2015 German Stata Users Group meeting
26 June 2015, Nürnberg

Multiprocess modeling with Stata

Tamás Bartus

Corvinus University of Budapest

The preparation of the presentation was supported by the Janos Bolyai Scholarship of the Hungarian Academy of Sciences.

Overview

1. A short overview of multiprocess models
2. Estimating systems of survival equations
3. Estimating survival models with dummy endogenous variables
4. Further extensions and discussion

1. A short overview of multiprocess models

2. Estimating systems of survival equations
3. Estimating survival models with dummy endogenous variables
4. Further extensions and discussion

Multiprocess models. Motivation

- Multiprocess models are
  - extensions of simultaneous equation models to survival processes
  - used by demographers who are concerned by issues of endogeneity and self-selection.
- In a series of influential papers, Lillard and his colleagues made a distinction between two forms of simultaneity (Lillard 1993, Lillard and Waite 1993)
  - the hazard of an event depends on the hazard of another event (for instance, women expecting their marriage to be short-lived should postpone motherhood)
  - the hazard of an event depends on the outcome of another related survival process (endogeneity - for instance, divorce risks depend on the presence or number of children; but having children is the outcome of the timing of births, which might depend on expected divorce risks)
- Lillard and Panis (2003) developed the aML software for the purpuse of estimating systems of multilevel equations with correlated random intercepts.

Multiprocess models

- Multiprocess models were originally invented as simultaneous equation models in which
  - at least one of the equations is a hazard equation;
  - all equations include a random intercept (or heterogeneity term)
  - the equation-specific random intercepts are correlated
- Note that survival models with shared frailty components are not multiprocess models, even when they have a multilevel structure.
  - The multilevel structure is not important if cross-equation correlation of residuals can be modeled without the help of random intercepts (Bartus and Roodman 2014).
  - Models with shared frailty for repeated events are of course important tools to control for sample-selection bias arising from the present of unobserved personality traits (Kravdal 2001)

Two classes of MLMP models

- Simultaneous equations for hazards:

$$\begin{aligned} \ln h_1 &= \lambda_1 \ln h_2 + \beta_1 X_1 + u_1 \\ \ln h_2 &= \lambda_2 \ln h_1 + \beta_2 X_2 + u_2 \end{aligned}$$

- Hazard models with endogenous dummy explanatory variable(s):

$$\begin{aligned} \ln h &= \alpha_1 y + \beta_1 X_1 + u_1 \\ y^* &= \beta_2 X_2 + u_2 \end{aligned}$$

  where y is the observed realization of the latent continuous variable $y*$

- In both models,
  - the random effects (the us) might be correlated (this will be discussed later)
  - values of X might change over spells within individuals, and the us are random effects (subscripts for individuals and spells are omitted)
  - identification of structural parameters require the presence of excluded instruments (this will be discussed later)

Why these two classes?

- The most general model of systems of equations including both observed qualitative or censored endogenous variables and the underlying latent endogenous variables is given by:

$$\begin{array}{lcccccc} \ln y_1^* & = & \lambda_1 y_2^* & + & \alpha\, y_2 & + & \ldots \\ \ln y_2^* & = & \lambda_2 y_1^* & + & \alpha\, y_1 & + & \ldots \end{array}$$

  which formalizes the idea that a latent outcome might depend on another latent outcome and the observed realization thereof.

- However, the general model is logically inconsistent. Logically consistent models satisfy the following restrictions (see Maddala 1983):

$$\begin{array}{ll} \text{R1:} & \lambda_1\, \alpha_2 = \lambda_2\, \alpha_1 = 0 \\ \text{R2:} & \alpha_1\, \alpha_2 = 0 \end{array}$$

- The simultaneous equation model obtains if the αs are restricted to zero.
- The other model obtains if the λs and one of the αs are restricted to zero.

Estimation with Stata

- The official **gsem**
  - allows one to estimate multilevel equations with correlated random intercepts
  - supports several parametric survival models
  - supports *logit*, *mlogit*, and *cloglog* links, which enable one to estimate discrete-time models, competingr-risk models, and models with endogenous qualitative predictors
- The user-written **cmp** command
  - allows one to estimate systems of seemingly unrelated recursive equations with jointly distributed Gaussian error terms
  - supports interval-censored regression models, which are just lognormal survival models
  - supports probit and multinomial probit models, which enable one to estimate discrete-time models and models with endogenous qualitative regressors

1. A short overview of multiprocess models

2. Estimating systems of survival equations

3. Estimating survival models with dummy endogenous variables
4. Further extensions and discussion

Simultaneous equations for hazards

- The structural model for two equations is given by:

$$\begin{aligned}\ln h_1 &= \lambda_1 \ln h_2 + \beta_1 X_1 + u_1\\ \ln h_2 &= \lambda_2 \ln h_1 + \beta_2 X_2 + u_2\end{aligned}$$

- Suppose there are excluded instruments z_1 and z_2 in X_1 and X_2. Then the structural model can be rewritten as

$$\begin{aligned}\ln h_1 &= \lambda_1 \ln h_2 + \beta_1 X + \gamma_1 z_1 + u_1\\ \ln h_2 &= \lambda_2 \ln h_1 + \beta_2 X + \gamma_2 z_2 + u_2\end{aligned}$$

- The reduced-form model, which can consistently be estimated, is

$$\begin{aligned}\ln h_1 &= \pi_{10} X + \pi_{11} z_1 + \pi_{12} z_2 + v_1\\ \ln h_2 &= \pi_{20} X + \pi_{21} z_1 + \pi_{22} z_2 + v_2\end{aligned}$$

- The *v*s are linear combination of all *u*s. Hence, the *v*s are correlated.

Identification of structural parameters

- Ideally, the estimation of the reduced-form model should be followed by the estimation of structural parameters.
- In the presence of excluded instruments, the effect of latent hazards can be estimated as follows:

$$\begin{aligned} \hat{\lambda}_1 &= \hat{\pi}_{12} / \hat{\pi}_{22} \\ \hat{\lambda}_2 &= \hat{\pi}_{21} / \hat{\pi}_{11} \end{aligned}$$

- Using these estimates, the structural parameters can be recovered as

$$\begin{aligned} \hat{\beta}_1 &= \hat{\pi}_{10} - \hat{\lambda}_1 \hat{\pi}_{20} \\ \hat{\beta}_2 &= \hat{\pi}_{20} - \hat{\lambda}_2 \hat{\pi}_{10} \end{aligned}$$

- These nonlinear combinations and the standard errors thereof can easily be computed with the **nlcom** command.

Example

- We will use a sample dataset on American women, which is shipped with the statistical software aML (Lillard and Panis 2003).
- The data contains information on marital births and marriage durations. The slightly modified and Stata-compatible version is obtained as follows:

```
use "http://web.uni-corvinus.hu/bartus/stata/divorce.dta"
```

- The data has a multilevel structure: conception episodes are nested within marriages, and marriages are nested within individuals.
- We select the second conception episode from within first marriages.
- Objective: joint modeling of conception and marital dissolution processes

Multispell data structure

- The hazard of conception and separation might change over conception episodes. We split conception episodes into smaller intervals within which these hazards might be assumed to be constant.
- Note that **mardur** measures the duration of the marriage at the beginning of each conception episode, and **time** is the duration of marriage when an event happens

```
gen dur = time-mardur
stset dur , fail(sep==1) id(id)
stsplit bdur , at(1 2 5 10)

// Corrections
replace dur = _t - _t0
replace mardur =  mardur + _t0
replace birth  = 0 if sep==.
replace sep    = 0 if sep==.

// rename sep, to avoid confusions
rename sep divorce
```

Multispell and multiprocess data structure

- *mardur* and *bdur* measures time at the beginning of each spell
- *dur* measures the length of the spell
- The process specific survival times are *mardur+dur* and *bdur+dur* for the divorce and birth processes, respectively

id	mardur	bdur	dur	birth	divorce
1164	3.083	0	1	0	0
1164	4.083	1	2	0	0
1164	5.083	2	2.083	1	0
1166	7.912	0	1	0	0
1166	8.912001	1	1.123	0	1

The model

- We study the following structural model:

$$\begin{aligned}
\ln h_{\text{Birth}} &= \lambda_1 \ln h_{\text{Divorce}} + \beta_1\, \mathit{hereduc} + \gamma_1\, \mathit{age} + u_1 \\
\ln h_{\text{Divorce}} &= \lambda_2 \ln h_{\text{Birth}} + \beta_2\, \mathit{hereduc} + \gamma_2\, \mathit{mardur} + u_2
\end{aligned}$$

- Variables
  - *hereduc* is women's level of education (computed from years of schooling)
  - *age* is the age at the beginning of a spell, centered around 30
  - *mardur* is the duration of the marriage at the beginning of a spell

Syntax of gsem I.

- Let *time* and *t*0 denote the surival time and the entry time. Let *y* be the failure variable indicating the occurrence of events. Finally, *d* denotes a distribution.
- The essence of gsem syntax for multilevel survival model is:

```
gsem ( time <- varlist U, family( d , lt(t0) fail(y) ) ) ///
    [ , options ]
```

- The gsem syntax for systems of multilevel survival models is then

```
gsem ///
( time1 <- varlist1 U1, family( d1 , lt(t02) fail(y1) ) ) ///
( time2 <- varlist2 U2, family( d2 , lt(t01) fail(y2) ) ) ///
[ , options ]
```

Estimation with gsem

- In this example, we chose the exponential distribution. (We thus assume that hazards are constants within the spells, after controlling for age and marriage duration)
- Exponential hazard models are just Poisson models of events, provided that the duration of the spell is added as an exposure variable.
- The gsem syntax for systems of exponential survival models is then

```
gsem ///
( y1 <- varlist1 U1, poisson exposure(dur1) ) ///
( y2 <- varlist2 U2, poisson exposure(dur2) ) ///
....
```

where *dur1*, *dur2*, …. measure the process-specific durations.

Estimation with gsem

- Model specification with the help of macros:

```
global xvars ib2.hereduc age mardur
global model poisson exposure(dur)
```

- First, we estimate the two equations separately, that is, we constraint the covariance of the random effects to zero:

```
gsem ( birth   <- $xvars U[id] , $model ) ///
     ( divorce <- $xvars V[id] , $model ) ///
     ,   vce(cluster id) cov( U[id]*V[id]@0 )
est store sep
```

- Then, we estimate the true multiprocess models wih correlated random effects:

```
gsem ( birth   <- $xvars U[id] , $model ) ///
     ( divorce <- $xvars V[id] , $model ) ///
     ,   vce(cluster id)
est store joint
```

gsem results I. Coefficients

```
------------------------------------------------
            Variable |    sep         joint
---------------------+--------------------------
birth                |
             hereduc |
          <12 years  |  -0.389***    -0.391***
          16+ years  |   0.066        0.068
                 age |  -0.095***    -0.095***
              mardur |  -0.073***    -0.076***
               _cons |  -1.969***    -1.972***
---------------------+--------------------------
divorce              |
             hereduc |
          <12 years  |  -0.331**     -0.336**
          16+ years  |  -0.184       -0.203
                 age |  -0.064***    -0.062***
              mardur |   0.085***     0.091***
               _cons |  -4.520***    -4.722***
---------------------+--------------------------

 legend: * p<0.05; ** p<0.01; *** p<0.001
```

This is an edited output. Coefficients of the latent variables are 1s and omitted

gsem results II. Random effects

```
------------------------------------------------
            Variable |    sep          joint
---------------------+--------------------------
var(U[id])           |
               _cons |   1.007***     1.018***
---------------------+--------------------------
var(V[id])           |
               _cons |   1.109***     1.515**
---------------------+--------------------------
     cov(V[id],U[id])|
               _cons |               -0.216*
------------------------------------------------

 legend: * p<0.05; ** p<0.01; *** p<0.001
```

- Random effects are negatively correlated
- The negative correlation suggests that the effects of the latent hazards have opposite signs…..

Estimation of the effect of latent variables

Effect of the separation hazard on the conception hazard

```
nlcom _b[birth:mardur] / _b[divorce:mardur]
```

```
----------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|
-------------+--------------------------------------------------------
       _nl_1 |  -.8309784   .2362941    -3.52   0.000
----------------------------------------------------------------------
```

Effect of the conception hazard on the separation hazard

```
nlcom _b[divorce:age] / _b[birth:age]
```

```
----------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|
-------------+--------------------------------------------------------
       _nl_1 |   .6587895   .1337735     4.92   0.000
----------------------------------------------------------------------
```

Estimation of structural coefficients

Structural effect of higher education on birth risks

```
nlcom _b[birth:3.hereduc] - ///
  ( _b[birth:mardur] / _b[sep:mardur] ) * _b[sep:3.hereduc]

--------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|
-------------+------------------------------------------------------------
       _nl_1 |  -.1000164   .1629748    -0.61   0.539
--------------------------------------------------------------------------
```

Structural effect of higher education on divorce risk

```
nlcom _b[sep:3.hereduc] - ///
  ( _b[divorce:age] / _b[birth:age] ) ) * _b[birth:3.hereduc]

-------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|
-------------+---------------------------------------------------------
       _nl_1 |  -.2478165    .184713    -1.34   0.180
--------------------------------------------------------------------------
```

Flexibility of gsem

- We could have estimated Weibull or gamma or lognormal survival models.
- These models require process-specific survival times as dependent variables. In our example, these variables are

```
gen tbirth   = bdur   + dur
gen tdivorce = mardur + dur
```

- A model in which lognormal and Weibull duration dependence characterizes the respective birth and separation processes would be:

```
global birth   family( lognormal , fail(birth)   lt(bdur)   )
global divorce family( weibull   , fail(divorce) lt(mardur) )
gsem ( tbirth   <- $xvars U[id] , $birth   ) ///
     ( tdivorce <- $xvars V[id] , $divorce ) ///
     ,   vce(cluster id)
```

System of lognormal survival models. cmp

- Lognormal survival models assume that the hazard first sharply increases then slowly decreases with survival time. Models of this sort can easily be estimated with **cmp** .
- Lognormal models are just interval-censored regressions. Interval regression models require two dependent variables, labeled the lower and upper limits, which define the intervals within which the true value of log duration lies.
- For the birth process, the lower and upper limits are generated as follows:

```
gen blo = ln(bdur+dur)
gen bhi = blo if birth==1
```

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

```
gen mlo = ln(mardur+dur)
gen mhi = mlo if divorce==1
```

Syntax of cmp. A selective intro I.

- Single equation lognormal survival model using single-spell data

```
cmp ( label : tlo thi = varlist  ) ///
    , indicators(7)   [ options ]
```

  - *label* **:** is optional but useful: it instructs **cmp** to use birth to label the equation.
  - *tlo* and *thi* indicate the lower and upper limits of survival time.
  - The **indicators(7)** option means that this equation is interval regression

- Single equation lognormal survival model using multi-spell data

```
cmp ( label : tlo thi = varlist , trunc(ln(t0) .) ) ///
    , indicators(7)   [ options ]
```

  - *t0* is the variable recoding the entry time and the **trunc( )** option handles left-truncation of survival times

Syntax of cmp. A selective intro II.

- The syntax for estimating two lognormal models jointly using multi-spell data is

```
cmp ( label1 : tlo1 thi1 = varlist1 , trunc(ln(t01) .) ) ///
    ( label2 : tlo2 thi2 = varlist2 , trunc(ln(t02) .) ) ///
    , indicators(7 7)  [ options ]
```

- The **indicators(**7 7**)** option specifies that the first and second equations are lognormal ones.
- The dependent and explanatory variables, as well as truncation limit experssions are equation-specific.

Syntax of cmp in our example

- First, we estimate the two equations separately, that is, we constraint the covariance of the random effects to zero:

```
cmp  (birth:   blo bhi = $xvars , trunc(ln(bdur) .)  ) ///
     (divorce: mlo mhi = $xvars , trunc(ln(mardur) .)) ///
     , ind(7 7) vce(cluster id)  cov(indep)
est store sep
```

- Then, we estimate the true multiprocess models wih correlated random effects:

```
cmp  (birth:   blo bhi = $xvars , trunc(ln(bdur) .)  ) ///
     (divorce: mlo mhi = $xvars , trunc(ln(mardur) .)) ///
     , ind(7 7) vce(cluster id)
est store joint
```

cmp results

```
------------------------------------------------
            Variable |    sep          joint
---------------------+--------------------------
birth                |
             hereduc |
           <12 years |  -0.049       -0.036
           16+ years |  -0.255***    -0.270***
                     |
                 age |   0.032***     0.033***
              mardur |   0.112***     0.115***
               _cons |   1.775***     1.794***
---------------------+--------------------------
divorce              |
             hereduc |
           <12 years |  -0.022        0.020
           16+ years |   0.210*       0.289*
                     |
                 age |   0.032***     0.036**
              mardur |   0.072***     0.072***
               _cons |   3.074***     3.442***
---------------------+--------------------------
(output omitted)
---------------------+--------------------------
atanhrho_12          |
               _cons |               -0.580***
------------------------------------------------
        legend: * p<0.05; ** p<0.01; *** p<0.001
```

Estimation of the effect of latent variables

Effect of latent time to divorce on the time to conception

```
. nlcom _b[birth:mardur] / _b[divorce:mardur]

--------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|
-------------+------------------------------------------------------------
       _nl_1 |   1.593011   .5752617     2.77   0.006
--------------------------------------------------------------------------
```

Effect of latent time to conception on the time to divorce

```
. nlcom _b[divorce:age] / _b[birth:age]

--------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|
-------------+------------------------------------------------------------
       _nl_1 |   1.076084   .4313306     2.49   0.013
--------------------------------------------------------------------------
```

Estimation of structural coefficients

Structural effect of higher education on the time to birth

```
nlcom _b[birth:3.hereduc] - ///
  ( _b[birth:mardur] / _b[divorce:mardur] ) * _b[divorce:3.hereduc]

----------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|
-------------+--------------------------------------------------------------
       _nl_1 |  -.7307283   .2926678    -2.50   0.013
----------------------------------------------------------------------------
```

Structural effect of higher education on the time to divorce

```
nlcom _b[divorce:3.hereduc] - ///
  ( _b[divorce:age] / _b[birth:age] ) * _b[birth:3.hereduc]

----------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|
-------------+--------------------------------------------------------------
       _nl_1 |   .5799594    .187107     3.10   0.002
----------------------------------------------------------------------------
```

Summary of cmp results

- Higher education reduces the time to second births, and increases the time to divorce.
- These effects are understated in the reduced-form models
- The correlation between the residuals is negative (like in the **gsem** output)
- There is a positive relationship between the latent waiting times
  - This is counter-intuitive, and cannot explain the negative correlation of the residuals
  - Remember these effects are estimates, based on the reduced form coefficients of marriage duration.
  - The problem is that marriage duration decreases the risk of divorce in the cmp model – in contrast, marriage duration increases divorce risks in the gsem model.
  - The negative effect of marriage duration on the hazard of divorce might be an artefact of imposing lognormal duration dependence on the divorce process.

What about estimating discrete-time survival models jointly?

- The example presented above makes use of parametric continuous-time models.
- In theory, both **gsem** and **cmp** are able to estimate discrete-time survival models: both support the probit link function, and **gsem** also supports the logit link function.
- However, the discrete-time modeling framework is not the best choice for simultaneous survival processes:
  - Different processes rarely or never terminate at the same time (empty cell problem)
  - The problem is that the estimated correlation between the residuals will be close to -1, whatever the true correlation is.
  - Even when there are no empty cells, some simulation evidence suggests that bivariate probit estimates are not numerically reliable if events are rare.

1. A short overview of multiprocess models
2. Estimating systems of survival equations

3. Estimating survival models with dummy endogenous variables

4. Further extensions and discussion

The model

- The model:

$$\begin{array}{lclclcl} \ln h & = & \alpha_1 y & + & \beta_1 X & + & u_1 \\ y^* & = & \gamma z & & \beta_2 X & + & u_2 \end{array}$$

  where

  - y is a dummy variable, which is the observed realization of the latent variable $y*$
  - the random effects (the us) are correlated
  - z is the excluded instrument, which enables identification, provided that the random effects are allowed to be correlated,

Example

- We use the child mortality dataset shipped with aML. Data has multilevel structure: observations about children are nested within mothers.
- Outcome we wish to study: death hazard
- Endogenous dummy variable: hospital delivery
- Common explanatory variables (the *X*s): mother's education.
- Excluded instrument in the equation explaining hospital delivery: distance to nearest hospital
- The slightly modified and Stata-compatible version is obtained as follows:

```
use "http://web.uni-corvinus.hu/bartus/stata/children.dta"
```

Relevance of this example for labor economists

When you read..	**… you might think of...**
Mother	Unemployed person
Child	Unemployment spell
Death	Finding employment
Time to death	Length of unemployment spell
Hospital delivery	Participation in a program
Distance to nearest hospital	Distance to the place of the program
Mother's education	Persons's education
The effect of hospital delivery on death risk should be **negative**	The effect of program participation on the hazard of finding a job should be **positive**

Estimating a Weibull model with gsem

- We wish to estimate a Weibull model of time to death. The model specification:

```
global death     hospital i.edu
global hospital  distance i.edu
global model     family( weibull  , fail(death)  )
```

- First, we fit the two equations separately :

```
gsem    (age      <-  $death    U[id] , $model  ) ///
        (hospital <-  $hospital V[id] , probit )  ///
        , vce(cluster id) cov( U[id]*V[id]@0 )
est store sep
```

- Then, we estimate the true multiprocess models wih correlated random effects:

```
gsem    (age      <-  $death    U[id] , $model  ) ///
        (hospital <-  $hospital V[id] , probit )  ///
        , vce(cluster id)
est store joint
```

gsem results I. Coefficients

```
------------------------------------------------
            Variable |    sep          joint    
---------------------+--------------------------
age                  |
            hospital |  -0.421*      -0.719**   
                educ |
        high school  |  -0.325       -0.247     
            college  |  -2.125**     -1.983**   
                     |
               U[id] |   1.000        1.000     
                     |
               _cons |  -2.471***    -2.441***  
---------------------+--------------------------
hospital             |
            distance |  -0.034       -0.034     
                educ |
        high school  |   1.051***     1.045***  
            college  |   1.639***     1.639***  
                     |
               V[id] |   1.000        1.000     
                     |
               _cons |  -1.091***    -1.089***  
---------------------+--------------------------
```

gsem results II. Ancillary parameters and random effects

```
------------------------------------------------
            Variable |    sep           joint
---------------------+--------------------------
age_ln_p             |
               _cons |  -1.235***    -1.230***
---------------------+--------------------------
var(U[id])           |
               _cons |   0.595*        0.666*
---------------------+--------------------------
var(V[id])           |
               _cons |   0.462**       0.471**
---------------------+--------------------------
     cov(V[id],U[id])|
               _cons |                 0.247
------------------------------------------------
        legend: * p<0.05; ** p<0.01; *** p<0.001
```

Interpretation

- Even when the correlation of random effects is not significant, hospital delivery has a larger negative effect in the joint model. Similar finding can be found in the aML manual.
- Interpretation:
  - Hospital delivery has a large negative effect on the hazard of death
  - Women who are aware that the baby has a high death risk have large propensity to chose hospital over home delivery.
  - The self-selection of problematic births into hospitals has the consequence of understating the negative effect of hospital delivery in the separate model.

Estimating a discrete-time model with cmp

- The main advantage of discrete-time models over continuous-time parametric models is that the functional form for duration dependence might be modeled.
- Changing the dataset into a discrete-time dataset. Each observation refers to a person-year. Age is age at the beginning of a person-year.

```
replace age = ceil(age)
gen double tid = _n
expand age
sort tid
qui by tid : replace age = _n-1
qui by tid : replace death = 0 if  _n<_N
```

Discrete-time model with an endogenous dummy. cmp

- We experiment with a curvilinear duration dependence. The model specification:

```
global death     hospital i.edu c.age##c.age
global hospital  distance i.edu
```

- First, we estimate the two equations separately, that is, we constraint the correlation of the underlying residuals to zero:

```
cmp   (death = $death )   (hospital = $hospital ) ///
      , ind(4 4) vce(cluster id)   cov(indep)
est store sep
```

Here the **indicator(**4 4**)** option means that both equations are probit

- Then, we estimate the true multiprocess models wih correlated residuals:

```
cmp   (death = $death )   (hospital = $hospital ) ///
      , ind(4 4) vce(cluster id)
est store joint
```

cmp results

```
------------------------------------------------
            Variable |    sep          joint
---------------------+--------------------------
death                |
            hospital |  -0.137       -0.068
                educ |
        high school  |  -0.124       -0.142
            college  |  -0.862***    -0.893*
                 age |  -0.233***    -0.233***
         c.age#c.age |   0.006***     0.006***
               _cons |  -1.407***    -1.418***
---------------------+--------------------------
hospital             |
            distance |  -0.043*      -0.043*
                educ |
        high school  |   0.814***     0.814***
            college  |   1.312***     1.312***
               _cons |  -0.785***    -0.785***
---------------------+--------------------------
atanhrho_12          |
               _cons |               -0.040
------------------------------------------------
        legend: * p<0.05; ** p<0.01; *** p<0.001
```

Interpretation

- Again, the correlation of the residuals lack statistical significance
- In contrast to the previous **gsem** results, we do not find evidence that hospital delivery would reduce the probability of dying in a given year.
- This presentation is not about a serious research into mortality, thus I do not discuss this problem further....

1. A short overview of multiprocess models
2. Estimating systems of survival equations
3. Estimating survival models with dummy endogenous variables

4. Further extensions and discussion

Survival models with sample selection

- Suppose we are interested in examining the relationship between mother's education and child mortality using a sample of children who were born in hospitals.
- The sample of those children is selective. Sample selection bias can be controlled by estimating the survival model jointly with the probit model of hospital choice.

```
gen ageh = age if hospital==1
gsem     (ageh     <-  $death    U[id] , $model  ) ///
         (hospital <-  $hospital V[id] , probit )


cmp   (death = $death )   (hospital = $hospital ) ///
      , ind("hospital*4" 4)
```

- Determination of the relevant estimation sample is automatic in gsem. In contrast, it is the task of the user with cmp. The **indicators()** option allows expressions; observations where an expression evaluates to zero will be not used when estimating the equation to which the expression belongs.

Survival models and panel attrition

- Suppose the data on survival of children is collected in a panel survey including three waves. You need the third wave to observe a sufficient number of deaths.
- There is panel attrition which is not random. Dummies w3 and w2 indicate participation in waves 2 and 3, respectively. Participation in that waves is predicted using variables from wave 1 and wave 2, respectively.
- Survival models might be estimated jointly with probit models of panel continuation (Lillard and Panis 1998). A model might be

```
gsem    (age <-  $death             U[id] , $model  ) ///
        (w3  <-  varlist_wave2 V2[id] , probit )  ///
        (w2  <-  varlist_wave1 V1[id] , probit )  ///

cmp     (death = $death )   ///
        (w3  = varlist_wave2 )  ///
        (w2  = varlist_wave1 )  ///
        , ind("w3*4" "w2*4" 4)
```

Models with endogenous qualitative variables

- Suppose there are both public and private hospitals. Now hospital has three categories: 0 if home delivery; 1 if delivery in a public hospital; and 2 if delivery in a private hospital.
- The mortality model with an endogenous multinomial variable has the following structure:

```
gsem (age          <-  $death    U[id]  , $model  ) ///
     (1.hospital <-  $hospital V1[id] , mlogit )  ///
     (2.hospital <-  $hospital V2[id] , mlogit )

cmp  (death = $death )  (hospital = $hospital , iia ) ///
     , ind(4 6)
```

- In the cmp sytax,
  - suboption **iia** enforces the independence of irrelevant alternatives assumption
  - 6 in the indicator options refers to multinomial probit.

Survival models with endogenous switching

- Suppose that the effect of explanatory variables depend on the type of delivery.
- The examples assume if hospital had three categories: home delivery (0), delivery in a public hospital (1), and delivery in a private hospital (2).
- Estimation of the swithing model with **gsem** would look like:

```
separate age , by(hospital)
gsem (age0 age1 age2  <-  $death U[id] , $model  ) ///
     (1.hospital  <-  $hospital V1[id] , mlogit )  ///
     (2.hospital  <-  $hospital V2[id] , mlogit )
```

- This model is very demanding computationally..... it might be the case that **gsem** will not find the ML solution.

Survival models with endogenous switching

- Estimation with **cmp**:

```
cmp  (death = $death ) (death = $death ) (death = $death ) ///
     (hospital = $hospital , iia )                           ///
,ind("(hospital==0)*4" "(hospital==1)*4" "(hospital==2)*4" 6)
```

- The first three equations seems to be the same – they are, but they will be estimated in three different samples, identified by the values of hospital (mind the **indicators()** option!)
- The three survival equations are estimated jointly with a multinomial probit of hospital choice.
- **cmp** can estimate the swithing model, but one should control the simulated likelihood estimation procedure, in general, and the number of GHK draws, in particular. (One should specify the **ghkdraws(#)** option, using a relatively small number.)

Conclusions

- Recently, Stata became able to estimate various forms of multiprocess models:
  - Both **cmp** and **gsem** in Stata 14 can handle truncated dependent variables
  - **gsem** in Stata 14 supports various parametric survival models
- There is, however, room for improvement
  - there are multiprocess models which include more than two equations
  - I experienced serious „initial values not feasible" and convergence issues when I tried to estimate such models with **gsem**.
  - **cmp** has less problems with systems including three or even more equations
- Can complicated models be estimated with the **bayesmh** command?

  Multiprocess models including several equations were successfuly estimated with MLwiN software, which implements MCMC (Steele etal. 2005)

References

Bartus, T. and D. Roodman. 2014. Estimation of multiprocess survival models with **cmp**. *Stata Journal 14: 756–777.*

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.

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.

Maddala, G.S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.

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

Steele, F., C. Kallis, H. Goldstein and H. Joshi. 2005. The relationship between Childbearing and Transitions from Marriage and Cohabitation in Britain. Demography 42: 647-673.