10th German Stata Users Group Meeting, Berlin, 1 June 2012

Can multilevel multiprocess models be estimated using Stata?

A case for the cmp command

Tamás Bartus

Corvinus University of Budapest
Demographic Research Institute, HCSO, Budapest

Intro

- Multilevel multiprocess models (MLMP henceforth) became popular among demographers who are concerned by issues of endogeneity and self-selection.
- In a nutshell, MLMP models consist of equations of hazards which include correlated heterogeneity components (Lillard 1993)
- The **cmp** command, written by David Roodman (2011) and available through the SSC archive, is a general framework to estimate models with various link functions jointly.
- I think **cmp** was not written for researchers interested in survival analysis
- In this talk, I hope I will show that **cmp** has the potential of enhancing the survival analysis capabilities of Stata

Overview

- Description of MLMP models
- How to estimate MLMP models using the **cmp** command
- Example: the effect of marriage and premarital cohabitation on the birth and the union dissolution processes

Description of MLMP models

What are MLMP models?

- Multilevel multiprocess models (MLMP henceforth) were invented to control for selection biases which arise from the fact that some unobserved personality traits affect several outcomes.
- Classic example: Does premarital cohabitation reduce the hazard of separation? (Lillard, Brian and Waite 1995)
- Theory suggests yes. Premarital cohabitation should reduce uncertainties about match quality, and thereby the hazard of marital disruption.
- Surprisingly, earlier studies were not able to support this theory.
- Lillard, Brian and Waite (1995) argue that this is due to the presence of a selection effect: premarital cohabitation is chosen by couples who have pessimistic expectations about the duration of the marriage.
- MLMP models solve the endogeneity problem as follows....

More details on MLMP models

- MLMP models consists of equations for hazards which include individual-specific heterogeneity terms:

$$\begin{aligned}\ln h_{ij}^{(1)} &= \beta^{(1)} X_{ij}^{(1)} + u_i^{(1)} \\ \ln h_{ij}^{(2)} &= \beta^{(2)} X_{ij}^{(2)} + u_i^{(2)}\end{aligned}$$

  where *i* indexes individuals and *j* indexes episodes. Vector *X* might include time-dependent variables as well as separate forms of duration dependencies.

- The equations are seemingly unrelated in the sense that the individual-specific residuals (the *u*s) are assumed to be correlated.
- Applied MLMP models usually fall into the following categories
  (1) Simultaneous equations for hazards
  (2) Hazard models with endogenous regressors
  (3) Hazard models with sample selection

(1) Simultaneous equations for hazards

- Hazard of an event depends not only on observed characteristics but also on the hazard of another event
- If we model the latter hazard using another set of variables, we arrive at the reduced-form model

$$
\begin{aligned}
\ln h_{ij}^{(1)} &= \beta^{(1)} X_{ij}^{(1)} + u_i^{(1)} \\
\ln h_{ij}^{(2)} &= \beta^{(2)} X_{ij}^{(2)} + u_i^{(2)}
\end{aligned}
$$

- Examples include the joint modeling of the hazards of
  - conceptions within marriages and marital disruption (Lillard 1993, Lillard and Waite 1993)
  - giving birth to first, second, and third child (Kravdal 2001)

(2) Hazard models with endogenous explanatory variables

- Hazard of the event under study is affected by an endogenous dummy variable. Therefore, the hazard in question and the occurrence of the endogenous dummy are modeled jointly:

$$\begin{aligned} \ln h_{ij} &= \alpha y_{ij} + \beta^{(1)} X_{ij}^{(1)} + u^{(1)} \\ y^*_{ij} &= \beta^{(2)} X_{ij}^{(2)} + u^{(2)} \end{aligned}$$

- Examples include the modeling of
  - the effect of premarital cohabitation (y_{ij}) on marital stability (h_{ij}) (Lillard, Brien and Waite 1995)
  - the effect of hospital delivery (y_{ij}) on child mortality (h_{ij}) (Lillard and Panis 2003)

(3) Hazard models with sample selection

- Hazard of the event under study can be examined using the sample of individuals who have experienced another event. That sample is not a random one, thus the hazard under study and the occurrence of the latter event are modeled jointly:

$$\begin{aligned} \ln h_{ij} &= \beta^{(1)} X_{ij}^{(1)} + u^{(1)} \quad \text{if } y_{ij}^{*} > 0 \text{ or } y_{ij} = 1 \\ y^{*}_{ij} &= \beta^{(2)} X_{ij}^{(2)} + u^{(2)} \end{aligned}$$

- Examples include the joint estimation of:
  - the hazard of second birth using the sample of mothers with a probit model of being a mother (Kreyenfeld 2002).
  - the effect of marital status on household income using the PSID, together with a model of panel attrition (Lillard and Panis 1998) (OK, this is not a hazard model with sample selection...)

Softwares for MLMP models

- MLMP models are usually estimated using specialized software like aML and MLwiN.
- aML was designed for MLMP modeling
  - supports continuous-time piecewise exponential hazard models, as well as several other models
  - cumbersome syntax, and lack of post-estimation tools
- MLwiN is for multilevel modeling
  - supports the estimation of multilevel discrete-time event-history models
  - recently, a stata command was developed to run MLwiN from within Stata (**runmlwin**, written by George Leckie and Chris Charlton)
- But can MLMP models be estimated using Stata?

Estimation of MLMP models using cmp

The case for the cmp command

- The user-written **cmp** command (Roodman 2011) allows one to estimate systems of seemingly unrelated recursive equations
  - the residuals must be normally distributed
  - the seemingly unrelated equations may include random intercepts
- How can MLMP models be estimated using **cmp**?
  - **cmp** supports, among others, interval-censored regressions.
  - the lognormal survival model is just an interval-censored regression of log failure times
  - The estimation of MLMP models therefore boils down to estimating lognormal survival models jointly with other lognormal survival models or probit models for endogenous regressors..

How to estimate MLMP models using cmp

- Reformulate the proportional hazard model as a model log failure times (*τ)*:

$$\begin{aligned} \ln t^{(1)} &= \beta^{(1)} X_{ij}^{(1)} + u_i^{(1)} + \epsilon_{ij}^{(1)} \\ \ln t^{(2)} &= \beta^{(2)} X_{ij}^{(2)} + u_i^{(2)} + \epsilon_{ij}^{(2)} \end{aligned}$$

  The signs of coefficients are the opposite to those of the hazard model.

- The multilevel error structure can be ignored since the correlation of the *u*s implies the correlation of the total errors

$$\begin{aligned} e_{ij}^{(1)} &= u_i^{(1)} + \epsilon_{ij}^{(1)} \\ e_{ij}^{(1)} &= u_i^{(2)} + \epsilon_{ij}^{(2)} \end{aligned}$$

- Why can the multilevel error structure be ignored? Well, estimation of multilevel models takes considerable time and convergence issues are more likely to occur. Besides, there are theoretical arguments...

Why can the multilevel error structure be ignored?

- In order to control for the presence of selection effects, seemingly unrelated equations must be estimated.
- Researchers often model processes using proportional hazard models, in general, and piecewise-exponential regression equations, in particular. Exponential regression equations are, however, unrelated, not seemingly unrelated. (This is a key implicit assumption of MLMP models)
- Repeated observations and the addition of a person-specific heterogeneity (or shared frailty) term turns a system of unrelated proportional hazard equations into a system of seemingly unrelated equations.
- Equations with Gaussian errors, in general, and lognormal failure time models, in particular, may constitute a system of seemingly unrelated equations.
- Repeated episodes and multilevel modeling are not a must. But if they are repeated episodes, the sandwhich variance estimator accounts for clustering.

Multiprocess data structure

- MPML modeling requires a multiprocess data structure. The starting point is the multispell data structure. Consider the life history of Lady Diana

Event	Date
marriage	july 1981
conception	october 1981
conception	january 1984
separation	august 1996

- Setting up the multiprocess data structure proceeds in two steps:

1. Create a dataset of durations, with optional failure indicators for each process; then
2. Define failure indicators and dependent variables for each process

Step 1. Dataset of durations, process-specific indicators

- Durations are just the differences between dates and lagged dates. (Duration will be missing for the first record, it can be replaced by making assumptions about the beginning of the risk periods)
- It might be useful to add separate indicators for process-specific failures and censoring. We consider the birth and the marital disruption processes.
- The following duration and indicator variables are obtained

Event	Date	Duration (in months)	birth	separation
marriage	july 1981	.	0	0
conception	october 1981	3	1	0
conception	january 1984	27	1	0
separation	december 1992	107	0	1

Step 2. Dependent variables

- Interval-censored regressions require two dependent variables, labeled the lower and upper limits, which define the intervals within which the true value of log duration lies.
- Due to a minor, and hopefully temporary, bug in cmp, right-censoring should be coded as a large positive number instead of infinity.
- For the birth process, the lower and upper limits are generated as follows:

```
gen bdurlo = cond(birth==1,ln(dur-.9),ln(dur))
gen bdurhi = cond(birth==1,ln(dur)   , 999    )
```

- For the marital disruption process, the lower and upper limits are

```
gen mdurlo = cond(separation==1,ln(dur-.9),ln(dur))
gen mdurhi = cond(separation==1,ln(dur)   , 999   )
```

- Now we are in a position to estimate a MLMP model....

Syntax of cmp. Intro

- The syntax for a single-equation interval regression of log duration to conceptions is

```
cmp ( birth : bdurlo bdurhi = varlist ) , indicators(7)
```

  - **birth :** is optional: it instructs **cmp** to use birth to label the equation
  - the dependent and independent variables must be separated by the equal sign. The lower limit comes first.
  - The **indicators(7)** option means that this equation is interval regression

- The syntax for a single-equation probit model of conception would be

```
cmp ( birth = varlist ) , indicators(4)
```

  - **indicators(4)** means that the equation is probit

cmp syntax for MLMP models

(1) Simultaneous equations for hazards

```
cmp ( birth      : bdurlo bdurhi = varlist_1 ) ///
    ( disruption : mdurlo mdurhi = varlist_2 ) , indicators(7 7)
```

- Two seemingly unrelated interval-censored equations will be estimated.
- Note that number 7 appears twice in the indicators() option

(2) Hazard models with endogenous regressors

```
cmp ( birth : bdurlo bdurhi = married varlist_1 ) ///
    ( married = varlist_2 ) , indicators(7 4)
```

- married is a dummy indicated married individuals
- **indicators(7 4)** means „the first equation is an interval-censored equation, the second equation is a probit one."

cmp syntax for MLMP models

(3) Hazard models with sample selection

```
cmp ( birth : bdurlo bdurhi = married varlist_1 )     ///
    ( married = varlist_2 ) , indicators(”married*7” 4)
```

- The indicator option allows expressions. Expressions should be enclosed in double quotes.
- Observations where the expression evaluates 0 are not used to estimate the equation
- Thus, **indicators(”married*7” 4)** means that the first equation is an interval-censored one to be estimated using the sample of married individuals, and the second equation is a probit one to be estimated using all individuals.

Example

Example

- To illustrate the use of cmp, I will present a sample research. The questions we will examine are:
  (1) Does marriage and premarital cohabitation reduce the waiting time to conceptions?
  (2) Does marriage and premarital cohabitation increase the waiting time to union dissolution?
- Data taken from the *Turning Points of the Life Course* panel survey, conducted by the Demographic Resarch Institute of the Hungarian Central Statistical Office.
- Three waves administered in 2001, 2004 and 2008.
- Retrospective event history data collected.
- Our sample include women born 1946-1983. We use episodes where respondents were either cohabiting or married.

Variables

- Dependent variables
  - bdurlo and bdurhi record log durations for the birth process
  - mdurlo and mdurhi record log durations for the dissolution process
- Independent variables
  - married (1 if married, 0 if consensual union)
  - premarital cohabitation (1 if marriage preceeded by cohabitation, 0 otherwise)
  - education (1=primary, 2=vocational, 3=secondary, 4=higher)
  - birth year
- The dataset should contain an ID variable for persons, idcode

Session 1. Effect of marriage on time to conception

- First, we define two global macros for later purposes

```
global mdur  mar cohab ib0.nchild byear ib4.edu
global sel   byear ib4.edu
```

- We begin with estimating the conception equation separately

```
cmp ( birth : bdurlo bdurhi = $mdur ) ///
    , ind(7) vce(cluster idcode)
```

- To control for the effect of unobserved factors, we proceed with estimating the conception equation jointly with the a model of union dissolution

```
cmp ( birth      : bdurlo bdurhi =  $mdur ) ///
    ( disruption : mdurlo mdurhi =  $mdur ) ///
    , ind(7 7) vce(cluster idcode)
```

Session 1. Effect of marriage on time to conception

- Another way of controlling for unobserved factors is to treat marriage as an endogenous regressor

```
cmp ( birth    : bdurlo bdurhi = $mdur ) ///
    ( married  : mar           = $sel  ) ///
    , ind(7 4) vce(cluster idcode)
```

- The two approaches may be combined resulting in our last model:

```
cmp ( birth      : bdurlo bdurhi  = $mdur ) ///
    ( disruption : mdurlo mdurhi  =  $mdur )  ///
    ( married    : mar            = $sel  ) ///
    , ind(7 7 4) vce(cluster idcode)
```

Waiting time to conceptions. Results

```
--------------------------------------------------------------------
             Variable |  Separate       Joint        Endog.     Joint+Endog
----------------------+---------------------------------------------
birth                 |
              Married |  -0.743***    -0.764***    -1.474***    -1.963***
Premar. cohabitation  |  -0.183***    -0.196***    -0.178***    -0.232***
               nchild |
                    1 |   0.778***     0.781***     0.779***     0.791***
                    2 |   1.634***     1.648***     1.641***     1.688***
 (Birth year - 1946)  |   0.009***     0.009***    -0.001       -0.006*
                  edu |
                    1 |  -0.434***    -0.435***    -0.447***    -0.458***
                    2 |  -0.264***    -0.265***    -0.221***    -0.198***
                    3 |  -0.146**     -0.146**     -0.126*      -0.108*
             Constant |   4.145***     4.174***     4.879***     5.411***
----------------------+---------------------------------------------
lnsig_1               |                 (output omitted)
disruption            |                 (output omitted)
lnsig_2               |                 (output omitted)
married               |                 (output omitted)
----------------------+---------------------------------------------
atanhrho_12           |               -0.094        0.252**     -0.198***
atanhrho_13           |                                          0.388***
atanhrho_23           |                                          0.921***
--------------------------------------------------------------------
                   legend: * p<.05; ** p<.01; *** p<.001
```

Interpretation

- Married women wait less to conceptions, but separate modeling underestimates the magnitude of the negative effect:

```
----------------------------------------------------------------------------
             Variable |   Separate       Joint        Endog.      Joint+Endog
---------------------+------------------------------------------------------
              Married |  -0.743***    -0.764***    -1.474***    -1.963***
---------------------+------------------------------------------------------
```

- In the Joint model, the correlation of residuals is not significant. In the Joint model with endogenous marriage, the correlation is negative and significant.
- The residuals of the birth and marriage equations are positively correlated.

```
---------------------+------------------------------------------------------
atanhrho_12           |                -0.094        0.252**     -0.198***
atanhrho_13           |                                           0.388***
atanhrho_23           |                                           0.921***
----------------------------------------------------------------------------
```

Interpretation

- There are unobserved characteristics
  - which favor childbearing (shorter time to conception) and stabilize unions (longer time to dissolution).
  - which increase the waiting time to conception and the probability of being married
- The second conclusion is a bit strange, it might be an artifact of specification error. (The small research reported here surely needs improvements)

Session 2. Effect marriage on union dissolution

- Again, we begin with a separate estimate of the separation process

```
cmp ( disruption  :  mdurlo mdurhi  = $mdur ) ///
    , ind(7) vce(cluster idcode)
```

- The next model is the joint estimation of the marriage and conception equations. This model is already estimated (see Session 1)
- In the third model, marriage is an endogenous regressor

```
cmp ( disruption   : mdurlo mdurhi  = $mdur ) ///
    ( married      : mar           = $sel  ) ///
    , ind(7 4) vce(cluster idcode)
```

- And our last model is the joint estimation of the marriage and conception equations together with the probit model of marriage. This model is already estimated.

Time to separation. Results

```
---------------------+-----------------------------------------------------------
Variable             |   Separate       Joint         Endog.       Joint+Endog
---------------------+-----------------------------------------------------------
disruption           |
             Married |   1.823***      1.886***     -0.036        -0.679**
Premar. cohabitation |   1.544***      1.584***      1.531***      1.648***
              nchild |
                   1 |   0.279***      0.258**       0.281***      0.182*
                   2 |  -0.220        -0.266*       -0.179        -0.351*
 (Birth year - 1946) |  -0.008*       -0.009*       -0.041***     -0.062***
                 edu |
                   1 |  -0.061        -0.051        -0.107        -0.051
                   2 |  -0.058        -0.051         0.091         0.222
                   3 |  -0.194*       -0.196*       -0.125        -0.080
            Constant |   4.173***      4.271***      6.214***      7.641***
---------------------+-----------------------------------------------------------
lnsig_1              |                (output omitted)
birth                |                (output omitted)
lnsig_2              |                (output omitted)
married              |                (output omitted)
---------------------+-----------------------------------------------------------
atanhrho_12          |                -0.094         0.706***     -0.198***
atanhrho_13          |                                             0.388***
atanhrho_23          |                                             0.921***
---------------------+-----------------------------------------------------------
                      legend: * p<.05; ** p<.01; *** p<.001
```

Interpretation

- The simple separate modeling shows that marriage stabilizes unions. But if marriage is assumed to be endogenous, marriage has either no significant effect or it surprisingly reduce the failure time to separation.

```
---------------------+----------------------------------------------------
Variable             |   Separate       Joint        Endog.     Joint+Endog
---------------------+----------------------------------------------------
              Married |   1.823***     1.886***    -0.036        -0.679**
---------------------+----------------------------------------------------
```

- This contradiction arises because women who tend to live in long-term relationships have higher chances of being married.

```
---------------------+----------------------------------------------------
atanhrho_12          |               -0.094       0.706***     -0.198***
atanhrho_13          |                                          0.388***
atanhrho_23          |                                          0.921***
---------------------+----------------------------------------------------
```

Conclusions

- David Roodman's **cmp** package allows one to estimate survival models for parallel processes with endogenous regressors.
- However, the MLMP models which can be estimated using Stata differ from MLMP models that were originally proposed by Lillard and others
  - The Lillard model includes the popular exponential survival models
  - **cmp** supports the less popular lognormal survival models
- This talk focused on the multiprocess feature of MLMP models, but ignored the multilevel aspect (although **cmp** allows for multilevel equations)
  - fitting systems of multilevel equations is painfully slow (and when I prepared this presentation, I was not patient enough....)
  - multilevel modeling is demanded only by models belonging to the exponential family

References

Kravdal, O. 2001. The high fertility of college educated women in Norway: An artefact of the separate modeling of each parity transition. *Demographic Research* 5: 187-216.

Kreyenfeld, M. 2002. Time-squeeze, partner effect or self-selection? An investigation into the positive effect of women's education on second birth risks in West Germany. *Demographic Research* 7: 15-48.

Lillard, L.A. 1993. Simultaneous equations for hazards: marriage duration and fertility timing. *Journal of Econometrics* 56: 189–217.

Lillard, L.A., M.J. Brien and L.J. Waite. 1995. Premarital cohabitation and subsequent marital dissolution: a matter of self-selection. *Demography* 32: 437–57.

Lillard, L.A. and L.J. Waite. 1993. A joint model of marital childbearing and marital disruption. *Demography* 30: 653–81.

Lillard, L. A. and C. W. A. Panis. 1998. Panel Attrition from the Panel Study of Income Dynamics: Household Income, Marital Status,and Mortality. *The Journal of Human Resources*, 33: 437-457.

Lillard, L. A. and Panis, C. W. A. 2003. aML Multilevel Multiprocess. Statistical Software, version 2.0. EconWare, Los Angeles, California.

Roodman, D. 2011. Estimating fully observed recursive mixed-process models with cmp. *Stata Journal* 11: 159-206.