Hour 7: Non-parametric mixed proportional hazard rate model

The essence of non-parametric maximum likelihood estimation of duration model is to avoid arbitrary assumptions on functional form of duration baseline and unobserved heterogeneity.

Because in reality, when we face observational data, we really don't know the underlying data generating process. Traditional methods of modeling are to use known parametric functional formulation on duration baseline, as well as unobserved heterogeneity, without sufficient evidences. Recent development of both methodology and computational power has invited more applications on using totally non-parametric modeling and estimation.

Unfortunately, due to the complexity of likelihood function and maximization routine, it is rare to see many programs that are capable of doing NPMLE on duration models with unobserved heterogeneity.

Stata has no such commands for do NPMLE out-of-the-box. There are very few user-written programs for this purpose either. Among available add-on programs for Stata, there is one hshaz written by Prof. Stephen Jenkins at University of Essex, and another one gllamm by Skrondal and Rabe-Hesketh at UC Berkeley. Note these are for single risk model only.

Using hshaz

The hshaz can be downloaded from http://ideas.repec.org/c/boc/bocode/s444601.html.

The hshaz is written as ado file, and using the Heckman and Singer (1984) method to do NPMLE of proportional hazard rate model with grouped hazard discrete time durations. The idea is to use probability mass points to characterize the unobserved heterogeneity distribution, instead of any parametric functional form. The mass points and associated probabilities are unknown. The maximization routine is based on evaluation of Gateaux derivative. To maximize the Gateaux derivative by adding on mass point, the algorithm is searching for a new linear search direction to maximize the likelihood function. The algorithm terminates when the Gateaux derivative cannot be improved further.

The following example shows the application of hshaz on a simulated data. The data is simulated with a genuine two points discrete distribution of unobserved heterogeneity
$\Pr(v = 1.2) = 0.5,\ \Pr(v = 0.8) = 0.5$
such that E(v)=1.

The data discrete_weibull_with_2pkt.dta has 2000 observation, and has the same distribution for x and y as well as discrete Weibull baseline.

We need to do exactly as we did with the discrete Weibull baseline hazard with gamma distributed unobserved heterogeneity case.

1. open discrete_weibull_with_2pkt.dta
2. des; sum;
3. expand t;
4. sort id;
5. gen epid=_n;
6. replace d=0 if _n ~=_N;
7. tab epid, gen(dur);

   the syntax for hshaz is similar to pgmhaz. For detailed options etc, see help hshaz. We have however some choices of maximization, such as start value, number of support points for unobserved heterogeneity etc.

8. hshaz x y dur2-dur12, id(id) dead(d) seq(epid) nmp(2)

   here, we have an option to specify the number of support points we want hshaz to estimate. Hshaz can estimate the non-parametric unobserved heterogeneity only by preset number of support points. Since in DGP we have exactly a two-point discrete distribution for unobserved heterogeneity, we can e.g. specify that hshaz estimates only two points.

Using gllamm

The gllamm is implemented as ado file. It is a very comprehensive package which is capable of estimate a large class of Generalized Linear Latent And Mixed Model. For detailed information please visit http://www.gllamm.org/.

The syntax and options are very complex. But we can fit the above discrete Weibull model with some simple command. The data structure is exact the same as for hshaz.

9. gllamm d x y dur1-dur12, i(id) ip(fn) nip(2) l(cll) allc , nocons
   i(id) is the id option
   ip(fn) means the discrete mass points can be estimated freely
   nip(2) specifies there are 2 mass points to be estimated
   f(bin) specifies the conditional density of hazard is binomial
   l(cll) specifies the link to be used for conditional density (on frailty) is cloglog
   allc means all output displays including random effect

Non-Parametric Maximum Likelihood Estimation: the "Frisch program"
http://www.frisch.uio.no/NPMLE.html

We at Frisch Centre have developed a program that is tailored for estimation of non-parametric maximum likelihood estimation of multivariate mixture models. The program was originally designed to estimate competing risks models with vector of unobserved heterogeneity which has unknown distribution. We use the non-parametric specification that originated by Heckman and Singer (1984).

The key features of the program:

- estimate competing risks models, with multivariate mixed unknown distribution for unobserved heterogeneity
- no functional form assumptions made on duration dependence and unobserved heterogeneity
- use large scale of dummy variables for piecewise constant duration dependence, use smart technique to code the dummies to reduce the dimension of covariates
- competing risks models allow transition-specific risk sets
- no predetermined number of mass points, start with 1 and add additional one as long as likelihood improves
- factor loading
- suitable for timing-of-event analysis
- written in Fortran, R, and Python, running on Unix/Linux environment
- optimized for multiple cores parallel processing through MPI, linear speedup up to 200 cores
- "No breaks, no safety belts, no fire extinguishers ("scientific quality") *Quote Simen Gaure the programmer*

Sample model specification:

```
*covars
  id          1
  x1          2
  x2          3
  x3          4
  lengde      5
  d           6
  alfa        7
*impdum
  ddur 1 100 1
*model cont
  transitions 2
  transit     d
  spellid     id
  duration    lengde
  risksets    alfa (1,2) (1)

x1
x2
x3
ddur
**transition 1
rc alfa
```

Examples of running THE program

discrete_weibull_with_uobs2.dta: 2000 obs, gamma distributed unobserved heterogeneity

test.dat: from out current Monte Carlo study on heterogeneous treatment effect in NPMLE of competing risks timing-of-event model.