BIOSTAT III: Survival Analysis

Examination

December 19, 2012

Time: 9:00–11.30

Exam room location: Wargentin room, MEB,
Nobels väg 12A, 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 ask the exam supervisor for additional 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 concepts in modelling epidemiological data whereas the second section covers more specific topics in survival analysis. The marks available for each part are indicated.
- A score of 6 marks or more out of 11 in the first section, and a score of 9 or more out of 18 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.

Section 1

1. The questions in this section test your knowledge of general concepts in statistical modelling of epidemiological data. You will recognise these questions from the self-assessment test.

   All questions are based on data from a cohort study designed to study risk factors for incidence of coronary heart disease (CHD). We will study three exposures of interest, body mass index (BMI), job type (3 categories) and energy intake (classified as high or low and where high is considered exposed). The Stata output shown on this page is not central to the question but is shown for completeness. The output below shows how a variable for BMI has been created and how job type and energy intake are coded.

   We have analysed the data using logistic regression, which is not completely appropriate given that these data are from a cohort study where individuals were at risk for different amounts of time. For the purpose of this exam you should interpret the results from the models as if logistic regression was appropriate.

```
. use http://biostat3.net/download/diet, clear
. /** Generate a variable containing BMI **/
. gen bmi=weight/(height/100)^2

. codebook bmi
          type:  numeric (float)

         range:  [15.875263,33.292957]        units:  1.000e-06
 unique values:  321                      missing .:  5/337

          mean:   24.1237
      std. dev:   3.21202

   percentiles:        10%       25%       50%       75%       90%
                   20.0605    21.584   24.1144   26.5157    28.206


. codebook job
                  type:  numeric (byte)
                 label:  job

                 range:  [1,3]                        units:  1
         unique values:  3                        missing .:  0/337

            tabulation:  Freq.   Numeric  Label
                           102         1  driver
                            84         2  conductor
                           151         3  bank

. codebook hieng
                  type:  numeric (float)
                 label:  hieng

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/337

            tabulation:  Freq.   Numeric  Label
                           155         0  low
                           182         1  high
```

We now estimate a logistic regression model where the outcome is CHD (0 = No CHD 1 = CHD) and the exposures are coded as described above.

```
. /*Model 1*/
. logistic chd i.hieng i.job bmi

Logistic regression                             Number of obs     =        332
                                                LR chi2(4)        =       7.77
                                                Prob > chi2       =     0.1003
Log likelihood = -127.84724                     Pseudo R2         =     0.0295
------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.hieng |   .4546316   .1532119    -2.34   0.019     .2348566    .8800685
             |
         job |
          2  |   1.793175   .7950121     1.32   0.188     .7520364    4.275695
          3  |   1.169097   .4660996     0.39   0.695     .5351687    2.553939
             |
         bmi |   1.082693   .0565679     1.52   0.128       .97731     1.19944
------------------------------------------------------------------------------
```

(a) (1 mark) Interpret the estimated odds ratio for BMI, including a comment on statistical significance.

(b) (1 mark) Both P-values for the parameters representing the effect of occupation (job type) are greater than 0.1. Can we conclude that there is no evidence of a statistically significant overall association between occupation and CHD risk? If not, how could you test whether there is an association between occupation and CHD risk?

We now fit another model (labelled model 2).

```
. /*Model 2*/
. logistic chd i.hieng bmi

Logistic regression                               Number of obs   =        332
                                                  LR chi2(2)      =       5.91
                                                  Prob > chi2     =     0.0522
Log likelihood =    -128.78                       Pseudo R2       =     0.0224
------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.hieng |    .468139   .1563834    -2.27   0.023     .2432362    .9009932
         bmi |   1.063526   .0535557     1.22   0.221     .9635722    1.173848
------------------------------------------------------------------------------
```

(c) (1 mark) Based on model 2, among individuals with a BMI of 24, what is the estimated odds ratio for individuals with a high energy compared to those with a low energy intake? You do not have to comment on statistical significane.

(d) (2 marks) Based on model 2, what is the estimated odds ratio for individuals with a BMI of 30 compared to individuals with a BMI of 25? Is the difference statistically significant?

(e) (2 marks) Is it possible to ascertain, using the output from models 1 and/or 2, whether the effect of high energy intake is confounded by job type? If so, comment on whether the effect of high energy intake is confounded by job type. If not, describe how you could study this.

We now fit another model, labelled model 3.

```
. /* Model 3 */
. logistic chd i.hieng##i.job bmi

Logistic regression                               Number of obs   =        332
                                                  LR chi2(6)      =       7.89
                                                  Prob > chi2     =     0.2461
Log likelihood = -127.78775                       Pseudo R2       =     0.0300
------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.hieng |   .3792746   .2469698    -1.49   0.137     .1058479    1.359018
             |
         job |
          2  |   1.588197   .9160756     0.80   0.423      .512778    4.919028
          3  |   1.074633   .5513115     0.14   0.888     .3931644    2.937286
             |
   hieng#job |
        1 2  |   1.342565   1.189766     0.33   0.740     .2363798    7.625359
        1 3  |   1.242141   1.018884     0.26   0.792     .2488634    6.199846
             |
         bmi |    1.08078   .0567668     1.48   0.139     .9750546     1.19797
------------------------------------------------------------------------------
```

(f) (2 marks) Based on model 3, what is the OR of high energy intake compared to low for each of the 3 different job types?

(g) (2 marks) Using information from any of the models fitted so far, is there evidence that the effect of high energy intake is modified by job type? Conduct a formal hypothesis test. You should state the null hypothesis, alternative hypothesis, value of a test statistic, assumed distribution of the test statistic under the null hypothesis, the name of the statistical test you are using, and a comment on statistical significance.

Section 2

2. The following table summarises the data from a cohort study designed to study the association between mortality (the outcome) and the exposures sex and age (grouped into two categories; 0=young, 1=old). The table shows the number of events (`deaths`) and person-years at risk (`pyears`) for each of the four categories of sex and age.

sex	age	deaths	pyears
male	0	30	2000
male	1	90	1500
female	0	20	2000
female	1	90	2000

(a) (2 marks) We fitted the Poisson regression model

$$\ln(\lambda) = \beta_0 + \beta_1 X_{\text{sex}} \qquad \text{(model 1)}$$

where X_{sex} is modelled as a continuous variable and coded as 1 for females and 0 for males. The output is not shown. What are the estimates for β_0 and β_1?

(b) (2 marks) Model 1 can be used to provide a prediction (or fitted value) for the number of deaths for each of the four rows in the table. What would be the predicted (fitted) number of deaths for 'old males' (i.e., the second row in the table)?

(c) (1 mark) The data in the table on the previous page were obtained by collapsing the individual-level data (i.e., with one observation per individual) and summing the number of deaths and person-time for individuals with the same values of age and sex. Would the parameter estimates change, compared to part (a), if we fitted a Poisson regression model to the individual-level data with sex as the only explanatory variable? That is, if we refitted model 1 to individual rather than grouped data.

(d) (1 mark) Would the parameter estimates change, compared to part (a), if we fitted a Cox model to the individual-level data with sex as the only explanatory variable?

(e) (3 marks) We now return to the grouped data. Compared to part (a), would the estimates of β_0 and β_1 change if we fitted the same model, but with X_{sex} modelled as a continuous variable coded as 2 for females and 1 for males? We will write this model as

$$\ln(\lambda) = \beta_0' + \beta_1' X_{\text{sex}}' \qquad \text{(model 1A)}$$

It is sufficient to state (and motivate) whether the parameter estimates would remain identical, become larger, or become smaller.

(f) (2 marks) We now extend the model to control for the effect of age. The model is

$$\ln(\lambda) = \beta_0 + \beta_1 X_{\text{sex}} + \beta_2 X_{\text{age}} \qquad \text{(model 2)}$$

where X_{sex} is coded as in part (a) and age is coded as in the table. Would the estimate of β_2 be less than zero, exactly zero, or greater than zero? Motivate your answer

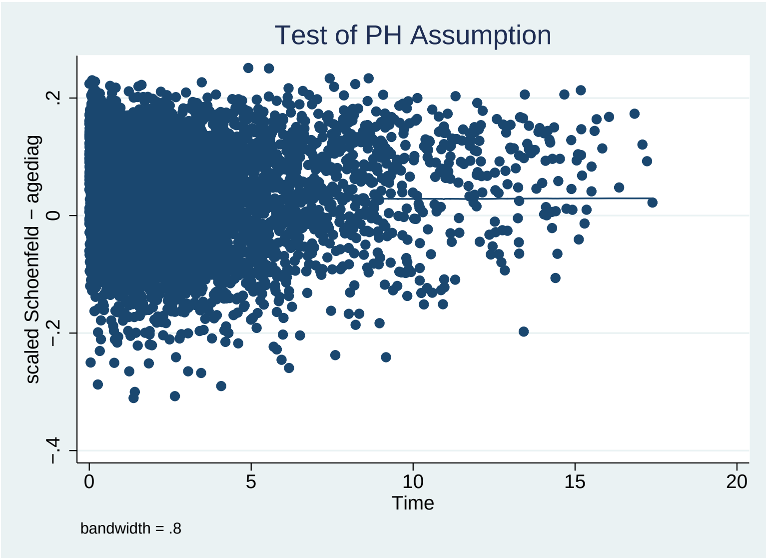
(g) (1 mark) We now further extend the model to

$$\ln(\lambda) = \beta_0 + \beta_1 X_{\text{sex}} + \beta_2 X_{\text{age}} + \beta_3 X_{\text{age*sex}} \qquad \text{(model 3)}$$

where $X_{\text{age*sex}}$ is coded as 1 for old females and 0 for the other three categories. Based on this model, what would be the predicted (fitted) number of deaths for 'old males' (i.e., the second row in the table)?

(h) (2 marks) Based on model 3, what is the estimated probability that a young male in the study survives 2 years? State any assumptions that you make.

3. (a) (2 marks) We continue with the study introduced in the previous question and now fit a Cox model with time since entry as the timescale. Covariates in the model were age at entry (in years) and sex. That is, we modeled age in years rather than age in two categories. Following is a plot, produced by Stata, of the scaled Schoenfeld residuals for the effect of age at entry. Under the proportional hazards assumption, what would you expect to see from this plot?



(b) (2 marks) Additional plots and tests suggest that a proportional hazards assumption is not appropriate for sex. A colleague suggests you fit a 'stratified Cox model' (stratified by sex) since that model does not require an assumption of proportional hazards for sex. Is that a sensible suggestion?

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