BIOSTAT III: Survival analysis for epidemiologists

Examination

4 December 2009

Code:

Please do not write your name; you have been allocated a code so the examiner is blinded to your identity

- Time allowed is 2 hours.
- Please try and write your answers on the exam sheet. You may use separate paper if absolutely necessary. Your working, not just the final answer, will be assessed when grading the examination.
- The exam contains 2 questions, each with several parts. The marks available for each part are indicated.
- A score of 12 marks or more out of a possible 23 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. You may not use a mobile phone or other communication device as a calculator or for any other purpose.
- 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) id(id) scale(12) noshow

                id:  id
     failure event:  dead_dtc != 0 & dead_dtc < .
obs. time interval:  (surv_mm[_n-1], surv_mm]
 exit on or before:  failure
    t for analysis:  time/12

------------------------------------------------------------------------------
     5554  total obs.
        0  exclusions
------------------------------------------------------------------------------
     5554  obs. remaining, representing
     5554  subjects
     1026  failures in single failure-per-subject 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 papillary i.period
i.sex             _Isex_1-2           (naturally coded; _Isex_1 omitted)
i.period          _Iperiod_1-3        (naturally coded; _Iperiod_1 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(4)      =    254.59
Log likelihood  =   -8487.7933                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Isex_2 |   .5635346   .0370315    -8.73   0.000     .4954337    .6409963
   papillary |   .5159256   .0323929   -10.54   0.000     .4561877    .5834862
  _Iperiod_2 |   .8086736   .0603834    -2.84   0.004      .698577    .9361214
  _Iperiod_3 |   .5590883   .0453575    -7.17   0.000     .4768968    .6554452
------------------------------------------------------------------------------

. *** 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) (4 marks) Based on models 1 and/or 2, is there evidence of an association between histological type and age group? If so, describe how the distribution of histological type varies by age.

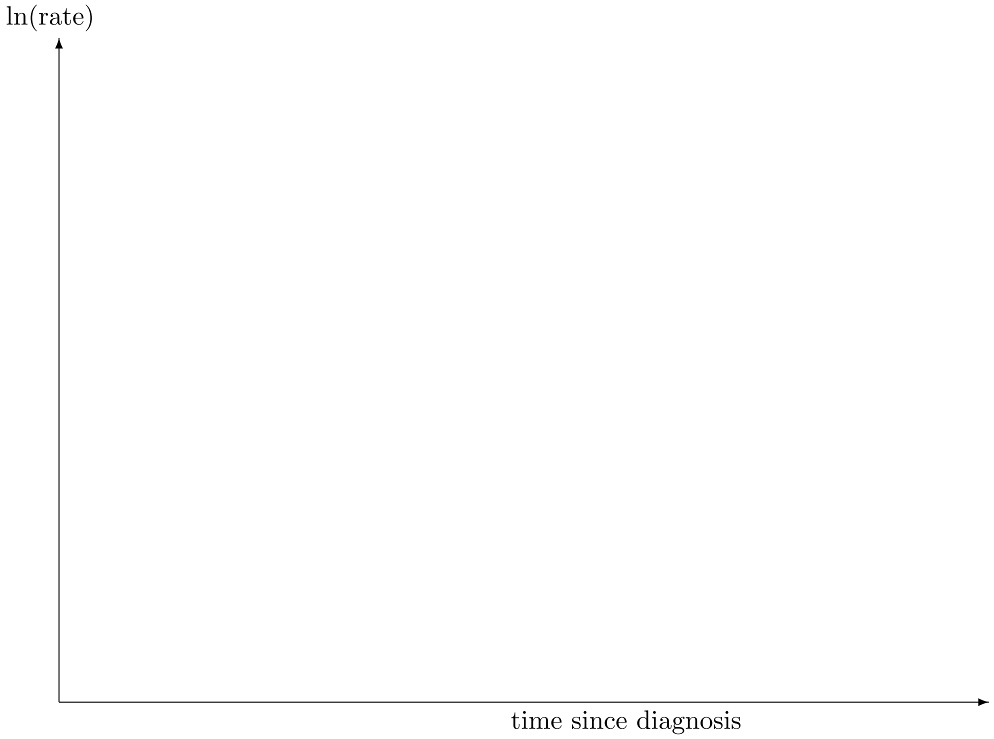
(b) (2 marks) Based on model 2, complete the 5 missing cells in the table below with the hazard ratio for each of the 5 categories compared to individuals diagnosed with follicular carcinoma in 1958–67. That is, the joint reference category is follicular carcinoma diagnosed in 1958–67. The hazard ratios you provide should be applicable for males aged 0–39.

	follicular	papillary
1958–67	1.00	
1968–77		
1978–87		

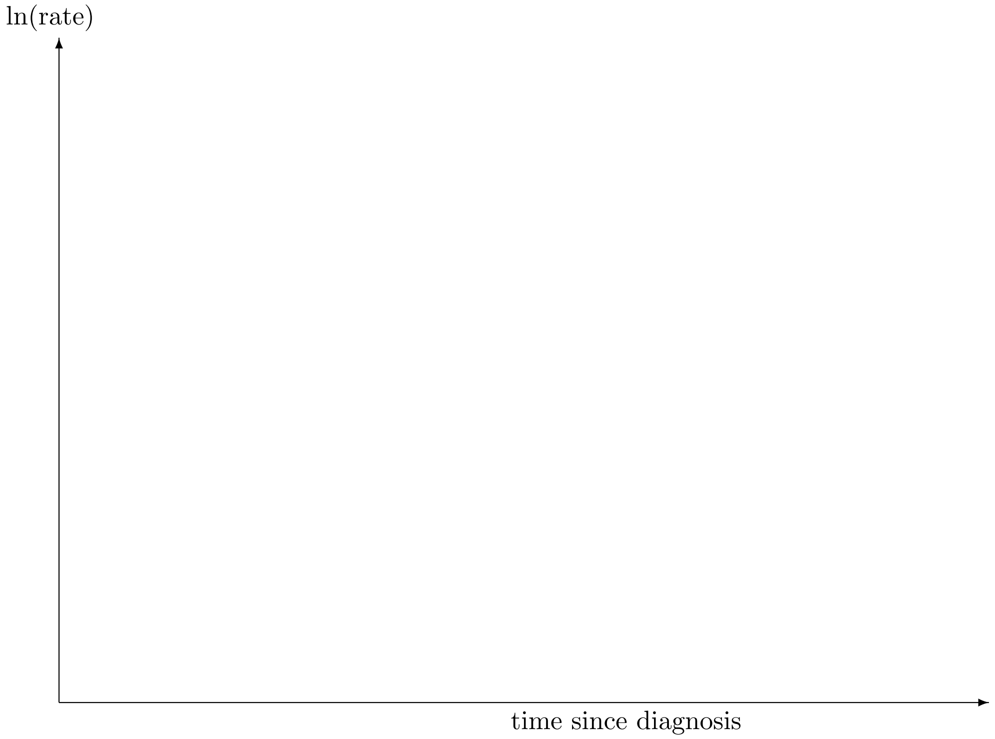
(c) (1 mark) How would the numbers in the table in the previous question change if you instead constructed the table for females aged 0–39?

(d) (2 marks) Based on model 1, it is possible to plot the predicted log-hazard as a function of time since diagnosis for each combination of sex and histology for patients diagnosed during the first calendar period. Illustrate below how such a graph might look.

You are not expected to label the values on the Y axis (the output tells you nothing about the magnitude of the log-hazard) but you are expected to indicate how the estimated hazard ratios are represented on the graph. Neither are you expected to know the exact functional form for how the log-hazard varies with follow-up time (i.e., you may choose any functional form). It is suggested that you also read the next part before completing this part.



(e) (2 marks) A colleague suggests an alternative to model 1. He suggests that rather than adjusting for sex you should fit a so called stratified Cox model where you stratify on sex. Repeat the previous question (i.e., plot the predicted log-hazards as a function of time since diagnosis for each combination of sex and histology) for the model stratified on sex. The aim is for you to demonstrate you understand the differences between a standard Cox model and a stratified Cox model.



We now split by time since diagnosis and fit a Poisson regression model, which we will call model 3. The splitting is done in annual intervals up to 15 years and the variable `fu` takes the values 0–14. That is, `fu=0` refers to the first year of follow-up. Part of the output has been omitted.

```
. stsplit fu, at(0(1)15) trim
(0 + 3053 obs. trimmed due to lower and upper bounds)
(61573 observations (episodes) created)

. *** MODEL 3 *** Poisson regression
. xi: streg i.fu i.sex papillary, dist(exp) nohr
i.fu              _Ifu_0-15           (naturally coded; _Ifu_0 omitted)
i.sex             _Isex_1-2           (naturally coded; _Isex_1 omitted)

No. of subjects =         5554       Number of obs   =     64074
No. of failures =          954
Time at risk    =  62777.79167
                                     LR chi2(16)     =    793.20
Log likelihood  =   -3932.5568       Prob > chi2     =    0.0000

--------------------------------------
          _t |      Coef.   Std. Err.
-------------+------------------------
      _Ifu_1 |  -.7809395   .1060801
      _Ifu_2 |  -1.072776   .1198709
      _Ifu_3 |  -1.018338   .1193797
      _Ifu_4 |  -1.208827    .130702
      _Ifu_5 |  -1.753111   .1666616
      _Ifu_6 |  -1.649412   .1616002
      _Ifu_7 |  -1.759166   .1723537
      _Ifu_8 |   -1.83878   .1811221
      _Ifu_9 |  -2.079463   .2045917
     _Ifu_10 |  -2.053329   .2045997
     _Ifu_11 |  -2.288301   .2310889
     _Ifu_12 |  -2.149023   .2210386
     _Ifu_13 |   -2.61796   .2834165
     _Ifu_14 |  -2.716724    .307103
     _Isex_2 |  -.5704134   .0680352
   papillary |  -.6991869    .064821
       _cons |  -2.094152   .0781802
--------------------------------------
```

(f) (3 marks) Based on model 3, what is the estimated hazard ratio (mortality rate ratio) and 95% confidence interval comparing papillary to follicular carcinoma?

(g) (3 marks) Based on model 3, perform a formal hypothesis test of the effect of sex. 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.

(h) (3 marks) Based on model 3, what is the estimated mortality rate (deaths due to DTC per 1000 person-years) during the third year of follow-up for females diagnosed with papillary carcinoma.

2. (3 marks) You have been asked to design a nested case-control study within the cohort analysed in the previous question. The aim is to study the effect of treatment on mortality due to DTC, where information on treatment will be abstracted from medical records. It is known that DTC mortality depends on calendar year of diagnosis and that treatment guidelines have changed with calendar time. Would you recommend matching on year of diagnosis in the nested case-control study? Motivate your answer and describe any possible pitfalls with matching.

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 $P(\chi^2 \geq c) = \alpha$.