Karolinska Institutet

BIOSTAT III: Survival Analysis

Examination

November 23, 2012

Time: 9:00–11.30

Exam room location: Lecture hall MTC,
Nobels väg 16, Karolinska Institutet

Code (please do not write your name):

- Time allowed is 2 1/2 hours.
- Please try and write your answers on the exam sheet. You may use separate paper if absolutely necessary. Your working and motivation for your answer, not just the final answer, will be assessed when grading the examination.
- The exam contains 2 sections; the first section tests your knowledge in general epidemiological concepts in a survival analysis framework whereas the second section focusses on more specific topics in survival analysis. Each section contains multiple questions (with several parts). The marks available for each part are indicated.
- A score of 6 marks or more out of 10 in the first section, and a score of 13 or more out of 22 in the second section 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.

Description of the data sets used in this exam

The recidivism data

For the first four questions of this exam we have used data from a study by Rossi, Berk, and Lenihan (1980) on recidivism (i.e., reoffending) of 432 prisoners during the first year after their release from Maryland state prisons. The aim of the research was to determine the efficacy of financial aid to released inmates as a means of reducing recidivism. Half of the inmates were randomly assigned to financial aid. They were followed for one year after their release and were interviewed monthly during that period. Data on arrests were taken from police and court records.

The following Stata output shows output from the `stset` command and frequency tables for some of the variables used in the analyses for this exam.

```
. /** stset the data using time since release from prison as the timescale
(in complete weeks) **/

. stset week, failure(arrest)

     failure event:  arrest != 0 & arrest < .
obs. time interval:  (0, week]
 exit on or before:  failure

------------------------------------------------------------------------------
      432  total obs.
        0  exclusions
------------------------------------------------------------------------------
      432  obs. remaining, representing
      114  failures in single record/single failure data
    20127  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        52



------------------------------------------------------------------------------
fin                            The inmate received financial aid after release
------------------------------------------------------------------------------

                  type:  numeric (double)
                 label:  fin_lab

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/432

            tabulation:  Freq.   Numeric  Label
                           216         0  No financial aid
                           216         1  Financial aid


-------------------------------------------------------------------------------
wexp             The inmate had full-time work experience before incarceration
-------------------------------------------------------------------------------

                  type:  numeric (double)
                 label:  wexp

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/432

            tabulation:  Freq.   Numeric  Label
                           185         0  No
                           247         1  Yes
```

```
-------------------------------------------------------------------------------
age                                Age in years at the time of release from prison
-------------------------------------------------------------------------------

                  type:  numeric (double)

                 range:  [17,44]                      units:  1
         unique values:  28                       missing .:  0/432

                  mean:   24.5972
              std. dev:   6.11338

-------------------------------------------------------------------------------
educ                                        Highest level of completed schooling
-------------------------------------------------------------------------------

                  type:  numeric (double)
                 label:  educ

                 range:  [2,6]                        units:  1
         unique values:  5                        missing .:  0/432

           tabulation:  Freq.   Numeric  Label
                           24         2  6th grade or less
                          239         3  7th to 9th grade
                          119         4  10th to 11th grade
                           39         5  12th grade
                           11         6  some college
```

The melanoma data

For questions five and six in this exam we analyse melanoma data from Finland. The aim is to study cause-specific survival from melanoma with respect to patient and disease characteristics such as age at diagnosis, year of diagnosis, sex and stage at diagnosis. The underlying time scale for the analysis is time since diagnosis.

The following Stata output shows output from the `stset` command and frequency tables for some of the variables used in the analysis.

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

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

------------------------------------------------------------------------------
     7775  total obs.
        0  exclusions
------------------------------------------------------------------------------
     7775  obs. remaining, representing
     7775  subjects
     1913  failures in single failure-per-subject data
 51269.71  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  20.95833




------------------------------------------------------------------------------
agegrp                                          Age in 4 categories
------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  agegrp

                 range:  [0,3]                        units:  1
         unique values:  4                        missing .:  0/7775

            tabulation:  Freq.   Numeric  Label
                          2046         0  0-44
                          2238         1  45-59
                          2280         2  60-74
                          1211         3  75+

------------------------------------------------------------------------------
year8594                                    Year of diagnosis 1985-94
------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  year8594

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/7775

            tabulation:  Freq.   Numeric  Label
                          3031         0  Diagnosed 75-84
                          4744         1  Diagnosed 85-94
```

```
-------------------------------------------------------------------------------
stage                                             Clinical stage at diagnosis
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  stage

                 range:  [0,3]                        units:  1
         unique values:  4                        missing .:  0/7775

            tabulation:  Freq.   Numeric  Label
                          1631         0  Unknown
                          5318         1  Localised
                           350         2  Regional
                           476         3  Distant

-------------------------------------------------------------------------------
sex                                                                         Sex
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  sex

                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  0/7775

            tabulation:  Freq.   Numeric  Label
                          3680         1  Male
                          4095         2  Female
```

Section 1

The following questions test your knowledge of general concepts in statistical modelling of epidemiological data.

We first fit a Cox regression model adjusted for time since release from prison (in weeks), age at the time of release, whether the inmate received financial aid after release, and whether the inmate had full-time work experience before incarceration (Model A).

```
Model A:
stcox age i.fin i.wexp

Cox regression -- Breslow method for ties

No. of subjects =          432                     Number of obs   =       432
No. of failures =          114
Time at risk    =        20127
                                                   LR chi2(3)      =     21.49
Log likelihood  =   -664.94013                     Prob > chi2     =    0.0001

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .944855   .0206276    -2.60   0.009     .9052785    .9861618
       1.fin |   .7142303   .1357226    -1.77   0.077     .4921388    1.036547
      1.wexp |    .694241   .1406016    -1.80   0.072     .4667885    1.032524
------------------------------------------------------------------------------
```

1. a: Interpret the estimated hazard ratio in the output that refers to the variable labelled 1.fin. You should also include a comment on statistical significance. (1 mark)

   b: Write down the null and alternative hypothesis for the z-test of the effect of age. What is the distribution of the test statistic under the null hypothesis? (2 marks) c: What is the estimated hazard ratio, comparing an inmate who was 40 years old at the time of the release, compared to someone who was 35 years? You can assume that all other covariates are fixed to the same level in the comparison. (1 mark)

d: A colleague suggests that the estimated effect of financial aid after release on the risk of getting re-arrested might be confounded by the social class of the inmate. The suggested motivation is that social class is likely to be strongly associated with criminal recidivism and the proposed solution is that you adjust the Cox model above for highest level of completed schooling. Do you agree with your colleague that the observed hazard ratio might be confounded by social class? If yes, explain how you would assess the degree of confounding. If no, motivate why. (2 marks)

2. We next fit another Cox model (Model B). In addition to the three main effects included in Model A, Model B also includes an interaction term between the the variables that represent whether financial aid was given and whether the inmate had full-time work experience prior to the incarceration. Parts of the Stata output from Model B are provided below.

```
 Model B
 stcox age i.fin##i.wexp

         failure _d:  arrest
   analysis time _t:  week


Cox regression -- Breslow method for ties

No. of subjects =          432                     Number of obs   =       432
No. of failures =          114
Time at risk    =        20127
                                                   LR chi2(4)      =     21.50
Log likelihood  =   -664.93151                     Prob > chi2     =    0.0003

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9449339   .0206233    -2.60   0.009     .9053654    .9862317
       1.fin |   .6980957   .1800201    -1.39   0.163     .4211268    1.157223
      1.wexp |   .6795773   .1765467    -1.49   0.137     .4084191    1.130763
             |
    fin#wexp |
        1 1  |   1.051319   .4005587     XXXX   XXXXX     XXXXXXXX    XXXXXXXX
------------------------------------------------------------------------------
```

a: Based on the output from Model B, what is the effect of receiving financial aid for each level of prior work experience? (2 marks) b: Perform a statistical hypothesis test to assess whether the effect of financial aid is modified by prior work experience? Remember to 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 marks)

Section 2

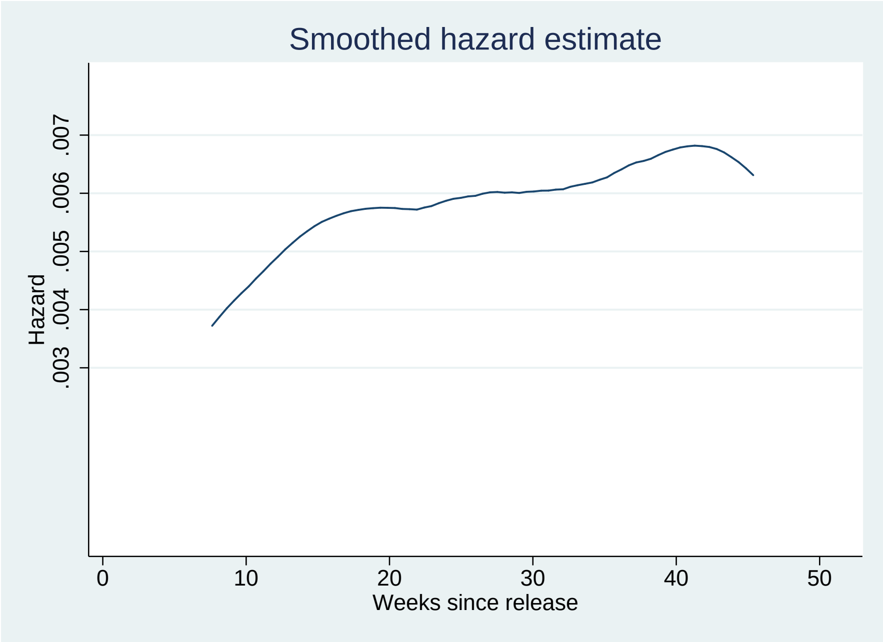
The following questions test your knowledge of concepts that are of special interest in survival analysis.

3. a: Fill in the Kaplan-Meier estimate for the part marked with X.XXXX in the output below. (1 mark)

```
sts list

         failure _d:  arrest
   analysis time _t:  week

           Beg.          Net            Survivor      Std.
  Time    Total   Fail   Lost           Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
     1      432      1      0             0.9977    0.0023     0.9837    0.9997
     2      431      1      0             0.9954    0.0033     0.9816    0.9988
     3      430      1      0             0.9931    0.0040     0.9786    0.9978
     4      429      1      0             0.9907    0.0046     0.9755    0.9965
     5      428      1      0             0.9884    0.0051     0.9724    0.9952
     6      427      1      0             0.9861    0.0056     0.9693    0.9937
     7      426      1      0             0.9838    0.0061     0.9663    0.9922
     8      425      5      0             0.9722    0.0079     0.9516    0.9841
     9      420      2      0             0.9676    0.0085     0.9459    0.9807
    10      418      1      0             0.9653    0.0088     0.9431    0.9789
    11      417      2      0             0.9606    0.0094     0.9375    0.9754
    12      415      2      0             X.XXXX    0.0099     0.9319    0.9717
```

b: State how you would interpret the Kaplan-Meier estimate that you filled in in part a). (1 mark)

Below is a graph showing the hazard function for the whole data set.



c: Explain what a hazard rate attempts to estimate.
Note: You do not have to provide the mathematical definition of the hazard to get full marks.
(1 mark)

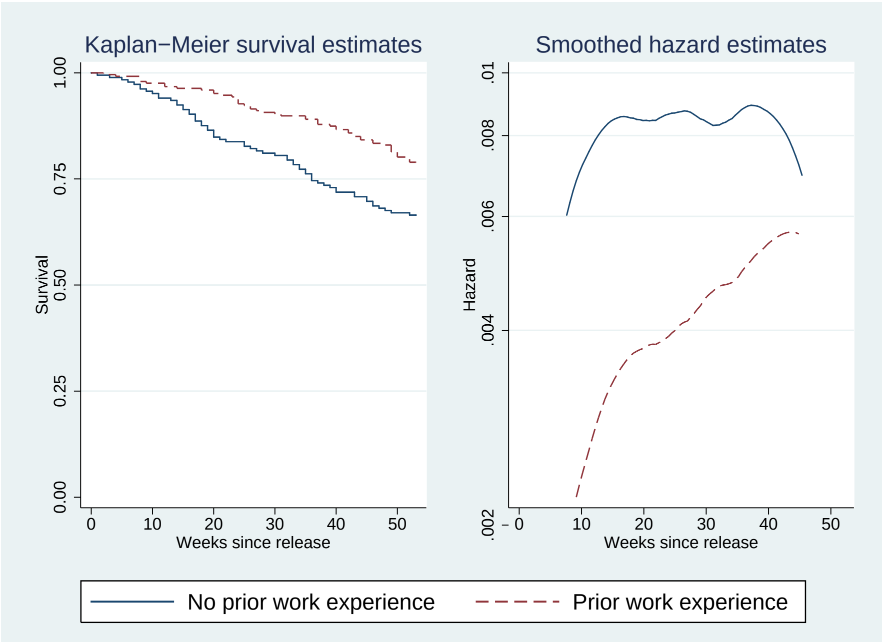
d: How would you characterize the association between criminal recidivism and time since release from prison based on what is shown in this graph? (1 mark)

4. We now fit a Cox model (Model C).

```
/*Model C*/
stcox wexp, nohr

No. of subjects =          432
No. of failures =          114
Time at risk    =        20127

LR chi2(1)      =       9.61
Log likelihood  =   -670.87678

-----------------------------------------------
          _t |      Coef.   Std. Err.      z
wexp         |
1: Yes       |  -.5824554   .1881361    -3.10
-----------------------------------------------
```

a: Based on the output from Model C write a short summary of this analysis (restrict your response to 2-3 sentences). Your response should include an estimate of the hazard ratio (including a 95% confidence interval), an interpretation of the point estimate as well as a comment on the statistical significance. (2 marks)

Below are the Kaplan-Meier survival estimates and the hazard functions for the recidivism data (by prior work experience).



b: What would you expect to see in these two graphs if the proportional hazards assumption for the effect of prior work experience was appropriate? (2 marks)

c: Describe two ways how you could formally test the appropriateness of the proportional hazards assumption in Model C. (2 marks)

5. We now switch to the melanoma cancer data set that has been used extensively during the course.

   a: Using the Stata output below give a point estimate of the mortality rate ratio comparing patients with regional metastasis at diagnosis to patients diagnosed with localised melanoma. (1 mark)

```
 strate stage, 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
(7775 records included in the analysis)

  +--------------------------------------+
  |     stage      D  person-time   Rate  |
  |---------------------------------------|
  |   Unknown    274   10.2671         X  |
  | Localised   1013   38.6266         X  |
  |  Regional    218    1.5002         X  |
  |   Distant    408    0.8758         X  |
  +--------------------------------------+
```

We now fit a Cox regression model including stage (Model D). The output is provided below.

```
 Model D
 stcox i.stage

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

Cox regression -- Breslow method for ties

No. of subjects =         7775                     Number of obs   =      7775
No. of failures =         1913
Time at risk    =  51269.70833
                                                   LR chi2(3)      =   1559.64
Log likelihood  =   -15614.364                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stage |
  Localised  |   1.018815   .0693932     0.27   0.784      .891494    1.164319
  Regional   |   5.116341   .4649793    17.96   0.000     4.281552    6.113891
  Distant    |   15.14297    1.20037    34.28   0.000     12.96394    17.68826
------------------------------------------------------------------------------
```

b: What is the mortality rate ratio comparing patients with regional metastasis at diagnosis to patients diagnosed with localised melanoma according to this model. Would you expect the hazard ratio from the Cox model to be the same to that from part a)? Motivate your answer. (2 marks)

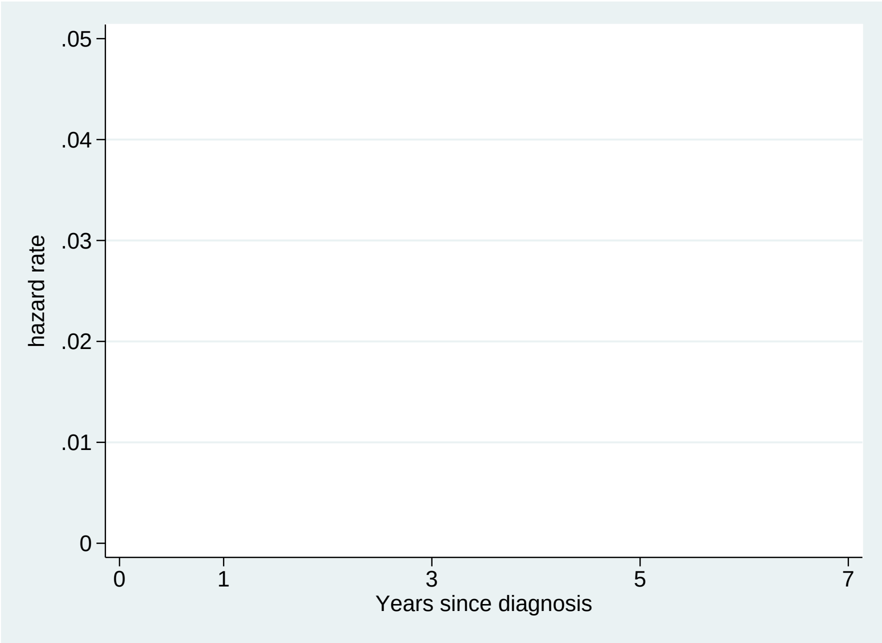
c: Explain how you could replicate the result (i.e., achieve identical hazard ratios) from the Model D by using Poisson regression instead of Cox regression. (2 marks)

6. We now split the follow-up for each patient into four categories (as shown in the Stata output below) and fit a Poisson model (Model E) adjusted for time since diagnosis, stage at diagnosis, age at diagnosis, sex and calendar period of diagnosis.

```
/*Split the data*/
 stsplit fup, at(1 3 5)
 . tab fup, missing

        fup |      Freq.     Percent        Cum.
------------+-----------------------------------
   0-1 year |      7,775       32.07       32.07
  1-3 years |      7,202       29.71       61.78
  3-5 years |      5,253       21.67       83.45
   >5 years |      4,011       16.55      100.00
------------+-----------------------------------
      Total |     24,241      100.00


/*Model E*/

streg i.fup i.stage i.agegrp i.sex i.year8694, distribution(exponential) nohr

Exponential regression -- log relative-hazard form

No. of subjects =         7775                   Number of obs   =      24241
No. of failures =         1913
Time at risk    =  51269.70833
                                                 LR chi2(11)     =    2465.49
Log likelihood  =   -5710.9538                   Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fup |
  1-3 years  |    .460638   .0613299     7.51   0.000     .3404336    .5808424
  3-5 years  |   .0546389   .0745306     0.73   0.463    -.0914384    .2007162
   >5 years  |  -.7717868   .0774943    -9.96   0.000    -.9236729   -.6199007
             |
       stage |
  Localised  |   .0395216   .0683156     0.58   0.563    -.0943745    .1734177
  Regional   |   1.589254   .0914024    17.39   0.000     1.410109      1.7684
  Distant    |   2.629674   .0797682    32.97   0.000     2.473331    2.786017
             |
      agegrp |
  45-59 yrs  |   .2612425   .0673349     3.88   0.000     .1292685    .3932166
  60-74 yrs  |   .5603511    .064849     8.64   0.000     .4332494    .6874529
  75+ yrs    |   1.018314   .0749736    13.58   0.000     .8713685    1.165259

         sex |
  Female     |  -.3609177   .0472235    -7.64   0.000     -.453474   -.2683613
             |
    year8594 |
  85-94      |  -.1680542   .0478656    -3.51   0.000    -.2618691   -.0742393
             |
  _cons      |  -3.688015   .0976335   -37.77   0.000    -3.879373   -3.496657
------------------------------------------------------------------------------
```

a: Based on the Stata output, write out the linear predictor from model E for a male patient diagnosed in 1980, at age 42, with unknown stage for the first year of follow-up. (1 mark) b: Based on the Stata output, write out the linear predictor from Model E for a female patient diagnosed in 1980, at age 42, with unknown stage for the first year of follow-up. (1 mark)

c: Write out an expression that shows how your responses in part a) and b) are related to the HR for the effect of sex. (1 mark) d: Complete the figure below by drawing the lines representing the cause-specific mortality rates at each and every point during follow-up for males and females, diagnosed in 1980, at age 42, with unknown stage. (3 marks)



e: Does Model E assume proportional hazards for the effect of sex? (1 mark)

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