

Survival analysis

Solutions to Exercises

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

<http://www.bioepi.org/>

Exercise solutions

1. The results are contained in the Excel file `exercise1.xls` and are also shown in the Stata output below.
2. `. ltable surv_mm csr_fail, interval(12)`

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

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

```

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

```

```
-----
35 total obs.
0 exclusions

```

```
-----
35 obs. remaining, representing
16 failures in single record/single failure data
1504 total analysis time at risk, at risk from t = 0
      earliest observed entry t = 0
      last observed exit t = 108

```

. sts list

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

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

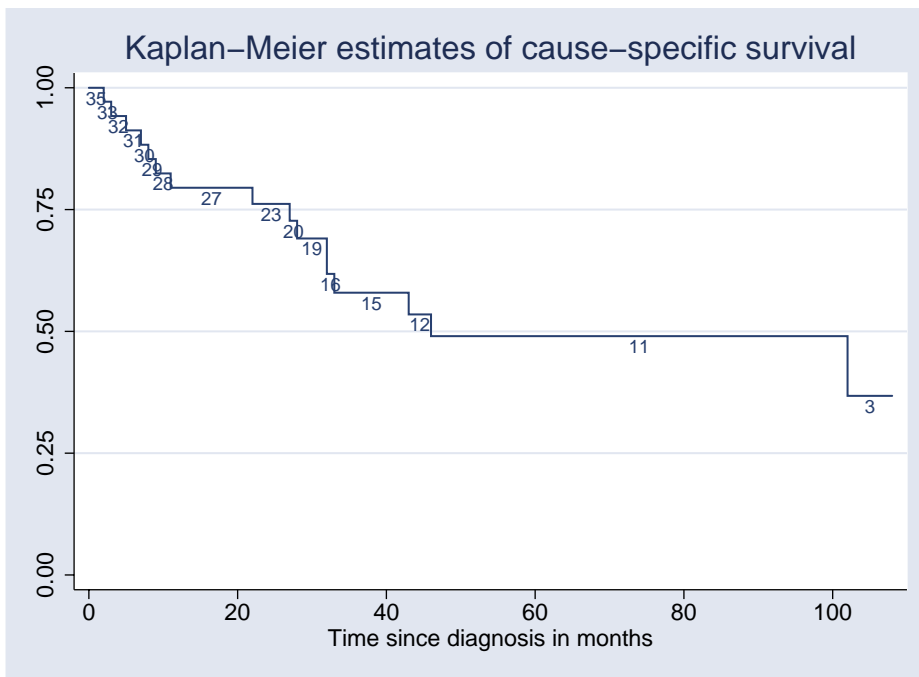


Figure 1: Kaplan-Meier plot of the cause-specific survivor function for sample of 35 patients diagnosed with colon carcinoma. The number at risk at each time point are shown on the curve.

```

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

. keep if stage == 1
   (2457 observations deleted)

. 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
460860.8 total analysis time at risk, at risk from t =      0
          earliest observed entry t =      0
          last observed exit t =      251

. sts graph, by(year8594)

```

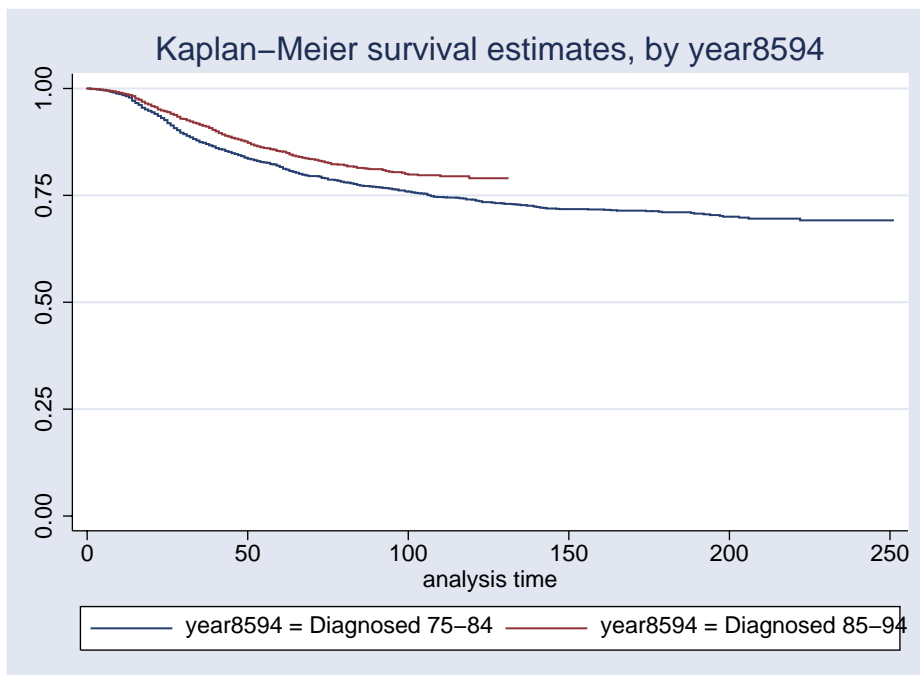


Figure 2: Skin melanoma. Kaplan-Meier plot of the cause-specific survivor function for each calendar period of diagnosis

- (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.

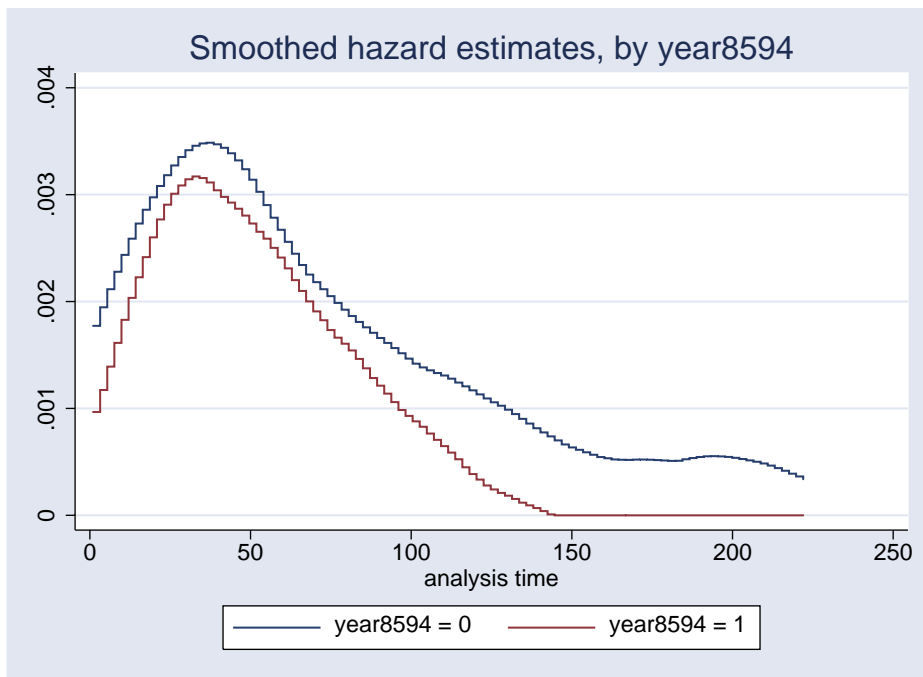


Figure 3: Skin melanoma. Plot of the cause-specific hazard for each calendar period of diagnosis

- (b) 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.

```
4. . sts test year8594
```

```
Log-rank test for equality of survivor functions
```

```
-----+-----
```

year8594	Events 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
```

```
-----+-----
```

year8594	Events observed	expected	Sum of 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.

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

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

```
-----+-----
```

	age	sex	stage	surv_mm	surv_yy	osr_fail
1.	81	Female	Localised	26	2	1
2.	75	Female	Localised	55	4	1
3.	78	Female	Localised	177	14	1
4.	75	Female	Unknown	29	2	1
5.	81	Female	Unknown	57	4	1

```
-----+-----
```

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
611349.3 total analysis time at risk, at risk from t = 0
          earliest observed entry t = 0
          last observed exit t = 251
```

```
. 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) xtitle(Time months)
```

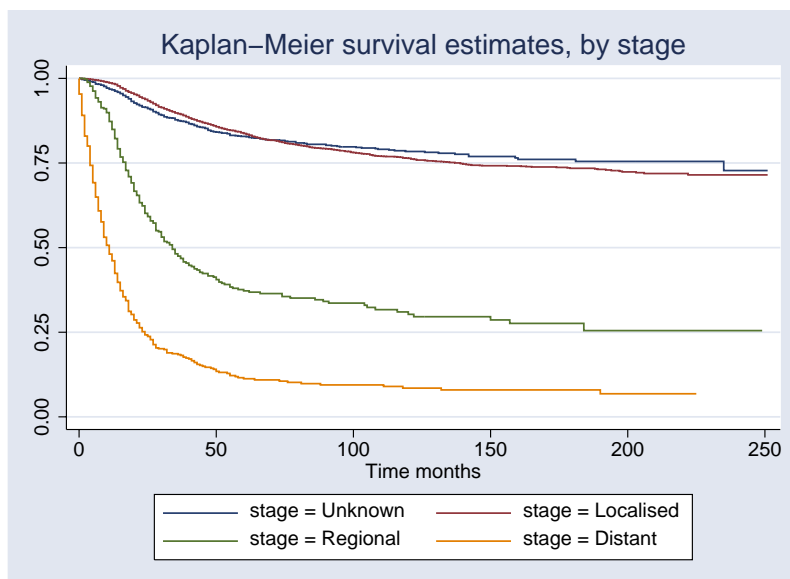


Figure 4: Skin melanoma. Kaplan-Meier estimates of cause-specific survival for each stage.

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

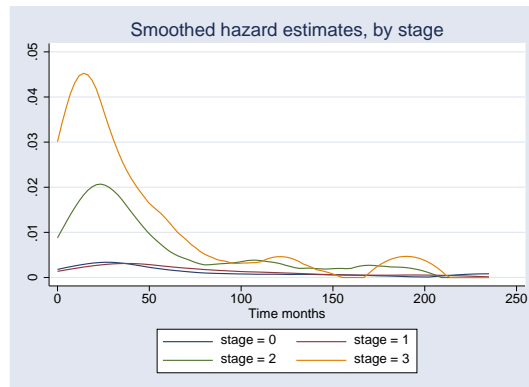


Figure 5: Skin melanoma. Estimates of the hazard (cause-specific mortality rate) for each stage.

```
(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	0.0022387	0.0019888	0.0025202
Localised	1013	4.6e+05	0.0021981	0.0020668	0.0023377
Regional	218	1.8e+04	0.0122280	0.0107079	0.0139638
Distant	408	1.0e+04	0.0397226	0.0360493	0.0437702

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% CI
(7775 records included in the analysis)

stage	D	Y	Rate	Lower	Upper
Unknown	274	1.0e+04	0.026865	0.023865	0.030242
Localised	1013	3.8e+04	0.026377	0.024801	0.028052
Regional	218	1.5e+03	0.146735	0.128494	0.167566
Distant	408	855.9350	0.476672	0.432592	0.525243

(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% CI
(7775 records included in the analysis)

stage	D	Y	Rate	Lower	Upper
Unknown	274	10.1992	26.865	23.865	30.242
Localised	1013	38.4050	26.377	24.801	28.052
Regional	218	1.4857	146.735	128.494	167.566
Distant	408	0.8559	476.672	432.592	525.243

(d) We see that the crude mortality rate is higher for males than females, a difference which is also reflected in the survival curves (Figure 6).

```
. strate sex, per(1000)
      failure _d: status == 1
      analysis time _t: surv_mm/12
```

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

sex	D	Y	Rate	Lower	Upper
Male	1074	21.8156	49.231	46.373	52.265
Female	839	29.1302	28.802	26.917	30.818

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

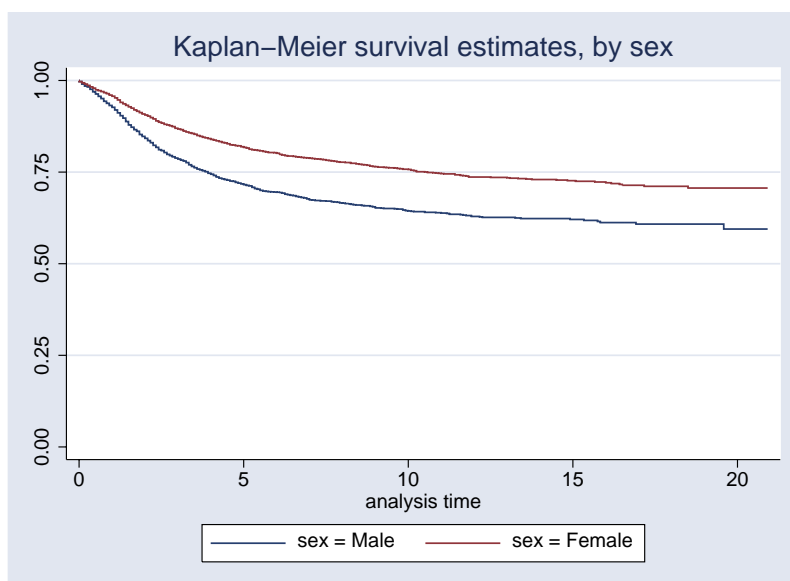


Figure 6: Skin melanoma (all stages). Kaplan-Meier estimates of cause-specific survival for each sex.

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

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

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals (337 records included in the analysis)

	hieng	D	Y	Rate	Lower	Upper
	low	28	2.0594	13.5960	9.3875	19.6912
	high	18	2.5442	7.0748	4.4574	11.2291

```
. display 7.0748/13.596
.52035893
```

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

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

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

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

- (c) A histogram (Figure 7) 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
```

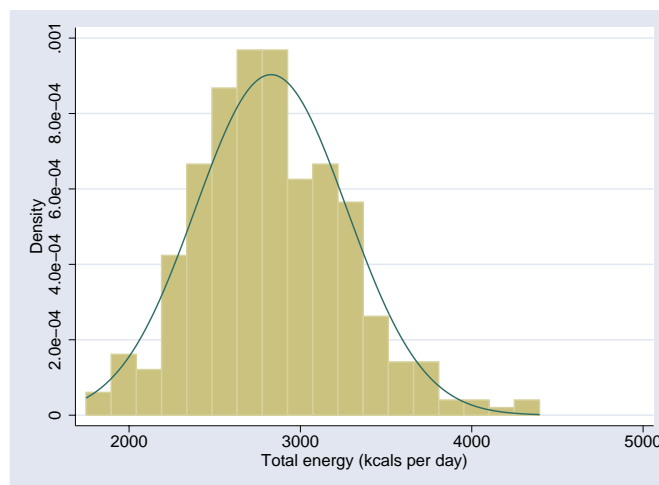


Figure 7: Histogram of energy with superimposed normal density curve (with the sample mean and variance).

```
. sum energy, detail
```

Total energy (kcal per day)				

	Percentiles	Smallest		
1%	1876.13	1748.43		
5%	2168.86	1854.02		
10%	2311.24	1858.8	Obs	337
25%	2536.69	1876.13	Sum of Wgt.	337
50%	2802.98		Mean	2828.872
		Largest	Std. Dev.	441.7528
75%	3109.66	4063.02		
90%	3366.61	4234.06	Variance	195145.5
95%	3595.05	4256.81	Skewness	.4430434
99%	4063.02	4395.75	Kurtosis	3.506768

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

```
. tabulate eng3
```

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

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

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

Estimated rates (per 1000) and lower/upper bounds of 95% Cis
(337 records included in the analysis)

+-----+					
eng3	D	Y	Rate	Lower	Upper

1500	16	0.9466	16.9020	10.3547	27.5892
2500	22	2.0173	10.9059	7.1810	16.5629
3000	8	1.6398	4.8787	2.4398	9.7555
+-----+					

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

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

```
(g) . set more off
     . list eng3 X1 X2 X3 if eng3==1500 in 1/100
```

```

+-----+
| eng3  X1  X2  X3 |
+-----+
1. | 1500  1   0   0 |
2. | 1500  1   0   0 |
3. | 1500  1   0   0 |
4. | 1500  1   0   0 |
5. | 1500  1   0   0 |
+-----+

```

```

     . list eng3 X1 X2 X3 if eng3==2500 in 1/100
```

```

+-----+
| eng3  X1  X2  X3 |
+-----+
76. | 2500  0   1   0 |
77. | 2500  0   1   0 |
78. | 2500  0   1   0 |
79. | 2500  0   1   0 |
80. | 2500  0   1   0 |
+-----+

```

```

     . list eng3 X1 X2 X3 if eng3==3000 in 200/300
```

```

+-----+
| eng3  X1  X2  X3 |
+-----+
226. | 3000  0   0   1 |
227. | 3000  0   0   1 |
228. | 3000  0   0   1 |
229. | 3000  0   0   1 |
230. | 3000  0   0   1 |
+-----+

```

```

     . set more on
```

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

```

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

```

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

```

```

-----+-----
chd |          IRR   Std. Err.      z    P>|z|   [95% Conf. Interval]
-----+-----
X2 |   .6452416   .2120034   -1.33  0.182   .3388815   1.228561
X3 |   .2886479   .1249882   -2.87  0.004   .1235342   .6744495
y | (exposure)
-----+-----

```

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

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

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

```
-----+-----
chd |          IRR   Std. Err.      z    P>|z|   [95% Conf. Interval]
-----+-----
X1 |   1.549807   .5092114    1.33   0.182   .8139601   2.950884
X3 |   .4473485   .1846929   -1.95   0.051   .1991671   1.004788
y | (exposure)
```

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

```
. xi: poisson chd i.eng3, e(y) irr
i.eng3          _Ieng3_1500-3000   (naturally coded; _Ieng3_1500 omitted)
```

```
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]
-----+-----
_Ieng3_2500 |   .6452416   .2120034   -1.33   0.182   .3388815   1.228561
_Ieng3_3000 |   .2886479   .1249882   -2.87   0.004   .1235342   .6744495
y | (exposure)
```

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

```
. summarize y chd
```

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

```
. display (337*.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 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
```

7. (a) `. stsplot fu, at(0(1)10) trim`
 (0 + 1452 obs. trimmed due to lower and upper bounds)
 (30206 observations (episodes) created)

(b) 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
      id: id
      note: fu>10 trimmed
```

Estimated rates (per 1000) and lower/upper bounds of 95% CIs
 (34072 records included in the analysis)

fu	D	Y	Rate	Lower	Upper
0	81	5.2507	15.4266	12.4077	19.1800
1	233	4.8317	48.2227	42.4119	54.8297
2	196	4.2112	46.5429	40.4626	53.5370
3	139	3.6915	37.6541	31.8870	44.4641
4	98	3.2484	30.1685	24.7497	36.7738
5	77	2.8489	27.0278	21.6176	33.7920
6	58	2.5113	23.0953	17.8548	29.8739
7	29	2.1766	13.3236	9.2589	19.1729
8	36	1.8735	19.2154	13.8606	26.6389
9	14	1.5722	8.9044	5.2737	15.0349

- (c) The pattern is similar. The plot of the mortality rates (Figure 8) could be considered an approximation to the ‘true’ functional form depicted in Figure 9. By estimating the rates for each year of follow-up we are essentially approximating the curve in Figure 9 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.

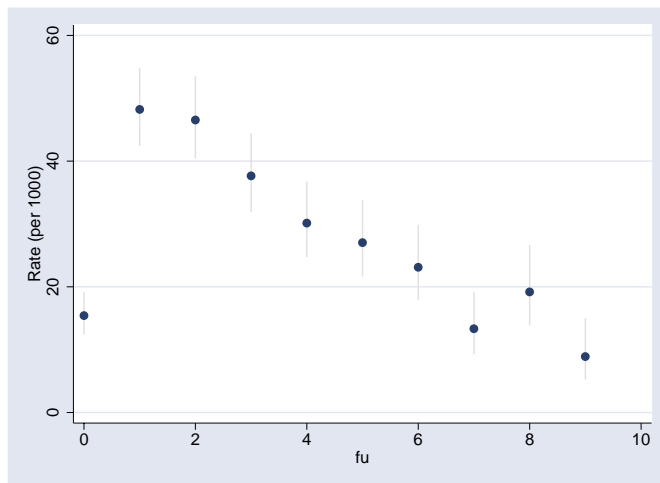


Figure 8: Localised melanoma. Disease-specific mortality rates as a function of time since diagnosis (annual intervals).

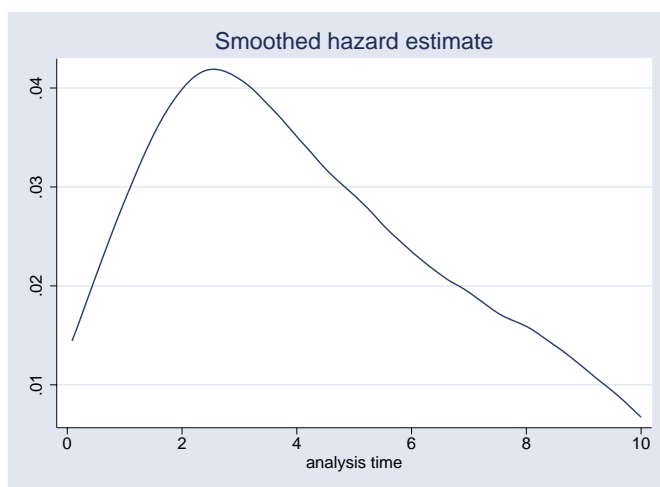


Figure 9: Localised melanoma. Disease-specific mortality rates as continuous function of time since diagnosis (using a smoother).

```
(d) . xi: streg i.fu, dist(exp)
      i.fu _Ifu_0-10      (naturally coded; _Ifu_0 omitted)
```

```
      failure _d: status == 1
      analysis time _t: surv_mm/12
      id: id
      note: fu>10 trimmed
```

```
note: _Ifu_10 dropped due to collinearity
```

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

```
No. of subjects =          5318          Number of obs   =          34072
No. of failures =           961
Time at risk    =  32216.08583
LR chi2(9)      =          201.23
Log likelihood  = -3283.9327          Prob > chi2      =          0.0000
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
_Ifu_1	3.125943	.4032046	8.84	0.000	2.427658	4.025081
_Ifu_2	3.017055	.3985223	8.36	0.000	2.328886	3.908574
_Ifu_3	2.440853	.3411953	6.38	0.000	1.855907	3.210162
_Ifu_4	1.955618	.2936669	4.47	0.000	1.45701	2.624855
_Ifu_5	1.752026	.2788568	3.52	0.000	1.282512	2.393426
_Ifu_6	1.497108	.257516	2.35	0.019	1.068659	2.097334
_Ifu_7	.8636789	.1868991	-0.68	0.498	.5651365	1.319931
_Ifu_8	1.2456	.2495041	1.10	0.273	.8411542	1.844511
_Ifu_9	.5772138	.1670674	-1.90	0.058	.3273158	1.017903

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 `_Ifu_1` is the rate ratio for the second year compared to the first year.


```
(e) . xi: streg i.fu i.agegrp year8594 sex, dist(exp)
i.fu          _Ifu_0-10      (naturally coded; _Ifu_0 omitted)
i.agegrp      _Iagegrp_0-3  (naturally coded; _Iagegrp_0 omitted)
```

```
No. of subjects =      5318          Number of obs   =      34072
No. of failures =      961
Time at risk    = 32216.08583

LR chi2(14)     =      412.72
Prob > chi2    =      0.0000
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
_Ifu_1	3.198105	.4125777	9.01	0.000	2.483601	4.118164
_Ifu_2	3.154676	.4170114	8.69	0.000	2.434647	4.08765
_Ifu_3	2.597297	.3636721	6.82	0.000	1.973953	3.417483
_Ifu_4	2.108206	.3174629	4.95	0.000	1.569407	2.831984
_Ifu_5	1.908097	.3049509	4.04	0.000	1.39496	2.60999
_Ifu_6	1.634282	.2825604	2.84	0.004	1.16455	2.293486
_Ifu_7	.9374623	.2038112	-0.30	0.766	.6122042	1.435527
_Ifu_8	1.33826	.2701805	1.44	0.149	.9009313	1.987875
_Ifu_9	.6112851	.1779332	-1.69	0.091	.345522	1.081464
_Iagegrp_1	1.320821	.1242054	2.96	0.003	1.0985	1.588137
_Iagegrp_2	1.850913	.1679404	6.79	0.000	1.549363	2.211152
_Iagegrp_3	3.370404	.3515811	11.65	0.000	2.747196	4.13499
year8594	.7189216	.0475506	-4.99	0.000	.6315121	.8184297
sex	.5892988	.038546	-8.08	0.000	.5183923	.6699041

There is no evidence that the effect of follow-up is confounded by age, sex, and period (the estimates for fu do not change a great deal when age, sex, and period are included in the model).

```
(f) . xi: streg i.fu i.agegrp year8594 i.sex i.sex*year8594, dist(exp)
i.fu          _Ifu_0-10      (naturally coded; _Ifu_0 omitted)
i.agegrp      _Iagegrp_0-3  (naturally coded; _Iagegrp_0 omitted)
i.sex         _Isex_1-2     (naturally coded; _Isex_1 omitted)
i.sex*year8594 _IsexYear8_# (coded as above)
```

```
Exponential regression -- log relative-hazard form
No. of subjects =      5318          Number of obs   =      34072
No. of failures =      961
Time at risk    = 32216.08583

LR chi2(15)     =      412.95
Prob > chi2    =      0.0000
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
.... [output omitted]						
_Iagegrp_1	1.31966	.1241227	2.95	0.003	1.097491	1.586804
_Iagegrp_2	1.849583	.1678451	6.78	0.000	1.548209	2.209624
_Iagegrp_3	3.369595	.35147	11.65	0.000	2.746578	4.133933
year8594	.7393188	.0653467	-3.42	0.001	.6217217	.8791592
_Isex_2	.6061075	.0533638	-5.69	0.000	.5100431	.7202652
_IsexYear~2	.9396257	.1226729	-0.48	0.633	.7274886	1.213622

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

```
8. . stcox year8594
```

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

```
Cox regression -- Breslow method for ties
```

```
No. of subjects =          5318          Number of obs =          5318
No. of failures =          1013
Time at risk    =         460860.75
Log likelihood   =        -8255.0613          LR chi2(1)    =          15.44
                                          Prob > chi2    =          0.0001
```

```
-----+-----
      _t |
      _d | Haz. Ratio   Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
year8594 |   .773003   .0507504   -3.922   0.000   .6796679   .8791553
-----+-----
```

(a) Patients diagnosed during 1985–94 experience only 77.3% of the cancer mortality experienced by those diagnosed 1975–84. That is, mortality due to skin melanoma has decreased by 22.7% in the latter period compared to the earlier period. This estimate is not adjusted for potential confounders. There is strong evidence of a statistically significant difference in survival between the two periods (based on the test statistic or the fact that the CI for the hazard ratio does not contain 1).

(b) The three test statistics are

log-rank 15.50

Wald $-3.922^2 = 15.38$

Likelihood ratio 15.44

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) . xi: stcox sex year8594 i.agegrp

```
Cox regression -- Breslow method for ties
```

```
No. of subjects =          5318          Number of obs =          5318
No. of failures =          1013
Time at risk    =         460860.75
Log likelihood   =        -8158.363          LR chi2(5)    =         208.83
                                          Prob > chi2    =          0.0000
```

```
-----+-----
      _t |
      _d | Haz. Ratio   Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
      sex |   .6060685   .0385507   -7.873   0.000   .5350306   .6865384
year8594 |   .7154353   .0472398   -5.071   0.000   .6285878   .8142818
Iagegr_1 |   1.297032   .1173508    2.875   0.004   1.086268    1.54869
Iagegr_2 |   1.830303   .1601515    6.908   0.000   1.541852   2.172716
Iagegr_3 |   3.282573   .3360783   11.610   0.000   2.685753   4.012015
-----+-----
```

- i. For patients of the same sex diagnosed in the same calendar period, those aged 60–74 at diagnosis have an estimated 83% higher risk of death due to skin melanoma than those aged 0–44 at diagnosis. The difference is statistically significant.

If this were an exam question the previous paragraph would be awarded full marks. It is worth noting, however, that the analysis is adjusted for the fact that mortality may depend on time since diagnosis (since this is the underlying time scale) and the mortality ratio between the two age groups is assumed to be the same at each point during the follow-up (i.e., proportional hazard).

- ii. No, there is no evidence of strong confounding, since the parameter estimate for period changes very little (from 0.77 to 0.71) when age and sex are added to the model.
- iii. Age (modelled as a categorical variable with 4 levels) is highly significant in the model.

```
. test _Iagegrp_1 _Iagegrp_2 _Iagegrp_3
      chi2( 3) = 153.65
      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) = 142.50
(Assumption: . nested in A) Prob > chi2 = 0.0000
```

- (e) i. Both models adjust for the same factors. When fitting the Poisson regression model we split time since diagnosis into annual intervals and explicitly estimated the effect of this factor in the model. The Cox model does not estimate the effect of ‘time’ but the other estimates are adjusted for ‘time’.
- ii. Since the two models are conceptually similar we would expect the parameter estimates to be similar, which they are.

	_t	Haz. Ratio	Std. Err.	z	P> z	[95% CI]
Cox regression						
sex		.6060685	.0385507	-7.87	0.000	.5350306 .6865384
year8594		.7154353	.0472398	-5.07	0.000	.6285878 .8142818
_Iagegrp_1		1.297032	.1173508	2.87	0.004	1.086268 1.54869
_Iagegrp_2		1.830303	.1601515	6.91	0.000	1.541852 2.172716
_Iagegrp_3		3.282573	.3360783	11.61	0.000	2.685753 4.012015
Poisson regression						
sex		.5892988	.038546	-8.08	0.000	.5183923 .6699041
year8594		.7189216	.0475506	-4.99	0.000	.6315121 .8184297
_Iagegrp_1		1.320821	.1242054	2.96	0.003	1.0985 1.588137
_Iagegrp_2		1.850913	.1679404	6.79	0.000	1.549363 2.211152
_Iagegrp_3		3.370404	.3515811	11.65	0.000	2.747196 4.13499

- 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.

```
9. . stpplot, by(year8594)
```

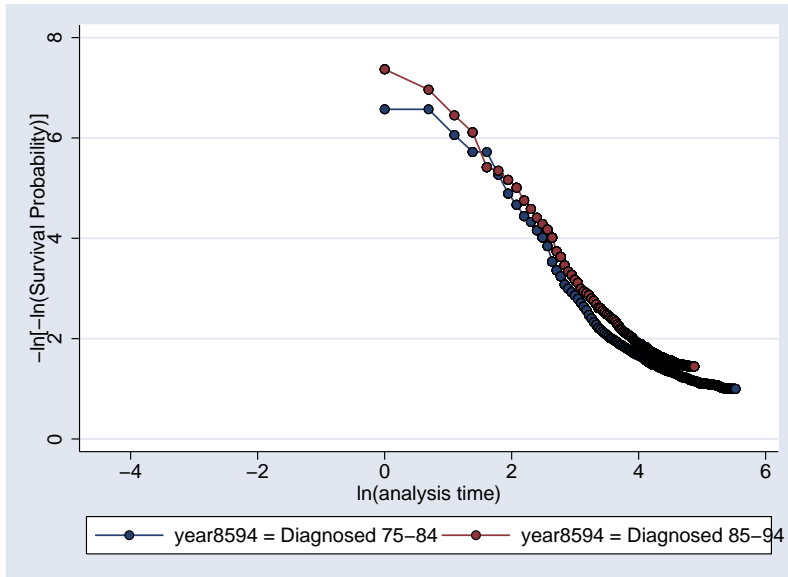


Figure 10: Localised skin melanoma. Plot of the log cumulative hazard function for each calendar period of diagnosis. Each plot symbol represents an event time.

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

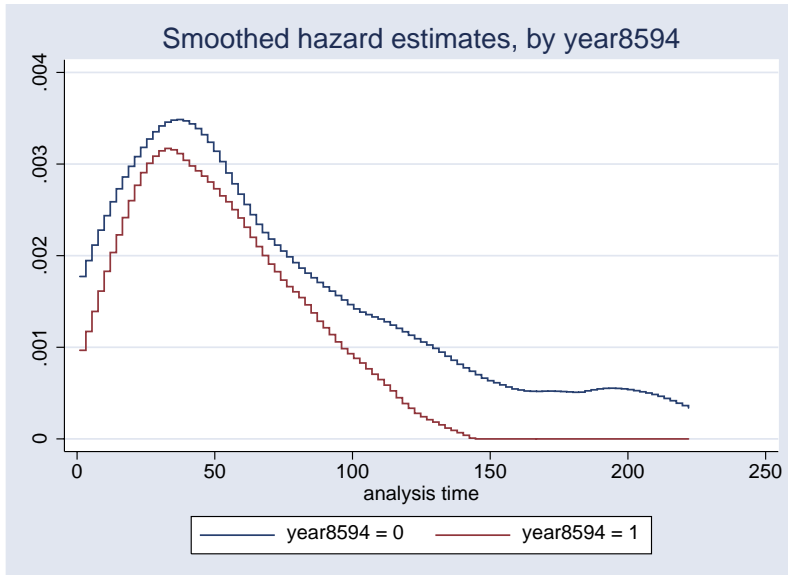


Figure 11: Localised skin melanoma. Plot of the estimated hazard function for each calendar period of diagnosis.

- (a) We know that patients diagnosed during 1984–95 have superior survival. Consequently, they have a lower cumulative hazard, a lower log cumulative hazard, and therefore a higher negative log cumulative hazard. The difference between the two curves is similar over time, except for one point where the curves cross. As such, there is no reason to reject an assumption of proportional hazards.

We should not pay too much attention to the curves for values up to 2 on the x axis. Noting that $\exp(2) = 7.4$ we see that the curves between 0 and 2 on the x axis are based on only 7 data points. We should give more weight to differences where we have more data.

- (b) The log rank test assumes that any difference in survival between the groups takes the form of proportional hazards. As such, the log rank test may fail to detect a difference in survival if the difference does not take the form of proportional hazards.
- (c) A rough estimate of the difference between the curves is 0.2, which is an estimate of the log hazard ratio. An estimate of the hazard ratio is therefore $\exp(-0.2) = 0.82$.
- (d) The estimated hazard ratio from the Cox model is 0.77 which is similar (as it should be) to the estimate made by looking at the difference in the plots of the log cumulative hazard.
- (e) It seems that there is evidence of non-proportional hazards, particularly for age and sex.

```
. stphtest, detail
```

```
Test of proportional hazards assumption
```

	rho	chi2	df	Prob>chi2
sex	0.07535	5.65	1	0.0175
year8594	0.03335	1.12	1	0.2896
_Iagegrp_1	-0.05293	2.84	1	0.0919
_Iagegrp_2	-0.07392	5.49	1	0.0192
_Iagegrp_3	-0.11356	12.40	1	0.0004
global test		17.94	5	0.0030

- (f) The differences in survival by age are most apparent early in the follow-up.

```
. stphplot, by(agegrp)
```

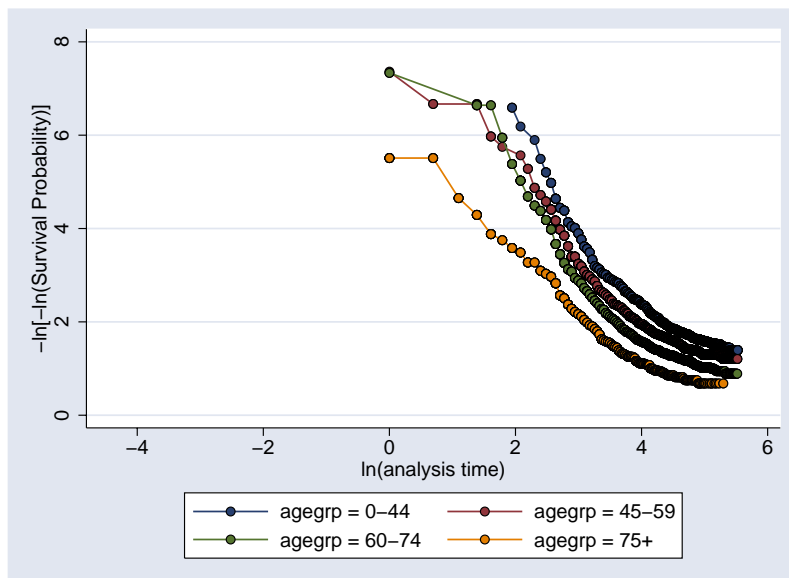


Figure 12: Localised skin melanoma. Plot of the log cumulative hazard function for each age group. Each plot symbol represents an event time.

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

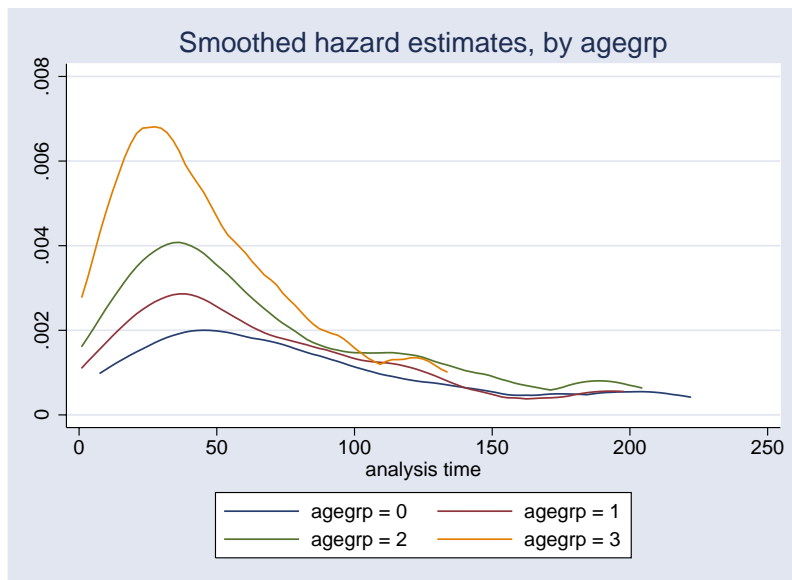


Figure 13: Localised skin melanoma. Plot of the estimated hazard function for each age group.

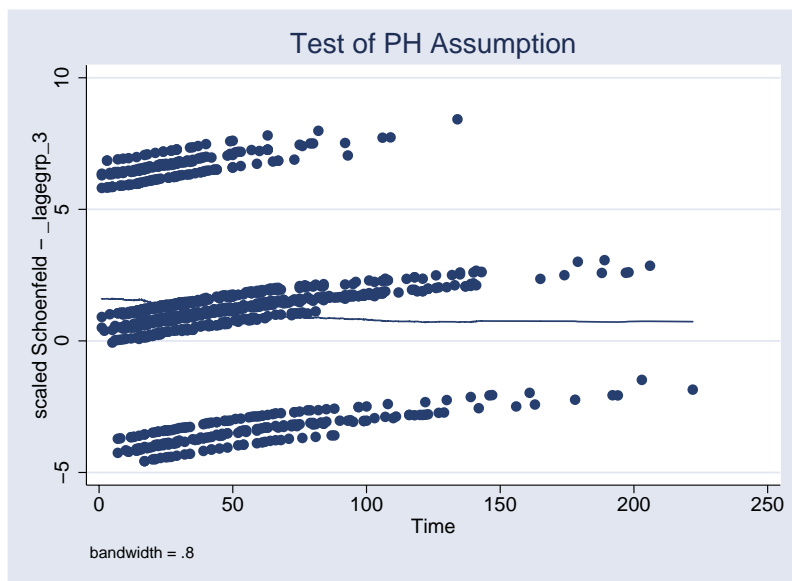


Figure 14: Localised skin melanoma. Plot of the scaled Schoenfeld residuals for age3.

If the proportional hazards assumption is appropriate then we should see parallel lines in Figure 12. This doesn't look too bad, apart from the line for the oldest age group being somewhat lower than the others in the early period. Note that these curves are not based on the estimated Cox model (i.e., they are unadjusted).

It's difficult to assess the PH hypothesis from Figure 13 although this figure does give us a good idea of the shape of the underlying hazards. Do we really expect hazards to be proportional over the entire 200 month follow-up period when the magnitude of mortality varies greatly over this interval? Note that these curves are not based on the estimated Cox model (i.e., they are unadjusted).

We saw evidence of non-proportional hazards by age, particularly in the eldest age group. The smooth line in Figure 14 shows the estimated hazard ratio as a function of time. We see that the estimated hazard ratio is highest immediately following diagnosis, decreases over the first 100 months and is then relatively constant.

```
(g) . stcox sex year8594 _Iagegrp_1 _Iagegrp_2 _Iagegrp_3,
      tvc( _Iagegrp_1 _Iagegrp_2 _Iagegrp_3) texp(_t>=24)
```

		_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
rh							
	sex		.6081152	.0386691	-7.82	0.000	.5368579 .6888305
	year8594		.7137211	.0471632	-5.10	0.000	.6270189 .8124123
	_Iagegrp_1		1.702949	.3343537	2.71	0.007	1.158987 2.502217
	_Iagegrp_2		2.461019	.4612078	4.81	0.000	1.704494 3.553322
	_Iagegrp_3		5.390195	1.033536	8.79	0.000	3.701621 7.849048
t							
	_Iagegrp_1		.7081921	.1567346	-1.56	0.119	.4589508 1.092788
	_Iagegrp_2		.6847885	.145389	-1.78	0.075	.4516853 1.03819
	_Iagegrp_3		.4801757	.1107029	-3.18	0.001	.3056036 .7544699

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 $5.39 \times 0.48 = 2.59$. Note that simply cutting the follow-up at 2 years probably does adequately capture the manner in which the effect of age varies as a function of time since diagnosis.

- (h) The approach is to split the data at 24 months since diagnosis and fit a model with an interaction between follow-up and age. We first need to create an ID variable.

```

. use melanoma, clear
. keep if stage == 1
. gen id=_n
. stset surv_mm, failure(status==1) id(id)
. list id surv_mm _st _d _t0 _t if id < 5

```

	id	surv_mm	_st	_d	_t0	_t
1.	1	26	1	0	0	26
2.	2	55	1	0	0	55
3.	3	177	1	0	0	177
4.	4	19	1	1	0	19

```

. stsplit fu, at(24)
. list id surv_mm _st _d _t0 _t fu if id < 5, sepby(id)

```

	id	surv_mm	_st	_d	_t0	_t	fu
1.	1	24	1	0	0	24	0
2.	1	26	1	0	24	26	24
3.	2	24	1	0	0	24	0
4.	2	55	1	0	24	55	24
5.	3	24	1	0	0	24	0
6.	3	177	1	0	24	177	24
7.	4	19	1	1	0	19	0

```

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

```

No. of subjects =	5318	Number of obs =	9809
No. of failures =	1013		
Time at risk =	460860.03	LR chi2(9) =	246.18
Log likelihood =	-3490.6961	Prob > chi2 =	0.0000

```

-----+-----
      _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      sex | .5931472   .0377298   -8.21   0.000   .5236221   .6719037
  year8594 | .9323015   .0612735   -1.07   0.286   .8196208   1.060473
  _Iagegrp_1 | 1.594928   .304503    2.45   0.014   1.097059   2.318741
  _Iagegrp_2 | 2.35879    .427177    4.74   0.000   1.654005   3.363891
  _Iagegrp_3 | 4.859066   .9021761    8.51   0.000   3.376847   6.991884
  _Ifu_24 | 1.133585   .1938742    0.73   0.463   .8107283   1.585014
  _IageXfu_1~4 | .7914908   .1716243   -1.08   0.281   .5174568   1.210648
  _IageXfu_2~4 | .7886481   .1633875   -1.15   0.252   .525456   1.183669
  _IageXfu_3~4 | .6953743   .1572626   -1.61   0.108   .4463905   1.083234
-----+-----

```

We have estimated an additional parameter compared to the Cox model (the hazard ratio between the two calendar periods `_Ifu_24`). The parameters `_Iagegrp_1`–`_Iagegrp_3` give the effect of age during the first 24 months of follow-up and are similar to the estimates from the Cox model. They are not identical since Poisson regression makes the assumption that the hazards are constant within this interval whereas the Cox model does not. There are greater differences in the estimates of the interaction effects since the assumption that the hazards are constant during the second interval is less sound (than the assumption that the hazards are constant during the first interval).


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

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

```
-----
5318 total obs.
0 exclusions
-----
```

```
5318 obs. remaining, representing
1795 failures in single record/single failure data
460860.8 total analysis time at risk, at risk from t = 0
earliest observed entry t = 0
last observed exit t = 251
```

```
. xi: stcox sex year8594 i.agegrp
i.agegrp Iagegr_0-3 (naturally coded; Iagegr_0 omitted)
```

```
failure _d: status == 1 2
analysis time _t: surv_mm
```

Cox regression -- Breslow method for ties

```
No. of subjects = 5318 Number of obs = 5318
No. of failures = 1795
Time at risk = 460860.75
Log likelihood = -13860.546 LR chi2(5) = 1052.43
Prob > chi2 = 0.0000
```

```
-----
_t |
_d | Haz. Ratio Std. Err. z P>|z| [95% Conf. Interval]
-----+-----
sex | .6251256 .0298843 -9.827 0.000 .5692137 .6865295
year8594 | .7489942 .0387021 -5.593 0.000 .6768541 .8288231
Iagegr_1 | 1.603024 .1304825 5.797 0.000 1.36664 1.880294
Iagegr_2 | 3.400802 .2544105 16.362 0.000 2.936999 3.937847
Iagegr_3 | 9.515058 .7574396 28.301 0.000 8.140519 11.12169
-----
```

- (a) For patients of the same sex diagnosed in the same period, those aged 60–74 at diagnosis have a 3.4 times higher risk of death *due to any causes* than those aged 0–44 at diagnosis. This difference is statistically significant.
- (b) Note that the previous model estimated cause-specific hazard ratios whereas the current model estimates all-cause hazard ratios. The estimated hazard ratios for sex and period are similar, whereas the estimated hazard ratios for age are markedly different. This is because non-cancer mortality is heavily dependent on age, but only lightly dependent on sex and calendar period.

11. (a) . stcox sex

Cox regression -- Breslow method for ties

```
No. of subjects =          7775          Number of obs   =          7775
No. of failures =          1913
Time at risk    =        611349.29
Log likelihood  =   -16342.555          LR chi2(1)       =          103.25
                                          Prob > chi2     =          0.0000
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
sex	.6273066	.0289338	-10.11	0.000	.573085	.6866581

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

(b) . xi: stcox sex year8594 i.agegrp i.subsite i.stage
i.agegrp _Iagegrp_0-3 (naturally coded; _Iagegrp_0 omitted)
i.subsite _Isubsite_1-4 (naturally coded; _Isubsite_1 omitted)
i.stage _Istage_0-3 (naturally coded; _Istage_0 omitted)

Cox regression -- Breslow method for ties

```
No. of subjects =          7775          Number of obs   =          7775
No. of failures =          1913
Time at risk    =        611349.29
Log likelihood  =   -15476.269          LR chi2(11)     =          1835.82
                                          Prob > chi2     =          0.0000
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
sex	.7490676	.036445	-5.94	0.000	.6809368	.8240153
year8594	.7867739	.0376881	-5.01	0.000	.7162681	.8642199
_Iagegrp_1	1.268542	.0855596	3.53	0.000	1.111459	1.447824
_Iagegrp_2	1.730767	.1126805	8.43	0.000	1.523427	1.966326
_Iagegrp_3	2.785848	.2128337	13.41	0.000	2.398431	3.235845
_Isubsite_2	1.393153	.0984179	4.69	0.000	1.213016	1.600041
_Isubsite_3	1.032021	.0767263	0.42	0.672	.8920829	1.19391
_Isubsite_4	1.305318	.133562	2.60	0.009	1.06812	1.59519
_Istage_1	1.038328	.0713262	0.55	0.584	.9075334	1.187972
_Istage_2	4.771515	.4363494	17.09	0.000	3.988549	5.70818
_Istage_3	13.48664	1.097917	31.96	0.000	11.49766	15.8197

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.

```
. gen fem0=(sex==2)*(agegrp==0)
. gen fem1=(sex==2)*(agegrp==1)
. gen fem2=(sex==2)*(agegrp==2)
. gen fem3=(sex==2)*(agegrp==3)

. xi: stcox i.agegrp fem0 fem1 fem2 fem3
No. of subjects =          7775          Number of obs =          7775
No. of failures =          1913
Time at risk    =    611349.29          LR chi2(7)      =    331.08
Log likelihood   =   -16228.639          Prob > chi2    =    0.0000
-----+-----
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
_Iagegrp_1	1.197101	.1017692	2.12	0.034	1.013369	1.414145
_Iagegrp_2	1.497299	.1267028	4.77	0.000	1.268466	1.767412
_Iagegrp_3	2.322161	.2401309	8.15	0.000	1.896142	2.843895
fem0	.4578165	.0478157	-7.48	0.000	.3730692	.5618151
fem1	.5526258	.0504729	-6.49	0.000	.4620494	.660958
fem2	.7132982	.0565997	-4.26	0.000	.6105607	.833323
fem3	.6750958	.0713516	-3.72	0.000	.5487834	.8304813

```
-----+-----
. test fem0=fem1=fem2=fem3
      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.

```
. xi: stcox i.agegrp year8594 i.subsite i.stage fem0 fem1 fem2 fem3
No. of subjects =          7775          Number of obs =          7775
No. of failures =          1913
Time at risk    =    611349.29          LR chi2(14)     =    1840.42
Log likelihood   =   -15473.971          Prob > chi2     =    0.0000
-----+-----
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
_Iagegrp_1	1.188947	.1014449	2.03	0.043	1.005855	1.405367
_Iagegrp_2	1.5508	.1318113	5.16	0.000	1.312827	1.831911
_Iagegrp_3	2.485421	.2605605	8.68	0.000	2.023782	3.052363
year8594	.7868595	.0376845	-5.01	0.000	.7163599	.8642973
_Isubsite_2	1.401988	.0992064	4.78	0.000	1.220428	1.610558
_Isubsite_3	1.039415	.0773326	0.52	0.603	.8983792	1.202593
_Isubsite_4	1.315538	.1349198	2.67	0.007	1.075983	1.608428
_Istage_1	1.036942	.0712433	0.53	0.598	.9063011	1.186414
_Istage_2	4.702828	.4312718	16.88	0.000	3.929161	5.628833
_Istage_3	13.38869	1.091144	31.83	0.000	11.41215	15.70757
fem0	.6251314	.0662091	-4.44	0.000	.5079472	.7693502
fem1	.7300673	.0678894	-3.38	0.001	.608428	.8760252
fem2	.8120201	.0653462	-2.59	0.010	.6935337	.9507494
fem3	.8068979	.086154	-2.01	0.044	.654537	.9947249

```
-----+-----
. test fem0=fem1=fem2=fem3
      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)

12. (a)

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

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

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

```
. /* Cox model with time in study as the scale */
. stset dox, id(id) fail(chd) origin(doe) scale(365.25)
. stcox hieng
```

```
Cox regression -- no ties
```

```
No. of subjects =          337          Number of obs   =          337
No. of failures =           46
Time at risk    =   4603.66872
Log likelihood  =  -253.32253          LR chi2(1)       =          4.73
                             Prob > chi2        =          0.0296
```

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

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

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

```
. /* Cox model with attained age as the scale */
. stset dox, id(id) fail(chd) origin(dob) entry(doe) scale(365.25)
. stcox hieng
```

```
Cox regression -- Breslow method for ties
```

```
No. of subjects =          337          Number of obs   =          337
No. of failures =           46
Time at risk    =   4603.66872
Log likelihood  =  -234.78217          LR chi2(1)       =          4.20
                             Prob > chi2        =          0.0405
```

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

13. (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(365.25) 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)/365.25
          origin: time dob
```

```
-----
399 total obs.
  0 exclusions
-----
```

```
399 obs. remaining, representing
399 subjects
278 failures in single failure-per-subject data
2435.641 total analysis time at risk, at risk from t = 0
          earliest observed entry t = 75.13758
          last observed exit t = 96.50376
```

. 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  135.026  116.721  156.202 |
|  2   97  1.0952   88.572   72.589  108.074 |
-----+-----
```

- i. The timescale is attained age, which would seem to be a reasonable choice.
- ii. Males have the higher mortality which is to be expected.
- iii. Age could potentially be a confounder.

```
. tabstat _t0, by(sex)
```

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

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

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

```
. streg sex, dist(exp) nolog
```

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

```
No. of subjects =          399          Number of obs = 399
```

```
No. of failures =          278
```

```
Time at risk    = 2435.641342
```

```
LR chi2(1)      = 11.64
```

```
Log likelihood  = 355.79411
```

```
Prob > chi2    = 0.0006
```

```
-----+-----
 _t | Haz. Ratio  Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
sex | .6559621    .0825422    -3.35  0.001    .5125885    .839438
-----+-----
```

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

```
(d) . streg brv, 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.

```
. gen brv_m=brv*(sex==1)
. gen brv_f=brv*(sex==2)
. streg sex brv_m brv_f, nolog
```

Exponential regression -- log relative-hazard form

```
No. of subjects =          399          Number of obs   =          555
No. of failures =          278
Time at risk    = 2435.641342
Log likelihood  =   358.60684          LR chi2(3)      =          17.26
                                          Prob > chi2    =          0.0006
```

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

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

```
. strate age
Estimated rates and lower/upper bounds of 95% confidence intervals
(1036 records included in the analysis)
```

-----+-----	age	D	Y	Rate	Lower	Upper
-----+-----	75	45	704.1123	0.063910	0.047718	0.085597
	80	123	1.2e+03	0.103831	0.087012	0.123902
	85	95	489.6099	0.194032	0.158687	0.237249
	90	12	55.0205	0.218100	0.123861	0.384041
	95	3	2.2868	1.311883	0.423110	4.067583
-----+-----						


```

. /* Poisson regression: effect of bereavement
                                controlled for attained age */
. xi: streg brv i.age, nolog

No. of subjects =          399          Number of obs =          1036
No. of failures =          278
Time at risk    = 2435.641342

Log likelihood = 378.36458          LR chi2(5) =          56.78
                                Prob > chi2 =          0.0000

```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
brv	.8591401	.1178292	-1.11	0.268	.6566342 1.124099
_Iage_80	1.667713	.2929561	2.91	0.004	1.181942 2.353132
_Iage_85	3.203792	.5989022	6.23	0.000	2.22098 4.621511
_Iage_90	3.621248	1.191405	3.91	0.000	1.900245 6.90092
_Iage_95	21.10446	12.59434	5.11	0.000	6.552538 67.97337

```

. /* Poisson regression: effect of bereavement
                                controlled for attained age and sex */
. xi: streg sex brv i.age, nolog

Log likelihood = 385.75207          LR chi2(6) =          71.55
                                Prob > chi2 =          0.0000

```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
sex	.6113709	.0798235	-3.77	0.000	.4733342 .7896627
brv	.9733576	.1364632	-0.19	0.847	.7394951 1.281178
_Iage_80	1.677381	.2946826	2.94	0.003	1.188756 2.366852
_Iage_85	3.177095	.5918009	6.21	0.000	2.205342 4.577037
_Iage_90	3.665054	1.205863	3.95	0.000	1.923184 6.984571
_Iage_95	28.02706	16.88249	5.53	0.000	8.606854 91.26632

```

(g) . /* Poisson regression: effect of bereavement for each
                                gender (controlled for attained age) */
. xi: streg sex brv_m brv_f i.age, nolog

```

```

                                LR chi2(7) =          73.39
Log likelihood = 386.66981          Prob > chi2 =          0.0000

```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
sex	.5367391	.0889089	-3.76	0.000	.3879402 .7426116
brv_m	.8235299	.1585253	-1.01	0.313	.5647126 1.200968
brv_f	1.199552	.2500957	0.87	0.383	.7971703 1.80504
_Iage_80	1.679325	.2950651	2.95	0.003	1.190076 2.369708
_Iage_85	3.135021	.5851488	6.12	0.000	2.174525 4.519771
_Iage_90	3.663357	1.205622	3.95	0.000	1.921968 6.982523
_Iage_95	28.97621	17.47942	5.58	0.000	8.883173 94.51808

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

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

Cox regression -- Breslow method for ties

```
No. of subjects =          399      Number of obs   =          1036
No. of failures =           278
Time at risk    = 2435.641342
Log likelihood  = -1379.1483      LR chi2(1)     =           2.25
                                      Prob > chi2    =           0.1333
```

```
-----+-----
   _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
brv |   .8134514   .1131032    -1.48  0.138     .6194119    1.068276
-----+-----
```

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

Cox regression -- Breslow method for ties

```
No. of subjects =          399      Number of obs   =          1036
No. of failures =           278
Time at risk    = 2435.641342
Log likelihood  = -1372.3656      LR chi2(2)     =          15.82
                                      Prob > chi2    =           0.0004
```

```
-----+-----
   _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
brv |   .9249887   .1317637    -0.55  0.584     .6996545    1.222895
sex |   .6233905   .0815085    -3.61  0.000     .4824643    .8054806
-----+-----
```

```
(j) . /* Cox: effect of brv for each gender (controlled for attained age) */
    . stcox sex brv_m brv_f, nolog
```

Cox regression -- Breslow method for ties

```
No. of subjects =          399      Number of obs   =          1036
No. of failures =           278
Time at risk    = 2435.641342
Log likelihood  = -1371.7342      LR chi2(3)     =          17.08
                                      Prob > chi2    =           0.0007
```

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

Splitting on two time scales and calculating SMRs

```
1. . use diet, clear
   (Diet data with dates)

   . stset dox, fail(chd) origin(dob) entry(doe) scale(365.25) id(id)

           id: id
   failure event: chd ~= 0 & chd ~= .
obs. time interval: (dox[_n-1], dox]
   enter on or after: time doe
   exit on or before: failure
   t for analysis: (time-origin)/365.25
   origin: time dob

-----
   337 total obs.
     0 exclusions
-----

   337 obs. remaining, representing
   337 subjects
   46 failures in single failure-per-subject data
4603.669 total analysis time at risk, at risk from t = 0
           earliest observed entry t = 30.07529
           last observed exit t = 69.99863

   . stsplot ageband, at(30(5)70) trim
   (no obs. trimmed because none out of range)
   (864 observations (episodes) created)
```

```
2. . stsplot period, after(time=d(1/1/1900)) at(50(5)80) trim
(no obs. trimmed because none out of range)
(933 observations (episodes) created)
```

```
. tab period
```

period	Freq.	Percent	Cum.
55	201	9.42	9.42
60	538	25.21	34.63
65	605	28.35	62.98
70	505	23.66	86.64
75	285	13.36	100.00
Total	2134	100.00	

```
.
. replace period=period+1900
period was byte now int
(2134 real changes made)
```

```
. tab period
```

period	Freq.	Percent	Cum.
1955	201	9.42	9.42
1960	538	25.21	34.63
1965	605	28.35	62.98
1970	505	23.66	86.64
1975	285	13.36	100.00
Total	2134	100.00	

```
. list id ageband period in 1/15
```

	id	ageband	period
1.	1	45	1960
2.	1	45	1965
3.	1	50	1965
4.	1	50	1970
5.	1	55	1970
6.	1	55	1975
7.	1	60	1975
8.	2	50	1960
9.	2	50	1965
10.	2	55	1965
11.	2	55	1970
12.	2	60	1970
13.	2	60	1975
14.	3	55	1965
15.	3	60	1965

```
3. . generate _y=_t-_t0 if _st==1
```

```
. table ageband period, c(sum _d)
```

```
-----
```

ageband	period				
	1955	1960	1965	1970	1975
30		0	0		
35	0	0	0	0	
40	0	0	0	1	0
45	1	3	1	0	0
50	1	4	2	1	1
55	0	2	4	2	1
60	3	1	3	5	2
65	0	0	3	3	2

```
-----
```

```
. table ageband period, c(sum _y) format(%5.1f)
```

```
-----
```

ageband	period				
	1955	1960	1965	1970	1975
30		19.3	1.3		
35	1.1	39.3	34.0	1.3	
40	27.8	130.3	54.5	36.2	1.3
45	82.2	324.8	181.1	53.3	15.6
50	80.9	361.1	374.2	180.9	27.3
55	39.1	240.2	385.2	338.7	79.5
60	7.7	96.4	340.7	364.9	148.1
65	3.4	24.6	90.7	303.0	113.8

```
-----
```

```
4. . generate obsrate=_d/_y*1000
```

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

```
-----
```

ageband	period				
	1955	1960	1965	1970	1975
30		0.0	0.0		
35	0.0	0.0	0.0	0.0	
40	0.0	0.0	0.0	27.6	0.0
45	12.2	9.2	5.5	0.0	0.0
50	12.4	11.1	5.3	5.5	36.7
55	0.0	8.3	10.4	5.9	12.6
60	387.7	10.4	8.8	13.7	13.5
65	0.0	0.0	33.1	9.9	17.6

```
-----
```

```
5. . sort ageband period
```

```
. merge ageband period using ref  
ageband was byte now int
```

```
. tab _merge
```

_merge	Freq.	Percent	Cum.
2	12	0.56	0.56
3	2134	99.44	100.00
Total	2146	100.00	

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

```
6. . tab refrate
```

refrate	Freq.	Percent	Cum.
11	2134	100.00	100.00
Total	2134	100.00	

```
. generate e=_y*refrate/1000
```

```
. list id e _d in 1/10
```

	id	e	_d
1.	82	.006859	0
2.	83	.0463265	0
3.	94	.0316147	0
4.	90	.0394449	0
5.	72	.0453929	0
6.	75	.0425168	0
7.	83	.0078453	0
8.	72	.005948	0
9.	153	.0014155	0
10.	152	.0014155	0

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

```
failure _d: chd  
analysis time _t: (dox-origin)/365.25  
origin: time dob  
enter on or after: time doe  
id: id
```

Estimated SMRs and lower/upper bounds of 95% confidence intervals
(2134 records included in the analysis)

_D	_E	_SMR	_Lower	_Upper
46	50.64	0.9084	0.6804	1.2127

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

```
      failure _d:  chd  
analysis time _t: (dox-origin)/365.25  
      origin:  time dob  
enter on or after:  time doe  
      id:  id
```

Estimated SMRs and lower/upper bounds of 95% confidence intervals
(2134 records included in the analysis)

hieng	_D	_E	_SMR	_Lower	_Upper
low	28	22.65	1.2360	0.8534	1.7901
high	18	27.99	0.6432	0.4052	1.0208