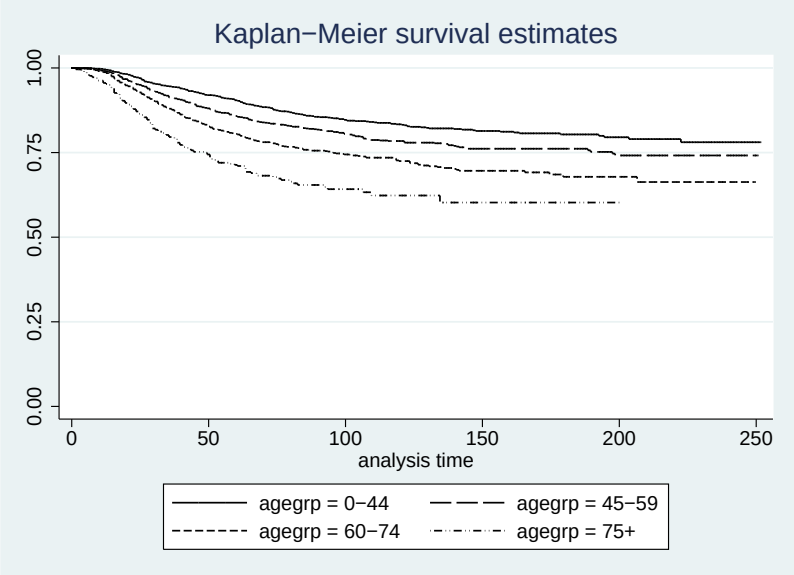


Figure 8: Skin melanoma. Plot of the cause-specific survival function for each age group

(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 (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) The model formulation for the previous poisson model can be written:

$$\ln(\lambda) = \beta_0 + \beta_1 \text{hieng}$$

(d) 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
```

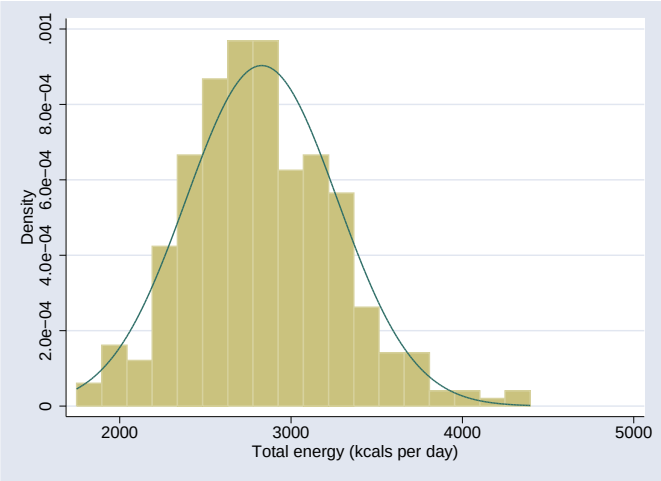


Figure 9: Histogram of energy with superimposed normal density curve (with the sample mean and variance).

```
. 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
```

(e)
```
. 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
```

(f) 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 |
  +-------------------------------------------------+
```

The incidence rate ratio for the second level (to the first) is:

```
. display 10.9059/16.9020
.64524317
```

The incidence rate ratio for the third level (to the first) is:

```
. display 4.8787/16.9020
.28864631
```

(g)
```
. 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	

(h)
```
. list energy eng3 X1 X2 X3 if eng3==1500 in 1/100
```

```
------------------------------+
     |  energy   eng3   X1   X2   X3 |
     |------------------------------|
  1. | 2023.25   1500    1    0    0 |
  2. | 2448.68   1500    1    0    0 |
  3. | 2281.38   1500    1    0    0 |
  4. | 2467.95   1500    1    0    0 |
  5. | 2362.93   1500    1    0    0 |
     |------------------------------|


. list energy eng3 X1 X2 X3 if eng3==2500 in 1/100
     +------------------------------+
     |  energy   eng3   X1   X2   X3 |
     |------------------------------|
 76. | 2664.64   2500    0    1    0 |
 77. | 2533.33   2500    0    1    0 |
 78. | 2854.08   2500    0    1    0 |
 79. | 2673.77   2500    0    1    0 |
 80. | 2766.88   2500    0    1    0 |
     |------------------------------|


. list energy eng3 X1 X2 X3 if eng3==3000 in 200/300
     +------------------------------+
     |  energy   eng3   X1   X2   X3 |
     |------------------------------|
226. | 3067.36   3000    0    0    1 |
227. | 3298.95   3000    0    0    1 |
228. |  3147.6   3000    0    0    1 |
229. | 3180.47   3000    0    0    1 |
230. | 3045.81   3000    0    0    1 |
     |------------------------------|
```

(i) 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)
------------------------------------------------------------------------------
```

(j) The model formulation for the previous poisson model can be written:

$$\ln(\lambda) = \beta_0 + \beta_1 \text{X2} + \beta_2 \text{X3}$$

(k) 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)
------------------------------------------------------------------------------
```

The model formulation is similar to the previous, but now X2 has been replaced by X1 indicating that X2 is now the reference.

$$\ln(\lambda) = \beta_0 + \beta_1 \text{X1} + \beta_2 \text{X3}$$

(l) 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)
------------------------------------------------------------------------------
```

(m) Somehow (there are many different alternatives) you 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 (85-94).

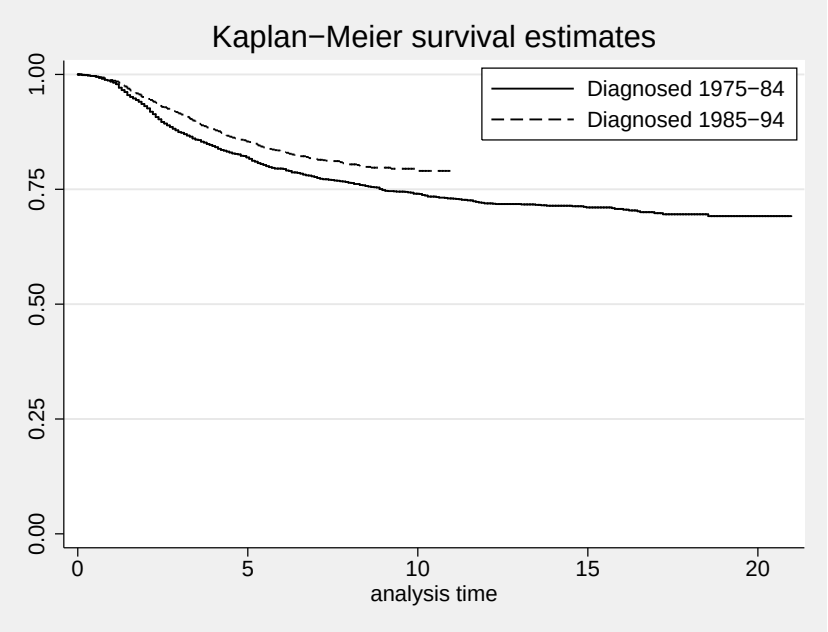


Figure 10: Localised melanoma. Kaplan-Meier estimates of cause-specific survival.

ii. Mortality is lower during the latter period.

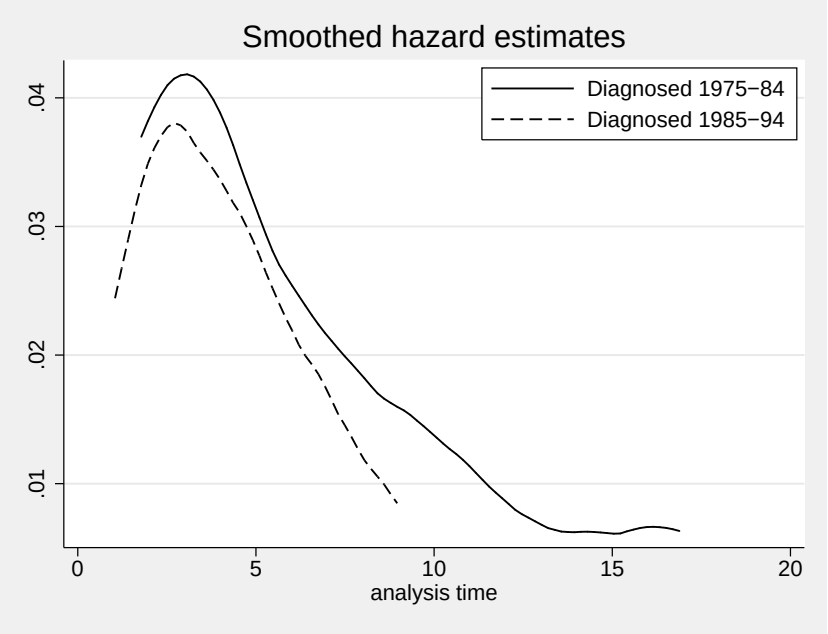


Figure 11: Localised melanoma. Smoothed cause-specific hazards (cause-specific mortality rates).

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. Patients diagnosed with localised melanoma in years 85-94 have approximately 12% lower mortality (due to melanoma) than those diagnosed in years 75-84.

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

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
    year8594 |
Diagnosed 85-94 |   .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.

iv. $\ln(\lambda) = \beta_0 + \beta_1 \text{year8594}$

(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.

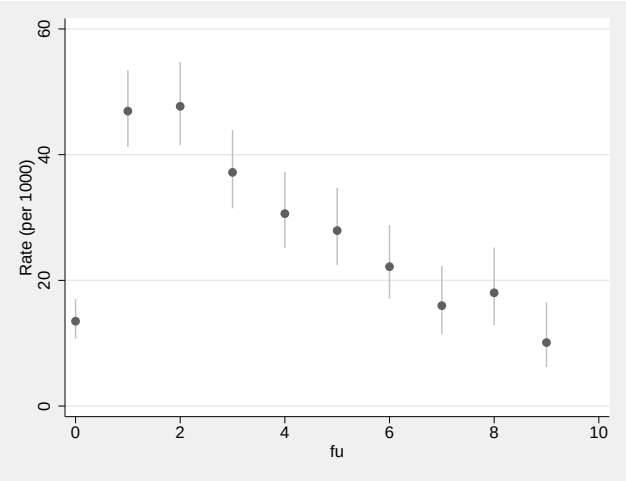


Figure 12: Localised melanoma. Disease-specific mortality rates as a function of time since diagnosis (annual intervals).

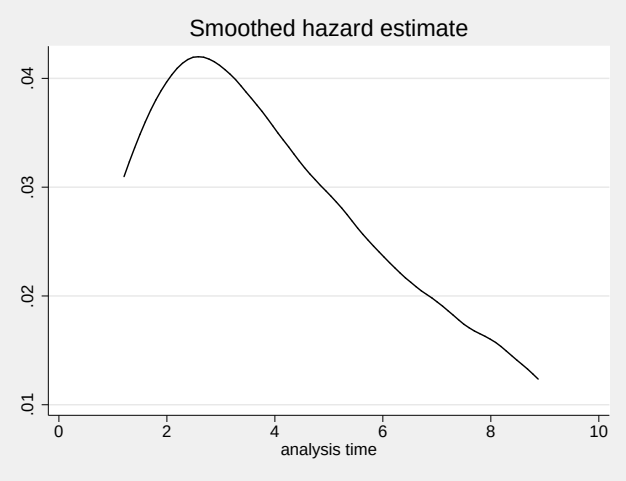


Figure 13: Localised melanoma. Disease-specific mortality rates as continuous function of time since diagnosis (using a smoother).

(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.

i. $\ln(\lambda) = \beta_0 + \beta_1 \text{fu}_{1-2} + \beta_2 \text{fu}_{2-3} + \beta_3 \text{fu}_{3-4} + \beta_4 \text{fu}_{4-5} + \beta_5 \text{fu}_{5-6} + \beta_6 \text{fu}_{6-7} + \beta_7 \text{fu}_{7-8} + \beta_8 \text{fu}_{8-9} + \beta_9 \text{fu}_{9-10}$, where fu_{1-2} indicates follow-up between years 1 and 2.

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

```
Exponential PH regression

No. of subjects =        5,318                  Number of obs    =      34,309
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 |
Diagnosed 85-94 |   .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 i.year8594 i.sex, dist(exp)

Exponential PH regression

No. of subjects =        5,318                 Number of obs    =      34,309
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 |
Diagnosed 85-94 |   .7224105   .0478125    -4.91   0.000     .6345233    .8224709
             |
         sex |
      Female |   .5875465   .0384565    -8.12   0.000     .5168076     .667968
       _cons |   .0126917   .0018177   -30.49   0.000     .0095854    .0168046
------------------------------------------------------------------------------
```

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 i.year8594##i.sex, dist(exp)`

```
Exponential PH regression

No. of subjects =        5,318                 Number of obs    =      34,309
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. The model formulation is:

$$\ln(\lambda) = \beta_0 + \beta_1 \text{fu}_{1-2} + \beta_2 \text{fu}_{2-3} + \beta_3 \text{fu}_{3-4} + \beta_4 \text{fu}_{4-5} + \beta_5 \text{fu}_{5-6} + \beta_6 \text{fu}_{6-7} + \beta_7 \text{fu}_{7-8} + \beta_8 \text{fu}_{8-9} + \beta_9 \text{fu}_{9-10} + \beta_{10} \text{age45-59} + \beta_{11} \text{age60-74} + \beta_{12} \text{age75+} + \beta_{13} \text{year8594} + \beta_{14} \text{female} + \beta_{15} \text{year8594} * \text{female}$$

(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 i.year8594 sex_early sex_latter, dist(exp)

Exponential PH regression

No. of subjects =        5,318                 Number of obs    =      34,309
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_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 i.year8594#i.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 have no clear pattern and 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)
```

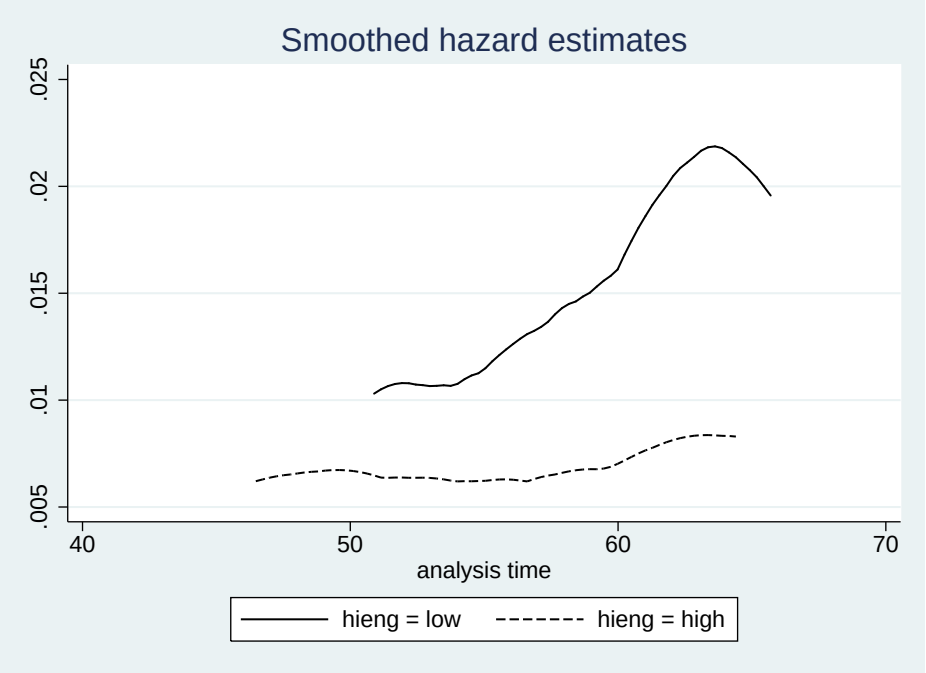


Figure 14: Diet data. Kaplan-Meier estimates of hazard rate for each energy intake level, with attained age as time scale.

```
* 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)
```

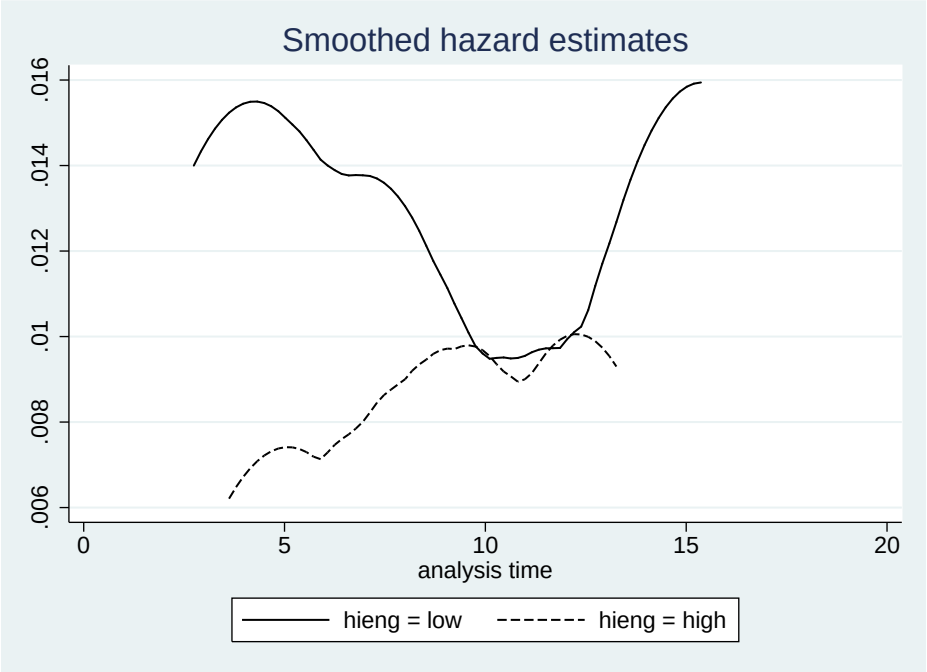


Figure 15: Diet data. Kaplan-Meier estimates of hazard rate for each energy intake level, with time since entry as time scale.

(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 i.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 |
       high  |   .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 i.hieng i.job bmi, e(y) irr

Poisson regression                              Number of obs     =        332
                                                LR chi2(4)        =       8.16
                                                Prob > chi2       =     0.0861
Log likelihood = -168.42784                     Pseudo R2         =     0.0236

------------------------------------------------------------------------------
         chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |
       high  |   .4868519   .1517623    -2.31   0.021     .2642767    .8968811
             |
         job |
  conductor  |   1.579581   .6422652     1.12   0.261     .7119336    3.504649
       bank  |   .8963158   .3315282    -0.30   0.767     .4341298    1.850557
             |
         bmi |   1.071483   .0521307     1.42   0.156     .9740289    1.178687
       _cons |   .0024302   .0030291    -4.83   0.000     .0002112    .0279646
       ln(y) |          1  (exposure)
------------------------------------------------------------------------------
```

$$\ln(\lambda) = \beta_0 + \beta_1 \text{hieng} + \beta_2 \text{conductor} + \beta_3 \text{banker} + \beta_4 \text{bmi}$$

(d) The y variable is not correct since it is kept for all split 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-_t0
. 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 split 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 i.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 |
       high  |   .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 i.hieng i.job bmi i.ageband, 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 |
       high  |   .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)
------------------------------------------------------------------------------
```

Our timescale in this model is attained age, since we have included this in our model using the variable `ageband`, we have made the assumption that the underlying rate is constant within each of the three agebands.

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

Poisson regression                              Number of obs     =      1,100
                                                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 |
       high  |    .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 i.hieng i.job bmi i.fuband, e(risktime) irr

Poisson regression                              Number of obs     =      1,084
                                                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 |
       high  |   .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. We can see this by comparing the models where we do not adjust for time-since-entry (IRR for `hieng`=0.52, see 112b) and the model where we adjust for time-since-entry (IRR for `hieng`=0.52). We can also see this by considering the graphs at the beginning of the exercise where we concluded that the rates were constant over time-since-entry. 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 i.year8594

Cox regression -- Breslow method for ties

No. of subjects =        5,318                 Number of obs    =       5,318
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 |
Diagnosed 85-94  |   .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 any potential confounders except time. 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 i.sex i.year8594 i.agegrp`

```
Cox regression -- Breslow method for ties
No. of subjects =        5,318                 Number of obs    =       5,318
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 |
         Female |   .5888144   .0385379    -8.09   0.000     .5179256    .6694059
                |
       year8594 |
Diagnosed 85-94 |   .7168836   .0474446    -5.03   0.000     .6296723    .8161739
                |
         agegrp |
          45-59 |   1.326397   .1249113     3.00   0.003     1.102841     1.59527
          60-74 |   1.857323   .1687866     6.81   0.000     1.554295     2.21943
            75+ |   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.
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. 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.

```
. stcox i.year8594 i.sex i.agegrp
. est store Cox

. stsplit fu, at(0(12)120) trim
. streg i.fu i.year8594 i.sex i.agegrp, dist(exp)
. est store Poisson
. est table Cox Poisson, eform equations(1)
```

```
-------------------------------------
    Variable |    Cox        Poisson
-------------+-----------------------
    year8594 |
Diagnosed..  |  .71688362    .72241051
             |
         sex |
      Female |  .58881445    .58754651
             |
      agegrp |
       45-59 |  1.3263971    1.3277947
       60-74 |  1.8573227    1.8623763
         75+ |  3.3726522    3.4002869
             |
          fu |
          12 |               3.5546847
          24 |               3.6934975
          36 |               2.9321966
          48 |               2.4477533
          60 |               2.2562326
          72 |               1.7974533
          84 |               1.2886666
          96 |               1.4394596
         108 |               .79615726
             |
       _cons |               .00105764
-------------------------------------
```

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) i.

$$\ln(\lambda) = \beta_0 + \beta_1 \text{fu}_{1-2} + \beta_2 \text{fu}_{2-3} + \beta_3 \text{fu}_{3-4} + \beta_4 \text{fu}_{4-5} + \beta_5 \text{fu}_{5-6} + \beta_6 \text{fu}_{6-7} + \beta_7 \text{fu}_{7-8} + \beta_8 \text{fu}_{8-9} + \beta_9 \text{fu}_{9-10} + \beta_{10} \text{age1} + \beta_{11} \text{age2} + \beta_{12} \text{age3} + \beta_{13} \text{year8594} + \beta_{14} \text{sex}$$

ii.

$$\text{Model (a): } \ln(\lambda(t)) = \ln(\lambda_0(t)) + \beta_1 \text{year8594}$$

$$\text{Model (c): } \ln(\lambda(t)) = \ln(\lambda_0(t)) + \beta_1 \text{year8594} + \beta_2 \text{sex} + \beta_3 \text{age1} + \beta_4 \text{age2} + \beta_5 \text{age3}$$

The intercept in the Poisson regression model β_0 is the log rate in the first timeband of followup (0-1 year since diagnosis), in the reference level of all variables X, i.e. males diagnosed 1975-84 in agegroup 0. The "intercept" in the Cox models (a) and (c) is the log baseline rate$\ln(\lambda_0(t))$, which is the rate among the persons at the reference level of all variables X, i.e. males diagnosed 1975-84 in agegroup 0. This intercept is not estimated, so it is not a parameter in the model. This Cox baseline rate corresponds, conceptually, to the intercept plus the linear predictor for fu_{1-2}, ...,fu_{9-10} in the Poisson model, $\beta_1 \text{fu}_{1-2} + \beta_2 \text{fu}_{2-3} + \beta_3 \text{fu}_{3-4} + \beta_4 \text{fu}_{4-5} + \beta_5 \text{fu}_{5-6} + \beta_6 \text{fu}_{6-7} + \beta_7 \text{fu}_{7-8} + \beta_8 \text{fu}_{8-9} + \beta_9 \text{fu}_{9-10}$.

iii. Rate of males diagnosed 1985-94 in agegroup 2:

$$\lambda(t|\text{sex} = 0, \text{year8594} = 1, \text{age2} = 1) = \lambda_0(t) \exp(\beta_1 * 1 + \beta_2 * 0 + \beta_3 * 0 + \beta_4 * 1 + \beta_5 * 0) = \lambda_0(t) \exp(\beta_1 + \beta 4)$$

Rate of females diagnosed 1985-94 in agegroup 2:

$$\lambda(t|\text{sex} = 1, \text{year8594} = 1, \text{age2} = 1) = \lambda_0(t) \exp(\beta_1 * 1 + \beta_2 * 1 + \beta_3 * 0 + \beta_4 * 1 + \beta_5 * 0) = \lambda_0(t) \exp(\beta_1 + \beta_2 + \beta 4)$$

Hazard ratio females to males diagnosed 1985-94 in agegroup2:

$$\text{HR} = (\lambda_0(t)\exp(\beta_1+\beta_2+\beta_4))/(\lambda_0(t)\exp(\beta_1+\beta_4)) = \exp(\beta_2)$$

Comment: The hazard ratio of females to males diagnosed 1985-94 in agegroup 2 is a constant, and so does not vary over time t. This is the definition of proportional hazards. Hence, the rates of females and males are assumed to be proportional over time in this model specification.

(g) `. 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.

(h) `. 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.

(i) 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)
```

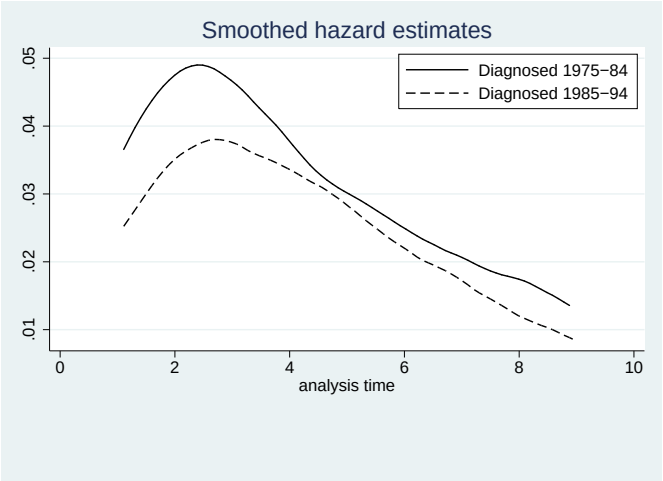


Figure 16: Localised skin melanoma. Plot of the estimated hazard function for each calendar period of diagnosis.

(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)
```

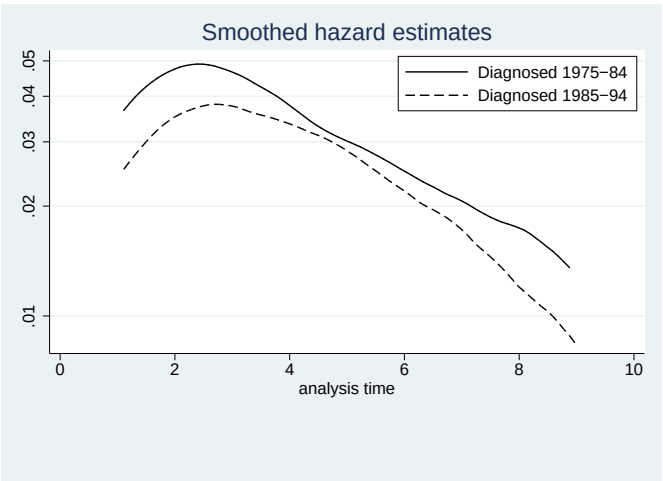


Figure 17: Localised skin melanoma. Plot of the estimated hazard function for each calendar period of diagnosis using a log scale for the y axis.

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

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

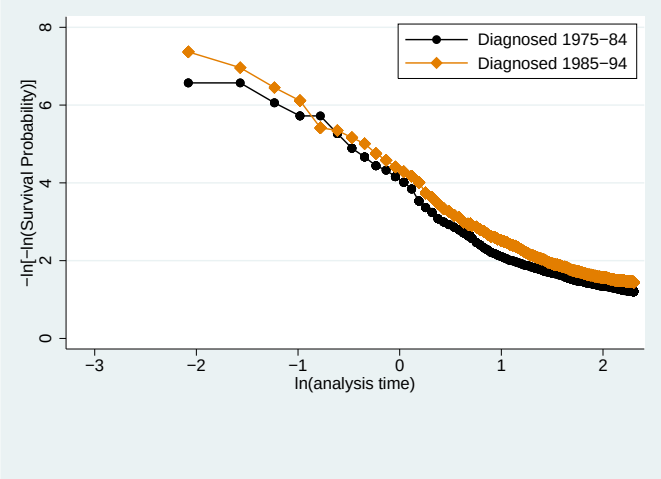


Figure 18: Localised skin melanoma. Plot of the log cumulative hazard function for each calendar period of diagnosis. Each plot symbol represents an event time. Note that the x axis is the natural logarithm of time in years, so a value of 0 corresponds to 1 year.

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

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

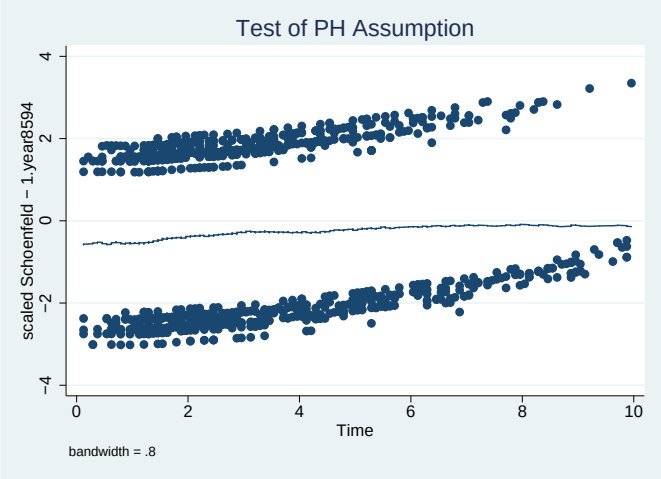


Figure 19: Localised skin melanoma. Plot of the scaled Schoenfeld residuals for calendar period 1985–94. The smooth line shows the estimated hazard ratio as a function of time.

(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 i.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 i.sex i.year8594 agegrp2 agegrp3 agegrp4, tvc(agegrp2 agegrp3 agegrp4) texp(_t>=2)

Cox regression -- Breslow method for ties

No. of subjects =        5,318                 Number of obs    =       5,318
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 |
      Female |   .5906795   .0386481    -8.05   0.000     .5195865    .6714998
             |
    year8594 |
Diagnosed 85-94 |   .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.

```
stsplit fuband, at(0,2)
list id _t0 _t fu in 1/10

stcox i.sex i.year8594 i.agegrp##i.fuband

. 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 i.sex i.year8594 i.fuband i.fuband#i.agegrp

Cox regression -- Breslow method for ties

No. of subjects =        5,318                 Number of obs    =       9,856
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 |
      Female |   .5906795   .0386481    -8.05   0.000     .5195865    .6714998
             |
    year8594 |
Diagnosed 85-94 |   .7153885   .0473797    -5.06   0.000     .6283005    .8145476
    2.fuband |   7.415862          .         .       .            .           .
             |
fuband#agegrp |
     0#45-59 |   1.698848   .3335545     2.70   0.007     1.156187    2.496208
     0#60-74 |   2.457673   .4605845     4.80   0.000     1.702171    3.548502
       0#75+ |   5.399496   1.035355     8.79   0.000      3.70796    7.862694
     2#45-59 |   1.232911   .1328384     1.94   0.052     .9982062    1.522802
     2#60-74 |   1.703178   .1784726     5.08   0.000     1.386961    2.091489
       2#75+ |   2.662634    .350343     7.44   0.000      2.05737    3.445963
------------------------------------------------------------------------------
```

	0-2 years	2+ years
Agegrp0	1.00	1.00
Agegrp1	1.70	1.23
Agegrp2	2.46	1.70
Agegrp3	5.40	2.66

(j)

i.

$$\lambda(t) = \lambda_0(t)\exp(\beta_1\text{sex}+\beta_2\text{year8594}+\beta_3\text{age1}+\beta_4\text{age2}+\beta_5\text{age3}+\beta_6\text{age1}*\text{fu}_2+\beta_7\text{age}_2*\text{fu}_2+\beta_8\text{age}_3*\text{fu}_2)$$

	0-2 years	2+ years
Agegrp0	$\lambda_0(t)$	$\lambda_0(t)$
Agegrp1	$\lambda_0(t)\exp(\beta_3)$	$\lambda_0(t)\exp(\beta_3)\exp[\beta_6)$
Agegrp2	$\lambda_0(t)\exp(\beta_4)$	$\lambda_0(t)\exp(\beta_4)\exp(\beta_7)$
Agegrp3	$\lambda_0(t)\exp(\beta_5)$	$\lambda_0(t)\exp(\beta_5)\exp(\beta_8)$

ii. Hazard ratio comparing agegrp3 to agegrp0, during 0-2y of followup:

$$\text{HR} = (\lambda_0(t)\exp(\beta_5))/(\lambda_0(t)) = \exp(\beta_5)$$

iii. Hazard ratio comparing agegrp3 to agegrp0, during 2+ years of followup:

$$\text{HR} = (\lambda_0(t)\exp(\beta_5)\exp(\beta_8))/(\lambda_0(t)) = \exp(\beta_5)\exp(\beta_8)$$

(k) 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.

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

(a) `. stcox i.sex`

```
Cox regression -- Breslow method for ties

No. of subjects =        7,775                 Number of obs    =       7,775
No. of failures =        1,913
Time at risk    =     615236.5
                                               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 |
      Female |   .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 i.sex i.agegrp i.stage i.subsite i.year8594`

```
Cox regression -- Breslow method for ties

No. of subjects =        7,775                 Number of obs    =       7,775
No. of failures =        1,913
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 |
           Female |   .7490676    .036445    -5.94   0.000     .6809368    .8240153
                  |
           agegrp |
            45-59 |   1.268542   .0855596     3.53   0.000     1.111459    1.447824
            60-74 |   1.730767   .1126805     8.43   0.000     1.523427    1.966326
              75+ |   2.785848   .2128337    13.41   0.000     2.398431    3.235845
                  |
            stage |
       Localised  |   1.038328   .0713262     0.55   0.584     .9075334    1.187972
        Regional  |   4.771515   .4363494    17.09   0.000     3.988549     5.70818
         Distant  |   13.48664   1.097917    31.96   0.000     11.49766     15.8197
                  |
          subsite |
            Trunk |   1.393153   .0984179     4.69   0.000     1.213016    1.600041
            Limbs |   1.032021   .0767263     0.42   0.672     .8920829     1.19391
Multiple and NOS  |   1.305318    .133562     2.60   0.009      1.06812     1.59519
                  |
         year8594 |
 Diagnosed 85-94  |   .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 i.year8594 i.subsite i.stage i.agegrp i.sex#i.agegrp

Cox regression -- Breslow method for ties

No. of subjects =        7,775                 Number of obs    =       7,775
No. of failures =        1,913
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 |
 Diagnosed 85-94 |   .7868595   .0376845    -5.01   0.000     .7163599    .8642973
             |
     subsite |
       Trunk |   1.401988   .0992064     4.78   0.000     1.220428    1.610558
       Limbs |   1.039415   .0773326     0.52   0.603     .8983792    1.202593
Multiple and NOS |   1.315538   .1349198     2.67   0.007     1.075983    1.608428
             |
       stage |
   Localised |   1.036942   .0712433     0.53   0.598     .9063011    1.186414
    Regional |   4.702828   .4312718    16.88   0.000     3.929161    5.628833
     Distant |   13.38869   1.091144    31.83   0.000     11.41215    15.70757
             |
      agegrp |
       45-59 |   1.188947   .1014449     2.03   0.043     1.005855    1.405367
       60-74 |     1.5508   .1318113     5.16   0.000     1.312827    1.831911
         75+ |   2.485421   .2605605     8.68   0.000     2.023782    3.052363
             |
  sex#agegrp |
 Female#0-44 |   .6251314   .0662091    -4.44   0.000     .5079472    .7693502
Female#45-59 |   .7300673   .0678894    -3.38   0.001      .608428    .8760252
Female#60-74 |   .8120201   .0653462    -2.59   0.010     .6935337    .9507494
  Female#75+ |   .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 i.sex i.year8594 i.agegrp i.subsite i.stage
estat phtest, detail
```

```
Test of proportional-hazards assumption

Time:  Time
------------------------------------------------------------------
             |       rho            chi2       df       Prob>chi2
-------------+----------------------------------------------------
1b.sex       |            .               .        1            .
2.sex        |      0.03157         1.93         1         0.1644
0b.year8594 |            .               .        1            .
1.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. It seems reasonable that the effect is higher in the first 2 years after diagnosis, so let's fit a model where the HR for stage differs before and after 2 years. 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.

```
. stsplit timeband, at(0,2,100)
(6,100 observations (episodes) created)

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

        Failure _d: status==1
  Analysis time _t: surv_mm/12
       ID variable: id

Iteration 0:   log likelihood = -16394.181
Iteration 1:   log likelihood =  -16061.47
Iteration 2:   log likelihood = -15461.021
Iteration 3:   log likelihood = -15410.585
Iteration 4:   log likelihood = -15409.514
Iteration 5:   log likelihood = -15409.514
Refining estimates:
Iteration 0:   log likelihood = -15409.514

Cox regression with Breslow method for ties

No. of subjects =       7,775                         Number of obs =  13,875
No. of failures =       1,913
Time at risk    = 51,269.7083
                                                      LR chi2(17)   = 1969.33
Log likelihood = -15409.514                           Prob > chi2   =  0.0000

------------------------------------------------------------------------------
          _t | Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
```

```
        sex#agegrp |
     Female#0-44   |   .6151956   .0651456    -4.59   0.000     .4998917    .7570952
    Female#45-59   |   .7381433    .068742    -3.26   0.001     .6149924     .885955
    Female#60-74   |   .7995144   .0643722    -2.78   0.005     .6827984    .9361815
      Female#75+   |   .8021172   .0855874    -2.07   0.039     .6507483    .9886956
                   |
             agegrp |
             45-59 |   1.172296   .1000479     1.86   0.063     .9917286    1.385741
             60-74 |   1.551673   .1318516     5.17   0.000     1.313622    1.832864
               75+ |   2.447432   .2566963     8.53   0.000      1.99266    3.005993
                   |
          year8594 |
 Diagnosed 85-94   |   .7901069   .0377861    -4.93   0.000     .7194124    .8677482
                   |
           subsite |
             Trunk |   1.363457   .0963669     4.39   0.000      1.18708    1.566041
             Limbs |    1.01201   .0752092     0.16   0.872     .8748355    1.170694
Multiple and NOS   |   1.284234   .1318631     2.44   0.015     1.050132    1.570522
                   |
             stage |
       Localised   |   .6945836   .0735206    -3.44   0.001     .5644509    .8547179
        Regional   |   4.786207   .6028838    12.43   0.000     3.739141    6.126482
         Distant   |   15.78975    1.66382    26.19   0.000     12.84344    19.41196
                   |
       2.timeband  |   3.377186          .        .       .            .           .
                   |
   stage#timeband  |
     Localised#2   |   1.900092   .2646924     4.61   0.000     1.446099    2.496613
      Regional#2   |   .9275423   .1698571    -0.41   0.681     .6478233    1.328039
       Distant#2   |   .4055014    .074699    -4.90   0.000     .2826111    .5818292
------------------------------------------------------------------------------
```

```
. 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) =    4.61
         Prob > chi2 =    0.2029
```

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

(e)

$$\text{Model in (a):}\lambda(t) = \lambda_0(t)\exp(\beta_1 \text{sex})$$

$$\text{Model in (b):}\lambda(t) = \lambda_0(t)\exp(\beta_1 \text{sex} + \beta_2 \text{age}_1 + \beta_3 \text{age}_2 + \beta_4 \text{age}_3 + \beta_5 \text{stage}_1 + \beta_6 \text{stage}_2 + \beta_7 \text{stage}_3 + \beta_8 \text{subsite}_1 + \beta_9 \text{subsite}_2 + \beta_{10} \text{subsite}_3 + \beta_{11} \text{year8594})$$

$$\text{Model in (c):}\lambda(t) = \lambda_0(t)\exp(\beta_1 \text{sex} + \beta_2 \text{age}_1 + \beta_3 \text{age}_2 + \beta_4 \text{age}_3 + \beta_5 \text{stage}_1 + \beta_6 \text{stage}_2 + \beta_7 \text{stage}_3 + \beta_8 \text{subsite}_1 + \beta_9 \text{subsite}_2 + \beta_{10} \text{subsite}_3 + \beta_{11} \text{year8594} + \beta_{12} \text{sex} * \text{age}_1 + \beta_{13} \text{sex} * \text{age}_2 + \beta_{14} \text{sex} * \text{age}_3)$$

i. Rate for females in agegroup3 while all other variables is at reference level:

$$\lambda(t) = \lambda_0(t)\exp(\beta_1 + \beta_4 + \beta_{14})$$

ii. Rate for males in agegroup3 while all other variables is at reference level:

$$\lambda(t) = \lambda_0(t)\exp(\beta_4)$$

Hazard ratio females to males:

$$\text{HR} = (\lambda_0(t)\exp(\beta_1 + \beta_4 + \beta_{14}))/(\lambda_0(t)\exp(\beta_4)) = \exp(\beta_1 + \beta_{14})$$

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

(a)
```
. poisson chd i.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 |
       high  |   .5203602   .1572055    -2.16   0.031     .2878382    .9407184
       _cons |    .013596   .0025694   -22.74   0.000     .0093875    .0196912
       ln(y) |          1  (exposure)
------------------------------------------------------------------------------

. stset dox, id(id) fail(chd) enter(doe) origin(doe) scale(365.25)
. stcox i.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 |
       high  |   .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) enter(doe) scale(365.24)
. stcox i.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 |
       high  |   .5426351   .1643032    -2.02   0.043     .2997606    .9822933
------------------------------------------------------------------------------
```

(c)

$$\text{Poisson model (a)}:\lambda = \exp(\beta_0 + \beta_1 \text{hieng})$$

$$\text{Cox model (a)}:\lambda(t) = \lambda_0(t)\exp(\beta_1 \text{hieng}), \text{ where t is time-since-diagnosis}$$

i. The Poisson model in (a) is not adjusting for the timescale time-since-diagnosis, but estimates the effect of high versus low energy for overall (averge) rates over followup. Thus, the β_1 from this Poisson model may be confounded by time-since-diagnosis.
The Cox model in (a) is adjusting for the timescale time-since-diagnosis automatically via the baseline hazard. Hence, the β_1 is the effect of high versus low energy intake at each point in time across followup.

ii.

$$\text{Cox model (b)}:\lambda(t_{\text{age}}) = \lambda_0(t_{\text{age}})\exp(\beta_1 \text{hieng}), \text{ where } t_{\text{age}} \text{ is attained age.}$$

The Cox models look similar in (a) and (b), since they only include one parameter β_1, but they are completely different since they timescales are different. In Cox model (a) the β_1 is adjusted for time-since-diagnosis, i.e. the β_1 is the effect of high versus low energy intake adjusted for time-since-diagnosis. While in Cox model (b), the β_1 is adjusted for age, i.e. the β_1 is the effect of high versus low energy intake adjusted for attained age.

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 entry (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
```

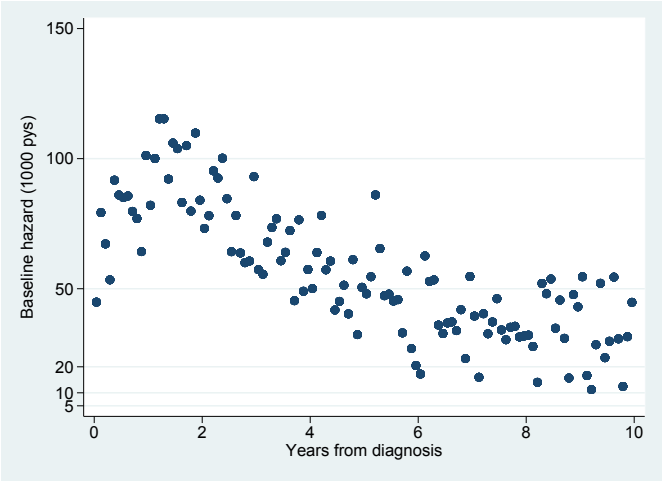


Figure 20: Localised skin melanoma. Plot of the estimated baseline hazard function for the piecewise model.

(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
```

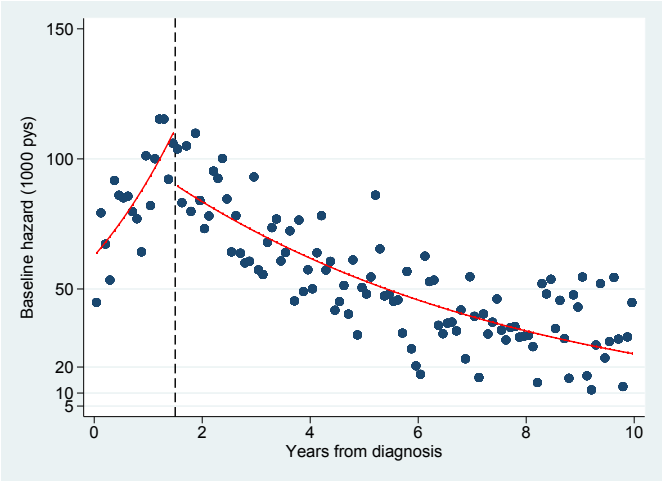


Figure 21: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and linear spline model.

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
```

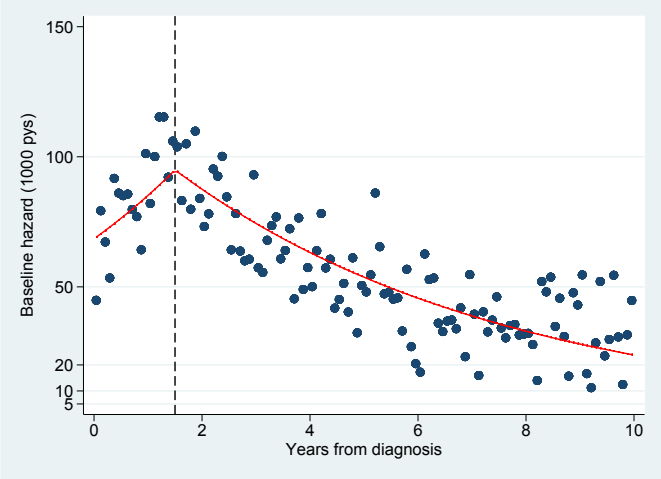


Figure 22: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and linear spline model.

```
. 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)
```

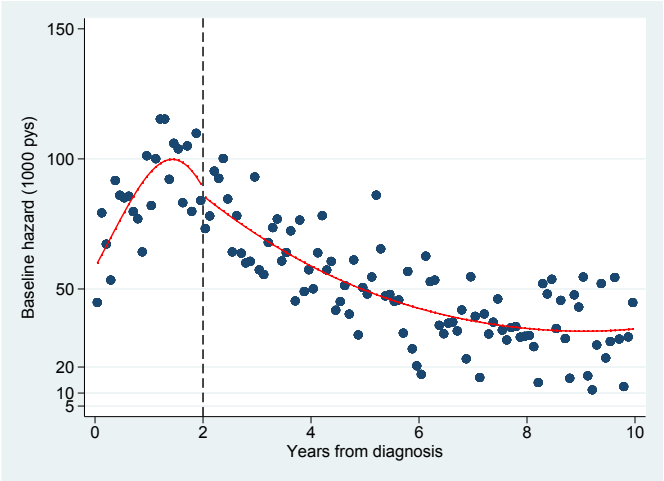


Figure 23: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and cubic spline model.

(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)
```

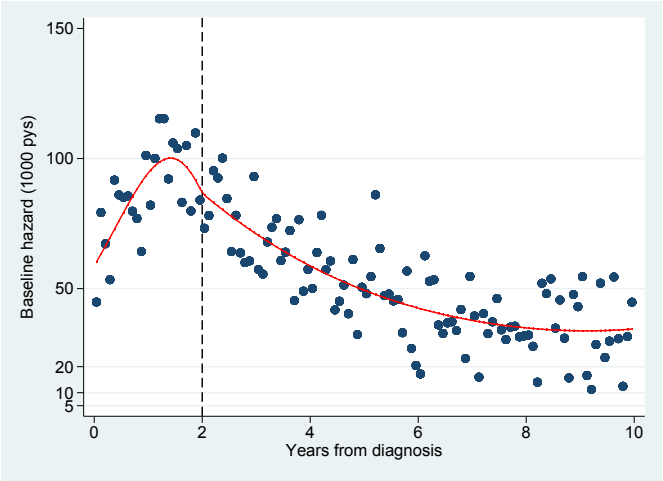


Figure 24: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and cubic spline model.

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)
```

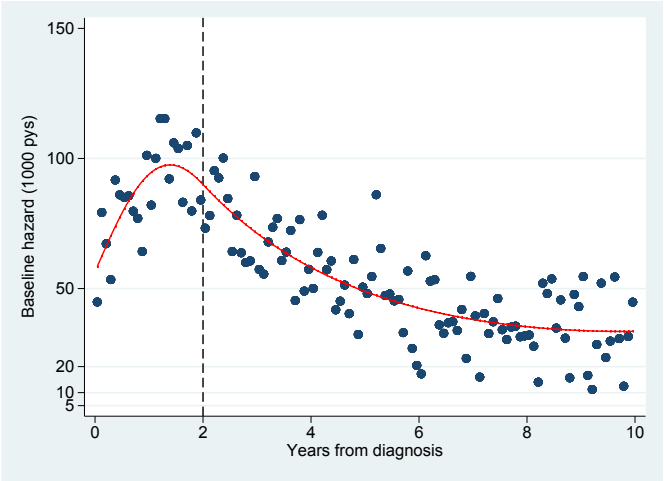


Figure 25: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and cubic spline model with continuous first derivatives.

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)
```

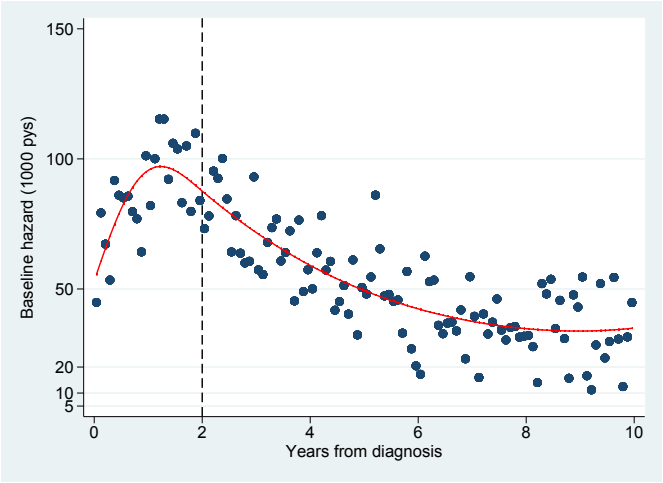


Figure 26: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and cubic spline model with continuous first and second derivatives.

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)
```

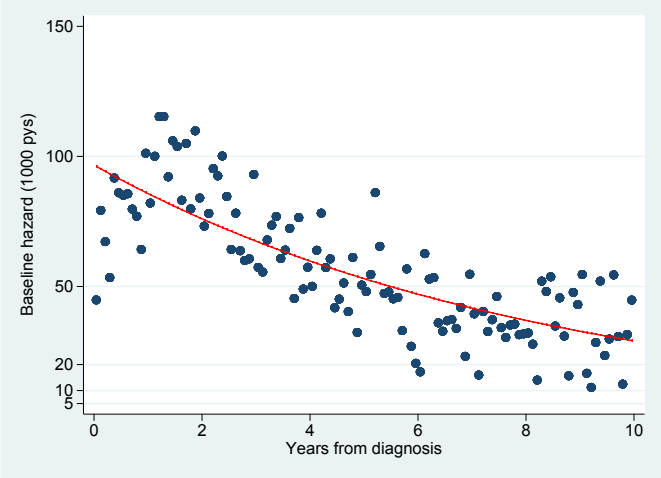


Figure 27: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and linear model.

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)
```

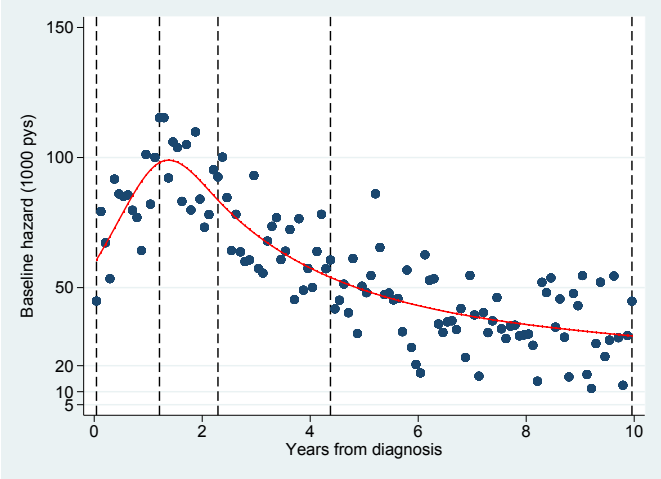


Figure 28: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and restricted cubic spline model.

(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)
```

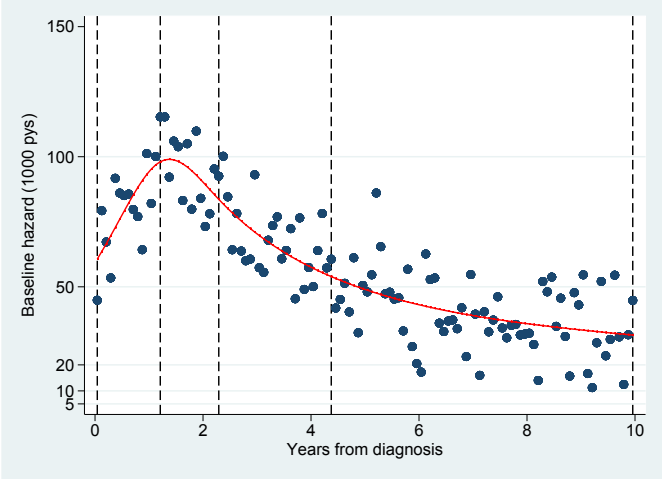


Figure 29: Localised skin melanoma. Plot of the estimated baseline hazard functions for the piecewise model and restricted cubic spline model with knots at 1, 2, and 3 years.

131. **Flexible Parametric Models for cause-specific mortality**

This exercise has no written solutions. A do-file is provided.

132. **Flexible Parametric Models with time-dependent effects**

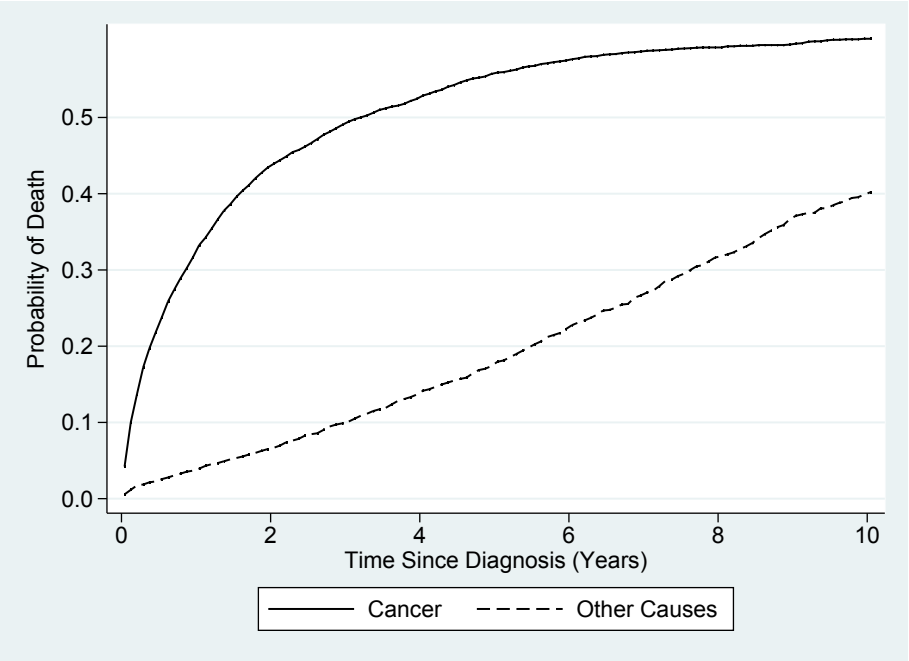
This exercise has no written solutions. A do-file is provided.

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

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

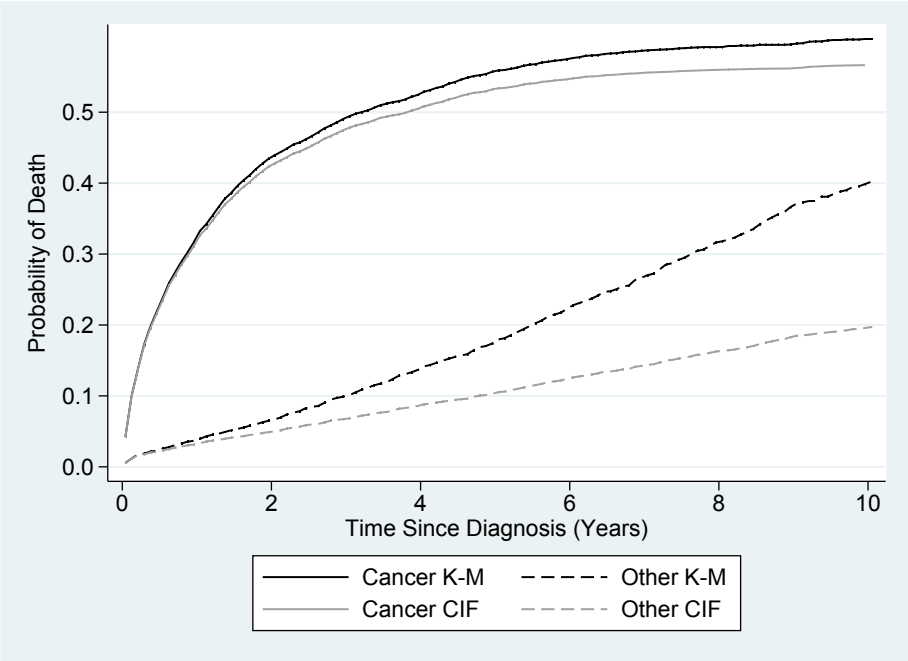
```
use colon, clear
drop if stage ==0
gen female = sex==2
```

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.

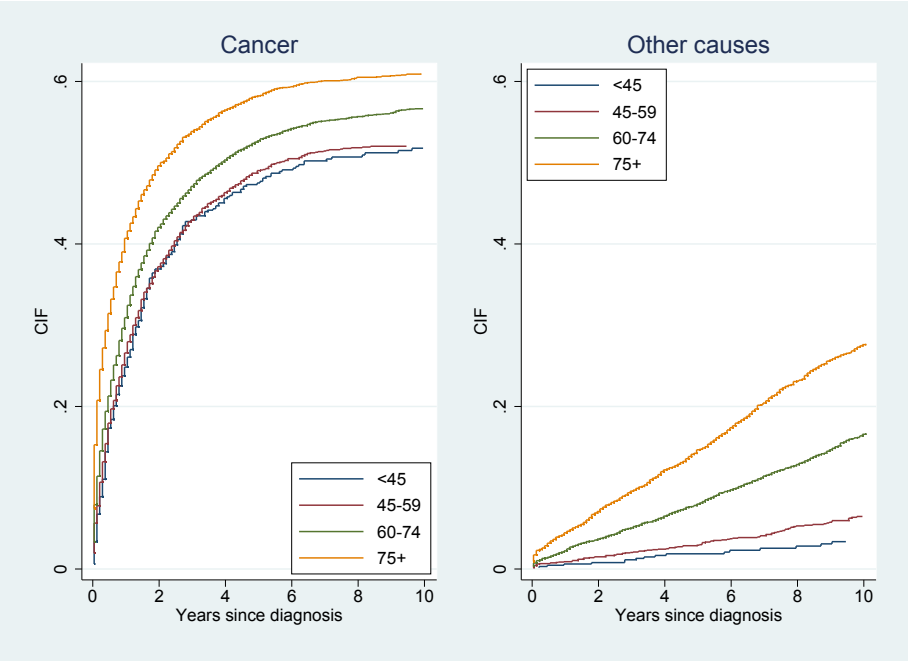
(c) Obtain estimates of the CIF for cancer and other causes by age group. Plot and interpret the curves.

```
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.

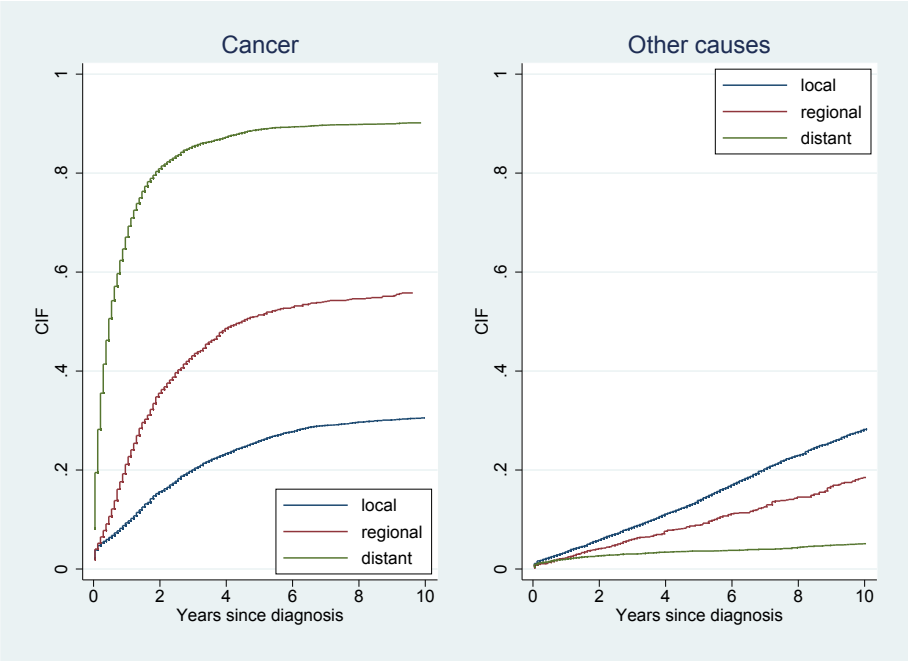
(d) Now obtain the CIF for cancer and other causes by stage group. Plot the results.

```
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.

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)