Hour 3: Continuous time duration estimation, partial likelihood, parametric baseline hazard

In this hour we will try some of the most common duration estimation methods.

1. Cox proportional hazard rate estimation

A central assumption in hazard rate model is the proportionality assumptioan, which means the baseline hazard (when all covariates take value 0) is proportional for all observations.

$$h(t \mid x) = b(t) \underbrace{\lambda(x)}_{\exp(x'\beta)}$$

This is due to Cox (1972). But Cox model, or Cox method, is a much misunderstood term. In fact it is actually meant the Cox partial likelihood estimation method, based on the proportionality assumption of hazard rate.

$$P(T_i = t \mid i \in N_t) = \frac{b(t)\exp(x_i'\beta)}{\sum_{j=N_t} b(t)\exp(x_j'\beta)} = \frac{\exp(x_i'\beta)}{\sum_{j=N_t} \exp(x_j'\beta)}$$

Here in the probability observation *i* contributing to the overall likelihood, the baseline is cancelled out. Thus the estimation on coefficients to covariates can be carried out without the consideration of duration baseline.

In Stata, a special command is dedicated to Cox partial likelihood estimation. stcox has many features, which is useful as the first exploration of model parameters.

Before use stcox, we have to stset the survival data.

Exercise 3-1: Interactive mode.

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

   You will see that there are two variables x and y. these two are structural covariates. In DGP, both x and y are standard normal distributed. The coefficients connected to x y are 1 and -1. t is the duration time (we from now on assume start time is 0), d=1 if there is a transition out, d=0 means censoring. id is the individual id.

4. stset data
5. do Kaplan-Meier plot to explore both the empirical survival function and empirical hazard.
6. do a Cox partial likelihood estimation on x and y: **stcox x y**;

   Note that stcox reports the hazard rate ratio for x and y. The hazard rate ratio is the proportional change on hazard for 1 unit change of the variable. Here the hazard rate

ration is just the exponential of estimated coefficients. If we want to have the exact estimated coefficient, we will need to specify option nohr: stcox x y, nohr; The estimates are not far from the DGP values, in fact the 95% confidence intervals cover the DGP values. So we can recover the interesting structural parameters without having to consider the functional form of baseline.

7. we can use stcox to estimate the hazard, the cumulative hazard rate and survival function.

stcox x y, basehc(basehc) nohr; /* this is to create variable basehc to contain baseline hazard. */
stcox x y , basechazard(basechaz) nohr; /* this is to create variable basechaz to contain cumulative baseline hazard. */
stcox x y, basesurv(basesurv) nohr; /* basesurv is the variable contains survival function */

8. once we have done stcox x y, basechazard(basechaz), we can use **stcurve** to produce graphs of cumulative baseline hazard, hazard rate, or survival functions. But remember, the stcurve must follow stcox immediately because stcurve uses the latest estimation results from stcox.

stcox x y, basehc(basehc) nohr;
stcurve, hazard;
stcox x y , basechazard(basechaz) nohr;
stcurve, cumhaz;
stcox x y, basesurv(basesurv) nohr;
stcurve, survival;

Also note that the stcurve plots the survival, hazard and cumulative hazard at baseline, meaning it evaluates the functions at mean value of covariates.

9. write a do file to do all these above, plus redo all these using continous_weibull.dta, save logs and/or graphs.

2. Parametric baseline and estimation.

The parametric baseline specification is to provide a convenient functional form for the baseline. The popular functional forms include: exponential, Weibull, log logistic, Gompetz etc.

For parametric estimation of duration model, Stata provides the powerful **streg** command. It is the similar syntax as in stcox, but you have the opportunity to specify which baseline distribution you wish to assume. The central parameters for each distribution you specify are estimated together with structural parameters.

Exercise 3-2: Interactive mode.

1. open data file continous_weibull.dta.

2. describe the variables
3. do a summarizing statistics

The continous_weibull.dta is simulated with the same distribution of x and y as in constant_hazard.dta, except that the baseline is simulated with a Weibull distribution, with scale parameter set to 1.

In fact, the full formulation of Weibull distribution is
$h(t) = b(t)\exp(X'\beta) = \lambda^{\alpha}\alpha t^{\alpha-1}\exp(X'\beta)$, where α is the shape parameter:
$\alpha > 1$ implies positive duration dependence; $\alpha < 1$ implies negative duration dependence.
If $\alpha = 1$, the Weibull distribution reduces to exponential distribution.

Set $\lambda = 1$, we get $h(t) = b(t)\exp(X'\beta) = \alpha t^{\alpha-1}\exp(X'\beta)$
The data is simulated with
$X = (x, y),\ where\ x \sim N(0,1),\ y \sim N(0,1),\ \alpha = 0.9, \beta = (1,-1)$

4. **stset t , id(id) failure(d);**
5. first let's do a Cox regression to get a feeling of what the estimated coefficients could be: **stcox x y;**
6. now run a formal streg with Weibull distribution

streg x y, distribution(weibull);

You will see that the streg reports the default estimates in hazard ratio form. If we wish to compare directly to DGP parameters, we can use nohr option.

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

It is interesting to see that the default streg , dist(weibull) estimates with a constant term. This constant term is in fact the estimate (log form) for the scale parameter. We can suppress the constant term by noconstant option, which is equivalent to assume $\lambda = 1$.

streg x y, dist(weibull) nohr nocons;

The estimates for x and y are both rather good. The estimated confidence intervals cover the true DGP value, and the estimates are almost the same as in Cox results.

7. Stata report two ancillary parameter estimates, p and 1/p. Actually p is the same as α in the model specification. The /ln_p is in fact the estimate we got from maximum likelihood estimation. So p=exp(/ln_p) is the estimate for α, the shape parameter in Weibull distribution. It is actually not bad.

8. again we can use **stcurve** to produce fitted hazard function (Remember that stcurve plots the hazard function at mean value of covariates) cumulative hazard and survival functions.

Let's do a streg with constant first.

streg x y, dist(weibull) nohr;

then

stcurve, hazard;

we can specify the stcurve to plot at specific value of covariates. Say at x=0 and y=0.

stcurve, hazard at1(x=0) at2(y=0);

To make figure looks nice, we can specify the range of time for plot. By reviewing statistics of duration t, we find most duration length less than 1.

stcurve, hazard range(0 1);

In fact since the x and y in data have approximately mean values at 0, this is the nice baseline hazard for Weibull distribution.

Note however, when nocons option is provided, stcurve does not seem to plot anything! Why?

streg x y, dist(weibull) nohr nocons;
stcurve, hazard range(0 1);

Because stcurve uses the mean values of covariates for calculate survival and hazard function. In our case, the mean values for x and y are ca 0. If no constant term, the hazard rate is approximately 0.

Compare graphs with that produced by Cox regression.

9. now suppose we don't know a priori the distribution of baseline in DGP, and we fit the data with a wrong distribution type. What would happen?

streg x y, dist(exponential) nohr nocons;

from the result you will see that the x is biased and the confidence interval does not cover the DGP value. This is not surprising since we have a misspecified model.

10. write a do file and do all these again with the constant_hazard.dta data. Note however that the exponential distribution is a special case of Weibull with p=1.

In Stata Reference menu for Survival Analysis and Epidemiological Tables, pp 237. there is an interesting discussion of model selection, based on information criteria. It is worth reading.