**8. Competing risks models**

1. **Aim**

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

1. **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.)

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 | 1,507

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)

1. **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)

1. **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.

logit leftui age famresp tyentry reg1-reg4 logt, nolog

Logistic regression Number of obs = 11,234

LR chi2(8) = 64.33

Prob > chi2 = 0.0000

Log likelihood = -2495.4163 Pseudo R2 = 0.0127

------------------------------------------------------------------------------

leftui | Coefficient Std. err. z P>|z| [95% conf. interval]

-------------+----------------------------------------------------------------

age | -.0034686 .0048337 -0.72 0.473 -.0129425 .0060052

famresp | .0681293 .0854925 0.80 0.426 -.0994328 .2356915

tyentry | .6964857 .0934628 7.45 0.000 .513302 .8796693

reg1 | .1435767 .2052072 0.70 0.484 -.258622 .5457754

reg2 | .0696173 .2055525 0.34 0.735 -.3332581 .4724928

reg3 | -.0692704 .2204173 -0.31 0.753 -.5012804 .3627396

reg4 | -.0150869 .2120189 -0.07 0.943 -.4306364 .4004625

logt | .1662141 .0452708 3.67 0.000 .077485 .2549431

\_cons | -3.440587 .2751854 -12.50 0.000 -3.979941 -2.901234

------------------------------------------------------------------------------

.

. \* estimate the components of the competing risk model

.

. logit job age famresp tyentry reg1-reg4 logt, nolog

Logistic regression Number of obs = 11,234

LR chi2(8) = 159.15

Prob > chi2 = 0.0000

Log likelihood = -1925.1314 Pseudo R2 = 0.0397

------------------------------------------------------------------------------

job | Coefficient Std. err. z P>|z| [95% conf. interval]

-------------+----------------------------------------------------------------

age | -.0023954 .0055883 -0.43 0.668 -.0133482 .0085575

famresp | .0497747 .1006708 0.49 0.621 -.1475364 .2470857

tyentry | 1.093551 .1241995 8.80 0.000 .8501247 1.336978

reg1 | .2316425 .2438092 0.95 0.342 -.2462147 .7094998

reg2 | .1339491 .2451883 0.55 0.585 -.3466112 .6145093

reg3 | -.1794071 .2654604 -0.68 0.499 -.6997001 .3408858

reg4 | -.0160927 .2515802 -0.06 0.949 -.5091809 .4769955

logt | -.2101334 .05197 -4.04 0.000 -.3119927 -.1082741

\_cons | -3.574223 .3239443 -11.03 0.000 -4.209142 -2.939304

------------------------------------------------------------------------------

logit other age famresp tyentry reg1-reg4 logt, nolog

Logistic regression Number of obs = 11,234

LR chi2(8) = 279.43

Prob > chi2 = 0.0000

Log likelihood = -778.78785 Pseudo R2 = 0.1521

------------------------------------------------------------------------------

other | Coefficient Std. err. z P>|z| [95% conf. interval]

-------------+----------------------------------------------------------------

age | -.0083874 .0095695 -0.88 0.381 -.0271432 .0103685

famresp | .0933012 .1599862 0.58 0.560 -.2202661 .4068685

tyentry | .1099328 .1666334 0.66 0.509 -.2166627 .4365282

reg1 | -.1315418 .3740342 -0.35 0.725 -.8646354 .6015518

reg2 | -.1124893 .3703368 -0.30 0.761 -.8383361 .6133575

reg3 | .2332233 .3885612 0.60 0.548 -.5283426 .9947893

reg4 | .0045288 .3873086 0.01 0.991 -.7545821 .7636397

logt | 1.939172 .1617527 11.99 0.000 1.622142 2.256201

\_cons | -8.282012 .6536204 -12.67 0.000 -9.563084 -7.000939

------------------------------------------------------------------------------

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).

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.

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 depvar age famresp tyentry reg1-reg4 logt, nolog baseoutcome(0)

Multinomial logistic regression Number of obs = 11,234

LR chi2(16) = 432.87

Prob > chi2 = 0.0000

Log likelihood = -2698.7776 Pseudo R2 = 0.0742

------------------------------------------------------------------------------

depvar | Coefficient Std. err. z P>|z| [95% conf. interval]

-------------+----------------------------------------------------------------

unemployed | (base outcome)

-------------+----------------------------------------------------------------

Exit\_job |

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 | -3.579488 .3239595 -11.05 0.000 -4.214437 -2.94454

-------------+----------------------------------------------------------------

Exit\_other |

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

------------------------------------------------------------------------------

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.

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.

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)