# Hour 5: Continuous and discrete time, single risk parametric unobserved heterogeneity, gamma frailty, pgmhaz

The treatment of unobserved heterogeneity (or frailty) is a common problem is duration analysis. The unobserved heterogeneity can be considered to stem from two sources:1. omitted variables, which can be generically unobservable, such as motivation, ability, determination etc. 2. the measurement errors. It is almost certain that any empirical analysis with observational data will suffer from the impact of unobserved heterogeneity.

It is a known fact that uncontrolled unobserved heterogeneity will base the duration dependence to negative duration dependence (a falling duration baseline). It is also proved by many that the uncontrolled unobserved heterogeneity will bias the structural parameters towards zero. So it is hardly possible to publish any paper in duration analysis without  control for the unobserved heterogeneity.

In duration analysis, it is a known class of model with the name mixed duration model, where we mix the distribution of unobserved heterogeneity with the distribution of duration. Under proportionality assumption, there has been showed that the mixture distribution model can be estimated, and conditional on the mixed distribution of unobserved heterogeneity, the estimates of structural parameters are consistent.

The treatment of unobserved heterogeneity can be difficult, since we must rely on statistical assumption on the underlying unobserved distributions. One often assumes some known distribution of the unobserved heterogeneity, and integrates them out of the likelihood function. That is equivalent to say that the estimation and inference are conditional on the distribution of unobserved heterogeneity. In this exercise, we will focus on the parametric assumptions of the distribution of unobserved heterogeneity.

Stata has limited support for estimation when the unobserved heterogeneity is present. There are a few other user-programmed ado files that can do some extended estimation of complexity model specification. But the built-in support of Stata for such tasks is mainly based on continuous time duration data with parametric distribution of unobserved heterogeneity.

Exercise 5-1:

In this exercise, we will use the continuous_weibull.dta data to do model  estimation with gamma distributed unobserved heterogeneity. The continous_weibull.dta data has the variable *t_uobs* to measure the duration with the presence of gamma distributed unobserved heterogeneity, as well as the corresponding censoring indicator *d_uobs*. The gamma distributed unobserved heterogeneity has the expectation equals to 1 and variance equal to 0.6475 (I made this in the similar way as in Zhang(2003)).

1. open continous_weibull.dta
2. do summarizing statistics on t_uobs d_uobs.
3. **stset** the data

4. the stcox command has the option to specify distribution for frailty (unobserved heterogeneity). One can simply specify option frailty(dist), where the dist is the parametric distribution of frailty, say gamma or Gaussian.
   Let's first run a Cox regression without control for frailty.
   **stcox x y, nohr;**
5. now let's assume the underlying distribution for frailty has the gamma distribution. In Stata, the gamma distribution is parameterized with a single nuisance parameter $\theta$, which is the variance of gamma distribution (with expectation 1). Notice however that when in Stata we require the frailty distribution, we have to specify the share() option as well. This means that the frailty (or unobserved heterogeneity) should be common for the shared(varname). This is further understood in this way: the shared frailty is the within group correlation. Observations within group have a common latent random effect (frailty, unobserved heterogeneity). If we specify shared(id), this means that all spells from the same id should have the same frailty. This is intuitively easy to understand, because if we have multiple spells for single individual, the individual specific unobserved heterogeneity (frailty) is of course the same for the same individual across spells.

   To require the gamma distribution for frailty for stcox, we simply do this:
   **stcox x y, nohr frailty(gamma) shared(id);**

   We see that comparing with the results from stcox with no control for frailty, the estimated x and y are closer to the values in DGP, which confirms the statement that uncontrolled heterogeneity will bias the estimates towards zero (notice the word towards!). But the estimates are still biased.

   Note also the maximization log. The stcox first fits a log profile likelihood, which means it first maximize the likelihood with respect to $\theta$. Then for the optimized $\theta$, the Stata maximize the (penalized) likelihood to get the coefficients for x and y. For further details, see Methods and Formulas for stcox.

   We can also do other conventional stuff with stcox, such as stcurve etc. But the main point in this exercise is to see how the stcox is dealing with unobserved heterogeneity, and verify the fact that uncontrolled unobserved heterogeneity biases the parameters towards zero.

   Let us do the similar exercise with streg dist(weibull)
   **streg x y, dist(weibull) nohr nocons;**

   See again how the estimates are biased towards zero comparing to those of DGP (1, -1), and the estimated duration baseline parameter p.

   if we run
   **streg x y, dist(weibull) nohr nocons frailty(gamma)**

we see that he estimates are closer to the true value in DGP, and in fact the confidence intervals cover the DGP value, and the duration baseline parameter now is closer to 0.9. The estimated theta, which is the variance of gamma distriubuted frailty is close to that of DGP. The Stata streg with weibull distribution and failty(gamma) seem to do a reasonably well job.

Exercise 5-2: discrete time grouped-hazard model with gamma frailty

If we have a discrete time duration model, we know that the model misspecification can derive from: 1. misspecify the duration baseline distribution 2. misspecification of unobserved heterogeneity. The Stata does not have a good tool for estimate discrete time duration model, yet of the mixed discrete time duration model estimation.

Fortunately there are quite a few experts out there who wrote the special ado files for Stata which can do the excellent job for this task. Professor Stephen Jenkins is the professor of sociology at Universities of Essex, and a senior Stata programmist. He wrote an excellent Stata program for estimation of grouped hazard single risk duration model with gamma distributed frailty: **pgmhaz**. In fact he has an excellent lecture on survival analysis with Stata maintained at [http://www.iser.essex.ac.uk/teaching/degree/stephenj/ec968/](http://www.iser.essex.ac.uk/teaching/degree/stephenj/ec968/) The **pgmhaz** can be downloaded from that link. To install, follow the instruction given above, or simply copy the pgmhaz.ado, pgm_ll.ado (the definition file for the log likelihood function) and pgmhaz.hlp to your installation of personal or site ado folder. Or any folder, and use e.g.

**adopath + c:\temp**

to add this folder to Stata system folders. You can check ado folder path by **sysdir** command.

The **pgmhaz** requires that the data is organized as such that each individual will have as many rows as that of the total length of spell. Each row corresponds to single period of time of spell duration. There has to be a seq variable which holds the id of subspells splitted from the original spell. This data structure is exactly we did in exercise 4-1. so the data preparation is exactly the same.

1. open data file discrete_weibull_with_uobs.dta.
2. describe the variables
3. do a summarizing statistics

we need to define 12 dummies for piecewise constant baseline. This is exactly the same as we did before. Each individual contribute different numbers of periods to the overall durations. Thus we need to split each individual's duration spell to subspells, so that they sum up to the original total length. Then for each subspell we define the corresponding dummy to reflect which subspell/period this is referring to. Each subspell will have the censoring indicator to be 0, meaning censored, except the last subspell for individual will retain its original censoring status.

4. episode splitting by expand: **expand t;** (duplicate each observation by t, the total duration period. Thus each subobservation generated reflects one period of the original duration, and they sum up to original t).
   ```
   sort id;
   by id: replace d =0 if _n!=_N;
   ```
5. we will need to define each subspell's id.
   ```
   by id: gen epid=_n;
   ```
   note that epid[_n]-epid[_n-1]=1 is the exact length of subspell.
6. Since we are assuming that for each subspell the duration baseline is constant (piecewise constant) we will have to define a set of dummies to indicate which period the subspell actually is corresponding to. There are 12 dummies in total.
   ```
   tabulate epid, gen(dur);
   ```
7. now we can do a cloglog to estimate the model ignoring the unobserved heterogeneity.
   cloglog depvar indepvar, options;
   here:
   ```
   cloglog d x y dur2-dur12; /* note here we drop dur1 to estimate with a constant,
   alternatively */
   cloglog d x y dur1-dur12, noconstant
   ```

   the pgmhaz syntax is: **pgmhaz indepvar, id() dead() seq(),** where id() is the individual id, dead() is the censoring indicator, where d=1 means a transition, and d=0 means censoring, seq() is the subspell id. Actually pgmhaz do a cloglog first to estimate model without unobserved heterogeneity.

   ```
   pgmhaz  x y dur2-dur12, id(id) dead(d) seq(epid);
   ```

   we see that the pgmhaz first fit a cloglog model without unobserved heterogeneity. This is exactly what cloglog itself in Stata does. Then the pgmhaz tries to fit the full model with gamma distributed unobserved heterogeneity, with the parameter $\theta$. we observe that the estimates for x and y are larger in absolute value comparing to cloglog, which confirms that the unobserved heterogeneity will bias the structural parameters towards zero (cloglog). We also see that the dur1-dur12 estimates are somewhat less negative, which also in accordance with the known theory that unobserved heterogeneity will bias the duration dependence towards negative duration dependence. However the likelihood ratio test did not seem to reject the null hypothesis of no difference between cloglog model and full model.

   This shows also that pgmhaz is useful when we wish to handle the gamma distributed unobserved heterogeneity in discrete duration time model. But the command is somehow not fully to the expectation. This might be due to the limitation of data (too few observations and too few variation of duration dummies.) and limitation of Stata's maximization routine. Then again, if we apply pgmhaz on large dataset, the computational cost might be overwhelming (it will take quite a long time to run).

To further explore the pgmhaz, let's do another exercise. This time we use the data discrete_weibull_with_nor.dta, where there are 2000 individuals, and the unobserved heterogeneity is simulated with a lognormal distribution.

Note the formulation of mixed proportional hazard rate, the unobserved heterogeneity *v* enters as multiplicative term in proportional hazard,

$$h(t \mid x, v) = b(t)\lambda(x)v$$

and in complementary loglog modeling,

$$p(t-1 < T < t \mid T > t-1) = 1 - \exp\left(-\exp(x_t'\beta + \lambda_t + \ln(v))\right)$$

So in DGP, we simulate

$$\delta = \ln(v) \sim N(\mu, \sigma^2)$$

and thus *v* has a lognormal distribution. We choose $\mu = -0.5, \sigma^2 = 1$ such that $E(v) = 1, Var(v) = 1.718$ (See e.g. Sydsæter, Strøm, and Berck (1999)) for lognormal distribution formula.

Now we can use pgmhaz to see if it performs in a misspecified case, where in DGP we have lognormal distributed unobserved heterogeneity, and we proximate it with gamma.

```
pgmhaz  x y dur2-dur12, id(id) dead(d) seq(epid);
```

Well, comparing with no control for unobserved heterogeneity, the pgmhaz moves the coefficients to x and y to towards the right direction. But it is far from recovering the DGP parameters.

Another interesting possibility is that, in fact the discrete data used for cloglog and pgmhaz above comprises a very nice random effect panel data structure, where we treat the censoring indicator d as dependent variable, and assume the (latent) random effect to have a normal distribution. We can use

```
xtcloglog d x y dur2-dur12, i(id) re;
```

this produces a somewhat similar estimates for x y and dur1-dur12 as in pgmhaz, except that xtcloglog returns the distributional parameters for normal distributed unobserved heterogeneity.

The above exercises show the possibility to handle the unobserved heterogeneity. It demonstrates: 1. uncontrolled unobserved heterogeneity biases the structural parameters towards zero, and duration baseline towards negative duration dependence. 2. with continuous time duration data, Stata's built-in option frailty()

can be used to fit models with gamma or normal distributed frailty. 3. with discrete time duration data, pgmhaz is a nice tool to try, or if one is willing to assume normality of frailty, xtcloglog can also be applied. However, one would need to do a serious Monte Carlo study to be able to say with some certainty which method is most suitable for empirical analysis. And we should note that in this exercise, the number of observations is very small. You can try with larger number of observations to improve both identifiably and significances of estimates.