The Stata Journal (2010)
10, Number 2, pp. 226–251

Analyzing longitudinal data in the presence of informative drop-out: The jmre1 command

Nikos Pantazis
Department of Hygiene, Epidemiology, and Medical Statistics
University of Athens Medical School
Athens, Greece
npantaz@med.uoa.gr

Giota Touloumi
Department of Hygiene, Epidemiology, and Medical Statistics
University of Athens Medical School
Athens, Greece
gtouloum@med.uoa.gr

Abstract. Many studies in various research areas have designs that involve repeated measurements over time of a continuous variable across a group of subjects. A frequent and serious problem in such studies is the occurrence of missing data. In many cases, missing data are caused by an event that leads to a premature termination of the series of repeated measurements on some subjects. When the probability of the occurrence of this event is related to the subject-specific underlying trend of the variable of interest, this missingness process is called informative censoring or informative drop-out. Standard likelihood-based methods (for example, linear mixed models) fail to give consistent estimates. In such cases, one needs to apply methods that simultaneously model the observed data and the missingness process. In this article, we review a method proposed by Touloumi et al. (1999, *Statistics in Medicine* 18: 1215–1233) to adjust for informative drop-out in longitudinal data analysis. We also present the `jmre1` command, which can be used to fit the proposed model. The estimation method combines the restricted iterative generalized least-squares method with a nested expectation-maximization algorithm. The method is implemented mainly using Stata's matrix programming language, Mata. Our example is derived from the epidemiology of the HIV infection.

Keywords: st0190, jmre1, jmre1_p, datajoint1, missing data, informative censoring, informative drop-out, longitudinal data

1 Introduction

In many research areas, some studies collect longitudinal data on a continuous response variable for each participating subject. One major drawback in such studies, where the main objective is to estimate the rate of change of the response variable, is that the series of repeated measurements is prematurely terminated for some subjects because of the occurrence of an event resulting in highly unbalanced datasets.

Several methods to analyze such unbalanced data are available, provided that missing data are missing at random. Following the terminology of Little and Rubin (2002), conditionally on covariates, missing data can be classified as missing completely at random, when the missingness mechanism does not depend upon the response vector $\mathbf{Y}$; missing at random (MAR), when the probability of nonresponse depends on the observed part of the response ($\mathbf{Y}_0$) but not on the unobserved part ($\mathbf{Y}_m$); and missing nonignorable, when the probability of nonresponse depends on the unobserved outcomes.

Wu and Carroll (1988) introduced the term *informative right-censoring* for the special case of nonignorable missing data where the hazard of the terminating event is related to the subject-specific underlying trend of the response variable. However, from now on, we will use the term *informative drop-out* instead of informative censoring to avoid confusion with the common use of the term *censoring* in the survival analysis literature. It should be noted though that in many cases, the event that causes the termination of longitudinal measurements in a nonignorable way is death or disease progression and not drop-out of some individuals from a study.

Laird (1988) discussed the impact of various missingness processes on inference about the response variable with the main conclusion being that likelihood-based approaches can provide valid inferences, ignoring the missingness process only when data are missing completely at random or MAR. However, serious biases can occur when missingness is nonignorable. In such cases, one needs to apply methods that simultaneously model the response variable and the missingness process. Little (1995) summarized various methods to jointly model the response variable and the informative drop-out process, and he outlined possible extensions. One specific class of such models was derived by extending the so-called parametric conditional linear model (Wu and Carroll 1988; Wu and Bailey 1989), assuming a linear random-effects model for the longitudinal measurements of the response variable and a lognormal model for the terminating event, linked by shared or jointly distributed random parameters (Schluchter 1992; Faucett and Thomas 1996; Touloumi et al. 1999).

To our knowledge, there is no official nor user-written Stata command for modeling longitudinal data subject to informative drop-out, with the exception of a model proposed by Diggle and Kenward (1994) and shared random-effects models, which can be fit using `gllamm` (Rabe-Hesketh, Skrondal, and Pickles 2002). In this article, we describe the joint multivariate random-effects (JMRE) model introduced by Touloumi et al. (1999), and we present the `jmre1` command. For illustrative purposes, we apply `jmre1` to a dataset with longitudinal measurements of an immunologic marker (CD4 cell count) taken over a sample of HIV infected individuals, where death or AIDS onset lead to premature termination of CD4 cell-count measurements in a nonignorable way (that is, informative drop-out).

2 Background

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^T$ be the sequence of measurements of the continuous response variable (called *marker* from now on) on subject i ($i = 1, 2, \ldots, m$). Let X_i be the $n_i \times p$ design matrix associated with the fixed effects ($\mathbf{b}$) of the marker, and let Z_i be the corresponding $n_i \times \pi$ design matrix associated with the random effects ($\boldsymbol{\beta}_i$) of the marker. We assume the following linear mixed model for the longitudinal marker's measurements:

$$\mathbf{Y}_i = X_i\mathbf{b} + Z_i\boldsymbol{\beta}_i + \mathbf{e}_i$$

where $\mathbf{e}_i$ are the within-subject residuals, assumed to be independent and normally distributed with mean $\mathbf{0}$ and variance–covariance matrix $\sigma^2 I_{n_i}$, where I_n denotes the $n \times n$ identity matrix. In addition, we assume that the random effects or between-subjects residuals $\boldsymbol{\beta}_i$ follow the joint multivariate normal distribution with mean $\mathbf{0}$ and variance–covariance matrix Σ^m (superscripts m and d will be used appropriately where needed to denote marker or informative drop-out related parts of the model, respectively). In a typical scenario, Z consists of a column of ones and a column taking the values of the times of measurement so that there is a random intercept and a random coefficient of time (random slope). Similarly, X consists of the same two columns corresponding to a fixed intercept and a fixed slope. There may be additional columns in X corresponding to fixed effects of other covariates on intercept and/or slope. This kind of random effects model is often referred to as the linear growth-curve model or the Laird and Ware model (Laird and Ware 1982). Under the MAR assumption, the mixed model can be fit using Stata's `xtmixed` command.

Repeated measurements of the response variable are terminated either because of development of a terminating event or because of loss to follow-up or study termination. In the former case, the terminating event could be, for example, death. If the probability of death is related to the underlying marker's evolution, the event is considered informative (that is, informative drop-out). In the latter case, it is usually assumed that the resulting missing measurements are at random and thus ignorable in a maximum likelihood analysis. In this case, the actual time of the terminating event is not observed either. We treat these censored event times as noninformative according to the usual survival analysis assumption.

For each subject i, let T_i^s denote event time (for example, death time), and let C_i denote the censoring time (for example, study termination before the occurrence of the event). Note that other kinds of termination of the study may be viewed as informative. In general, T refers to the time to the event whose occurrence is assumed to be informative, whereas censoring refers to reasons for termination that are not considered as informative. The observed "survival" data consist of $T_i = \min(T_i^s, C_i)$ and an indicator variable, δ_i, taking the value of 1 if $T_i = T_i^s$ (that is, the terminating event is observed) and 0 otherwise (that is, the terminating event is not yet observed). To model event times, we used an accelerated failure-time lognormal model of the form

$$\log(T_i^s) = \mathbf{x}_i^{d^T}\mathbf{b}^d + e_i^d$$

where $\mathbf{x}_i^d$ is the design vector of ν explanatory variables for the event time and e_i^d are the errors assumed to follow the normal distribution with mean 0 and standard deviation σ^d. This model, independently of the marker, can be estimated using Stata's `tobit` or `streg` command.

To allow for informative drop-out in the marker's measurements, we assume that subject-level random coefficients $\boldsymbol{\beta}_i$ (for example, subject-specific baseline value and rate-of-change deviations from the corresponding population average) and the survival model's residuals, e_i^d (deviations from mean log event time), jointly follow the multivariate normal distribution:

$$\boldsymbol{\beta}_i^{\text{joint}} = \begin{pmatrix} \boldsymbol{\beta}_i \\ e_i^d \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \Sigma^{\text{joint}} \right\} \text{where } \Sigma^{\text{joint}} = \left\{ \begin{array}{cc} \Sigma^m & \\ \boldsymbol{\sigma}^{md^T} & \left(\sigma^d\right)^2 \end{array} \right\}$$

$\boldsymbol{\sigma}^{md}$ is a $\pi \times 1$ vector of the covariances of the marker's random coefficients ($\boldsymbol{\beta}_i$) with the survival model's residuals. If $\boldsymbol{\sigma}^{md} = \mathbf{0}$, then the marker's missing measurements due to the occurrence of the terminating event are at random; otherwise, the marker's missing data are informative.

2.1 Fitting procedure

The model can be written as

$$\mathbf{Y}_i^{\text{joint}} = X_i^{\text{joint}} \mathbf{b}^{\text{joint}} + Z_i^{\text{joint}} \boldsymbol{\beta}_i^{\text{joint}} + \mathbf{e}_i^{\text{joint}}$$

where

$$\mathbf{Y}_i^{\text{joint}} = \left\{ \mathbf{Y}_i^T, \ln(T_i^s) \right\}^T$$

$$\mathbf{b}^{\text{joint}} = \left(\mathbf{b}^T, \mathbf{b}^{d^T} \right)^T, \boldsymbol{\beta}_i^{\text{joint}} = \left(\boldsymbol{\beta}_i{}^T, e_i^d \right)^T, \mathbf{e}_i^{\text{joint}} = \left(\mathbf{e}_i^T, 0 \right)^T$$

$$X_i^{\text{joint}} = \begin{pmatrix} X_i & 0 \\ \mathbf{0}^T & \mathbf{X}_i^d \end{pmatrix} \text{and } Z_i^{\text{joint}} = \begin{pmatrix} Z_i & 0 \\ \mathbf{0}^T & 1 \end{pmatrix}$$

Note that the last element of $\boldsymbol{\beta}_i^{\text{joint}}$ includes the error term of the survival model (e_i^d), and thus the corresponding one in $\mathbf{e}_i^{\text{joint}}$ is set to zero.

For uncensored event-times data (that is, $T_i = T_i^s$ for all subjects), all model parameters could be estimated using standard software for mixed models, provided that it has the required flexibility in the definition of the distribution of the random effects and level-1 errors. For example, MLn or MLwiN (for more information, see http://www.cmm.bris.ac.uk/MLwiN/index.shtml) implementing the iterative generalized least-squares (IGLS) and restricted iterative generalized least-squares (RIGLS) algorithms could be used. For a detailed description of IGLS and RIGLS, see Goldstein (2003).

With censored survival data, however, the standard RIGLS method has to be modified to obtain unbiased estimates of the model parameters. To deal with censored survival data, a version of the expectation-maximization (EM) algorithm was applied, considering censored survival observations as missing data. At each iteration, the survival part of the response variable and the survival component of the residuals cross-product are replaced for the censored observations by their conditional expectations given the observed data and the current parameter estimates (E step), and then new parameters are obtained via RIGLS (M step). The estimation of these conditional expectations is based on multivariate normal theory (Johnson and Wichern 2007) and properties of the truncated normal distribution (Johnson et al. 1970). Initial values of the model parameters were obtained by treating censored survival times ($T_i = C_i$) as known survival times and applying the RIGLS method.

The null hypothesis $\boldsymbol{\sigma}^{md} = \mathbf{0}$ (that is, noninformative drop-out) can be tested via the likelihood-ratio test, estimating the likelihood of the joint model and that of the model in which $\boldsymbol{\sigma}^{md}$ (or some of its elements) are constrained to $\mathbf{0}$.

Standard errors of the fixed-effects parameters (as obtained from RIGLS in the M step of the algorithm) can be underestimated because of the replacement of censored survival times by their conditional expectations during the fitting procedure, ignoring the uncertainty of these expectations. A modified multiple-imputation (MI) method can be used to adjust the standard errors of the estimated fixed-effects parameters of the markers' trajectories for missing survival data (Rubin 1978; Touloumi et al. 2003).

The method consists of creating N pseudo-complete datasets where censored survival times have been replaced by random draws from the truncated normal distribution with mean and variance conditional on the marker's data and based on parameters' estimates at convergence. Fitting the model to each pseudo-complete dataset, one can estimate the empirical between-datasets variance for each fixed parameter. This between-datasets variance, weighted by $(N+1)/N$ to adjust for a finite number of samples, is then added to the within variance estimated from the previously fit joint model at its convergence to obtain the adjusted variance–covariance matrix of the fixed-effects parameters. It has been shown (Touloumi et al. 2003) that estimates obtained by this version of the MI method are similar to the ones obtained by the Louis (1982) method, with the MI method being less computationally intensive.

Most steps of the estimation procedure, including the whole RIGLS algorithm, have been implemented using Stata's matrix language, Mata.

3 The jmre1 command

3.1 Syntax

The `jmre1` command for Stata fits the JMRE model, as described in section 2.1. The syntax of `jmre1` is the following:

`jmre1` *depvar indepvars* [*if*] , `remark(`*varlist*`)` `dropout(`*varlist*`)` `id(`*varlist*`)`
`timevar(`*varname*`)` [`redrop(`*varname*`)` `l1(`*varname*`)` `maxi(`*#*`)` `tol(`*#*`)`
`burnin(`*#*`)` `trace(`*#*`)` `restr`
`corr0(`*varname1 varname2* [*varname3 varname4* ...] `)` `level(`*#*`)` `mi(`*#*`)`]

depvar is the joint response variable $\mathbf{Y}_i^{\text{joint}}$, comprising a column of the marker's measurements and the logarithm of the observed or censored event time for each level-2 unit.

indepvars are covariates for the evolution of the marker (X_i). `jmre1` does not automatically introduce a constant term in the regressors (intercept); thus a constant term should be explicitly included among *indepvars*.

We highly recommend that you use the accompanying `datajoint1` command (see section 3.5) to generate the joint response variable and to appropriately manipulate remaining data. In this case, the name of the response variable will be `_JY`, and all covariates' names start with `_M` or `_D` for the marker or drop-out model, respectively, with just `_M` and `_D` being the corresponding intercepts.

Finally, because of extensive manipulation of the data during the fitting procedure, `jmre1` uses a `preserve` and a `restore` command; thus `jmre1` cannot be used if data are already `preserved`.

3.2 Options

`remark(`*varlist*`)` defines the random effects for the marker's model. At least one variable is required, which in most cases will be the intercept term (`_M` if the data are generated by `datajoint1`; see section 3.5). This is a required option.

`dropout(`*varlist*`)` defines the structure of the lognormal drop-out model. The first variable should be the event indicator variable (1 means observed, 0 means censored in the usual survival analysis sense). Remaining variables define the independent covariates of the survival model. `jmre1` does not add an intercept automatically as most Stata commands do, so the user should (almost always) explicitly enter the drop-out model's constant variable (`_D` if the data are generated by `datajoint1`; see section 3.5). This is a required option.

`id(`*varlist*`)` defines ID variables for identifying subjects (level-2 units). This is a required option.

`timevar(`*varname*`)` indicates the constant term in the drop-out model. If left blank, the default is `timevar(_D)`, which is the corresponding variable created by `datajoint1`. This is a required option.

`redrop(`*varname*`)` indicates the drop-out model's constant variable (`_D` if the data are generated by `datajoint1`; see section 3.5). The default is `redrop(_D)`.

`l1(`*varname*`)` indicates the constant term in the marker model. The default is `l1(_M)`, which is the corresponding variable created by `datajoint1` (see section 3.5).

`maxi(`*#*`)` denotes the maximum number of iterations for the EM algorithm. The default is `maxi(40)`.

`tol(`*#*`)` denotes the tolerance criterion for declaring convergence. The default is `tol(0.01)`, which means that the model has converged when relative differences in all model parameters between successive iterations are less than 0.01 (that is, 1%).

`burnin(`*#*`)` specifies the number of "burn-in" iterations. The default is `burnin(2)`, which means that the command will perform two simple RIGLS or IGLS iterations before invoking the nested EM algorithm.

`trace(`*#*`)` specifies to display all model parameters (fixed-effects parameters and random parameters) after each iteration. Random parameters are shown as a column vector created by the elements of the estimated variance–covariance matrix of the random effects (in lower triangular representation) and the estimated level-1 variance in the marker model.

`restr` specifies to use the RIGLS method instead of the default IGLS method. For normally distributed data, RIGLS is equivalent to restricted maximum likelihood and IGLS to maximum likelihood estimation.

`corr0(`*varname1 varname2* [*varname3 varname4* `...`]`)` indicates pairs of variables that denote that the corresponding elements in the variance–covariance matrix are constrained to zero. The variance–covariance matrix includes the random effects of the marker model and the residuals of the lognormal drop-out model.

`level(`*#*`)` denotes the level of significance for the confidence intervals.

`mi(`*#*`)` requests corrected standard errors of fixed effects by using the modified MI method. The suggested number of MIs is five (that is, `mi(5)`). This correction is required to account for the uncertainty when replacing censored survival times with their expectations. Given that this procedure can be time consuming, it can be invoked only once a final model has been chosen and not during all stages of a model selection procedure. If not specified or if specified as `mi(0)`, standard errors remain uncorrected.

3.3 Saved results

`jmre1` saves the following in `e()`:

Scalars			
`e(N)`	number of records	`e(ll)`	(restricted) log likelihood
`e(var_eij)`	level-1 variance in marker model	`e(Nid)`	number of subjects
Macros			
`e(cmd)`	`jmre1`	`e(id)`	ID variable(s)
`e(method)`	`RIGLS` or `IGLS`	`e(depvar)`	name of joint dependent variable
`e(re)`	varlist with drop-out model's constant and marker's random effects	`e(properties)`	`b V`
`e(predict)`	program used to implement `predict`		
Matrices			
`e(b)`	coefficient vector (fixed effects)	`e(V)`	variance–covariance matrix of fixed effects
`e(cov_re)`	variance–covariance matrix of random effects	`e(corr_re)`	correlation matrix of random effects
Functions			
`e(sample)`	marks estimation sample		

3.4 Syntax for predict

As with all Stata estimation commands, `jmre1` supports the postestimation command `predict` (see [R] **predict**) to compute fitted values, empirical Bayes estimates of the marker's random effects (best linear unbiased predictions [BLUPs]), and residuals. The syntax for `predict` following `jmre1` is

`predict` [*type*] *newvarname* [*if*] [*in*] [`,` *statistic*
`equation(Drop_Out|Marker)`]

statistic	Description
`xb`	linear predictor for the fixed portion of the model; the default
`stdp`	standard error of the fixed-portion linear prediction xb
`reffects`	predictions of the marker's random effects (BLUPs)
`fitted`	linear predictor of the fixed portion plus contributions based on predicted random effects (fitted values)
`residuals`	residuals, response minus fitted values

(*Continued on next page*)

`equation(Drop_Out|Marker)` specifies the desired part of the model (only for the `xb` or `stdp` options). The default is `equation(Marker)`.

All predicted statistics, except for `xb` and `stdp`, are calculated for the marker part of the model and the corresponding observations. For `xb` and `stdp`, one can use the `equation()` option to specify the desired part of the model (drop-out or marker).

3.5 The datajoint1 utility command

The `datajoint1` command makes the appropriate manipulation of the data, which is required before fitting a `jmre` model.

`datajoint1 using` *marker_file*`, dfile(`*drop-out_file*`) markervalue(`*varname*`)`
`markertime(`*varname*`) dropevent(`*varname*`) droptime(`*varname*`) id(`*varlist*`)`
`[covariates(`*varlist*`) clear saving(`*filename*`) replace]`

The `datajoint1` command takes the name of a Stata data file (*marker_file*), which contains longitudinal measurements of a continuous variable (marker) along with the time the measurement was taken (`markertime()`). The file should also include an identifier variable for each subject. Each subject should typically have more than one marker's measurements, although subjects with only one measurement are allowed.

A second file (*drop-out_file*), provided in `dfile()`, contains the same identifier variable and has one record per subject. Required variables are a binary variable indicating whether the informative censoring event (`dropevent()`) has occurred and the time (`droptime()`) of the occurrence of the event (if the event has occurred) or the censoring time (here the word *censoring* has the usual survival analysis meaning). This file can also contain other baseline, non–time-dependent variables. The command checks the data for some basic requirements, but it is the user's responsibility to correctly prepare *marker_file* and *drop-out_file*.

Options

`dfile(`*drop-out_file*`)` specifies the name of the file that contains data on the informative drop-out event along with other time-constant covariates. This is a required option.

`markervalue(`*varname*`)` specifies the name of the variable that contains the longitudinal values of the marker. This is a required option.

`markertime(`*varname*`)` specifies the name of the variable that contains the time each marker's measurement was taken. This is a required option.

`dropevent(`*varname*`)` specifies the indicator variable for the occurrence of the informative drop-out event. This is a required option.

`droptime(`*varname*`)` specifies the event or censoring time. This is a required option.

`id(`*varlist*`)` specifies the subject's identifier variables. This is a required option.

`covariates(`*varlist*`)` specifies other baseline, non–time-dependent variables that will probably be used to model the marker's trend or the informative drop-out mechanism. These variables should only be included in the *drop-out_file*.

`clear` specifies to clear the data in memory, even though they have not yet been saved to disk.

`saving(`*filename*`)` specifies the name of the file to be created.

`replace` specifies that the file specified in the `saving()` option may be replaced if it already exists.

4 Example

The data for this example are derived from an HIV/AIDS cohort study (CASCADE Collaboration 2003). This study had two main objectives: 1) to estimate the rate of decline of the immunological marker called CD4 cell count after infection with the HIV-1 virus and 2) to evaluate the effect of potential prognostic factors (such as age, mode of infection-risk group, sex, etc.) on the initial levels of this marker and its subsequent rate of decline. The period of interest starts at the date of infection or, more precisely, the date of seroconversion and stops at the date a patient started antiretroviral treatment (ART), developed clinical AIDS, or died. Censoring the series of longitudinal CD4 cell-count measurements at the date of ART initiation should probably be considered a MAR mechanism because physicians' decisions regarding treatment are usually based on observed values of CD4 cell count. On the other hand, premature termination of measurements due to AIDS onset or death could be informative because the probability of AIDS onset or death is likely related to an individual's underlying trend of CD4 cell-count evolution. Hence, in the following analysis, death or AIDS onset will be treated as an informative (nonignorable) event, whereas all other causes of measurements' termination (end of study, loss to follow-up, censoring due to ART initiation) will be considered noninformative (ignorable) events.

The original dataset consisted of data on 5,739 individuals, but here we are using a random sample of 400 individuals. We have prepared two datasets: one containing multiple records per individual with longitudinal CD4 cell-count measurements and the other containing one record per individual with just two covariates (age at seroconversion and risk group) along with information about the drop-out mechanism. In the following output, we describe the variables of these two files and list their contents for the first three individuals.

(*Continued on next page*)

```
. use cd4
(CD4 sample file)

. describe

Contains data from cd4.dta
  obs:         4,677                          CD4 sample file
 vars:             5                          13 Mar 2009 12:31
 size:       140,310 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
study_id        byte   %8.0g                  Study ID
patient_id      str11  %11s                   Patient ID
cd4             int    %9.0g                  CD4 count
sqrt_cd4        float  %9.0g                  CD4 count(sq.root)
time            float  %9.0g                  CD4 time (yrs.)
-------------------------------------------------------------------------------
Sorted by:  study_id  patient_id  time

. list in 1/32, abbreviate(10) noobs sepby(study_id patient_id)

  +-----------------------------------------------------+
  | study_id   patient_id    cd4   sqrt_cd4        time |
  |-----------------------------------------------------|
  |        1         1531    383   19.57038    4.284737 |
  |        1         1531    455   21.33073    4.476386 |
  |        1         1531    423   20.56696    4.744695 |
  |        1         1531    449   21.18962    4.988364 |
  |        1         1531    548    23.4094    5.054072 |
  |        1         1531    548    23.4094    5.108829 |
  |        1         1531    399   19.97499    5.256673 |
  |        1         1531    576         24    5.631759 |
  |        1         1531    429   20.71231    6.105407 |
  |        1         1531    435   20.85665    6.324435 |
  |        1         1531    437   20.90454    6.387406 |
  |-----------------------------------------------------|
  |        1         1534   1096   33.10589     5.24846 |
  |        1         1534    614   24.77902    6.146475 |
  |        1         1534    586   24.20744    7.022587 |
  |        1         1534    479   21.88607    7.457906 |
  |        1         1534    563   23.72762    7.939767 |
  |        1         1534    510   22.58318    8.687201 |
  |        1         1534    631   25.11971     8.91718 |
  |        1         1534    449   21.18962    9.138946 |
  |        1         1534    464   21.54066    9.399042 |
  |        1         1534    374   19.33908    9.744011 |
  |-----------------------------------------------------|
  |        1         1552    445   21.09502     .276523 |
  |        1         1552    540    23.2379     .506502 |
  |        1         1552    524   22.89105     .793977 |
  |        1         1552    726   26.94439    1.062286 |
  |        1         1552    608   24.65766    1.774127 |
  |        1         1552   1098   33.13608    2.020534 |
  |        1         1552    823   28.68798    2.269678 |
  |        1         1552    969   31.12877    2.557153 |
  |        1         1552    827   28.75761      2.8282 |
  |        1         1552   1352   36.76955    3.271732 |
  |        1         1552    683   26.13427     3.63039 |
  +-----------------------------------------------------+
```

Variables `study_id` and `patient_id` are study identifiers and patient identifiers, respectively (data are derived from a multicenter study), and are both required to uniquely identify a patient. `cd4` is the CD4 cell-count measurement in its original scale (cells/μL), and `sqrt_cd4` is the corresponding square-root–transformed values. Finally, `time` is the time in years of the CD4 cell-count measurement since seroconversion.

```
. use surv
(AIDS/Death + covariates sample file)

. describe

Contains data from surv.dta
  obs:           400                          AIDS/Death + covariates sample file
 vars:             6                          13 Mar 2009 12:31
 size:        10,800 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
study_id        byte   %8.0g                  Study ID
patient_id      str11  %11s                   Patient ID
AD              byte   %9.0g       ynlbl      AIDS or Death
ADtime          float  %9.0g                  AIDS/Death or censoring time
agesero         byte   %9.0g       agesero    Age at SC
expo            byte   %17.0g      expo       Risk group
-------------------------------------------------------------------------------
Sorted by:  study_id  patient_id

. list in 1/3, abbreviate(15) noobs sep(0)

  +-------------------------------------------------------------------+
  | study_id   patient_id   AD     ADtime   agesero              expo |
  |-------------------------------------------------------------------|
  |        1         1531   No   6.387408     20-29   Heteros.(male) |
  |        1         1534   No   9.744013     20-29            Homos. |
  |        1         1552   No   4.246408     20-29            Homos. |
  +-------------------------------------------------------------------+

. label list agesero
agesero:
           1 15-19
           2 20-29
           3 30-39
           4 40+

. label list expo
expo:
           1 Homos.
           2 Heteros.(male)
           3 Heteros.(fem.)
           4 Drug users (male)
           5 Drug users (fem.)
           6 Haemoph.
```

`AD` is a binary variable having the value of 1 if an individual developed AIDS or died and 0 otherwise. `ADtime` contains the time of AIDS development or death when `AD` equals 1 or contains the time of end of follow-up or censoring due to ART initiation if `AD` equals 0. `agesero` and `expo` are categorical variables containing information about age at seroconversion and risk group, respectively.

4.1 Preparing the "joint" dataset

We will now use the datajoint1 command to prepare our "joint" dataset for the final analysis

```
. datajoint1 using cd4, dfile(surv) markervalue(sqrt_cd4) markertime(time)
> id(study_id patient_id)  dropevent(AD) droptime(ADtime)
> covariates(agesero expo) saving(jointdata) clear replace
file jointdata.dta saved
```

The command creates two indicator variables named _M and _D to separate records that correspond to marker models and to drop-out models, respectively. For each covariate specified in the covariates() option, two new variables with the same names but prefixed with _M and _D have been created. The new variables whose names start with _M have the same values as the original ones but only for records corresponding to the marker model (that is, those where _M equals one). New variables whose names start with _D have been created in a similar way for records corresponding to the drop-out model. Variable _JY is the joint response variable. Within each patient's block of records, _JY has the corresponding value of the dependent variable of the marker model (that is, the square root of CD4 cell count, sqrt_cd4) for the first _N-1 records and has the logarithm of the event or censoring time for the _Nth patient's record. Finally, a new variable (named _Mtime in this case) has been created containing the time of marker's measurements only for records corresponding to the marker model. The following list shows the actual data for the first three patients (some variables have been omitted):

```
. list patient_id _M _D _JY _Mtime _Magesero _Dagesero in 1/35, abbreviate(10)
> noobs sepby(study_id patient_id)

  +--------------------------------------------------------------------+
  | patient_id   _M   _D        _JY     _Mtime   _Magesero   _Dagesero |
  |--------------------------------------------------------------------|
  |       1531    1    0   19.57038   4.284737       20-29           . |
  |       1531    1    0   21.33073   4.476386       20-29           . |
  |       1531    1    0   20.56696   4.744695       20-29           . |
  |       1531    1    0   21.18962   4.988364       20-29           . |
  |       1531    1    0    23.4094   5.054072       20-29           . |
  |       1531    1    0    23.4094   5.108829       20-29           . |
  |       1531    1    0   19.97499   5.256673       20-29           . |
  |       1531    1    0         24   5.631759       20-29           . |
  |       1531    1    0   20.71231   6.105407       20-29           . |
  |       1531    1    0   20.85665   6.324435       20-29           . |
  |       1531    1    0   20.90454   6.387406       20-29           . |
  |       1531    0    1   1.854329          .           .       20-29 |
  |--------------------------------------------------------------------|
  |       1534    1    0   33.10589    5.24846       20-29           . |
  |       1534    1    0   24.77902   6.146475       20-29           . |
  |       1534    1    0   24.20744   7.022587       20-29           . |
  |       1534    1    0   21.88607   7.457906       20-29           . |
  |       1534    1    0   23.72762   7.939767       20-29           . |
  |       1534    1    0   22.58318   8.687201       20-29           . |
  |       1534    1    0   25.11971    8.91718       20-29           . |
  |       1534    1    0   21.18962   9.138946       20-29           . |
  |       1534    1    0   21.54066   9.399042       20-29           . |
  |       1534    1    0   19.33908   9.744011       20-29           . |
  |       1534    0    1   2.276653          .           .       20-29 |
  |--------------------------------------------------------------------|
  |       1552    1    0   21.09502    .276523       20-29           . |
  |       1552    1    0    23.2379    .506502       20-29           . |
  |       1552    1    0   22.89105    .793977       20-29           . |
  |       1552    1    0   26.94439   1.062286       20-29           . |
  |       1552    1    0   24.65766   1.774127       20-29           . |
  |       1552    1    0   33.13608   2.020534       20-29           . |
  |       1552    1    0   28.68798   2.269678       20-29           . |
  |       1552    1    0   31.12877   2.557153       20-29           . |
  |       1552    1    0   28.75761     2.8282       20-29           . |
  |       1552    1    0   36.76955   3.271732       20-29           . |
  |       1552    1    0   26.13427    3.63039       20-29           . |
  |       1552    0    1   1.446073          .           .       20-29 |
  +--------------------------------------------------------------------+
```

Note that for patient 1,531, for example, the last value of `_JY` is the logarithm of his event or censoring time [that is, log(6.387408) = 1.854329]. It is highly recommended that the user not alter the file created at all. In case new variables (or recoding of existing variables) are required, these should be created on the original `surv` file followed by a new run of the `datajoint1` command.

4.2 Fitting the JMRE model using the jmre1 command

We will now fit a JMRE model to the data described above. The marker model will include a random intercept and a random slope (that is, time coefficient). Fixed effects will include intercept and slope and their interactions with age at seroconversion and risk group. Consequently, the model will allow for the estimation of the average initial marker levels and their subsequent rate of change for the reference category (fixed intercept and slope, respectively), and the average deviations of these quantities according to age at seroconversion categories and risk group. The inclusion of a random intercept and a random slope implies that each individual is allowed to have initial marker values and a subsequent rate of change that deviate from the corresponding average values for the covariate pattern the individual belongs to. The random-effects variance–covariance matrix Σ^m is unstructured (there are no constraints defined via the `corr0()` option), allowing the variances of the random intercept and slope, along with their correlation, to be freely estimated.

The drop-out model will have the same prognostic factors as the marker model, allowing the average (log-transformed) time to death or AIDS onset to be different according to age at seroconversion and risk group. In general, we recommend that the user fit various simple lognormal survival models, using the `streg` command, to select a set of significant predictors. Then the user should use these predictors when fitting the JMRE model.

```
. xi: jmre1 _JY _M i._Magesero*_Mtime i._Mexpo*_Mtime, remark(_M _Mtime)
> dropout(AD _D i._Dagesero i._Dexpo) timevar(_Mtime) id(study_id patient_id)
> restr
  (output omitted)
```

The part following the `jmre1` command follows the usual Stata conventions with two exceptions: 1) the intercept term (`_M`) is explicitly entered and 2) the *depvars* part corresponds only to the marker model. The command uses the old `xi:` syntax and does not yet support factor notation. The drop-out model is practically declared within the `dropout()` option, where the first variable is the binary variable for the drop-out event (`AD`) and the remaining variables denote the corresponding covariates. As in the marker model, the intercept term (`_D`) is also entered explicitly. Options `redrop()` and `l1()` have been omitted; thus they have their default values (`_D` and `_M`, respectively). The required `timevar()` option declares the variable containing the time of the marker's measurements. Finally, we added the `restr` option to use the RIGLS estimation algorithm.

The results of the previous command are shown in the following output:

```
Response variable: _JY, 400 subjects, 4677 marker measurements
Restricted Log-likelihood = -13074.39
Fixed effects
------------------------------------------------------------------------------

------------------------------------------------------------------------------
         _JY |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Drop_Out     |
          _D |   2.439713    .107579    22.68   0.000     2.228863    2.650564
_I_Dageser~2 |  -.0308937   .1114061    -0.28   0.782    -.2492456    .1874583
_I_Dageser~3 |  -.2716915   .1206316    -2.25   0.024     -.508125    -.035258
_I_Dageser~4 |  -.4275789   .1248543    -3.42   0.001    -.6722889    -.182869
  _I_Dexpo_2 |   .4460437   .1672487     2.67   0.008     .1182423     .773845
  _I_Dexpo_3 |   .1914361   .1316591     1.45   0.146     -.066611    .4494832
  _I_Dexpo_4 |  -.0339984   .1170231    -0.29   0.771    -.2633595    .1953627
  _I_Dexpo_5 |   .0656182   .1397301     0.47   0.639    -.2082478    .3394843
  _I_Dexpo_6 |   .1245609   .1817095     0.69   0.493    -.2315832    .4807049
-------------+----------------------------------------------------------------
Marker       |
          _M |   24.84069   .9066342    27.40   0.000     23.06372    26.61766
_I_Mageser~2 |    .296751   .9524523     0.31   0.755    -1.570021    2.163523
_I_Mageser~3 |  -.4586846   1.018263    -0.45   0.652    -2.454444    1.537075
_I_Mageser~4 |  -.9201732   1.035369    -0.89   0.374    -2.949459    1.109112
      _Mtime |  -1.117109   .2200785    -5.08   0.000    -1.548455    -.685763
_I_MaX_Mti~2 |  -.3308813   .2240275    -1.48   0.140    -.7699671    .1082046
_I_MaX_Mti~3 |  -.4605507   .2498886    -1.84   0.065    -.9503234    .0292219
_I_MaX_Mti~4 |  -.5433615   .2559171    -2.12   0.034     -1.04495   -.0417731
  _I_Mexpo_2 |  -2.135086   1.326851    -1.61   0.108    -4.735666    .4654932
  _I_Mexpo_3 |   .2751593   1.054363     0.26   0.794    -1.791355    2.341674
  _I_Mexpo_4 |   .5212439   1.005444     0.52   0.604    -1.449389    2.491877
  _I_Mexpo_5 |   3.521577   1.213382     2.90   0.004     1.143392    5.899763
  _I_Mexpo_6 |    5.28702   1.568754     3.37   0.001     2.212318    8.361721
_I_MeX_Mti~2 |   .8563533   .3350725     2.56   0.011     .1996232    1.513083
_I_MeX_Mti~3 |   .0470733   .2670043     0.18   0.860    -.4762455    .5703921
_I_MeX_Mti~4 |  -.3032204   .2449095    -1.24   0.216    -.7832342    .1767934
_I_MeX_Mti~5 |   -.411662    .290144    -1.42   0.156    -.9803337    .1570098
_I_MeX_Mti~6 |  -.6364512    .348253    -1.83   0.068    -1.319015    .0461122
------------------------------------------------------------------------------
Random effects (Covariance matrix)
------------------------------------------------------------------------------
symmetric e(cov_re)[3,3]
              _D         _M     _Mtime
    _D  .52415435
    _M  .06505558  31.417146
_Mtime   .6146643  -3.164154  1.6422581
------------------------------------------------------------------------------

------------------------------------------------------------------------------
Level-1 variance (Marker):   9.31419
------------------------------------------------------------------------------
```

The output first lists the name of the joint response variable, the number of individuals, the number of marker measurements in the dataset, and the restricted (because we used the `restr` option) log likelihood. Then the usual Stata estimation-command table follows, which includes results for the drop-out model and the marker model fixed effects. Finally, the estimates for the random parameters are given: first, the variance–

covariance matrix for the random effects of the marker model's and the drop-out model's residuals, and then the estimate of the variance of the level-1 marker's residuals. One could list the correlation matrix of the random effects instead of the variance–covariance matrix, as shown below:

```
. matrix list e(corr_re)

symmetric e(corr_re)[3,3]
                _D          _M      _Mtime
    _D           1
    _M   .01603141           1
_Mtime   .66250274  -.44050773           1
```

The drop-out model's residuals have a relatively high positive correlation with the marker model's random slopes [Corr(_D, _Mtime) = 0.663]. This means that individuals with higher (less steep) slopes tend to develop AIDS or die later. The correlation with the initial values of the marker is still positive but much lower. To evaluate the significance of these findings, we could fit a JMRE model with these two correlations constrained to zero and compare with the previous unconstrained model via the likelihood-ratio test. The test will be valid because the two models share the same fixed-effects structure (thus the fitting through the restricted algorithm does not cause any problems), and because the parameters being tested are correlations, thus the null hypothesis does not lie on the boundary of the parameter space. Before proceeding with the fit of the constrained model, we are storing the estimation results of the current model:

```
. estimates store FullModel
```

Now we can fit the constrained model. The only change compared with the previous `jmre1` command is the addition of the `corr0(_D _M _D _Mtime)` option. This option specifies that the correlation of the drop-out residuals with the marker's random intercept (`_D` and `_M`) as well as with the marker's random slope (`_D` and `_Mtime`) should be constrained to zero. This is equivalent to fitting the marker and drop-out models separately (that is, ignoring the informative drop-outs). The age effect on the marker's slope (coefficients `_I_MaX_Mti~2`, `_I_MaX_Mti~3`, `_I_MaX_Mti~4`) is now attenuated and not significant (compare with the output of the unconstrained model)

```
. xi: jmre1 _JY _M i._Magesero*_Mtime i._Mexpo*_Mtime, remark(_M _Mtime)
> dropout(AD _D i._Dagesero i._Dexpo) timevar(_Mtime) id(study_id patient_id)
> restr corr0(_D _M _D _Mtime)
  (output omitted)
Response variable: _JY, 400 subjects, 4677 marker measurements
Restricted Log-likelihood = -13199.23

Fixed effects
------------------------------------------------------------------------------
------------------------------------------------------------------------------
         _JY |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Drop_Out     |
          _D |   2.461083   .1168558    21.06   0.000      2.23205    2.690116
_I_Dageser~2 |  -.0261404   .1211109    -0.22   0.829    -.2635135    .2112327
_I_Dageser~3 |  -.2622229   .1309363    -2.00   0.045    -.5188534   -.0055925
_I_Dageser~4 |  -.4619534   .1351883    -3.42   0.001    -.7269177   -.1969892
  _I_Dexpo_2 |   .4568295   .1851469     2.47   0.014     .0939482    .8197108
  _I_Dexpo_3 |   .1958888   .1443903     1.36   0.175    -.0871109    .4788886
  _I_Dexpo_4 |  -.0221913   .1263402    -0.18   0.861    -.2698135    .2254309
  _I_Dexpo_5 |   .0939474   .1517552     0.62   0.536    -.2034874    .3913823
  _I_Dexpo_6 |    .133236   .1942086     0.69   0.493    -.2474059    .5138779
-------------+----------------------------------------------------------------
Marker       |
          _M |   24.71954   .9150865    27.01   0.000     22.92601    26.51308
_I_Mageser~2 |   .1971617   .9625667     0.20   0.838    -1.689434    2.083758
_I_Mageser~3 |  -.5378366   1.028217    -0.52   0.601    -2.553104    1.477431
_I_Mageser~4 |  -1.021058   1.044502    -0.98   0.328    -3.068243    1.026128
      _Mtime |  -1.081002   .2237515    -4.83   0.000    -1.519547   -.6424574
_I_MaX_Mti~2 |  -.2355486   .2265238    -1.04   0.298     -.679527    .2084299
_I_MaX_Mti~3 |   -.318824   .2551211    -1.25   0.211    -.8188521    .1812041
_I_MaX_Mti~4 |  -.3664625   .2604899    -1.41   0.159    -.8770134    .1440883
  _I_Mexpo_2 |  -2.087981   1.336893    -1.56   0.118    -4.708242    .5322803
  _I_Mexpo_3 |   .3305508   1.063717     0.31   0.756    -1.754297    2.415398
  _I_Mexpo_4 |   .4232636   1.015625     0.42   0.677    -1.567326    2.413853
  _I_Mexpo_5 |   3.274163   1.229877     2.66   0.008     .8636475    5.684678
  _I_Mexpo_6 |   5.309659    1.58036     3.36   0.001     2.212211    8.407107
_I_MeX_Mti~2 |    .845122   .3432575     2.46   0.014     .1723496    1.517894
_I_MeX_Mti~3 |  -.0199391    .272227    -0.07   0.942    -.5534942    .5136161
_I_MeX_Mti~4 |  -.1980574   .2522235    -0.79   0.432    -.6924065    .2962916
_I_MeX_Mti~5 |  -.3552085   .2958411    -1.20   0.230    -.9350465    .2246295
_I_MeX_Mti~6 |  -.6829444   .3431703    -1.99   0.047    -1.355546   -.0103431
------------------------------------------------------------------------------
------------------------------------------------------------------------------

Random effects (Covariance matrix)
------------------------------------------------------------------------------

symmetric e(cov_re)[3,3]
               _D          _M      _Mtime
    _D  .58751149
    _M          0   31.641699
_Mtime          0  -3.3338155   1.5358076
------------------------------------------------------------------------------
------------------------------------------------------------------------------

Level-1 variance (Marker):   9.28585
```

We can now save the new estimates and compare the last two models using the likelihood-ratio test, as follows:

```
. estimates store ConstrModel
. estimates restore FullModel
(results FullModel are active now)
. local FMll = e(ll)
. estimates restore ConstrModel
(results ConstrModel are active now)
. local CMll = e(ll)
. display -2*(`CMll´-`FMll´)
249.68
. display chi2tail(2, -2*(`CMll´-`FMll´))
6.063e-55
```

The unconstrained model has a significantly higher likelihood (likelihood-ratio chi-squared = 249.68, degrees of freedom = 2, p-value < 0.001), indicating that the correlations being tested do significantly differ from zero. The significance implies that premature termination of the marker's measurements due to AIDS or death is informative (informative drop-out mechanism).

Note that the `lrtest` command is not fully supported; it does not work when the models under comparison differ in random-effects parameters. When it is used for testing fixed-effects parameters, it is the user's responsibility to make sure both models have been fit without the `restr` option. (Likelihood-ratio tests between models that differ in their fixed-effects structure are not valid when the models have been fit with the RIGLS algorithm.)

We can also use the `mi()` option to adjust the standard errors of the fixed-effects estimates to account for the uncertainty when replacing censored survival times (due to study termination, loss to follow-up, or ART initiation) with their expectations. We will use five imputed datasets because this is the minimum recommended value.

```
. xi: jmre1 _JY _M i._Magesero*_Mtime i._Mexpo*_Mtime, remark(_M _Mtime)
> dropout(AD _D i._Dagesero i._Dexpo) timevar(_Mtime) id(study_id patient_id)
> restr mi(5)
  (output omitted)
Fixed effects
------------------------------------------------------------------------------
------------------------------------------------------------------------------
         _JY |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Drop_Out     |
          _D |   2.439713   .1248479    19.54   0.000     2.195016    2.684411
_I_Dageser~2 |  -.0308937   .1290179    -0.24   0.811    -.2837641    .2219767
_I_Dageser~3 |  -.2716915   .1437134    -1.89   0.059    -.5533647    .0099817
_I_Dageser~4 |  -.4275789   .1398424    -3.06   0.002    -.7016651   -.1534928
  _I_Dexpo_2 |   .4460437   .2210933     2.02   0.044     .0127088    .8793785
  _I_Dexpo_3 |   .1914361   .1438282     1.33   0.183     -.090462    .4733342
  _I_Dexpo_4 |  -.0339984   .1405277    -0.24   0.809    -.3094277    .2414308
  _I_Dexpo_5 |   .0656182   .1651509     0.40   0.691    -.2580715     .389308
  _I_Dexpo_6 |   .1245609   .1947826     0.64   0.523    -.2572061    .5063278
-------------+----------------------------------------------------------------
Marker       |
          _M |   24.84069   .9128988    27.21   0.000     23.05144    26.62994
_I_Mageser~2 |    .296751   .9595332     0.31   0.757      -1.5839    2.177401
_I_Mageser~3 |  -.4586846   1.025669    -0.45   0.655    -2.468958    1.551589
_I_Mageser~4 |  -.9201732    1.04611    -0.88   0.379    -2.970511    1.130164
      _Mtime |  -1.117109   .2291318    -4.88   0.000    -1.566199   -.6680188
_I_MaX_Mti~2 |  -.3308813   .2386036    -1.39   0.166    -.7985358    .1367733
_I_MaX_Mti~3 |  -.4605507   .2613587    -1.76   0.078    -.9728045     .051703
_I_MaX_Mti~4 |  -.5433615   .2703591    -2.01   0.044    -1.073256   -.0134674
  _I_Mexpo_2 |  -2.135086   1.330645    -1.60   0.109    -4.743103      .47293
  _I_Mexpo_3 |   .2751593   1.059339     0.26   0.795    -1.801108    2.351426
  _I_Mexpo_4 |   .5212439    1.00704     0.52   0.605    -1.452517    2.495005
  _I_Mexpo_5 |   3.521577   1.216155     2.90   0.004     1.137957    5.905197
  _I_Mexpo_6 |    5.28702   1.570501     3.37   0.001     2.208894    8.365146
_I_MeX_Mti~2 |   .8563533   .3386069     2.53   0.011     .1926961    1.520011
_I_MeX_Mti~3 |   .0470733   .2795631     0.17   0.866    -.5008604     .595007
_I_MeX_Mti~4 |  -.3032204    .247707    -1.22   0.221    -.7887173    .1822765
_I_MeX_Mti~5 |   -.411662   .2959009    -1.39   0.164    -.9916171    .1682931
_I_MeX_Mti~6 |  -.6364512   .3512918    -1.81   0.070    -1.324971    .0520682
------------------------------------------------------------------------------
  (output omitted)
```

The inflation of the standard errors is almost negligible in the marker model and relatively small in the drop-out model.

4.3 Postestimation

We will now use the `predict` postestimation command to informally (that is, graphically) examine the distributions of the marker's random effects (the marker's level-1 residuals) and graphically compare the previously fit constrained and unconstrained models. We start by obtaining empirical Bayes estimates of the marker's random intercept and random slope (also known as BLUPs):

```
. estimates restore FullModel
(results FullModel are active now)

. predict eb, ref
(5077 real changes made)

. describe eb*

              storage  display     value
variable name   type   format      label      variable label
───────────────────────────────────────────────────────────────────────────────

eb_M            float  %9.0g                  BLUP of Constant(M)
eb_Mtime        float  %9.0g                  BLUP of CD4 time (yrs.)(M)
```

Note the names and the variable labels of the newly created variables. To examine their distributions, we will `preserve` the dataset and keep only the first record of each individual (see figure 1). (The BLUPs are constant within each patient's block of records and missing at the last record, which corresponds to the drop-out model.)

```
. preserve

. by study_id patient_id: keep if _n==1
(4677 observations deleted)

. histogram eb_M, scheme(s2mono)
(bin=20, start=-17.303648, width=1.6328218)

. restore
```

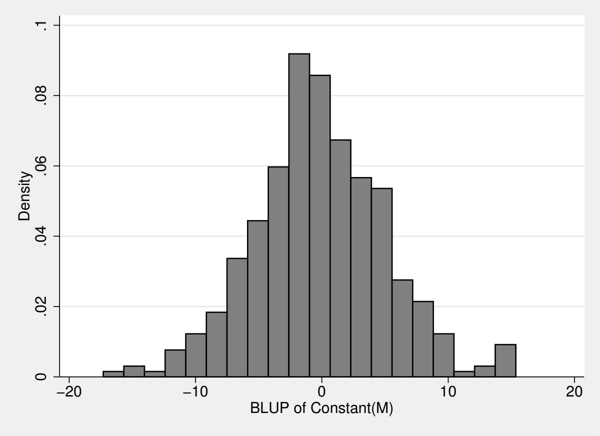

Figure 1. Histogram of BLUPs for the marker's random intercept

Similarly, we can have a histogram for the BLUPs of the marker's random slope:

```
. preserve

. by study_id patient_id: keep if _n==1
(4677 observations deleted)

. histogram eb_Mtime, scheme(s2mono)
(bin=20, start=-3.5219173, width=.33129596)

. restore
```

Or for the BLUPs of the marker's level-1 residuals:

```
. predict resid, residuals
(5077 real changes made)

. histogram resid, scheme(s2mono)
(bin=36, start=-16.156229, width=1.0067631)
```

Now we will use predicted values based on fixed-effects estimates from both the constrained and the unconstrained (full) models to graphically display (see figure 2) the predicted average evolution of CD4 cell count by age at seroconversion groups. Risk group will be kept constant to keep the graph simple. We are also using the `fillin` command to add some missing covariate patterns. Finally, note that we are back-transforming the predictions to the original scale:

```
. preserve

. drop _I*

. replace _Mtime = round(_Mtime,0.5)
(4658 real changes made)

. fillin _M _Mexpo _Magesero _Mtime

. xi  i._Magesero*_Mtime i._Mexpo*_Mtime
i._Magesero       _I_Magesero_1-4     (naturally coded; _I_Magesero_1 omitted)
i._Mag~o*_Mtime   _I_MaX_Mtim_#       (coded as above)
i._Mexpo          _I_Mexpo_1-6        (naturally coded; _I_Mexpo_1 omitted)
i._Mexpo*_Mtime   _I_MeX_Mtim_#       (coded as above)

. estimates restore FullModel
(results FullModel are active now)

. predict fe_fitted_full, xb
(7240 real changes made)

. replace fe_fitted_full = fe_fitted_full^2
(5956 real changes made)

. label var fe_fitted_full "Full model"

. estimates restore ConstrModel
(results ConstrModel are active now)

. predict fe_fitted_constr, xb
(7240 real changes made)

. replace fe_fitted_constr = fe_fitted_constr^2
(5956 real changes made)

. label var fe_fitted_constr "Constrained model"

. sort _Mexpo _Magesero _Mtime

. scatter fe_fitted_full _Mtime if _Mexpo==2 & _M, msymbol(i) connect(l)
> lpattern(solid) lcolor(gs0)|| scatter fe_fitted_constr _Mtime if _Mexpo==2
> & _M, msymbol(i) connect(l) lpattern(solid) lcolor(gs10) by(_Magesero)
> ytitle("Predicted CD4 cell count" "(cells/microL)")
> xtitle("Years since Seroconversion") ylabel(0(100)600,angle(hori))
> scheme(s2mono)

. restore
```

(Continued on next page)

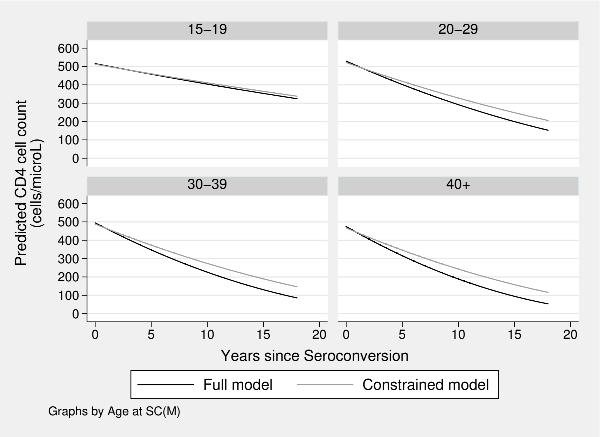

Figure 2. Average predicted CD4 cell-count evolution by age at seroconversion (risk group = heterosexuals/males); predictions based on models with ("full") or without ("constrained") adjustment for informative drop-outs

You can see that the constrained model, which does not make any adjustment for the informative drop-out, estimates less steep CD4 cell-count declines than does the unconstrained model. The differences between the two models are more pronounced at older ages, where informative drop-out is more severe (that is, more persons die or develop AIDS).

5 Conclusion

We presented the new Stata `jmre1` command, which implements the joint multivariate random-effects model (Touloumi et al. 1999). The model adjusts the estimates for the longitudinal evolution of a continuous marker when the series of its measurements are subject to informative drop-out. That is, the model adjusts the estimates when the series of the marker's measurements is prematurely terminated by an event whose probability is related to subject-specific parameters of the underlying marker's trend.

The estimation procedure is based on a combination of the EM algorithm with the RIGLS or IGLS method and is implemented mainly using Stata's matrix language, Mata. It has been shown that the model itself, if correctly specified, yields estimates of minimal bias with close to nominal coverage probabilities (Touloumi et al. 1999; Touloumi et al. 2003). It has also been shown that an expanded bivariate version of the same model (that is, two continuous markers modeled simultaneously in the presence of informative drop-out) is fairly robust to at least moderate deviations of its distributional assumptions (Pantazis and Touloumi 2007); therefore, we believe that the current model will be equally robust.

The Stata implementation we presented in this article is much more user friendly compared with the initial implementation (see the appendix in Touloumi et al. [2002]) in the MLn statistical software package. We have also provided a utility command (`datajoint1`) that facilitates the required data preparation.

During the development of these commands, we compared `jmre1` with the prior implementation of the same model in MLn; differences in the estimates were in the order of the convergence tolerance. We also performed simulations where we observed that if the drop-out is not informative, the results for the marker model from `jmre1` are coinciding with those obtained by `xtmixed`. Similarly, when we constrained the correlations between the random effects of the marker model and the residuals of the drop-out model to zero in the `jmre1` command (that is, forcing the model to ignore the informative drop-out effects), the obtained results for the marker model were the same as those produced by a corresponding random-effects model fit through the `xtmixed` command (results not shown).

However, our implementation of the RIGLS algorithm that was required in the estimation procedure is not fully optimized; thus convergence with big datasets can be slow. For example, the fit of the full model in the previously given example dataset (400 individuals with 4,677 marker measurements in total) required approximately 45 seconds using an Intel Core 2 Quad Q9450 PC clocked at 2.66 GHz. We plan to fully optimize the code in an updated version using all the computational details given in Goldstein and Rasbash (1992). We also plan to expand our implementation to the bivariate version of the JMRE model (Pantazis and Touloumi 2005) to allow simultaneous modeling of two continuous markers in the presence of informative drop-out.

6 Acknowledgments

We would like to thank the steering committee and all participants of the CASCADE collaboration for allowing us to use part of the CASCADE data for the illustration of `jmre1`'s use.

7 References

CASCADE Collaboration. 2003. Differences in CD4 cell counts at seroconversion and decline among 5739 HIV-1-infected individuals with well-estimated dates of seroconversion. *Journal of Acquired Immune Deficiency Syndromes* 34: 76–83.

Diggle, P., and M. G. Kenward. 1994. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, Series C* 43: 49–93.

Faucett, C. L., and D. C. Thomas. 1996. Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine* 15: 1663–1685.

Goldstein, H. 2003. *Multilevel Statistical Models*. 3rd ed. London: Arnold.

Goldstein, H., and J. Rasbash. 1992. Efficient computational procedures for the estimation of parameters in multilevel models based on iterative generalised least squares. *Computational Statistics & Data Analysis* 13: 63–71.

Johnson, N. L., S. Kotz, and N. Balakrishnan. 1970. *Continuous Univariate Distributions*, vol. 1. New York: Wiley.

Johnson, R. A., and D. W. Wichern. 2007. *Applied Multivariate Statistical Analysis*. 6th ed. Upper Saddle River, NJ: Prentice Hall.

Laird, N. M. 1988. Missing data in longitudinal studies. *Statistics in Medicine* 7: 305–315.

Laird, N. M., and J. H. Ware. 1982. Random-effects models for longitudinal data. *Biometrics* 38: 963–974.

Little, R. J. A. 1995. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 90: 1112–1121.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: Wiley.

Louis, T. A. 1982. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* 44: 226–233.

Pantazis, N., and G. Touloumi. 2005. Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs. *Journal of the Royal Statistical Society, Series C* 54: 405–423.

———. 2007. Robustness of a parametric model for informatively censored bivariate longitudinal data under misspecification of its distributional assumptions: A simulation study. *Statistics in Medicine* 26: 5473–5485.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2002. Multilevel selection models using gllamm. Combined Dutch and German Stata Users Group meeting proceedings. http://www.stata.com/meeting/2dutch/select.pdf.

Rubin, D. B. 1978. Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20–34.

Schluchter, M. D. 1992. Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine* 11: 1861–1870.

Touloumi, G., A. G. Babiker, M. G. Kenward, S. J. Pocock, and J. H. Darbyshire. 2003. A comparison of two methods for the estimation of precision with incomplete longitudinal data, jointly modelled with a time-to-event outcome. *Statistics in Medicine* 22: 3161–3175.

Touloumi, G., S. J. Pocock, A. G. Babiker, and J. H. Darbyshire. 1999. Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. *Statistics in Medicine* 18: 1215–1233.

———. 2002. Impact of missing data due to selective dropouts in cohort studies and clinical trials. *Epidemiology* 13: 347–355.

Wu, M. C., and K. R. Bailey. 1989. Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics* 45: 939–955.

Wu, M. C., and R. J. Carroll. 1988. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 44: 175–188.

About the authors

Nikos Pantazis has a PhD in biostatistics and works as a research collaborator in the Department of Hygiene, Epidemiology, and Medical Statistics at the University of Athens Medical School, Greece.

Giota Touloumi is an associate professor of biostatistics in the same department. Both share an interest on longitudinal data modeling, missing-data problems, and HIV epidemiology.