UNIVERSITY OF ESSEX
INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH
Professor Stephen P. Jenkins <stephenj@essex.ac.uk>

Essex Summer School course 'Survival Analysis'
and
EC968. Part II: Introduction to the analysis of spell duration data

Lesson 8. Competing risks models

Contents

1 Aim

The aim of this Lesson is to show how to estimate independent competing risk models.

2 Introduction

Up until now, we have modelled time-to-event data and only a single type of event has been distinguished: 'failure'. Models in which there are different types of events – multiple destinations – are also of interest. For example, in a model of unemployment duration, we may wish to know about not only time until exit from unemployment by whatever route, but also about time to exit from unemployment *to a job*, and compare this with this time to exit from unemployment *to economic inactivity*. Competing risk models provide a method of addressing such issues. We shall only consider the simplest case – the independent competing risk model. (See the Lecture Notes for a discussion of how this model may be generalized to allow for correlated risks.)

As explained in the Lectures, one supposes that there is a number of latent survival times, one for each different destination, and the actual destination entered (observed) is the minimum of the latent survival times. (Right censoring can also be interpreted as a competing risk.) Correlations between unobservable factors affecting each destination-specific hazard are assumed away – hence the label 'independent competing risks' model.

For continuous time models, the log-likelihood for a model with multiple destinations can be partitioned into a sum of sub-contributions, each of which is a function of the parameters of a single destination-specific hazard only. The separability property means that one can estimate a multiple-destination survival model by estimating a number of single-destination models separately, one for each destination (competing risk). And to estimate a given destination-specific hazard, one treats spell endings to destinations other than the one in question as right censored at the point of exit.

For models of competing risks in which the time scale is discrete, then the separability property does not hold, and modelling is more complex, as the Lecture Notes show. One notable exception is the case when time is intrinsically discrete. In this case, one may assume a 'multinomial logit' model of competing risks that is easily estimated with existing software.

If one needs to use a discrete time model because one has interval-censored data (continuous survival times are available only in grouped form), then modelling is rather complex, and one needs special programs to estimate the models. There is one exception to this, when transitions to the various destinations can only occur at the boundaries of the intervals. With this assumption, the likelihood for the competing risk model factors in a manner exactly analogous to that for a continuous time competing risk model, and estimates may be derived using a standard single-risk program. This is the only situation that we shall consider in this Lesson, but be aware that it may not be appropriate in practice. On the other, and more positively, observe also that the Lecture Notes demonstrated that, if the interval hazard is relatively small, then the 'multinomial logit' model of competing risks provided a close approximation to a proportional hazards model for interval-censored data for which one assumed that the continuous time hazard rate was constant within each interval.

To illustrate the statements above, we shall use the unemployment data (unemp.dta). This provides information about unemployment duration for a sample of Unemployment Insurance recipients. There is a variable **status** which tells us, not just whether an individual left unemployment, but what the destination was: whether UI entitlement was exhausted (and if so whether followed by Unemployment Assistance receipt, i.e. unemployment continued) or if the man got a job or if he left UI for other reasons (e.g. military service). If we were only interested in whether the UI spell had ended (a single destination), then we would treat UI spells ending in exhaustion as 'right censored' and exits for whatever reason as a 'failure'. (The variable **exit** is the censoring variable defined in this way.)

The survival times are discrete (interval censored, with intervals of length one month). (Ex 8_1 asks you to repeat the illustration below, pretending instead that time is continuous.) Before creating the required censoring variables, there are some other preliminaries such as reorganising the data in person-month form and creating some additional variables (cf. Lessons 3 and 6):

```
. use unemp, clear
(Spanish UI entrants sample Feb 1987, men 18-54)

. ta status exit

   general |
    status |    UI spell ended?
  variable |  censored       exit |     Total
-----------+----------------------+----------
   Exh-NoA |       432          0 |       432
  Exh-YesA |       409          0 |       409
   ExitJob |         0        487 |       487
   ExitOth |         0        179 |       179
-----------+----------------------+----------
     Total |       841        666 |      1507
```

Let’s create a new status variable called status2 that combines into the censored category the two types of UI exhaustion.

```
. ge status2 = 0

. replace status2 = 1 if status == 2
(487 real changes made)

. replace status2 = 2 if status == 3
(179 real changes made)

. lab def status2 0 "censored" 1 "Exit-job" 2 "Exit-other"

. lab val status2 status2
```

Now we do the episode-splitting to re-organise the data into person-month form.

```
. expand conmths
(9727 observations created)
```

We shall consider first a discrete time proportional hazards (cloglog) model applied to interval-censored data, and assume that exits from unemployment can only occur at the boundaries of the monthly intervals. (This is not true in reality!) Second, we estimate the multinomial logit competing risks model.

We shall suppose that the baseline hazard has the log(time) form, so let us create that and another set of covariates from the variable summarising region of residence (there are five categories):

```
. bysort newid: ge t = _n

. ge logt = ln(t)

. ta groupreg, ge(reg)

region,grou |
        ped |      Freq.     Percent        Cum.
------------+-----------------------------------
      North |       3231       28.76       28.76
     Centre |       3543       31.54       60.30
   North-Ea |       1648       14.67       74.97
      South |       2319       20.64       95.61
    Islands |        493        4.39      100.00
------------+-----------------------------------
      Total |      11234      100.00
```

3 Creating the relevant censoring variables

Now we create the destination-specific censoring indicators to be used with the cloglog model. The variable summarising whether persons have left UI at all is exit (see above) – but we need to manipulate this to create a new censoring indicator for the expanded data set in person-month form. If we were to assume a single detination state, the relevant monthly event variable is ‘leftui’. We also need similar indicator variables recording for each month whether there is an exit to a job or an exit to other destinations (I call them ‘cex_job’ and ‘cex_oth’). Let us create them:

```
. * single destination censoring vble
. by newid: ge leftui = exit == 1 & _n==_N

. lab var leftui "1=Exit UI"

. * multiple destination censoring vbles

. bysort newid (t): ge cex_job = status == 2 & _n == _N if status ~= .

. lab var cex_job "1=Exit UI to job"

bysort newid (t): ge cex_oth = status == 3 & _n == _N if status ~= .

. lab var cex_oth "1=Exit UI to other dest."
```

For the MNL model, we also use the data organised in person-month form, but we have to construct a new dependent variable, as follows. This variable, that I label deadml, has three categories corresponding to the occurrence of events in each spell month – whether there was an exit from UI in that month to a job or an exit for some other reason, or whether there was no exit (the right-censored case). If there was a job-related UI exit in the last month observed, deadmnl = 1, if there was another type of UI exit in the last month observed, deadmnl = 2 and, in all other cases, deadmnl = 0.

```
. ge deadmnl = 0

bysort newid (t): replace deadmnl = 1 if status2==1 & _n==_N
(487 real changes made)

bysort newid (t): replace deadmnl = 2 if status2==2  & _n==_N
(179 real changes made)

. ta deadmnl status2

           |              status2
   deadmnl |  censored   Exit-job  Exit-othe |     Total
-----------+---------------------------------+----------
         0 |     5,556      2,038      2,974 |    10,568 
         1 |         0        487          0 |       487 
         2 |         0          0        179 |       179 
-----------+---------------------------------+----------
     Total |     5,556      2,525      3,153 |    11,234 
```

4 Estimation

Now it is simply a matter of running the models. First we’ll look at the model for the overall risk of exit, and then at the component competing risk models.

```
. * estimate a single destination PH model with log(time) hazard
. cloglog leftui age famresp tyentry reg1-reg4 logt, nolog

Complementary log-log regression                Number of obs     =      11234
                                                Zero outcomes     =      10568
                                                Nonzero outcomes  =        666

                                                LR chi2(8)        =      64.10
Log likelihood = -2495.5318                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
  leftui |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |  -.0032994   .0046837     -0.704   0.481      -.0124792    .0058804
 famresp |   .0651581     .08273      0.788   0.431      -.0969897     .227306
 tyentry |   .6736559    .090667      7.430   0.000       .4959519    .8513598
    reg1 |    .138366   .1984373      0.697   0.486       -.250564    .5272961
    reg2 |   .0674445    .198876      0.339   0.735      -.3223453    .4572344
    reg3 |  -.0665157   .2133666     -0.312   0.755      -.4847067    .3516752
    reg4 |  -.0141535   .2051373     -0.069   0.945      -.4162152    .3879083
    logt |   .1595557   .0437978      3.643   0.000       .0737135    .2453978
   _cons |  -3.450175   .2665017    -12.946   0.000      -3.972509   -2.927841
------------------------------------------------------------------------------
```

In this (near-fictional) example, it appears that the type of employment contract is the only covariate with a statistically significant association with UI exit rates: those who entered UI from a temporary employment contract have hazard rates almost twice as high as those entering from a permanent job (exp(.67) ≈ 2). The coefficient on log(time), also statistically significant, indicates that the baseline hazard increases with time spent receiving UI. (In terms of Lesson 6, the coefficient is the estimate of q–1, so the estimated value for q is 1.16.)

Now consider the component sub-models, first for exits to a job and the exits to other destinations:

```
. cloglog cex_job age famresp tyentry reg1-reg4 logt, nolog

Complementary log-log regression                Number of obs     =      11234
                                                Zero outcomes     =      10747
                                                Nonzero outcomes  =        487

                                                LR chi2(8)        =     158.85
Log likelihood = -1925.2828                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
 cex_job |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |  -.0022752   .0054375     -0.418   0.676      -.0129325    .0083821
 famresp |   .0485768   .0978946      0.496   0.620       -.143293    .2404467
 tyentry |   1.074283   .1224166      8.776   0.000       .8343505    1.314215
    reg1 |   .2247948   .2373437      0.947   0.344      -.2403902    .6899798
    reg2 |   .1324284    .238803      0.555   0.579      -.3356169    .6004737
    reg3 |  -.1732715   .2588485     -0.669   0.503      -.6806053    .3340623
    reg4 |  -.0139021   .2450342     -0.057   0.955      -.4941603    .4663561
    logt |  -.2023275   .0503794     -4.016   0.000      -.3010693   -.1035857
   _cons |  -3.600207   .3158823    -11.397   0.000      -4.219325   -2.981089
------------------------------------------------------------------------------
```

```
. cloglog cex_oth age famresp tyentry reg1-reg4 logt, nolog

Complementary log-log regression                  Number of obs     =      11234
                                                  Zero outcomes     =      11055
                                                  Nonzero outcomes  =        179

                                                  LR chi2(8)        =     281.71
Log likelihood = -777.64841                       Prob > chi2       =     0.0000

------------------------------------------------------------------------------
 cex_oth |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |  -.0083294    .009366     -0.889   0.374      -.0266865    .0100277
 famresp |   .0911671   .1564844      0.583   0.560      -.2155366    .3978709
 tyentry |   .1095504   .1627213      0.673   0.501      -.2093775    .4284782
    reg1 |  -.1321078   .3651062     -0.362   0.717      -.8477028    .5834871
    reg2 |  -.1107218   .3614129     -0.306   0.759      -.8190781    .5976345
    reg3 |   .2203409   .3786543      0.582   0.561      -.5218079    .9624897
    reg4 |  -.0023397   .3780095     -0.006   0.995      -.7432248    .7385454
    logt |   1.942257   .1612489     12.045   0.000       1.626215    2.258299
   _cons |  -8.304859   .6451134    -12.873   0.000      -9.569258    -7.04046
------------------------------------------------------------------------------
```

There appear to be clear differences between the two processes. The risk of exiting to a job is strongly associated with the type of prior employment contract and the hazard rate declines with time spent receiving UI (the estimate of *q* in this case = 1–0.20 = 0.8 < 1). By contrast the risk of exiting to other destinations is not associated with any of the covariates and the hazard rate rises with UI receipt duration (the estimate of *q* is 2.94).

The Lecture Notes argued that the cloglog model may not be an appropriate one, because it corresponds to an assumption that transitions can only occur at the boundary of the intervals. We considered a number of other, more plausible, assumptions about the hazard rate within the intervals when we had interval-censored data. Unfortunately these models require special programs, which are beyond the scope of this course. More positively, the Lectures also showed that if the interval hazard rate was relatively 'small', a 'multinomial logit' model, originally developed for intrinsically discrete data, may provide estimates that are a close approximation to a model for interval-censored data that assumed that the (continuous) hazard was constant within intervals. Here's how we can estimate this model (the estimation method is due to Allison, *Sociological Methodology 1992*):

First, note that we use the 'expanded' person-month data, as for the earlier model, and use the three-category deadmnl variable as the outcome variable. We estimate the model using the **mlogit** command, and use the **baseoutcome()** option to specify which category is treated as the reference one – it is deadmnl = 0 (right-censored).

```
. mlogit deadmnl age famresp tyentry reg1-reg4 logt, nolog baseoutcome(0)

Multinomial logistic regression                   Number of obs   =      11234
                                                  LR chi2(16)     =     432.87
                                                  Prob > chi2     =     0.0000
Log likelihood = -2698.7776                       Pseudo R2       =     0.0742

------------------------------------------------------------------------------
     deadmnl |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
         age |  -.0024503   .0055886    -0.44   0.661    -.0134037    .0085031
     famresp |   .0514627   .1006845     0.51   0.609    -.1458754    .2488008
     tyentry |   1.092581   .1242046     8.80   0.000     .8491446    1.336018
        reg1 |   .2291151   .2438589     0.94   0.347    -.2488395    .7070698
        reg2 |   .1313109   .2452359     0.54   0.592    -.3493426    .6119644
        reg3 |  -.1784552   .2655201    -0.67   0.502    -.6988649    .3419546
        reg4 |  -.0178443   .2516282    -0.07   0.943    -.5110265     .475338
        logt |  -.1947983    .052208    -3.73   0.000    -.2971241   -.0924725
       _cons |  -3.579489   .3239595   -11.05   0.000    -4.214437    -2.94454
-------------+----------------------------------------------------------------
2            |
         age |  -.0084695   .0095718    -0.88   0.376    -.0272299     .010291
     famresp |   .0951781   .1600171     0.59   0.552    -.2184496    .4088058
     tyentry |   .1424349   .1666566     0.85   0.393     -.184206    .4690757
        reg1 |  -.1239563   .3741337    -0.33   0.740    -.8572449    .6093324
        reg2 |  -.1084249   .3704304    -0.29   0.770    -.8344551    .6176052
        reg3 |   .2285222   .3886337     0.59   0.557    -.5331859    .9902302
        reg4 |   .0045579   .3873852     0.01   0.991    -.7547031    .7638189
        logt |   1.931943   .1618052    11.94   0.000      1.61481    2.249075
       _cons |  -8.246268   .6537316   -12.61   0.000    -9.527559   -6.964978
------------------------------------------------------------------------------
((deadmnl==0 is the base outcome)
```

It turns out that the MNL estimates and cloglog estimates provide very similar estimates! One plausible explanation for this is that the exit rate from UI for Spanish men is relatively small, and so the model original developed in a discrete time context approximates the model for the continuous time context well.

A second issue is whether each of the regressors has the same effect on the two destination-specific hazard rates, and whether there is a similar pattern of duration dependence in each of the hazards. One could test these hypotheses formally by Wald or likelihood-ratio tests applied to these models. For one formal statistical test that can be implemented relatively straightforwardly for continuous time models, see Exercise 8_1.

Eyeball econometrics suggests that there are two main differences between the two equations. The first is in their duration dependence: the hazard for exits to a job is declining with time on UI, whereas the hazard for other types of exits is rising with time on UI. Second, we see that those men who had a temporary employment contract in their last job before UI receipt (tyentry = 1) are much more likely to exit to a job than are men who had a permanent employment contract in their last job (tyentry = 0). On the other hand, the type of employment contract has no significant association with the hazard of exit from UI for other reasons.

One implication of these results is that estimating a single-exit-type hazard regression model (i.e. not differentiating between the different types of exit) may not provide a sufficiently rich picture about the impact of different covariates on UI exit hazards or about duration dependence. For the record, here is the cloglog single destination state model:

```
. cloglog leftui age famresp tyentry reg1-reg4 logt, nolog

Complementary log-log regression                  Number of obs     =      11234
                                                  Zero outcomes     =      10568
                                                  Nonzero outcomes  =        666

                                                  LR chi2(8)        =      64.10
Log likelihood = -2495.5318                       Prob > chi2       =     0.0000

------------------------------------------------------------------------------
      leftui |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0032994   .0046837    -0.70   0.481    -.0124792    .0058804
     famresp |   .0651581     .08273     0.79   0.431    -.0969897     .227306
     tyentry |   .6736559    .090667     7.43   0.000     .4959519    .8513598
        reg1 |    .138366   .1984373     0.70   0.486     -.250564    .5272961
        reg2 |   .0674445    .198876     0.34   0.735    -.3223453    .4572344
        reg3 |  -.0665157   .2133666    -0.31   0.755    -.4847067    .3516752
        reg4 |  -.0141535   .2051373    -0.07   0.945    -.4162152    .3879083
        logt |   .1595557   .0437978     3.64   0.000     .0737135    .2453978
       _cons |  -3.450175   .2665017   -12.95   0.000    -3.972509   -2.927841
------------------------------------------------------------------------------
```

This illustrates features of both of the regressions for the separate destination hazards. Observe how in the single-destination model, the coefficient of tyentry is positive (as for the exit-to-a-job hazard), and the hazard apparently declines with time (as for the exit-for-other-reasons hazard).

5 Exercise 8.1

(i) Estimate a model with the same covariates and duration dependence specification as in the text, first for the overall UI exit hazard, and then for exits to a job and exits to other destinations – but now use a logistic hazard model rather than a cloglog one.

(ii) Now pretend that the survival times in the unemployment data set are measured in continuous time, and estimate a continuous time model using the same covariates as were used in the discrete time examples above. As there are no time-varying covariates, episode splitting is not required. Just **stset** your data and away you go (do it three times, once for each destination-specific risk). Assume a Weibull PH model – this will facilitate comparisons with the illustration in the main text and the exercise above.

(iii) Drawing on your results from (ii), implement the formal test of proportionality of risks proposed by W. Narendranathan and M.B. Stewart (1991), 'Testing the proportionality of cause-specific hazards in competing risk models', *Oxford Bulletin of Economics and Statistics* 53, 331–40. They show that for continuous time PH models, a test of whether exits to different states are behaviourally distinct (rather than simply incidental) corresponds to a particular set of restrictions: equality of all parameters except intercepts in the models for the destination-specific hazards. (In the Weibull model, the number of restrictions = #(covariates) – 1 (i.e. intercept) + 1 (i.e. shape parameter) = #(covariates).)

The test statistic is

$$2[\ln(L_{\mathrm{CR}}) - \ln(L_{\mathrm{SR}}) - \textstyle\sum_j n_j.\ln(p_j)]$$

where $\ln(L_{\mathrm{CR}})$ is the maximised log-likelihood from the competing risk model (the sum of those from the component models), $\ln(L_{\mathrm{SR}})$ is the maximised log-likelihood from the single-risk model, n_j = number of exits to state j and p_j =

$n_j/\sum_j n_j$, where there are $j = 1,\ldots, J$ destination states. The test statistic is distributed Chi-squared with degrees of freedom equal to the number of restrictions.