# The Stata Journal



The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

**Subscriptions** are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

**Subscription rates** listed below include both a printed and an electronic copy unless otherwise mentioned.

| U.S. and Canada | | Elsewhere | |
|---|---|---|---|
| **Printed & electronic** | | **Printed & electronic** | |
| 1-year subscription | $115 | 1-year subscription | $145 |
| 2-year subscription | $210 | 2-year subscription | $270 |
| 3-year subscription | $285 | 3-year subscription | $375 |
| 1-year student subscription | $ 85 | 1-year student subscription | $115 |
| 1-year institutional subscription | $345 | 1-year institutional subscription | $375 |
| 2-year institutional subscription | $625 | 2-year institutional subscription | $685 |
| 3-year institutional subscription | $875 | 3-year institutional subscription | $965 |
| **Electronic only** | | **Electronic only** | |
| 1-year subscription | $ 85 | 1-year subscription | $ 85 |
| 2-year subscription | $155 | 2-year subscription | $155 |
| 3-year subscription | $215 | 3-year subscription | $215 |
| 1-year student subscription | $ 55 | 1-year student subscription | $ 55 |

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.




The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2014)
**14**, Number 4, pp. 756–777

# Estimation of multiprocess survival models with cmp

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

David Roodman
Freelance Public Policy Consultant
Washington, DC
david@davidroodman.com

**Abstract.** Multilevel multiprocess hazard models are routinely used by demographers to control for endogeneity and selection effects. These models consist of multilevel proportional hazards equations, and possibly probit equations, with correlated random effects. Although Stata currently lacks a specialized command for fitting systems of multilevel proportional hazards models, systems of seemingly unrelated lognormal survival models can be fit with the user-written `cmp` command (Roodman 2011, *Stata Journal* 11: 159–206). In this article, we describe multiprocess survival models and demonstrate theoretical and practical aspects of estimation. We also illustrate the application of the `cmp` command using examples related to demographic research. The examples use a dataset shipped with the statistical software aML.

**Keywords:** st0358, survival analysis, multilevel analysis, multilevel multiprocess hazard model, simultaneous equations, SUR estimation, cmp

## 1 Introduction

Multilevel multiprocess hazard models are routinely used by demographers to adjust regression estimates for endogeneity and selection effects. Originally, multilevel multiprocess models were developed as systems of proportional hazards models with correlated individual-level random effects (Lillard 1993). The multilevel multiequation modeling framework also accommodates the joint estimation of hazard and probit equations to account for the endogeneity of dummy explanatory variables that appear in the hazard equation of primary interest (Lillard, Brien, and Waite 1995; Impicciatore and Billari 2012). The joint estimation accounts for the correlation of the random effects and allows researchers to control for the effects of unobserved personality traits.

To date, no official Stata commands are devoted to estimating systems of survival models. Therefore, it is not surprising that multiprocess hazard models are fit using other statistical packages, including aML (Lillard and Panis 2003) and MLwiN (Rasbash et al. [2012]; see also Leckie and Charlton [2013]). With the recently introduced `gsem` command, Stata can now fit systems of survival models with correlated random effects. In this article, we demonstrate how Stata users can fit multiprocess models with the user-written `cmp` command (Roodman 2011). `cmp` is a flexible tool to estimate systems of equations with various link functions and with normally distributed

 

and correlated errors. Although most of the bivariate models fit in the article could also be fit with `gsem`, `cmp` offers two advantages over `gsem` for survival modeling. First, `cmp` is structured to allow cross-equation correlations in modeling errors, even when equations have probit, interval, or other kinds of censoring. Doing the same with `gsem` is cumbersome because it requires a user to create latent variables, impose constraints upon them, and mathematically transform the results for intuitive interpretation. Second, `cmp` naturally deals with the truncated outcomes, which is necessary when using multispell survival data.

In this article, we explain how `cmp` can be used to fit systems of lognormal survival models or systems that may also include probit models. We also show how recent additions to the `cmp` package enable researchers to fit systems of multilevel models with random effects. This article is organized as follows. In section 2, we describe multiprocess hazard models, as developed by Lillard (1993). In section 3, we describe the `cmp` compatible models that we label multiprocess survival models. We give the syntax for estimation in section 4. We then present examples in section 5 using a dataset shipped with aML, and we conclude in section 6.

# 2 Multiprocess hazard models

## 2.1 Motivation

Multiprocess modeling was motivated by the insight that explanatory variables are often endogenous because of selection mechanisms. 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 because of the presence of two forms of endogeneity (Lillard and Waite 1993). First, having children is the outcome of a process of timing of births. Second, the latter process may depend on the latent propensity to end the marriage: if couples expect their marriage to be short-lived, they may postpone the first (or higher-order) births. Therefore, the latent expectation of marriage dissolution creates a spurious negative relationship between the number of children and the hazard of marital dissolution. If children negatively affect the hazard of separation, the relationship between children and marital dissolution might be both positive and negative.

The classic method of eliminating the endogeneity bias is to estimate systems of equations with joint normally distributed disturbances (Heckman 1978). Let $y_1^*$ and $y_2^*$ denote the endogenous latent variables under study; for instance, the former might denote the hazard of conception, and the latter might denote the hazard of marital dissolution. The dependence of each latent variable on the other, as well as on other explanatory variables, is described with the following structural equations

$$y_1^* = \boldsymbol{\alpha}_1' \mathbf{X}_1 + \lambda_1 y_2^* + \varepsilon_1$$
$$y_2^* = \boldsymbol{\alpha}_2' \mathbf{X}_2 + \lambda_2 y_1^* + \varepsilon_2$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are vectors of observed variables, and $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are vectors of coefficients. Observed realizations of the latent variables, like marital status or the number of children, may be included in the explanatory variables.

Endogeneity arises from the presence of latent variables on the right-hand side of the structural equations. Hence, one should estimate the system of reduced-form equations. Estimation must consider that the residuals in the reduced-form equations are probably correlated, even when the disturbances in the structural equations are independent of each other. If the latter error terms were normally distributed, the following system, supplemented with appropriate link functions, must be estimated:

$$
\begin{aligned}
y_1^* &= \boldsymbol{\beta}_{11}'\mathbf{X}_1 + \boldsymbol{\beta}_{12}'\mathbf{X}_2 + v_1 \\
y_2^* &= \boldsymbol{\beta}_{21}'\mathbf{X}_1 + \boldsymbol{\beta}_{22}'\mathbf{X}_2 + v_2 \\
\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} &\sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right)
\end{aligned}
$$

The estimation of the variance–covariance matrix of the residuals also makes it easier to interpret the results and separate the causal and selection effects (Heckman 1978). The reduced-form coefficients depend on the corresponding structural parameters and the selection parameters $\lambda_1$ and $\lambda_2$. The estimation of the covariance matrix of the reduced-form residuals enables one to estimate the selection effect, because the elements of this matrix depend on the selection parameters and not on the structural parameters. One can then use the available estimates of the selection effects to identify the structural parameters. For instance, suppose that $\lambda_2$ were constrained to be 0 and the variable $x_j$ appears in both $\mathbf{X}_1$ and $\mathbf{X}_2$. Here $\boldsymbol{\beta}_{21} = \boldsymbol{\alpha}_2$ and $\boldsymbol{\beta}_{22} = 0$, which implies that $\beta_{2j} = \alpha_{2j}$. Hence, the reduced-form coefficient of $x_j$ in the first equation, $\beta_{1j} = \alpha_{1j} + \lambda_1\beta_{2j}$, and the estimated covariance of the residuals is $\lambda_1\sigma_2^2$. Therefore, the structural coefficient can be recovered as

$$
\alpha_{1j} = \beta_{1j} - \lambda\beta_{2j} = \beta_{1j} - \frac{\sigma_{12}}{\sigma_2^2}\beta_{2j} \tag{1}
$$

## 2.2 Multilevel multiprocess hazard models

The estimation strategy outlined above cannot be applied to survival analysis without further modifications or extensions. In the popular proportional hazards models, the log of the hazard rate equals the linear combination of variables and coefficients. This model can be restated as a latent-variable model in which the random component of the latent variable follows an exponential distribution. Without a widely accepted multivariate exponential distribution, the correlation of the underlying residuals cannot be modeled, and the seemingly unrelated estimation strategy seems to be infeasible.

To solve this problem, Lillard (1993) suggested the joint estimation of hazard models including normally distributed and possibly correlated random effects. Including jointly normally distributed random effects allows one to adjust estimates for the correlation of the total underlying residuals, and it allows one to estimate the covariance matrix of residuals and, hence, the selection effects. Note that the assumption of joint normality applies only to the random effects. Note also that identifying the random effects requires

repeated occurrences of outcomes. This data requirement is fortunately easy to meet because demographic events such as marriage, divorce, and giving birth are recurrent.

The resulting model, often labeled as a multilevel multiprocess model, can be stated as follows

$$\begin{aligned} y^*_{1j} &= \boldsymbol{\beta}'_1 \mathbf{X}_{1j} + u_1 + v_{1j} \\ y^*_{2j} &= \boldsymbol{\beta}'_2 \mathbf{X}_{2j} + u_2 + v_{2j} \\ \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} &\sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{1.2} & \sigma_{12.2} \\ \sigma_{12.2} & \sigma^2_{2.2} \end{bmatrix}\right) \end{aligned} \tag{2}$$

where $j$ indexes the recurrent observations, and for simplicity, $\mathbf{X}_{1j}$ and $\mathbf{X}_{2j}$ encompass all explanatory variables in the equations. The subscript $j$ expresses that the latent outcomes as well as the explanatory variables may change over time, in general, and over spells, in particular. All variances and covariances are indexed by 2, indicating that the matrix expresses these covariances among level-2 residuals.

The model must be supplemented with appropriate link functions to reflect the incomplete observability of $y^*_{1j}$ and $y^*_{2j}$. In addition to the link function of proportional hazards models, other link functions can be accommodated. For instance, the multilevel hazard model of marital dissolution, which includes premarital cohabitation as an explanatory variable, may be fit jointly with a probit model of premarital cohabitation to control for a possible selection effect that arises because individuals who are less willing or less able to live in marriages may choose cohabitation or are at an above-average risk of divorce upon marriage (Lillard, Brien, and Waite 1995).

## 2.3 Fitting multilevel multiprocess models with gsem

To date, no official Stata command is explicitly devoted to estimating systems of hazard equations with correlated random effects. The official `gsem` command, however, could potentially fit such models. Because the focus of our article is not `gsem`, we just outline the procedure. The starting point is the equivalence of exponential hazard models and Poisson models (Holford 1980): if survival time $t$ follows an exponential distribution with parameter $h$, then the expected number of failures follows a Poisson distribution with parameter $ht$. Relying on this result, Skrondal and Rabe-Hesketh (2004) show that exponential hazard models can be restated as Poisson regression models, in which the dependent variable is the failure-indicator variable and the explanatory variables include the natural log of the duration of the current spell. These results suggest that a user can use `gsem` to estimate (2) when the user specifies the Poisson family. The user should also include correlated latent variables in the equations to represent the correlation of the random effects (the $u$'s) across the equations.

# 3 Multiprocess modeling with cmp

## 3.1 The multiprocess lognormal survival model

In this article, we argue that the user-written `cmp` command (Roodman 2011) allows one to fit systems of lognormal survival models with jointly normally distributed error terms. The `cmp` command supports interval regression models, even ones with truncated dependent variables. Note that the lognormal survival model is just an interval-censored regression of log failure-times. Note also that the lognormal model is formulated exclusively in the accelerated failure-time metric, so it cannot be formulated in the proportional hazards metric (Cleves et al. 2010). In short, multiprocess modeling boils down to fitting lognormal survival models jointly with other lognormal survival models or with probit models for endogenous regressors.

The seemingly unrelated system of log failure-times for $P$ interdependent survival processes (that is, the multiprocess lognormal survival model) can be defined as

$$\begin{aligned}
\log T_1 &= \boldsymbol{\beta}_1' \mathbf{X}_{1j} + \varepsilon_{1j} \\
&\cdots \\
\log T_P &= \boldsymbol{\beta}_P' \mathbf{X}_{Pj} + \varepsilon_{Pj} \\
\boldsymbol{\varepsilon} &\sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1P} \\ \vdots & \ddots & \vdots \\ \sigma_{1P} & \cdots & \sigma_P^2 \end{bmatrix}\right)
\end{aligned} \tag{3}$$

Subscript $j$ expresses that the explanatory variables and the residuals change over time and over the observation periods indexed by $j$. The process-specific equations are seemingly unrelated because the process-specific error terms can be correlated.

Unlike the classic multilevel multiprocess models, the current model does not include individual-specific random effects. Once the process-specific errors are assumed to be normal, they can easily be modeled as intercorrelated (the normal distribution easily generalizes to multiple dimensions). Therefore, it also becomes less essential to introduce individual-level random effects when we want cross-equation correlation. Nevertheless, we can add higher-level residuals to the structural equations, which are also assumed to be jointly normally distributed.

$$
\begin{aligned}
\log T_{1j} &= \boldsymbol{\beta}_1' \mathbf{X}_{1j} + u_1 + \varepsilon_{1j} \\
&\cdots \\
\log T_{Pj} &= \boldsymbol{\beta}_P' \mathbf{X}_{Pj} + u_P + \varepsilon_{Pj} \\
\boldsymbol{\varepsilon} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1) \\
\mathbf{u} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_2) \\
\boldsymbol{\Sigma}_1 &= \begin{bmatrix} \sigma_{1.1}^2 & \cdots & \sigma_{1P.1} \\ \vdots & \ddots & \vdots \\ \sigma_{1P.1} & \cdots & \sigma_{P.1}^2 \end{bmatrix} \\
\boldsymbol{\Sigma}_2 &= \begin{bmatrix} \sigma_{1.2}^2 & \cdots & \sigma_{1P.2} \\ \vdots & \ddots & \vdots \\ \sigma_{1P.2} & \cdots & \sigma_{P.2}^2 \end{bmatrix}
\end{aligned} \tag{4}
$$

## 3.2 Maximum likelihood estimation

We can estimate the parameters of (2) and (3) using maximum likelihood. We must write the formula for the likelihood, which we do at the individual level because the random effects are correlated (identical) across the observations that constitute each individual's history.

We consider multispell data. Multiprocess modeling often requires multispell data because qualitatively different events rarely happen simultaneously. That is, the exogenous or endogenous events throughout individual $i$'s life divide it into $J$ episodes or spells. (For simplicity, subscript $i$ is omitted.) For each process $p$, the outcome of the process in episode $j$ is characterized by two variables: the time variable $t_{pj}$ and the failure-indicator variable $y_{pj}$. The former measures the time to the occurrence of the event (or censoring). Termination of the process $p$ in episode $j$ is indicated by $y_{pj} = 1$; censoring is indicated by 0 values.

We begin with the log likelihood for a process with no random effect. We can construct the log-likelihood contribution of any individual as follows. Log time to event is either observed or censored. With the exception of the first spell, log time to event is left-truncated because waiting times in spell $j$ must be larger than the time elapsed until the beginning of that spell. The linear combination of the explanatory variables and the parameters is denoted by $\theta_j = \boldsymbol{\beta}' \mathbf{X}_j$. Explanatory variables might change over time but are assumed to be constant within each spell $j$. The log likelihood for individual $i$ is given by

$$
\log L = \sum_{j=1}^{J} \left\{ y_j \log f_j + (1 - y_j) \log S_j - \log S_j^0 \right\} \tag{5}
$$

where

$$
\begin{aligned}
f_j &= \phi\left(\log t_j - \theta_j; \sigma\right) \\
S_j &= \Phi\left(\frac{\theta_j - \log t_j}{\sigma}\right) \\
S_j^0 &= \Phi\left(\frac{\theta_j - \log t_{j-1}}{\sigma}\right)
\end{aligned}
$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the normal probability density and cumulative density functions and $t_0 = 0$. (See, for instance, Klein and Moeschberger [2003].) Following Roodman (2011), (4) can be rewritten as

$$
\log L_i = \sum_{j=1}^{J} \log \frac{\int_{A_j}^{B_j} f(\varepsilon) d\varepsilon}{\int_{C_j}^{\infty} f(\varepsilon) d\varepsilon}
$$

where

$$
\begin{aligned}
f(\varepsilon) &= \phi(\varepsilon; \sigma) \\
A_j &= \log t_j - \theta_j \\
B_j &= y_j(\log t_j - \theta_j) + (1 - y_j)\infty \\
C_j &= \log t_{j-1} - \theta_j
\end{aligned}
$$

The interpretation of the first integrand requires the conventions that $\int_a^a f(\varepsilon) d\varepsilon = f(\varepsilon)$, and $0\infty = 0$.

The generalization of the log-likelihood expression to multiprocess models is straightforward. For simplicity, we consider only two simultaneous processes. The log-likelihood expression involves the two-dimensional integral

$$
\log L_i = \sum_{j=1}^{J} \log \frac{\int_{A_{1j}}^{B_{1j}} \int_{A_{2j}}^{B_{2j}} f(\varepsilon_{1j}, \varepsilon_{2j}) d\varepsilon_{2j} d\varepsilon_{1j}}{\int_{C_{1j}}^{\infty} \int_{C_{2j}}^{\infty} f(\varepsilon_{1j}, \varepsilon_{2j}) d\varepsilon_{2j} d\varepsilon_{1j}}
$$

where

$$
\begin{aligned}
A_{pj} &= \log t_{pj} - \theta_{pj} \\
B_{pj} &= y_{pj}(\log t_{pj} - \theta_{pj}) + (1 - y_{pj})\infty \\
C_{pj} &= \log t_{p(j-1)} - \theta_{pj}
\end{aligned}
$$

The log-likelihood formula extends to $P$ processes analogously with order-$P$ integrals. Adding an equation to model a dummy endogenous variable introduces further complications, though the principles remain the same. Users can refer to Roodman (2011) for more details on formulation and practical computation of the integrals.

To add a random effect to the two-process model, as an example, we redefine $\theta_{pj} = \boldsymbol{\beta}_{p_j}{}'\mathbf{X_{pj}} + u_p$, where $u$ is the random effect defined earlier. Because $u$ is unobserved, to compute the individual-level likelihood, we must integrate it out.

$$\log L_i = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left\{\sum_{j=1}^{J}\log\frac{\int_{A_{1j}}^{B_{1j}}\int_{A_{2j}}^{B_{2j}} f(\varepsilon_{1j},\varepsilon_{2j})d\varepsilon_{2j}d\varepsilon_{1j}}{\int_{C_{1j}}^{\infty}\int_{C_{2j}}^{\infty} f(\varepsilon_{1j},\varepsilon_{2j})d\varepsilon_{2j}d\varepsilon_{1j}}\right\}\phi\{(u_1,u_2)';\boldsymbol{\Sigma}_2\}du_2du_1$$

In general, the outer integrals can only be computed using numerical methods. The dominant methods are adaptive quadrature and simulation. Roughly speaking, adaptive quadrature computes the integrand at 6–24 points and averages the results with special weights (Rabe-Hesketh, Skrondal, and Pickles 2002). Simulation computes the integrand at 50, 100, or 500 points and takes the simple average of the results (Haan and Uhlendorff 2006; Train 2009). The `cmp` command offers both methods as part of a multilevel modeling extension that was added following Roodman (2011).

# 4 Syntax

The description of the syntax is restricted to components of the `cmp` command that are specific to multiprocess survival modeling. For the full syntax of `cmp`, consult Roodman (2011) and the help file for `cmp`.

## 4.1 Single-equation survival models

We begin with the syntax for ordinary (or single-process single-level) lognormal survival models. Lognormal survival models are interval-censored regression models of log failure-times. We can fit interval-censored regression models with `cmp` as follows. Like the official `intreg` command, `cmp` expects two dependent variables that indicate the lower and upper bounds of failure-time. For observed failures (or uncensored observations), these limits are the same and are equal to the observed log failure-time. For censored durations, the lower limit is the log time of the interview and the upper limit is infinity, meaning that the event will occur somewhere in the future. We let *lo* and *hi* denote the lower- and upper-limit variables, respectively. An interval regression model of log failure-times is estimated with `cmp` using the following syntax

`cmp` ([ *label*: ] *lo hi* `=` *indepvars*, [ noconstant ]) [ *if* ] [ *in* ] [ *weight* ],
indicators(7) [ *options* ]

where *label* is used instead of *lo* in the output, the mandatory `indicators(7)` tells `cmp` that the equation to be estimated is an interval-censored regression, and *options* is just shorthand for any other options incorporated in `cmp`. The `indicators(7)` option can be replaced by `indicators($cmp_intreg)` if the `cmp setup` command is issued at the beginning of the session.

When using multispell data, researchers should account for the left-truncation of the outcome as well as the interdependence of residuals within individuals. Suppose the variable *id* identifies the individuals. Let *durvar* be the variable that records the log of entry time, which is the log duration of a state measured at the beginning of the spell. To account for the truncation of survival time as well as the dependence of residuals within individuals, the basic syntax is extended as follows:

```
cmp ([ label: ] lo hi = indepvars, truncpoints(durvar .) [noconstant]) [ if ]
    [ in ] [ weight ], indicators(7) vce(cluster id) [ options ]
```

Another way to account for the correlation of residuals within individuals' histories is to use a multilevel survival model. To accommodate the individual-level random effect, the basic syntax is modified as follows:

```
cmp ([ label: ] lo hi = indepvars, [noconstant] || id:) [ if ] [ in ] [ weight ],
    indicators(7) [ options ]
```

## 4.2 Syntax for multiprocess models

Next, we will fit systems of lognormal survival models. For simplicity, the exposition considers two processes. We let *lo_1* and *hi_1* denote the respective lower and upper limits for the first process, and we let *lo_2* and *hi_2* denote the respective lower and upper limits for the second process. Multiprocess modeling requires multispell data; we denote the respective entry times for the processes by *durvar_1* and *durvar_2*. The syntax is

```
cmp ([ label_1: ] lo_1 hi_1 = indepvars_1, truncpoints(durvar_1 .)
    [noconstant]) ([ label_2: ] lo_2 hi_2 = indepvars_2, truncpoints(durvar_2 .)
    [noconstant]) [ if ] [ in ] [ weight ], indicators(7 7) vce(cluster id)
    [ options ]
```

Note that the `indicators()` option now has two arguments, one for each equation. `indicators(7 7)` means that both equations are interval-censored. For clarity, you may instead type `indicators($cmp_intreg $cmp_intreg)`, provided that you issued the `cmp setup` command at the beginning of the session.

`cmp` also allows users to fit lognormal survival models jointly with probit models. This is useful if the survival models of primary interest include endogenous dummy variables. For simplicity, we consider one survival process and one probit equation. We let *lo* and *hi* denote the respective lower and upper limits for the survival process. The endogenous dummy variable is denoted by *dvar*. The syntax is

cmp ([ *label*: ] *lo hi* = *dvar indepvars_1*, truncpoints(*durvar* .) [noconstant])
(*dvar* = *indepvars_2*, [noconstant]) [*if*] [*in*] [*weight*], indicators(7 4)
vce(cluster *id*) [*options*]

The 4 (alternatively, $cmp_probit) requests a probit model for the second equation.

Heckman-type modeling to control for sample selection bias is also possible. Suppose that the survival model applies to a subsample that is identified by the indicator variable *sample*. Survival estimates can be adjusted for sample selection bias by using the following syntax:

cmp ([ *label*: ] *lo hi* = *indepvars_1*, truncpoints(*durvar* .) [noconstant])
(*sample* = *indepvars_2*, [noconstant]) [*if*] [*in*] [*weight*],
indicators("sample*7" 4) vce(cluster *id*) [*options*]

The second probit equation applies to the sample defined by the optional if, in, and weight syntax elements. The survival model, however, is fit using observations where *sample* equals 1.

# 5 Examples

## 5.1 Introduction: The research problem and the dataset

Our examples consider the relationship between education and the timing of births. Evidence suggests that highly educated women who postpone the transition to motherhood space the first and the second births closer together. We use a sample dataset that comes with the statistical software aML (Lillard and Panis 2003). The dataset contains information on marital births and marriage durations for American women. The slightly modified and Stata-compatible version is obtained as follows:

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

The data have a multilevel structure; conception episodes are nested within marriages, and marriages are nested within individuals. Marriages within individuals are identified with the variable marnum, and conception episodes within marriages are identified with the variable numkids, which measures the number of kids at the beginning of conception episodes. Each row records the duration of a conception episode; the duration is the difference between two variables, time and mardur. mardur measures the duration of the marriage at the beginning of each conception episode, and time measures the date of separation (or interview date).

We use data on only the first two conception episodes within the first marriages. For convenience, we recode the `numkids` variable to indicate parity or birth order. The Stata commands are as follows:

```
. keep if marnum==1 & numkids<2
(4210 observations deleted)
. replace numkids = numkids+1
(5446 real changes made)
```

Next, we generate the limit variables `lo` and `hi`. For observed failures (or uncensored observations), these limits are the same and are equal to the observed log failure-time. For censored durations, the lower limit is the log time of the interview and the upper limit is infinity, meaning that the event will occur in the future. In our example, failure time is the time to conception, which is the difference between `time` and `mardur`. The Stata commands are as follows:

```
. generate lo = ln(time-mardur)
. generate hi = cond(birth==1, lo, .)
(1857 missing values generated)
```

We want to know the effect of education on the spacing of second births. To find our answer, we regress the log of waiting time to the second birth on education and other control variables in the sample of mothers of one child. We use the key explanatory variable `hereduc`, which is a categorical variable with three categories: primary, secondary, and higher education. (Actually, these variables are computed from years of schooling.) We then use secondary education as a reference category. We do this with the help of the factor-variable notation `ib2.hereduc`, which forces Stata to treat the second category as a reference. For simplicity, we use only the age at the beginning of the conception spell (that is, the age when the first child was born) and its square as control variables. We centered the age around 30 to minimize the correlation between age and age-squared. After centering the age variable manually, the age-squared variable is created with the help of the factor-variable notation. Note that the `numkids==2` condition identifies mothers of one child.

```
. replace age = age + mardur - 30
(5446 real changes made)
. cmp (birth2 : lo hi = ib2.hereduc c.age##c.age) if numkids==2, indicators(7)
  (output omitted)
Mixed-process regression                        Number of obs   =       2121
                                                LR chi2(4)      =      84.05
Log likelihood = -2745.3068                     Prob > chi2     =     0.0000
```

| | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| birth2 | | | | | | |
| hereduc | | | | | | |
| <12 years | -.025006 | .0566602 | -0.44 | 0.659 | -.1360579 | .0860459 |
| 16+ years | -.2767705 | .0861557 | -3.21 | 0.001 | -.4456325 | -.1079085 |
| age | .041695 | .0056703 | 7.35 | 0.000 | .0305815 | .0528086 |
| c.age#c.age | -.0012928 | .0004843 | -2.67 | 0.008 | -.0022421 | -.0003436 |
| _cons | 1.782701 | .0491509 | 36.27 | 0.000 | 1.686367 | 1.879035 |
| /lnsig_1 | .101572 | .0193556 | 5.25 | 0.000 | .0636357 | .1395082 |
| sig_1 | 1.10691 | .0214249 | | | 1.065704 | 1.149708 |

The third level of the `hereduc` variable has a negative and significant coefficient. This suggests that the duration of the second conception episode is shorter among highly educated women than among women with lower education. In other words, highly educated women appear to space the first and second births closer together than women with secondary education.

Next, we control for several forms of endogeneity and sample selection to check whether the estimate of $-0.277$ is robust.

## 5.2 Multilevel modeling of recurrent events

Our first concern with the previous outcome is that it might result from the following selection effect: education negatively affects the transition to first birth, hence, education is positively correlated with unobserved causes of fertility in samples of mothers (Kravdal 2007). Therefore, the comparison of the fertility outcomes across educational categories measures not only the true effect of education but also the effect of unobserved preferences or personality traits (Kravdal 2001). We can control for this selection effect by modeling the parity-specific transitions jointly.

One way to model the parity-specific transitions jointly is multilevel modeling. We consider the waiting times to only the first two births. Note that the origin for the second birth is set when the first birth happens. For simplicity, we consider the first and second transitions. The dataset is already in long format and ready for multilevel analysis: episodes within the same person appear in different records. The unobserved person-specific characteristics that affect the transition to births are captured with a

random effect at the level of individuals. The fixed part of the model is extended so that the effects of education and age will be conditional on the number of children previously born.

The new syntax element `|| id:` specifies the random intercepts at the level of individuals. By default, `cmp` uses adaptive quadrature with 12 integration points when fitting the model. One can change the default behavior to simulation by using the `redraws()` option; see the `cmp` help file for details. To easily compare this model with the previous model, we use the second birth as the reference category so that the main effects of education and age are indeed effects conditional on already having a child.

```
. cmp (birth: lo hi = ib2.numkids##(ib2.hereduc c.age##c.age) || id:),
> indicators(7)
```

(*output omitted*)

```
Mixed-process multilevel regression              Number of obs   =       5446
                                                 LR chi2(9)      =     374.78
Log likelihood = -8025.5353                      Prob > chi2     =     0.0000
```

| | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| birth | | | | | | |
| 1.numkids | -.080781 | .1041972 | -0.78 | 0.438 | -.2850038 | .1234418 |
| hereduc | | | | | | |
| <12 years | .0857769 | .0915342 | 0.94 | 0.349 | -.0936268 | .2651807 |
| 16+ years | -.2325627 | .1386793 | -1.68 | 0.094 | -.5043692 | .0392439 |
| age | .0321114 | .0093012 | 3.45 | 0.001 | .0138814 | .0503413 |
| c.age#c.age | -.0010855 | .0008032 | -1.35 | 0.177 | -.0026597 | .0004887 |
| numkids# hereduc | | | | | | |
| 1#<12 years | .3124741 | .114095 | 2.74 | 0.006 | .088852 | .5360961 |
| 1#16+ years | .494463 | .1744228 | 2.83 | 0.005 | .1526007 | .8363253 |
| numkids#c.age | | | | | | |
| 1 | .0890989 | .014383 | 6.19 | 0.000 | .0609089 | .117289 |
| numkids# c.age#c.age | | | | | | |
| 1 | .003612 | .0012324 | 2.93 | 0.003 | .0011965 | .0060274 |
| _cons | 2.061027 | .0791532 | 26.04 | 0.000 | 1.90589 | 2.216165 |
| /lnsig_1_1 | -.1440281 | .058177 | -2.48 | 0.013 | -.2580528 | -.0300033 |
| /lnsig_1 | .2537051 | .0222605 | 11.40 | 0.000 | .2100753 | .2973349 |

| Random-effects Parameters | Estimate | Std. Err. | [95% Conf. | Interal] |
|---|---|---|---|---|
| Level: id | | | | |
| Standard deviations | | | | |
| _cons | .8658635 | .0503733 | .7725544 | .9704424 |
| Level: Residuals | | | | |
| Standard deviation | 1.288792 | .0286892 | 1.233771 | 1.346266 |

The standard deviation of the individual-level random effect is about two-thirds of the standard deviation of the residual. Thus the interclass correlation is about 0.31. The timing of second births is, therefore, not interdependent of the timing of first births. The estimated effect of higher education is negative and has a similar magnitude as in the previous example, but it lacks statistical significance. Now we have no evidence to conclude that higher-educated women would space the first and second births closer together.

## 5.3 Simultaneous equations for recurrent events

An alternative modeling strategy is to fit the two survival models jointly as seemingly unrelated equations. Estimation requires a wide data structure. After reloading the data and selecting the relevant cases, we drop the unnecessary variables and transform the data into wide format. Then, we place educational levels and age at the beginning of the conception episode in parity-specific global macros. The commands are

```
. use http://web.uni-corvinus.hu/bartus/stata/divorce, clear
(Data on marriages (source: divorce4.raw, shipped with aML))
. keep if marnum==1 & numkids<2
(4210 observations deleted)
. replace numkids = numkids+1
(5446 real changes made)
. generate lo = ln(time-mardur)
. generate hi = cond(birth==1, lo, .)
(1857 missing values generated)
. replace age = age + mardur - 30
(5446 real changes made)
. drop mardur time sep
. reshape wide birth age lo hi, i(id) j(numkids)
  (output omitted)
```

Two survival models are fit jointly as follows. The command lists two equations, labeled `birth2` and `birth1`. The mandatory `indicators(7 7)` option tells `cmp` that both equations are interval-censored regression models.

```
. cmp (birth2: lo2 hi2 = ib2.hereduc c.age2##c.age2)
> (birth1: lo1 hi1 = ib2.hereduc c.age1##c.age1), indicators(7 7)
```

(*output omitted*)

```
Mixed-process regression                        Number of obs   =       3325
                                                LR chi2(8)      =     299.69
Log likelihood = -7874.9256                     Prob > chi2     =     0.0000
```

| | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| birth2 | | | | | | |
| hereduc | | | | | | |
| <12 years | .0093223 | .0577222 | 0.16 | 0.872 | -.1038111 | .1224557 |
| 16+ years | -.2500341 | .0867196 | -2.88 | 0.004 | -.4200015 | -.0800668 |
| age2 | .0378136 | .0058145 | 6.50 | 0.000 | .0264175 | .0492098 |
| c.age2#c.age2 | -.0012528 | .0004896 | -2.56 | 0.011 | -.0022124 | -.0002932 |
| _cons | 1.82599 | .0510606 | 35.76 | 0.000 | 1.725913 | 1.926067 |
| birth1 | | | | | | |
| hereduc | | | | | | |
| <12 years | .4280959 | .0722 | 5.93 | 0.000 | .2865864 | .5696053 |
| 16+ years | .2532307 | .1111068 | 2.28 | 0.023 | .0354653 | .4709962 |
| age1 | .13542 | .0123288 | 10.98 | 0.000 | .111256 | .1595841 |
| c.age1#c.age1 | .0027916 | .001029 | 2.71 | 0.007 | .0007749 | .0048084 |
| _cons | 2.122898 | .075198 | 28.23 | 0.000 | 1.975513 | 2.270283 |
| /lnsig_1 | .1070316 | .0197925 | 5.41 | 0.000 | .0682389 | .1458243 |
| /lnsig_2 | .5709637 | .0163749 | 34.87 | 0.000 | .5388695 | .6030578 |
| /atanhrho_12 | .1329566 | .0398857 | 3.33 | 0.001 | .054782 | .2111312 |
| sig_1 | 1.112969 | .0220285 | | | 1.070621 | 1.156993 |
| sig_2 | 1.769972 | .0289831 | | | 1.714068 | 1.827699 |
| rho_12 | .1321787 | .0391889 | | | .0547273 | .208049 |

To interpret the findings, we explore a similarity between the joint modeling of survival processes and the Heckman-type selection modeling. The first-birth equation implicitly defines the probability of being a mother at time $t$ as $P(\theta_s + \sigma\varepsilon_1 \leqslant \log t) = \Phi\{(\log t - \theta_s)/\sigma\}$, where $\theta_s$ denotes the linear combination of explanatory variables and coefficients appearing in the survival model of first births. The probit equation of a Heckman model defines the probability of being a mother as $\Phi(\theta_p)$, where $\theta_p$ denotes the linear combination of explanatory variables and coefficients of the probit equation. Hence, the coefficients of the first-birth equation in the joint survival model divided by the estimated standard-deviation coefficient $\sigma$ implicitly define the coefficients of the probit selection equation. More precisely, because $\theta_p$ corresponds to $-\theta_s$, the survival coefficients define the probability of not being in the sample. This is not surprising,

because a large coefficient in the survival model corresponds to a small hazard of becoming a mother. To summarize, the first-birth equation in the above joint survival model has the same function as the probit model of being a mother in the Heckman selection model, except the survival equation of first births models the probability of not being in the sample, that is, of being childless.

Our primary interest is to explain why women with higher education space births closer together (see section 5.1). In the joint model, higher education in the second conception equation has a significant negative effect. The difference between the naïve estimate of −0.277, reported in section 5.1, and the joint estimate of −0.250 is the selection effect. The selection effect is small and negative (−0.027). The selection effect is negative because the cross-equation correlation of the residuals is positive, and higher education has a negative effect on the implicit sample inclusion probability because it has a positive effect on the waiting time to first births.

The joint estimation of survival models on sequentially constructed samples is a computationally attractive alternative to a Heckman-type selection modeling. Nevertheless, the `cmp` command accommodates the Heckman model (see section 4.2). In our example, the estimation of the survival model of the timing of second births requires the sample of mothers, while the estimation of becoming a mother uses the sample of all women. In our dataset, this condition can be expressed as `(birth1==1)`. The appropriate syntax is

```
. cmp (birth2: lo2 hi2 = $birth2) (sel1: birth1 = $birth1),
> indicators("(birth1==1)*7" 4)
```

(*output omitted*)

## 5.4 Simultaneous equations for different processes

We now turn to the joint model of the timing of second births and the timing of marital dissolutions. The motivation is that the timing of births depends on the quality of the marriage. Because match quality is unobserved, we must control for this omitted variable bias by fitting our survival model jointly with another survival model of marital dissolution. For simplicity, suppose that marital stability depends on education, age at the beginning of the conception episode, and the duration of the marriage. Given this specification, the timing of births should also depend on the duration of the marriage.

The example presented in this section also illustrates the use of the `cmp` command with multispell datasets. So far, we have used single-spell data. We turn to multispell data to accommodate for duration dependence, that is, the effect of the duration of the marriage on the timing of births and divorce. We split conception episode durations into smaller intervals by using the `stsplit` command. We use only first marriages and women who are at the risk of a second delivery. We use the following Stata commands to load the data and to create the multispell data structure, as well as the time-dependent variables for marriage duration and age at the beginning of a spell:

```
. use http://web.uni-corvinus.hu/bartus/stata/divorce, clear
(Data on marriages (source: divorce4.raw, shipped with aML))
. keep if marnum==1 & numkids==1
(7535 observations deleted)
. generate dur = time-mardur
. generate double id2 = _n
. stset dur, fail(sep==1) id(id)
  (output omitted)
. stsplit bdur, at(1 2 5 10)
(4201 observations (episodes) created)
. replace mardur =  mardur + _t0
(4201 real changes made)
. replace birth  = 0 if sep==.
(2454 real changes made)
. replace sep    = 0 if sep==.
(4201 real changes made)
. replace age = age + mardur - 30
(6322 real changes made)
```

Note that both marriage duration (`mardur`) and age are measured at the beginning of a spell and not at the time when events happen. Originally, the age variable measures age at the beginning of the marriage. The last command changes this variable into age at the beginning of a spell, centered around 30 years.

Next, we generate the dependent variables. We study two parallel processes (the timing of marriage dissolution and the timing of births), therefore, we have to generate two pairs of limit variables. The respective limit variable names for marriage dissolution and birth processes will begin with letters `m` and `b`. The Stata codes to create the dependent variables are

```
. generate mlo = ln(mardur+dur)
. generate mhi = cond(sep==1, mlo, .)
(6076 missing values generated)
. generate blo = ln(bdur+dur)
. generate bhi = cond(birth==1, blo, .)
(4854 missing values generated)
```

The only new explanatory variable is `mardur`, that is, the duration of the marriage at the beginning of the spell. The regression equations and the joint model are defined as follows. The new syntax elements include the request of clustered standard errors and the truncation option within both models. The latter accounts for the fact that times to event are left-truncated in multispell datasets, with the exception of the first spell.

```
. cmp (birth: blo bhi = ib2.hereduc c.age##c.age mardur, trunc(ln(bdur) .))
> (divorce: mlo mhi = ib2.hereduc mardur, trunc(ln(mardur) .))
> if numkids==1, vce(cluster id) indicators(7 7)
```

(*output omitted*)

```
Mixed-process regression                        Number of obs   =       6322
                                                Wald chi2(8)    =     683.92
Log pseudolikelihood = -3648.3458               Prob > chi2     =     0.0000
                                (Std. Err. adjusted for 2121 clusters in id)
```

| | Coef. | Robust Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| birth | | | | | | |
| hereduc | | | | | | |
| <12 years | .0075443 | .052122 | 0.14 | 0.885 | -.0946129 | .1097016 |
| 16+ years | -.2922881 | .073531 | -3.98 | 0.000 | -.4364062 | -.14817 |
| age | .0275883 | .0079738 | 3.46 | 0.001 | .0119599 | .0432166 |
| c.age#c.age | -.0023959 | .0009563 | -2.51 | 0.012 | -.0042702 | -.0005216 |
| mardur | .1214816 | .0139052 | 8.74 | 0.000 | .0942278 | .1487353 |
| _cons | 1.823491 | .0978358 | 18.64 | 0.000 | 1.631736 | 2.015245 |
| divorce | | | | | | |
| hereduc | | | | | | |
| <12 years | -.0163363 | .0729228 | -0.22 | 0.823 | -.1592624 | .1265897 |
| 16+ years | .3624115 | .1364096 | 2.66 | 0.008 | .0950536 | .6297693 |
| mardur | .1109086 | .0131804 | 8.41 | 0.000 | .0850754 | .1367417 |
| _cons | 3.11875 | .1964164 | 15.88 | 0.000 | 2.733781 | 3.50372 |
| /lnsig_1 | -.0309254 | .0341747 | -0.90 | 0.366 | -.0979065 | .0360558 |
| /lnsig_2 | -.0764664 | .1162889 | -0.66 | 0.511 | -.3043883 | .1514556 |
| /atanhrho_12 | -.5666712 | .1151736 | -4.92 | 0.000 | -.7924072 | -.3409352 |
| sig_1 | .9695479 | .033134 | | | .9067337 | 1.036714 |
| sig_2 | .9263841 | .1077281 | | | .7375744 | 1.163527 |
| rho_12 | -.5129104 | .084874 | | | -.6597706 | -.328312 |

To interpret the results, recall that the birth equation is not a structural equation but a reduced-form equation (see section 2). The structural effect of higher education must be recovered using (2). Because the covariance is the product of the displayed correlation and standard deviations, the structural coefficient in question can be estimated as

$$\widehat{\alpha}_{1j} = \widehat{\beta}_{1j} - \frac{\sigma_1 r_{12}}{\sigma_2}\widehat{\beta}_{2j} = -0.292 - \frac{0.969(-0.513)}{0.926}0.363 = -0.097$$

To assess the significance of this estimate, it is more useful to do the calculation with the official `nlcom` command. The resulting linear combination, labeled `_nl_1`, is as follows:

```
. nlcom _b[birth:3.hereduc] - _b[divorce:3.hereduc]*
> tanh(_b[atanhrho_12:_cons])*exp(_b[lnsig_1:_cons])/exp(_b[lnsig_2:_cons])
       _nl_1:  _b[birth:3.hereduc] - _b[divorce:3.hereduc]*
> tanh(_b[atanhrho_12:_cons])*exp(_b[lnsig_1:_cons])/exp(_b[lnsig_2:_cons])
```

| | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| _nl_1 | -.0977424 | .1021125 | -0.96 | 0.338 | -.2978793 | .1023945 |

The resulting nonlinear combination lacks statistical significance. Thus there is no evidence that education would have a direct effect on the timing of births. The negative net association between education and the time to second births is indirect; it is mediated through the latent satisfaction with marriage. The positive coefficient of higher education in the divorce equation suggests that highly educated women tend to live in relatively stable marriages. The negative correlation of the residuals suggests that women who are satisfied with their marriage tend to give birth to the second child earlier.

# 6 Conclusion

Seemingly unrelated systems of multilevel proportional hazards equations, often labeled multilevel multiprocess models, are routinely fit by demographers to adjust regression estimates for endogeneity and selection effects. In this article, we make a first step toward estimating systems of survival equations with Stata. We argue that systems of lognormal survival models can easily be fit with the user-written `cmp` command. After discussing the difference between multilevel multiprocess hazard models and multiprocess lognormal survival models, we demonstrate both the theoretical and the practical aspects of fitting models of the latter kind. Our exposition is restricted to the joint estimation of survival models, or the joint estimation of a survival and a probit model. We show how to fit these models and how to interpret the coefficients of interest. However, we do not consider systems including more than two equations and equations with error components.

Although no official Stata command is explicitly designed to estimate multiprocess survival models, the new `gsem` command in Stata 13 seems to enable users to fit systems of proportional hazards models with correlated random effects. Continuous-time exponential regression models can be restated as Poisson models (Skrondal and Rabe-Hesketh 2004), and `gsem` supports the Poisson distribution. However, `gsem` cannot handle the left-truncation of survival times, so it cannot be used to fit systems of continuous-time survival models on multispell data. In addition, `cmp` works in Stata 10.1 and later. In sum, the `cmp` command is a useful tool for survival modeling in Stata.

## 7 Acknowledgments

We thank an anonymous reviewer for helpful comments and suggestions.

Bartus received financial support from the TÁMOP project 4.2.1/B-09/1/KMR-2010-0005 of the Corvinus University of Budapest, as well as from the Hungarian Scientific Research Fund (OTKA) within the research project *Mapping Family Transitions: Causes, Consequences, Complexities, and Context* (grant number K109397).

## 8 References

Cleves, M., W. Gould, R. G. Gutierrez, and Y. V. Marchenko. 2010. *An Introduction to Survival Analysis Using Stata*. 3rd ed. College Station, TX: Stata Press.

Haan, P., and A. Uhlendorff. 2006. Estimation of multinomial logit models with unobserved heterogeneity using maximum simulated likelihood. *Stata Journal* 6: 229–245.

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.

Klein, J. P., and M. L. Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer.

Kravdal, O. 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.

Leckie, G., and C. Charlton. 2013. runmlwin—a program to run the MLwiN multilevel modelling software from within Stata. *Journal of Statistical Software* 52: 1–40.

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.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2002. Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal* 2: 1–21.

Rasbash, J., F. Steele, W. J. Browne, and H. Goldstein. 2012. *A User's Guide to MLwiN, v2.26*. Bristol, UK: Centre for Multilevel Modelling, University of Bristol.

Roodman, D. M. 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.

Train, K. E. 2009. *Discrete Choice Methods with Simulation*. 2nd ed. Cambridge: Cambridge University Press.

### About the authors

Tamás Bartus is an associate professor of sociology at the Corvinus University of Budapest and a senior research fellow at the Hungarian Demographic Research Institute in Budapest, Hungary.

David Roodman is a freelance public policy consultant in Washington, DC.

## Appendix. Multiprocess modeling with gsem

This appendix illustrates the capabilities of the `gsem` command. We thank the anonymous reviewer for the syntax and the permission to include the syntax in this appendix.

### ▷ Example 1. Introduction: The research problem and the dataset

```
. cmp (birth2: lo hi = ib2.hereduc c.age2##c.age2) if numkids==2, indicators(7)
. gsem (lo <- ib2.hereduc c.age2##c.age2 if numkids==2,
> family(normal, udepvar(hi)))
```

◁

### ▷ Example 2. Multilevel modeling of recurrent events

```
. cmp (birth: lo hi = ib2.numkids##(ib2.hereduc c.age2##c.age2) || id:),
> indicators(7)
. gsem (lo <- ib2.numkids##(ib2.hereduc c.age2##c.age2) M[id],
> family(normal, udepvar(hi)))
```

◁

**Example 3. Simultaneous equations for recurrent events**

```
. cmp (birth2: lo2 hi2 = ib2.hereduc c.age2##c.age2)
> (birth1: lo1 hi1 = ib2.hereduc c.age1##c.age1), indicators(7 7)
. gsem (lo2 <- ib2.hereduc c.age2##c.age2 M@(1), family(normal, udepvar(hi2)))
> (lo1 <- ib2.hereduc c.age1##c.age1 M, family(normal, udepvar(hi1)))
```

**Example 4. Simultaneous equations for different processes (this example cannot be replicated with gsem because it does not support truncated outcomes)**

```
. cmp (birth: blo bhi = ib2.hereduc c.age##c.age mardur, truncpoints(ln(bdur) .))
> (divorce: mlo mhi = ib2.hereduc mardur, truncpoints(ln(mardur) .)),
> vce(cluster id) indicators(7 7)
```