Hour 6: Competing risk model, independent competing risks, continuous time vs discrete time

Although many commands in Stata seemingly have support for multiple outcomes, the estimation in multiple outcomes model is only possible with independent competing risk model. That is to say when the competing survival processes are not correlated by unobserved heterogeneity, we can treat the one survival process as normal one outcome survival model, while the others are treated as censored. Note however this is only valid for continuous time survival model. For discrete time model, only under some assumptions can we do it similarly. Otherwise, with discrete time, and /or unobserved heterogeneity, one has to resort to special program (either user-written Stata ado files or other program packages) for estimation.

Exercise 6-1: continuous time independent competing risk model

In this exercise, we will have a look for a dataset simulated with two competing outcomes. The two competing risks are simulated with continuous Weibull distributed baselines, and the same univariate normal distributed x and y , with different coefficients. The DGP can be summarized as:

Outcome 1	Weibull with $\alpha = 0.9,\ \lambda = 1,\ x \sim N(0,1),\ y \sim N(0,1),\ \beta = (1,-1)$
Outcome 2	Weibull with $\alpha = 1.2,\ \lambda = 1,\ x \sim N(0,1),\ y \sim N(0,1),\ \beta = (-1,1)$

Recall that the Weibull distribution is formally

$$h(t) = b(t)\exp(X'\beta) = \lambda^{\alpha}\alpha t^{\alpha-1}\exp(X'\beta) = \alpha t^{\alpha-1}\exp(X'\beta + \ln(\lambda^{\alpha}))$$

1. open competing_risk_weibull.dta
2. do summarizing statistics:
   sum;
   tab d;

you see that d now has two values, 1 and 2 to indicate 2 outcomes. There is no censoring in this data.

If we assume that outcome 1 and 2 are independent, it is straight forward to apply standard commands for stset and estimation.

3. **stset t, failure(d==1) id(id);**

Note here, before we only specify failure(d) and Stata treat every non negative value of d as single outcome. But in competing risks, we need to be explicit about which transition we are looking at for each stset. We first treat d==1 as single outcome, and thus d==2

will be treated as censored. Actually this implies that in single risk model, censoring can be considered as a competing risk to event.

4. What if we don't specify failure(d==1) and failure(d==2)?
   stset t, failure(d) id(d);
   If we do that, the Stata will treat all d !=0 and d!=. as single outcome. We can run some sts to see the data at this point. Then the sts only report the aggregated survival and hazard rate, but not transition specific one.

5. corresponding to this stset, we can do familiar stcox and streg

   stcox x y, nohr;
   streg x y, dist(weibull) nohr nocons;

we see that both stcox and streg in this case can perform reasonably well, and the x and y for transition 1 (outcome 1) can be estimated well. Also the parameter for Weibull baseline can be estimated well.

6. now we can re stset the data for outcome 2.
   stset t, failure(d==2) id(id);
   stcox x y, nohr;
   streg x y, dist(weibull) nohr nocons;
   Again these central parameters can be reasonably well estimated.

Exercise 6-2: Discrete time independent competing risks model

The estimation for discrete time competing risks model is different. In the grouped hazard setting, if we assume the transition happens on the boundary of internals (Why we need this assumption?), we can use the cloglog method for estimation each transition separately, treating the competing transition as censored. This is analogous to that of continuous time competing risks model.

We can look at a sample data created with two underlying continuous Weibull distributions. The two Weibull distributed baselines are simulated as in continuous case above, but with exception that, 1) to make duration length longer for competing risk estimation, the scale parameter is changed to 0.05 for both transitions; 2) the total number of individual observations are increased to 2000 to make out enough observations.

Outcome 1	Weibull with $\alpha = 0.9,\ \lambda = 0.05,\ x \sim N(0,1),\ y \sim N(0,1),\ \beta = (1,-1)$	Scale factor: -2.696159046
Outcome 2	Weibull with $\alpha = 1.2,\ \lambda = 0.05,\ x \sim N(0,1),\ y \sim N(0,1),\ \beta = (-1,1)$	Scale factor: -3.594878728

The hazard rate model for each transition is then of the grouped-hazard rate form.

$$
\begin{aligned}
&p(t-1<T<t, K=k \mid T>t-1) \\
&\quad=\int_{t-1}^{t}\left(h_{kt}\exp\left(-\sum_{k}\int_{t-1}^{u}h_{kt}\,ds\right)\right)du \\
&\quad=\int_{t-1}^{t}\left(h_{kt}\exp\left(-\sum_{k}(u-(t-1))h_{kt}\right)\right)du \\
&\quad=\left(1-\exp\left(-\sum_{k}h_{kt}\right)\right)\frac{h_{kt}}{\sum_{k}h_{kt}},
\end{aligned}
$$

for k=1,2. where $h_{kt}=\exp(x'\beta_t+\lambda_t+const_k)$.

1. **use the competing_risk_discrete_weibull.dta**
2. **sum; tab d;**

we see that the maximum length for t is 12, there are almost evenly distributed numbers for d=1 and d=2.

We need to expand the data as we did in single risk case to split spells into as many subspells as t.

3. **expand t;**

Don't forget to replace d=0 for all subspells except the last one.

4. **sort id;**
   by id: replace d=0 if _n ~= _N;

Now we can generate the subspell id/length variable epid

5. **by id: gen epid=_n;**

Generating dummies to represent each subspells according to epid

6. **tab epid, gen(dur);**

at this point, we have created the necessary variables for cloglog estimation. before we move on, we need to preserve the data at current stage, such that when we have made changes, i.e. change d=2 to censoring, we can restore the data later to change d=1 for estimation of outcome 2, without destroy the data in memory.

7. **preserve;**

Now we need to replace d=2 to 0, that is treat the outcome 2 as censored to estimate outcome 1.

8. **replace d=0 if d==2;**
9. **cloglog d x y dur2-dur12;**

We have then got the estimates for coefficients of transition 1, treating transition 2 as censored. We can see that x and y are very nicely estimated. We observe a somewhat

falling baseline, but not clear. The scale factor, which is the estimated constant is also not bad.

We can use restore to recover data to its original state for estimation of transition 2. Each preserve should be paired with one restore command. If we wish to preserve again,

10. restore;
11. preserve;
12. replace d=0 if d==1;
13. replace d=1 if d==2; /* for cloglog estimation, we need dummy value for d */
14. cloglog d x y dur2-dur12;

We also see that the estimation for transition 2 is not bad. A more visible increasing baseline hazard (dur2-dur12) can be seen.

The reason we preserve again before estimation for the outcome 2 is that, even though cloglog and logit are different model specification, when the discrete time interval is "small" enough, they are quite similar. In this exercise, since in DGP, the scale parameter is set to 0.05, which implies that the hazard rate in discrete time internval is rather small. Therefore we can check whether cloglog and logit, and in this case mlogit are similar. We can thus try a multinomial logit estimation, with censoring state as basecategory.

15. restore;
16. mlogit d x y dur2-dur12, basecategory(0);

It turns out that the multinomial logit (estimate 2 transitions together) and cloglog (estimate separately, treating the other transitions as censored) produce very similar estimates. This is indeed interesting to note.