BIOSTAT III: Survival analysis for epidemiologists

Examination

November 14, 2008

Code:

- Time allowed is 2 hours.
- Please try and write your answers on the exam sheet. You may use separate paper if absolutely necessary.
- The exam contains 2 questions, each with several parts. The marks available for each part are indicated.
- A score of 18 marks or more out of a possible 30 will be required to obtain a passing grade.
- The questions may be answered in English or Swedish (or a combination thereof).
- A non-programmable scientific calculator (i.e., with ln() and exp() functions) will most probably be useful.
- The exam is not 'open book' but each student will be allowed to bring one A4 sheet of paper into the exam room which may contain, for example, hand-written notes or photocopies from textbooks/lecture notes etc. Both sides of the page may be used.
- The exam supervisors have been advised not to answer any questions you may have regarding the content of the exam. If you believe a question contains an error or is ambiguous then please write a note with your answer indicating how you have interpreted the question.
- Tables of critical values of the χ^2 distribution are provided on the last page.

1. In this question we will study survival of 5554 patients diagnosed with thyroid cancer in Sweden during the period 1958-1987. Our analysis is restricted to two histological types, papillary and follicular, which we will collectively call differentiated thyroid cancer (DTC). Our aim is to study how mortality due to DTC depends on age at diagnosis, calendar period of diagnosis, sex, and histology (papillary or follicular). We commence by studying the coding of relevant variables.

```
. codebook  sex dead_dtc papillary period agegrp

-----------------------------------------------------------------------
sex                                                                 Sex
-----------------------------------------------------------------------
            tabulation:  Freq.   Numeric  Label
                          1377         1  male
                          4177         2  female

-----------------------------------------------------------------------
dead_dtc                                  Indicator for death due to DTC
-----------------------------------------------------------------------
            tabulation:  Freq.   Numeric  Label
                          4528         0  Censored
                          1026         1  Dead due to DTC

-----------------------------------------------------------------------
papillary                   Histology papillary (otherwise follicular)
-----------------------------------------------------------------------
            tabulation:  Freq.   Numeric  Label
                          1966         0  Follicular
                          3588         1  Papillary

-----------------------------------------------------------------------
period                                                  Calendar period
-----------------------------------------------------------------------
            tabulation:  Freq.   Numeric  Label
                          1280         1  1958-67
                          1997         2  1968-77
                          2277         3  1978-87

-----------------------------------------------------------------------
agegrp                                           Age at diagnosis group
-----------------------------------------------------------------------
            tabulation:  Freq.   Numeric  Label
                          1419         0  0-39
                           960        40  40-49
                          1044        50  50-59
                          1110        60  60-69
                          1021        70  70+
```

We now stset the data with time since diagnosis as the timescale and death due to DTC as the outcome variable.

```
. stset surv_mm, fail(dead_dtc) scale(12)

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

-------------------------------------------------------------------
     5554  total obs.
        0  exclusions
-------------------------------------------------------------------
     5554  obs. remaining, representing
     1026  failures in single record/single failure data
 91292.33  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  41.95833
```

We now fit two Cox models, which we will refer to as models 1 and 2.

```
*** MODEL 1 ***
. xi: stcox i.sex i.period i.agegrp
i.sex         _Isex_1-2       (naturally coded; _Isex_1 omitted)
i.period      _Iperiod_1-3    (naturally coded; _Iperiod_1 omitted)
i.agegrp      _Iagegrp_0-70   (naturally coded; _Iagegrp_0 omitted)

Cox regression -- Breslow method for ties

No. of subjects =         5554         Number of obs   =      5554
No. of failures =         1026
Time at risk    =  91292.33333
                                       LR chi2(7)      =   1328.20
Log likelihood  =     -7950.99         Prob > chi2     =    0.0000

------------------------------------------------------------------------
         _t |Haz. Ratio   Std. Err.     z    P>|z|  [95% Conf. Interval]
------------+-----------------------------------------------------------
    _Isex_2 |  .5821111    .038319   -8.22   0.000  .5116505    .6622751
 _Iperiod_2 |  .7023412   .0524847   -4.73   0.000  .6066517    .8131242
 _Iperiod_3 |  .3994451   .0324388  -11.30   0.000  .3406679    .4683634
_Iagegrp_40 |  3.747874   .8146385    6.08   0.000  2.447754    5.738552
_Iagegrp_50 |   10.6226   2.091051   12.00   0.000  7.222252     15.6239
_Iagegrp_60 |  21.81348   4.192533   16.04   0.000  14.96665    31.79253
_Iagegrp_70 |  49.50315   9.508677   20.31   0.000  33.97286    72.13293
------------------------------------------------------------------------

*** MODEL 2 ***
. xi: stcox i.sex papillary i.period i.agegrp
i.sex          _Isex_1-2        (naturally coded; _Isex_1 omitted)
i.period       _Iperiod_1-3     (naturally coded; _Iperiod_1 omitted)
i.agegrp       _Iagegrp_0-70    (naturally coded; _Iagegrp_0 omitted)

Cox regression -- Breslow method for ties

No. of subjects =         5554           Number of obs   =      5554
No. of failures =         1026
Time at risk    =  91292.33333
                                         LR chi2(8)      =   1357.76
Log likelihood  =   -7936.2099           Prob > chi2     =    0.0000

------------------------------------------------------------------------
         _t |Haz. Ratio   Std. Err.    z    P>|z|   [95% Conf. Interval]
------------+-----------------------------------------------------------
    _Isex_2 |  .5908307   .0389297  -7.99   0.000   .5192512    .6722774
  papillary |  .7096868   .0447071  -5.44   0.000    .627256    .8029503
 _Iperiod_2 |  .7047072   .0526805  -4.68   0.000   .6086631    .8159065
 _Iperiod_3 |  .4093778   .0333114 -10.98   0.000   .3490289    .4801614
_Iagegrp_40 |  3.695118   .8032647   6.01   0.000   2.413179    5.658054
_Iagegrp_50 |  10.22584   2.014832  11.80   0.000   6.949989    15.04576
_Iagegrp_60 |  20.64451   3.975999  15.72   0.000   14.15365    30.11206
_Iagegrp_70 |  46.17276    8.89396  19.90   0.000   31.65369     67.3515
------------------------------------------------------------------------
```

(a) (2 marks) From model 1 we can obtain an estimate of the effect of sex. For which other variables is this estimate adjusted?

(b) (2 marks) Based on model 2, what is the estimated mortality rate ratio comparing papillary to follicular tumours for patients in the oldest age group (aged 70+ at diagnosis)? In other words, what is the effect of histology for the oldest age group? You do not need to comment on statistical significance.

(c) (2 marks) From model 1 we see that there is evidence that females experience lower DTC mortality than males. Is there evidence that this difference may be explained (fully or partly) by differences in the distribution of histological type between males and females (e.g., that females are more likely to be diagnosed with the less aggressive histological type)? Refer to the output from models 1 and/or 2 to support your answer.

(d) (4 marks) Is it possible, using results from models 1 and/or 2, to assess whether the effect of histology is modified by sex? If yes, comment on the magnitude of effect modification and whether it is statistically significant. If no, state how you would assess this.

We now extend our analysis to include an interaction between age group and histology.

```
*** MODEL 3 ***
.  xi: stcox i.sex papillary i.period i.agegrp i.agegrp*papillary

No. of subjects =         5554          Number of obs   =      5554
No. of failures =         1026
Time at risk    =  91292.33333
                                        LR chi2(12)     =   1384.49
Log likelihood  =   -7922.8418          Prob > chi2     =    0.0000
------------------------------------------------------------------------
          _t |Haz. Ratio   Std. Err.     z    P>|z|  [95% Conf. Interval]
-------------+----------------------------------------------------------
      _Isex_2 |  .5892885   .0388488   -8.02   0.000  .5178603    .6705687
   papillary |  .4987139   .1838953   -1.89   0.059  .2420918     1.02736
   _Iperiod_2 |  .7015236   .0525063   -4.74   0.000  .6058055    .8123653
   _Iperiod_3 |  .4094816   .0334009  -10.95   0.000  .3489819    .4804695
  _Iagegrp_40 |  3.578673   1.167643    3.91   0.000  1.887964    6.783448
  _Iagegrp_50 |  9.653023   2.847013    7.69   0.000  5.415198    17.20728
  _Iagegrp_60 |  18.33657   5.295007   10.07   0.000  10.41162    32.29368
  _Iagegrp_70 |  30.72466   8.895207   11.83   0.000  17.42018    54.19029
_IageXpap~40 |  1.031002   .4513082    0.07   0.944   .437176    2.431433
_IageXpap~50 |  1.036731   .4107889    0.09   0.927  .4768595    2.253937
_IageXpap~60 |  1.155012   .4452726    0.37   0.709  .5425467    2.458873
_IageXpap~70 |  2.107964   .8047936    1.95   0.051  .9974364    4.454931
------------------------------------------------------------------------
```

(e) (3 marks) Interpret the estimated hazard ratio for the variable labelled `_Iagegrp_60`, including a comment on statistical significance.

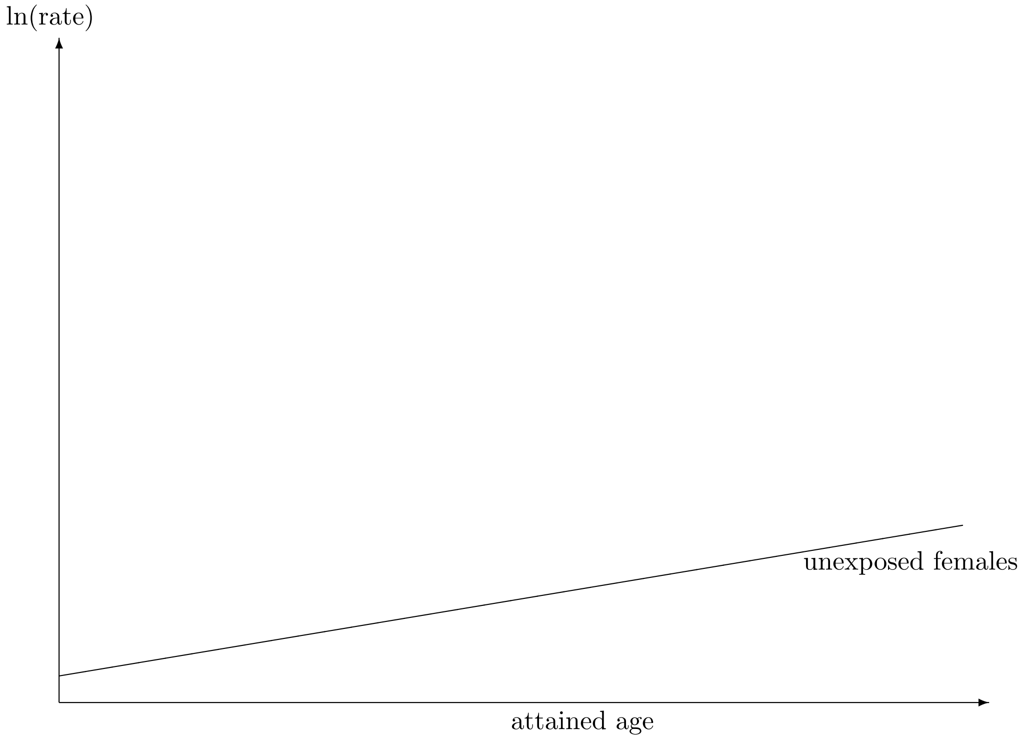
(f) (3 marks) What do we mean (conceptually, not mathematically), when we state that an effect is 'statistically significant'? If a result is statistically significant does it mean there is a 'real' or 'true' association?

(g) (3 marks) Based on model 3, what is the estimated mortality rate ratio (2 decimal places are sufficient) comparing papillary to follicular tumours for each and every age group? You do not need to present confidence intervals or comment on statistical significance.

(h) (3 marks) Is there evidence of a statistically significant interaction between histology and age group? If you choose to perform a formal hypothesis test you should state the null hypothesis, alternative hypothesis, value of the test statistic, assumed distribution of the test statistic under the null hypothesis, and a comment on statistical significance.

2. (a) (4 marks) It is known that the incidence rate of a certain disease depends on attained age and gender. A cohort study is conducted to determine whether the incidence rate depends on a binary exposure of interest. A Cox regression model is fitted to the data with attained age as the timescale. The estimated hazard ratio for gender (males/females) was 4 and the estimated hazard ratio for exposure (exposed/unexposed) was 2. Assume that there was no evidence of interaction between any of the variables.

Imagine that, for unexposed females, the association between the natural logarithm of the incidence rate and attained age has the form shown in the figure below. Assuming that the assumptions of the Cox model are appropriate, complete the figure below by drawing lines for the other 3 combinations of gender and exposure. The aim of this exercise is for you to demonstrate that you understand the assumptions of the Cox model. You should indicate how the estimated hazard ratios are represented on the graph.



(b) (4 marks) Now imagine that we split attained age into three categories and fit a Poisson regression model (with attained age, exposure, and gender as explanatory variables) to these same data. On the graph below, plot the fitted (i.e. predicted by the model) value of the natural logarithm of the incidence rate as a function of attained age for each of the four combinations of gender and exposure. The aim of this question is for you to demonstrate that you understand the fundamental difference between Cox regression and Poisson regression. You may assume that the estimated incidence rate ratio for gender (males/females) was 4 and the estimated incidence rate ratio for exposure (exposed/unexposed) was 2. You should indicate how these estimates are represented on the graph.

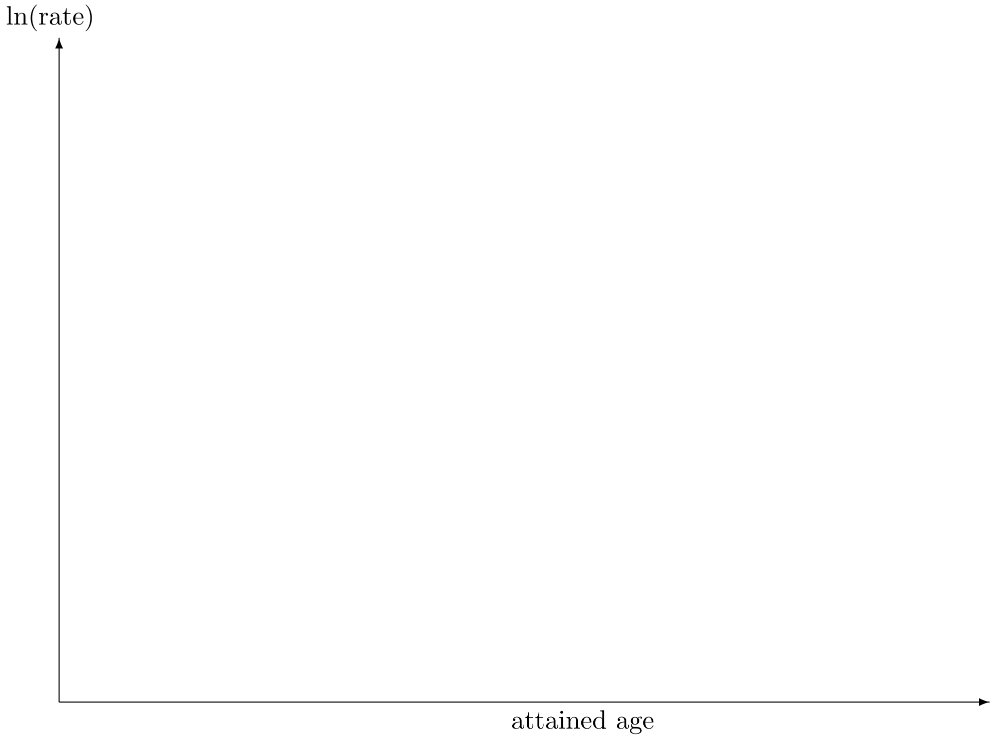


Table A3 Critical Values of Chi-Square

df	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
1	2.706	3.841	6.635
2	4.605	5.991	9.210
3	6.251	7.815	11.345
4	7.779	9.488	13.277
5	9.236	11.070	15.086
6	10.645	12.592	16.812
7	12.017	14.067	18.475
8	13.362	15.507	20.090
9	14.684	16.919	21.666
10	15.987	18.307	23.209
11	17.275	19.675	24.725
12	18.549	21.026	26.217
13	19.812	22.362	27.688
14	21.064	23.685	29.141
15	22.307	24.996	30.578
16	23.542	26.296	32.000
17	24.769	27.587	33.409
18	25.989	28.869	34.805
19	27.204	30.144	36.191
20	28.412	31.410	37.566
21	29.615	32.671	38.932
22	30.813	33.924	40.289
23	32.007	35.172	41.638
24	33.196	36.415	42.980
25	34.382	37.652	44.314
30	40.256	43.773	50.892
35	46.059	49.802	57.342
40	51.805	55.758	63.691
45	57.505	61.656	69.957
50	63.167	67.505	76.154
60	74.397	79.082	88.379
70	85.527	90.531	100.425
80	96.578	101.879	112.329
90	107.565	113.145	124.116
100	118.498	124.432	135.807

The value tabulated is c such that $P(\chi^2 \geq c) = \alpha$.