The Stata Journal (2017)
17, Number 2, pp. 442–461

Multilevel multiprocess modeling with gsem

Tamás Bartus
Corvinus University of Budapest
Budapest, Hungary
tamas.bartus@uni-corvinus.hu

Abstract. Multilevel multiprocess models are simultaneous equation systems that include multilevel hazard equations with correlated random effects. Demographers routinely use these models to adjust estimates for endogeneity and sample selection. In this article, I demonstrate how multilevel multiprocess models can be fit with the `gsem` command. I distinguish between two classes of multilevel multiprocess models: nonrecursive systems of hazard equations without observed endogenous variables and recursive systems that include a hazard equation with observed endogenous qualitative variables. I illustrate the estimation of both classes of models using sample datasets shipped with the statistical software aML. I pay special attention to identifying structural coefficients in nonrecursive simultaneous systems.

Keywords: st0481, survival analysis, multilevel multiprocess models, multilevel analysis, simultaneous equations, endogeneity, gsem

1 Introduction

Multilevel multiprocess models were developed as systems of proportional hazard models with correlated individual-level random effects. These models adjust estimates of the parameters of hazard equations for two forms of simultaneity (Lillard 1993; Lillard and Waite 1993). Suppose a researcher examines the impact of children on marital stability. Estimates of ordinary survival models of the hazard of divorce are likely to be biased; the first form of simultaneity is the endogeneity of the presence of children, because it is the outcome of a related process of timing of births. Furthermore, the conception hazard might depend on the latent dissolution hazard; if couples expect that their marriage will be short lived, they may decide to postpone the first (or higher-order) births. The second form of simultaneity arises because the latent hazard of marriage dissolution is an unobservable (endogenous) variable in the conception hazard equation.

The multilevel multiequation modeling framework has advantages. First, some of the explanatory variables in hazard models are endogenous, and estimation of the hazard model of substantive interest jointly with probit models explaining the endogenous variables eliminates the endogeneity bias (Lillard, Brien, and Waite 1995; Impicciatore and Billari 2012). Second, the multilevel multiprocess modeling framework easily deals with selection bias. Consider the estimation of the effect of education on second-birth hazards (Kravdal 2001). Finding a positive effect of higher education can be explained in terms of a selection effect. Because educated women postpone

first births, unmeasured factors that also affect the timing of births will be correlated with education in the sample of mothers, even when those unmeasured factors are independent of education in the population of childless women. The selection effect is appropriately controlled if the hazard models explaining first, second, and higher-order births are jointly fit.

In this article, I show how multilevel multiprocess models can be fit with the `gsem` command, which is a natural choice for two reasons: it allows one to estimate multilevel equations with correlated latent variables, and it supports survival equations in Stata 14. My endorsement of the `gsem` command contrasts an earlier suggestion of fitting systems of survival models with the user-written `cmp` command (Roodman 2011; Bartus and Roodman 2014). The advantage of the `cmp` command is that the correlation of residuals can be modeled without including random effects. This strategy has an additional computational advantage because systems including two equations can be estimated without numerically approximating two-dimensional integrals. However, the `cmp` command forces researchers to impose lognormal duration dependence on the data, an unrealistic assumption in several applications. Additionally, the computational advantage of the `cmp` command might have been overstated because numerical integration procedures seem to be substantially faster in Stata 14 than in older versions.

I begin by identifying two classes of multilevel multiprocess models: nonrecursive systems of hazard equations without observed endogenous variables and recursive systems that include hazard equations with observed endogenous qualitative variables. Afterward, I detail how both classes of models can be fit using the `gsem` command. The examples use sample datasets shipped with the statistical software aML, which was explicitly developed for multilevel multiprocess modeling (Lillard and Panis 2003). I pay special attention to identifying structural parameters in nonrecursive systems of hazard equations, an issue often neglected in empirical applications.

2 Multilevel multiprocess hazard models

2.1 Motivation

Multilevel multiprocess modeling addresses the problem that explanatory variables are often endogenous because of selection mechanisms. Consider the classic example of estimating the impact of children on marital stability. Estimates from a separate hazard model of divorce suffer from two forms of simultaneity biases (Lillard 1993; Lillard and Waite 1993). First, the presence of children is endogenous because it is the outcome of a process of timing of births. Second, the latent birth hazard might depend on the latent dissolution hazard as well. Similar biases arise if the researcher is also interested in examining the effect of marriage on childbearing. Marriage is the outcome of the partnership formation process, which may depend on the latent propensity of becoming a parent.

The aforementioned simultaneity problems can easily be studied within the framework of simultaneous equations with qualitative variables (Heckman 1978). Let y_{1t}^{*} and

y^*_{2t} denote the endogenous latent hazards under study; for instance, the former might be the hazard of conception, and the latter might denote the hazard of marital dissolution. Subscript t expresses the possible time dependence of the hazards. y_{1t} and y_{2t} are observed realizations of the latent variables. The dependence of each latent variable on the other, as well as on other (possibly time-varying) explanatory variables $\mathbf{x}_{1t}$ and $\mathbf{x}_{2t}$, is described with the structural equations

$$\begin{aligned} y^*_{1t} &= \alpha_1 y_{2t} + \lambda_1 y^*_{2t} + \boldsymbol{\beta}_1' \mathbf{x}_{1t} + \varepsilon_{1t} \\ y^*_{2t} &= \alpha_2 y_{1t} + \lambda_2 y^*_{1t} + \boldsymbol{\beta}_2' \mathbf{x}_{2t} + \varepsilon_{2t} \end{aligned} \tag{1}$$

The two forms of simultaneity are related to the presence of latent variables and observed realizations on the right-hand side of the equations. First, the error terms are correlated with the exogenous explanatory variables because of the presence of an unobserved hazard on the right-hand side and the dependence of that hazard on the same exogenous variables. Second, the expected value of the residual is not constant across the categories of the observed realizations (Lee 1979).

Joint estimation of the system is viewed as a method for eliminating both sources of endogeneity bias. I will discuss the method separately for two classes of the model. It is well known that the parameters of the model defined by (2) are not identified without further restrictions. Using classic results on logical consistency and identification (Maddala 1983), we see that the model exists only if $\lambda_1\alpha_2 = \lambda_2\alpha_1 = 0$ and $\alpha_1\alpha_2 = 0$. This condition implies that there are two forms of estimable systems. The first form is nonrecursive systems without observed endogenous variables ($\alpha_1 = \alpha_2 = 0$):

$$\begin{aligned} y^*_{1t} &= \lambda_1 y^*_{2t} + \boldsymbol{\beta}_1' \mathbf{x}_{1t} + \varepsilon_{1t} \\ y^*_{2t} &= \lambda_2 y^*_{1t} + \boldsymbol{\beta}_2' \mathbf{x}_{2t} + \varepsilon_{2t} \end{aligned}$$

The second form is recursive systems with observed endogenous variables ($\lambda_1 = \lambda_2 = 0$ and $\alpha_1 = 0$):

$$\begin{aligned} y^*_{1t} &= \boldsymbol{\beta}_1' \mathbf{x}_{1t} + \varepsilon_{1t} \\ y^*_{2t} &= \alpha_2 y_{1t} + \boldsymbol{\beta}_2' \mathbf{x}_{2t} + \varepsilon_{2t} \end{aligned} \tag{2}$$

I will now discuss these models briefly.

2.2 Nonrecursive systems without observed endogenous variables

In these systems, endogeneity bias emerges because unobserved endogenous hazards appear on the right-hand sides of both equations. The dependence of hazards on other hazards disappears in the reduced-form system. However, the reduced-form parameters are not equal to the structural parameters of interest. In this section, I focus on identifying these parameters via excluded instruments. To emphasize the presence of excluded instruments, we rewrite the structural model as

$$\begin{aligned} y^*_{1t} &= \lambda_1 y^*_{2t} + \boldsymbol{\beta}'_1 \mathbf{x}_t + \gamma_1 z_{1t} + \varepsilon_{1t} \\ y^*_{2t} &= \lambda_2 y^*_{1t} + \boldsymbol{\beta}'_2 \mathbf{x}_t + \gamma_2 z_{2t} + \varepsilon_{2t} \end{aligned}$$

where $\mathbf{x}$ is a vector of exogenous variables common to both equations and the z's are excluded instruments. The system of reduced-form equations is

$$\begin{aligned} y^*_{1t} &= \boldsymbol{\pi}'_{10} \mathbf{x_t} + \pi_{11} z_{1t} + \pi_{12} z_{2t} + v_{1t} \\ y^*_{2t} &= \boldsymbol{\pi}'_{20} \mathbf{x_t} + \pi_{21} z_{1t} + \pi_{22} z_{2t} + v_{2t} \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\pi}_{j0} &= (1 - \lambda_1 \lambda_2)^{-1} (\boldsymbol{\beta}'_j + \lambda_j \boldsymbol{\beta}'_k) \\ \pi_{jj} &= (1 - \lambda_1 \lambda_2)^{-1} \gamma_j \\ \pi_{jk} &= (1 - \lambda_1 \lambda_2)^{-1} \lambda_j \gamma_k \\ v_j &= \varepsilon_j + \lambda_j \varepsilon_k \end{aligned} \tag{3}$$

where $j = \{1, 2\}$ indexes the equations and $k = 3 - j$. Estimation must account for the residuals in the reduced-form equations being generally correlated, even when the disturbances in the structural equations are independent of each other. If the latter error terms are normally distributed, the correlation of the residuals can easily be modeled using the multivariate normal distribution. In proportional hazard models, however, the error terms are exponentially distributed. Hence, the correlation of the underlying residuals should be modeled with the help of jointly normally distributed random intercepts (Lillard 1993). The resulting multilevel multiprocess model can be stated as follows:

$$\begin{aligned} y^*_{1t} &= \boldsymbol{\pi}'_{10} \mathbf{x_t} + \pi_{11} z_{1t} + \pi_{12} z_{2t} + u_1 + \eta_{1t} \\ y^*_{2t} &= \boldsymbol{\pi}'_{20} \mathbf{x_t} + \pi_{21} z_{1t} + \pi_{22} z_{2t} + u_2 + \eta_{2t} \\ \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} &\sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right) \end{aligned} \tag{4}$$

In the presence of excluded instruments, the structural coefficients can be recovered as follows. First, notice from (3) that the selection coefficient λ_j can easily be estimated as follows:

$$\lambda_j = \pi_{jk} / \pi_{kk} \tag{5}$$

Second, use the estimated selection parameters to solve the system:

$$\begin{aligned} \boldsymbol{\pi}_{10} &= (1 - \lambda_1 \lambda_2)^{-1} (\boldsymbol{\beta}'_1 + \lambda_1 \boldsymbol{\beta}'_2) \\ \boldsymbol{\pi}_{20} &= (1 - \lambda_1 \lambda_2)^{-1} (\boldsymbol{\beta}'_2 + \lambda_2 \boldsymbol{\beta}'_1) \end{aligned}$$

The solution is a simple nonlinear combination of reduced-form coefficients:

$$\boldsymbol{\beta}_j = \boldsymbol{\pi}_{j0} - \lambda_k \boldsymbol{\pi}_{k0} = \boldsymbol{\pi}_{j0} - (\pi_{jk} / \pi_{kk}) \boldsymbol{\pi}_{k0} \tag{6}$$

Both the nonlinear combination and its standard error can be calculated using the `nlcom` command.

2.3 Recursive systems with observed endogenous variables

In recursive systems, all coefficients are structural. Endogeneity arises because the expected value of ε_1 differs in groups $y_{1t} = 1$ and $y_{1t} = 0$. The problem is the same as the problem of sample selection. The endogeneity bias is eliminated if the residuals are allowed to be correlated and the seemingly unrelated system is jointly estimated. In the case of two equations with normally distributed residuals, this boils down to the estimation of a bivariate probit model. However, the second equation is a proportional hazard model, and the joint estimation requires the inclusion of random intercepts. The multilevel multiprocess model is

$$\begin{aligned} y^*_{1t} &= \boldsymbol{\beta}_1'\mathbf{x}_{1t} + u_1 + \eta_{1t} \\ y^*_{2t} &= \alpha_2 y_{1t} + \boldsymbol{\beta}_2'\mathbf{x}_{2t} + u_2 + \eta_{2t} \\ \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} &\sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right) \end{aligned} \tag{7}$$

To identify the correlation of the random effects, one should include in the first-stage equation at least one variable not included in the second-stage hazard equation.

In empirical applications, the latent variable y^*_{1t} is often a time-constant latent propensity to experience an event. The classic example is the propensity to form a cohabiting union before marriage that in turn will affect the (time-varying) hazard of marital dissolution (Lillard, Brien, and Waite 1995).

3 Fitting multilevel multiprocess models with gsem

The official Stata `gsem` command can fit multiprocess hazard models because it supports multiequation survival models with correlated latent variables. The description of the syntax is restricted to components of the `gsem` command specific to our purposes. We also assume a multispell data structure where each record corresponds to an episode nested within an individual.

3.1 Multilevel hazard models

In the multispell dataset, *idvar* identifies the individuals, *timevar* records the survival time, *t0var* records entry time, and *event* is a dummy variable recording the occurrence of the event under study. The inclusion of the random intercept requires specification of a latent variable at the level of individuals. This latent variable might be specified as `U[`*idvar*`]`. Instead of `U`, one can choose any word beginning with a capital letter. However, specifying `[`*idvar*`]` after the chosen word is mandatory; this syntax element tells Stata that the latent variable is random intercept, which is constant within the individuals.

To fit a multilevel hazard model assuming distribution *family*, one should type

```
gsem (timevar <- indepvars U[idvar], family(family, fail(event) lt(t0var)))
```

3.2 Piecewise-constant multilevel hazard models

Most of the empirical applications in demography use piecewise-linear exponential hazard models. Thus I will focus on simple exponential hazard models. Exponential hazard models can easily be fit as Poisson models of events, provided that the explanatory variables include the natural log of the duration of the current spell (Skrondal and Rabe-Hesketh 2004). The reason is that if survival time t follows an exponential distribution with parameter h, the expected number of failures follows a Poisson distribution with parameter ht (Holford 1980).

Define *durvar* as *timevar* minus *t0var*. *durvar* thus measures the duration of the current spell. The multilevel piecewise-constant exponential hazard model can be fit as follows:

```
gsem (event <- indepvars U[idvar], poisson exposure(durvar))
```

Duration dependence is allowed if *indepvars* includes *t0var*, other variables generated from *t0var*, or indicator variables capturing the rank order of the current spell.

3.3 Fitting nonrecursive systems without observed endogenous variables

Systems of piecewise-constant exponential models require separate latent variables for the equations. Let `U1[`*idvar*`]` and `U2[`*idvar*`]` be the equation-specific latent variables (random intercepts). Two equations can be jointly estimated as follows:

```
gsem
    (event_1 <- indepvars_1 U1[idvar], poisson exposure(durvar))
    (event_2 <- indepvars_2 U2[idvar], poisson exposure(durvar))
```

`gsem` automatically estimates the variance–covariance matrix of the random effects. The loadings of the latent variables will be constrained to 1.

event_1 and *event_2* might refer to recurrent events of the same kind. (For simplicity, the outcomes of sequential choices, like the timing for first, second, and higher-order births, are also treated as recurrent events.) The practice of multilevel modeling suggests that equations for recurrent events should share the same latent variable. Even if K different equations are used to model recurrent events of the same kind, the equations should include a single latent variable, not K different latent variables. This strategy of modeling first, second, and higher-order births is present in Lillard's (1993) seminal

article. The reason for using one instead of K different latent variables is computational: integrating out one latent variable takes less time than integrating out K jointly distributed latent variables. The syntax for recurrent events is

```
gsem
    (event_1 <- indepvars_1 U[idvar], poisson exposure(durvar))
    (event_2 <- indepvars_2 U[idvar], poisson exposure(durvar))
```

`gsem` automatically constrains the loading of the latent variable to 1 in the first equation but estimates the loadings in the other equations and the variance of the latent variable.

Lillard (1993) modeled recurrent occurrences of births jointly with marital dissolution. The joint modeling of recurrent events nested within another process can be implemented as follows: Variables *event_11* and *event_12* capture the occurrences of the recurrent events. Variable *event_2* measures the termination of another process within which the occurrences of *event_11* and *event_12* are nested. The syntax, which combines the previous syntax elements, is

```
gsem
    (event_11 <- indepvars_12 U1[idvar], poisson exposure(durvar))
    (event_12 <- indepvars_12 U1[idvar], poisson exposure(durvar))
    (event_2 <- indepvars_2 U2[idvar], poisson exposure(durvar))
```

3.4 Fitting recursive systems with observed endogenous variables

For simplicity, consider one survival process and one probit equation. Again, *event* is the variable indicating failures. *xvar* is the endogenous dummy variable in the hazard equation. *indepvars* includes the exogenous variables appearing in both the hazard and the probit equations. Finally, *zvars* contains the excluded instrument (or the list of excluded instruments), which appears only in the probit equation. The syntax is

```
gsem
    (event <- xvar indepvars U[idvar], poisson exposure(durvar))
    (xvar  <- zvars indepvars V[idvar], probit)
```

gsem also allows one to estimate more complicated systems. Consider a hazard model that includes two endogenous dummy variables. (For an example, see Impicciatore and Billari [2012].) The syntax for fitting models of this kind is

```
gsem
    (event <- xvar_1 xvar_2 indepvars U[idvar], poisson exposure(durvar))
    (xvar_1  <- zvars_1 indepvars V1[idvar], probit)
    (xvar_2  <- zvars_2 indepvars V2[idvar], probit)
```

One can also fit a hazard model with an endogenous qualitative variable jointly with a multinomial selection model:

```
gsem
    (event <- xvar indepvars U[idvar], poisson exposure(durvar))
    (xvar  <- zvars indepvars V[idvar], mlogit)
```

4 Example 1. Nonrecursive simultaneous equations for hazards

4.1 Introduction: The research problem and the dataset

Our first example considers the relationship between education and second-birth rates. We hypothesize that higher education has a positive effect on second-birth hazards (even when higher education has a negative effect on first births). We use a sample dataset on married American women that was shipped with the statistical software aML (Lillard and Panis 2003). The original dataset was converted into a multispell dataset. You can obtain the data as follows:

```
. use "http://web.uni-corvinus.hu/bartus/stata/divorce2.dta"
(Data on marriages (source: divorce4.raw, shipped with aML))
```

The data have a multilevel structure: spells are nested within conception episodes, and conception episodes are nested within individuals. Our sample data include the first two conception episodes within the first marriage. Conception episodes within marriages are identified with the variable **numkids**, measuring the number of children at the beginning of conception episodes. The duration of a conception episode is the difference between two variables, **time** and **mardur**. **mardur** measures the duration of the marriage at the beginning of each spell, while **time** measures the date of separation (or interview date).

We begin by creating separate dummies for first and second conceptions. We use the **separate** command to separate the samples for the study of first and second births. Then, we define the model. The key explanatory variable is **hereduc**, which is a categor-

ical variable with three categories: primary, secondary, and higher education. (Actually, these variables are computed from years of schooling.) I chose secondary education as the reference category. To keep matters simple, I used only the age at the beginning of the conception spell (that is, the mother's age when the first child was born) as a control variable. We place the independent variables and the model definition in global macros. The commands are

```
. separate birth, by(numkids)
  (output omitted)
. global xvars ib2.hereduc age
. global model poisson exposure(dur)
```

We begin our analyses with fitting the model separately. We fit a multilevel model because records within the multispell dataset are nested within individuals. The command and result are

```
. gsem (birth2 <- $xvars U[id], $model)
  (output omitted)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
birth2 <-						
hereduc						
<12 years	-.0389349	.0727403	-0.54	0.592	-.1815032	.1036334
16+ years	.4029357	.1093571	3.68	0.000	.1885998	.6172716
age	-.0914562	.0062165	-14.71	0.000	-.1036404	-.079272
U[id]	1	(constrained)				
_cons	-1.86012	.0506612	-36.72	0.000	-1.959414	-1.760826
ln(dur)	1	(exposure)				
var(U[id])	.6216596	.068415			.5010442	.7713105

The third level of the `hereduc` variable (16+ years of education) has a positive and statistically significant coefficient. This suggests that second-birth rates are relatively high among educated women. In the rest of this section, we control for sample selection and endogeneity to check whether the estimate of 0.403 is robust.

4.2 Joint model for first and second births

Our first concern with the previous result is it might arise because of a selection effect. Education has a negative effect on the transition to first birth, so education will be positively correlated with unobserved causes of fertility in samples of mothers (Kravdal 2007). Therefore, the comparison of the fertility outcomes across educational categories in the sample of mothers measures not only the true effect of education but also the effect of unobserved preferences or personality traits (Kravdal 2001). This selection effect can be controlled for if the parity-specific transitions are modeled jointly by adding person-

specific random intercepts to both the second- and first-birth equations. The command is

```
. gsem (birth2   <- $xvars U[id], $model)
>      (birth1   <- $xvars U[id], $model)
```

(*output omitted*)

Results are not shown because the coefficient of higher education is again positive and statistically significant and, more importantly, the exact value of the estimate, 0.381, is close to the previous estimate. This finding suggests that the positive effect of higher education cannot be explained in terms of sample selection.

4.3 Joint model for second births and marital dissolutions

Our second concern is the dependence of the birth process on the latent hazard of marital dissolution; pessimistic expectations regarding the duration of the marriage are likely to affect second births. To eliminate the bias arising from simultaneity, we now turn to fitting a joint model of the timing of second births and the timing of marital dissolutions. Because the joint model includes reduced-form equations, identifying the structural parameters requires excluded instruments. We assume that age affects only second-birth rates, while the hazard of marital disruption depends exclusively on marriage duration. In other words, age and marital duration are the excluded instruments in the respective birth and dissolution equations. The reduced-form equations include all variables appearing in all structural equations. We again specify the model using a global macro:

```
. global xvars ib2.hereduc age mardur
```

We restrict the analysis to married mothers of one child. The reduced-form system is estimated as follows:

```
. gsem (birth2  <- $xvars U[id], $model)
>      (divorce <- $xvars V[id], $model)
>      if numkids==2
  (output omitted)
Generalized structural equation model           Number of obs     =      5,100
  (output omitted)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
birth2 <-						
hereduc						
<12 years	-.0028288	.068905	-0.04	0.967	-.1378801	.1322225
16+ years	.3128652	.1036965	3.02	0.003	.1096238	.5161066
age	-.0396747	.0087697	-4.52	0.000	-.056863	-.0224863
mardur	-.1071725	.0138669	-7.73	0.000	-.1343512	-.0799939
U[id]	1	(constrained)				
_cons	-1.180374	.0971953	-12.14	0.000	-1.370873	-.9898746
ln(dur)	1	(exposure)				
divorce <-						
hereduc						
<12 years	-.1479639	.1482184	-1.00	0.318	-.4384667	.1425389
16+ years	-.4958348	.2942543	-1.69	0.092	-1.072563	.0808931
age	-.0970939	.0212381	-4.57	0.000	-.1387197	-.0554681
mardur	.0857423	.0290449	2.95	0.003	.0288152	.1426693
V[id]	1	(constrained)				
_cons	-4.391784	.3508382	-12.52	0.000	-5.079414	-3.704153
ln(dur)	1	(exposure)				
var(U[id])	.4275274	.0636296			.3193583	.5723342
var(V[id])	.478624	.3789376			.1014095	2.258969
cov(V[id], U[id])	-.0836689	.1472204	-0.57	0.570	-.3722154	.2048777

Introducing the latent variables implies that five additional parameters should be estimated: the loadings and the variances of the latent variables and the covariance of the latent variables. We can identify three of these parameters because the variance–covariance matrix of the dependent variables includes the variances and the covariance of the outcomes. To identify these parameters, Stata constrains the loadings to unity.

To interpret the results, recall that the birth equation is not a structural equation but a reduced-form equation (see section 2.2). The structural effect of higher education must be recovered using (1). The structural effect is a nonlinear combination of four reduced-form coefficients. This nonlinear combination can easily be computed with the `nlcom` command:

```
. nlcom _b[birth2:3.hereduc] - (_b[birth2:mardur] / _b[divorce:mardur]) *
> _b[divorce:3.hereduc]
       _nl_1:  _b[birth2:3.hereduc] - (_b[birth2:mardur] / _b[divorce:mardur])
> * _b[divorce:3.hereduc]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	-.3068977	.4508039	-0.68	0.496	-1.190457	.5766616

The structural effect of higher education in the second-birth equation, labeled `_nl_1` in the output, is small and lacks statistical significance. This suggests that the partial correlation between higher education and second-birth rates is not direct but might be mediated by the latent separation hazard. This conjecture can easily be tested. Using (2.2), we can compute the structural effect of higher education on the dissolution hazard as follows:

```
. nlcom _b[divorce:3.hereduc] - (_b[divorce:age] / _b[birth2:age]) *
> _b[birth2:3.hereduc]
       _nl_1:  _b[divorce:3.hereduc] - (_b[divorce:age] / _b[birth2:age]) *
> _b[birth2:3.hereduc]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	-1.261495	.4305668	-2.93	0.003	-2.10539	-.4175992

Using (2.2), we see that the effect of the dissolution hazard on the second-birth hazard is

```
. nlcom _b[birth2:mardur] / _b[divorce:mardur]
       _nl_1:  _b[birth2:mardur] / _b[divorce:mardur]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	-1.249938	.4477062	-2.79	0.005	-2.127426	-.3724502

These linear combinations support the hypothesis that the positive effect of higher education on second births is mediated by the latent hazard of marital separation: highly educated women tend to live in relatively stable marriages, and marital stability has a positive effect on second-birth rates.

4.4 Joint model for first births, second births, and marital dissolution

In the previous subsections, we first modeled first- and second-birth processes, then modeled second-birth and marital dissolution processes jointly. The respective concerns were sample selection bias and endogeneity bias. We can address these concerns at the same time and estimate the first-birth, second-birth, and marital dissolution equations jointly. Indeed, this model is very close to the classic multilevel multiprocess model

presented in Lillard (1993). As described in section 3.3, we specify two correlated latent variables for the respective conception and marital dissolution processes. The syntax is

```
. gsem (birth2  <- $xvars U[id], $model)
>      (birth1  <- $xvars U[id], $model)
>      (divorce <- $xvars V[id], $model)
  (output omitted)
```

We omit the output because the ultimate interest lies in the structural coefficients. These can be recovered by computing the appropriate nonlinear combination:

```
. nlcom _b[birth2:3.hereduc] - (_b[birth2:mardur] / _b[divorce:mardur]) *
> _b[divorce:3.hereduc]
       _nl_1:  _b[birth2:3.hereduc] - (_b[birth2:mardur] / _b[divorce:mardur])
> * _b[divorce:3.hereduc]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	.1031687	.1907404	0.54	0.589	-.2706757	.4770131

Again, there is no evidence that higher education would have a direct effect on second-birth rates. By contrast, there is evidence that the positive effect of higher education is an indirect one, mediated by the latent dissolution hazard. The respective nonlinear combinations that estimate the direct effect of higher education on the dissolution hazard and the effect of the dissolution hazard on the second-birth hazard are as follows:

```
. nlcom _b[divorce:3.hereduc] - (_b[divorce:age] / _b[birth2:age]) *
> _b[birth2:3.hereduc]
       _nl_1:  _b[divorce:3.hereduc] - (_b[divorce:age] / _b[birth2:age]) *
> _b[birth2:3.hereduc]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	-.5290925	.2280571	-2.32	0.020	-.9760761	-.0821088

```
. nlcom _b[birth2:mardur] / _b[divorce:mardur]
       _nl_1:  _b[birth2:mardur] / _b[divorce:mardur]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	-.8249733	.2544315	-3.24	0.001	-1.32365	-.3262967

5 Example 2. Hazard models with endogenous dummy variables

5.1 Introduction: The research problem and the dataset

Our second example is about examining the impact of hospital delivery on child mortality. We hypothesize that children delivered in hospitals have a lower death hazard than similar children delivered at home. We use a modified and Stata-compatible version of the children dataset shipped with the statistical software aML to replicate one of the examples in the aML manual (Lillard and Panis 2003). You can obtain the data as follows:

```
. use "http://web.uni-corvinus.hu/bartus/stata/children1.dta", clear
(Child mortality data (source: aML))
```

The data have a multispell and multilevel structure: spells are nested within children, identified with the variable `bid`, and children are nested within mothers, identified with the variable `id`. For simplicity, the survival process is split into two spells; the first spell lasts three months (or less in case of early death). Episode splitting is motivated by the observation that child mortality is relatively large in the first three months. The survival process is described by three variables: `death` indicates deaths, `dur` records the duration of the spell (in months), and `age0` is the age of the child (in months) at the beginning of the spell.

We begin with estimating a simple multilevel child mortality hazard equation. The explanatory variables include the hospital dummy, education, and an indicator for being aged three months at the beginning of the current spell. For simplicity, we assume that the mortality hazard is constant within the spells. We use a random intercept at the level of mothers to model the interdependence of spells within mothers. We place the independent variables and the model definition in global macros. The commands are

```
. global death hospital i.edu  i.age0
. global model poisson exposure(dur)
. gsem (death <- $death  U[id], $model)
```

(*output omitted*)

Estimates are not shown. The coefficient of the hospital dummy is negative (the estimate is -0.382) but statistically not significant ($p = 0.064$). The robustness of this estimate is examined in the next subsection.

5.2 Joint estimation of hazard and probit equations

Finding no statistically significant negative effect of hospital delivery might be due to a selection effect. Mothers are aware of their health status and form an expectation about the mortality of their child. Hospital delivery is chosen by mothers who fear losing their baby and believe hospitals reduce this risk. By contrast, home delivery is chosen by women with a low risk of losing their baby. In short, hospital delivery is correlated with

factors affecting child mortality. To control for this endogeneity bias, one can fit the hazard model jointly with a probit model of hospital delivery on education and distance to the nearest hospital, the latter being the excluded instrument. The joint model is fit as follows:

```
. global hospital distance i.edu
. gsem (death    <- $death    U[id], $model)
>      (hospital <- $hospital V[id], probit)
  (output omitted)
Generalized structural equation model           Number of obs     =      2,002
  (output omitted)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
death <-						
hospital	-.5131628	.2411954	-2.13	0.033	-.9858971	-.0404285
educ						
high school	-.2625067	.1909157	-1.37	0.169	-.6366945	.1116811
college	-2.021169	.7341519	-2.75	0.006	-3.46008	-.5822573
3.age0	-4.920847	.1656668	-29.70	0.000	-5.245548	-4.596146
U[id]	1	(constrained)				
_cons	-3.12697	.1432276	-21.83	0.000	-3.407691	-2.846249
ln(dur)	1	(exposure)				
hospital <-						
distance	-.0231453	.0175738	-1.32	0.188	-.0575894	.0112987
educ						
high school	2.01218	.2895358	6.95	0.000	1.4447	2.57966
college	3.148736	.5114086	6.16	0.000	2.146393	4.151078
V[id]	1	(constrained)				
_cons	-2.209737	.2767038	-7.99	0.000	-2.752066	-1.667407
var(U[id])	.4091622	.2339894			.1333875	1.255093
var(V[id])	4.149642	1.079965			2.491617	6.910987
cov(V[id], U[id])	.2157169	.1885667	1.14	0.253	-.1538671	.5853009

The coefficient of the hospital delivery variable is now statistically significant. The estimate of -0.513 is larger than that appearing in the separate model. This suggests that hospital delivery has the expected negative effect on mortality, but this effect was partially suppressed by the aforementioned selection effect.

The present example assumes that the hazard of death is constant within the spells. Descriptive analyses, not reported in this article, suggest this assumption is unrealistic. The hazard is monotonically decreasing in the first three months, while it is approximately constant after surviving the first three months. A more realistic model specification would be a Weibull model, which can be fit as follows:

```
. global model family(weibull, fail(death) lt(age0))
. gsem (month    <- $death    U[id], $model)
>      (hospital <- $hospital V[id], probit)
```

(*output omitted*)

Note that the dependent variable in the hazard equation is the variable recording the survival time. The output is not reported because the coefficient of the hospital delivery dummy is statistically significant, and the size of the coefficient is very close to the previously estimated -0.513.

5.3 Joint estimation of hazard and multinomial logit equations

Suppose that children can be delivered in public hospitals, in private hospitals, and at home. Suppose further that hospital delivery improves life expectancy, but the negative effect of hospital delivery on child mortality differs between private and public hospitals. Women are expected to select the delivery form, which minimizes the risks but also economizes on (travel and other) costs. Again, the chosen form of delivery will be correlated with factors affecting the health of the child. To eliminate the endogeneity bias, one must fit the hazard model jointly with a multinomial model of delivery choice. The `gsem` command allows one to fit hazard models jointly with multinomial selection equations. To illustrate, we use a modified version of the child mortality dataset. The specification of the hazard and the selection equations is not changed. The only change is that we use a multinomial logit selection model instead of a probit model. The commands are

```
. use "http://web.uni-corvinus.hu/bartus/stata/children2.dta", clear
(Child mortality data (source: aML))
. global death i.hospital i.edu  i.age0
. global hospital distance i.edu
. global model poisson exposure(dur)
```

```
. gsem (death    <- $death    U[id], $model)
>      (hospital <- $hospital V[id], mlogit)
```

(*output omitted*)

```
Generalized structural equation model           Number of obs     =      2,002
```

(*output omitted*)

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
death <-						
hospital						
1	-.5085083	.2527484	-2.01	0.044	-1.003886	-.0131305
2	-3.024408	.5896255	-5.13	0.000	-4.180053	-1.868763
educ						
high school	-.1914267	.1871073	-1.02	0.306	-.5581502	.1752968
college	-1.965382	.7309886	-2.69	0.007	-3.398093	-.5326702
3.age0	-4.836307	.1661777	-29.10	0.000	-5.16201	-4.510605
U[id]	1	(constrained)				
_cons	-2.885532	.1392459	-20.72	0.000	-3.158449	-2.612615
ln(dur)	1	(exposure)				
0.hospital	(base outcome)					
1.hospital <-						
distance	-.0530196	.0337748	-1.57	0.116	-.1192171	.0131779
educ						
high school	3.239981	.4779804	6.78	0.000	2.303157	4.176805
college	4.849681	.7717843	6.28	0.000	3.337011	6.36235
V[id]	1	(constrained)				
_cons	-3.793961	.4544799	-8.35	0.000	-4.684725	-2.903196
2.hospital <-						
distance	-.0060001	.0212732	-0.28	0.778	-.0476949	.0356946
educ						
high school	1.074023	.1810285	5.93	0.000	.7192138	1.428833
college	1.707243	.3263577	5.23	0.000	1.067594	2.346892
V[id]	.2732314	.0500081	5.46	0.000	.1752174	.3712454
_cons	-1.293195	.1406849	-9.19	0.000	-1.568932	-1.017458
var(U[id])	.2992881	.2126223			.0743658	1.204497
var(V[id])	13.02677	2.624814			8.776573	19.3352
cov(V[id], U[id])	.3196824	.3756769	0.85	0.395	-.4166308	1.055996

Again, we find that hospital delivery reduces child mortality compared with home delivery. The effect is larger in hospitals coded with 2 than in hospitals coded with 1. (It is up to the reader to interpret the 1 code as a private or as a public hospital.)

6 Concluding remarks

Demographers routinely use multilevel multiprocess models to adjust estimates for endogeneity and sample selection. In this article, I showed how multilevel multiprocess models could be fit with the `gsem` command. I provided two examples to illustrate the estimation of nonrecursive systems without observed endogenous variables and recursive systems with observed endogenous variables. The examples used sample datasets shipped with the statistical software aML, explicitly developed for multiprocess multilevel modeling (Lillard and Panis 2003). I paid special attention to identifying structural effects in nonrecursive systems.

Most of the examples in this article illustrate the estimation of systems with two equations. In some empirical applications, however, more than two equations are estimated jointly (Upchurch, Lillard, and Panis 2002; Steele et al. 2005). As the number of equations increases, the number of correlated random intercepts increases. Fitting models with a large number of random effects is slow and may have convergence problems. Referencing the classic article on multilevel multiprocess modeling (Lillard 1993), I suggested a simple rule to avoid or minimize numerical problems: the number of latent variables must be equal to the number of processes under study, but separate equations for recurrent (or sequential) occurrences of events of the same kind should share the same latent variable.

For simplicity, I used (piecewise-constant) exponential hazard models for the purpose of survival modeling. As shown in section 3.1, `gsem` supports a large class of parametric survival models. Recently, multilevel multiprocess models often rely on discrete-time (binary and multinomial) logistic regression models (Steele et al. 2005). However, the Poisson model is flexible enough to model duration dependence and represent discrete-time event-history models. In theory, systems of logit and multinomial logit models can easily be estimated with `gsem`. In conclusion, the `gsem` command is a powerful tool to fit various forms of multilevel multiprocess models. I believe the examples shown in this article will help researchers solve complicated research problems.

Acknowledgments

An earlier version of this article was presented at the 2015 German Stata Users Group Meeting in Nuremberg on June 26, 2015. The research in this article was financially supported by the János Bolyai Scholarship of the Hungarian Academy of Sciences as well as by the Hungarian Scientific Research Fund (OTKA) within the research project *Mapping Family Transitions: Causes, Consequences, Complexities, and Context* (grant number K109397).

7 References

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

Heckman, J. J. 1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46: 931–959.

Holford, T. R. 1980. The analysis of rates and of survivorship using log-linear models. *Biometrics* 36: 299–305.

Impicciatore, R., and F. C. Billari. 2012. Secularization, union formation practices, and marital stability: Evidence from Italy. *European Journal of Population* 28: 119–138.

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

———. 2007. Effects of current education on second- and third-birth rates among Norwegian women and men born in 1964: Substantive interpretations and methodological issues. *Demographic Research* 17: 211–246.

Lee, L. 1979. Identification and estimation in binary choice models with limited (censored) dependent variables. *Econometrica* 47: 977–996.

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: 437–457.

Lillard, L. A., and C. W. A. Panis. 2003. *aML Multilevel Multiprocess Statistical Software, Version 2.0*. Los Angeles, CA: EconWare.

Lillard, L. A., and L. J. Waite. 1993. A joint model of marital childbearing and marital disruption. *Demography* 30: 653–681.

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

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

Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

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.

Upchurch, D. M., L. A. Lillard, and C. W. A. Panis. 2002. Nonmarital childbearing: Influences of education, marriage, and fertility. *Demography* 39: 311–329.

About the author

Tamás Bartus is an associate professor of sociology at the Corvinus University of Budapest. His research interests include the impact of education on partnership stability and fertility.