

# Longitudinal Data with Informative Dropout

This document contains the MATLAB m-file documentation for performing the analyses described in:

1. Guo W., Ratcliffe S.J. and Ten Have T.R. (2004) "A Random pattern-mixture model for longitudinal data with dropouts." Journal of the American Statistical Association.
2. Ratcliffe S.J., Guo G. and Ten Have T.R. (2004) "Joint modelling of longitudinal and survival data via a common frailty." Biometrics.

These m-files were written by Sarah Ratcliffe. Any questions or comments should be emailed to [sratclif@cceb.upenn.edu](mailto:sratclif@cceb.upenn.edu). The programs provided are as is and may be used provided the use is referenced.

The two main m-files that should be run are `rpat0` (for paper 1) and `surv` (for paper 2). The detailed description for each function is organised as follows:

- name of function
- a statement of purpose
- a synopsis of the function's syntax
- a description of what the function does

and, for selected functions,

- examples
- a detailed description of the mathematical procedure/equations comprising the function.

The remaining functions are utilities that are used by these two main functions.

---

**rpat0**

**Purpose** Calculates the random pattern-mixture model for longitudinal data with informative dropout.

**Syntax** `[out,stderr] = rpat(data,X,Z,W,patt,drop);`  
`[out,stderr] = rpat(data,X,Z,W,patt,drop,g);`  
`[out,stderr] = rpat(data,X,Z,W,patt,drop,g,maxit,tol,ind);`

**Description** `[out,stderr] = rpat(data,X,Z,W,patt,drop);` for longitudinal response data given in `data`, with corresponding fixed effects and subject level random effects design matrices `X` and `Z`, pattern level random effects design matrix `W`, pattern/grouping information given in `patt` and dropout information given in `drop`, returns the parameter estimates for the model in `out` and their corresponding standard errors in `stderr`.

The `data = [i Y]`, where `i` is the subject number and `Y` is the response variable, contains the responses for each subject stacked on top of one another. For each row in `data`, the corresponding fixed effects, subject level random effects and pattern level random effects design vectors are given in the same row in `X`, `Z`, and `W`, respectively. The grouping/pattern information for each subject is given in `patt`. This must be ordered in the same order as the subjects appear in `data`. The information for determining the survival time model is given in `drop = [delta Ti x2]`, where `delta` indicates if the subject informatively dropped out of the study (1) or was censored (0), `Ti` is the time at which this happened, and `x2` contains the fixed effect design vector for each subject's survival time (one row per subject).

`out` contains the parameter estimates for  $\alpha_1$  (`out.alpha1`),  $\Sigma_\beta$  (`out.SigBeta`),  $\Sigma_u$  (`out.Sig2u1`),  $\sigma_e^2$  (`out.sig2e`),  $\alpha_2$  (`out.alpha2`),  $b$  (`out.b`), and  $s^2$  (`out.sig2R`). The estimates are found using the EM algorithm. The corresponding standard errors for the fixed effects ( $\alpha_1$ ,  $\alpha_2$ ) are given in `stderr` and are calculated using the method of Louis (1982).

The initial parameter values should be stored in the MATLAB database file `rpatinit.mat`. If this file doesn't exist, then initial values will be set to ones and identity matrices.

`[out,stderr] = rpat(data,X,Z,patt,drop,g)` optionally specifies a link function for the times `Ti`. The choices are: 1 for a linear link function, and 2 for a log link function (default).

Further, `maxit` specifies the maximum number of iterations (default is 100), and `tol` the tolerance used for determining convergence (default is  $1e-5$ ). `ind` optionally specifies the output to be displayed; 0 = estimates at each iteration (default), 1 = iteration number only, 2 = iteration number plus final estimates for different values of `tol`, 3 = no output.

**Examples** First, create a MATLAB database file containing the initial values.

```
Sigu1 = 0.4602; SigBeta = 18.7549; sig2e=29.0325; sig2u = 0.01;
alpha1 = [16.4075; -0.5674; -0.2364; -2.0739];
alpha2 = [3.74096; -0.53108]; sig2R = 0.69749; b = -0.0423;
save rpatinit sig2u sig2R alpha2 alpha1 SigBeta b sig2u sig2e;
```

Set up the data as needed for the analysis with a random subject level intercept and pattern level intercept and a log link function. Then run the analysis.

```
Y = [ids outcome];
X = [ones(length(outcome),1) x11 x12 x13];
Z = [ones(length(outcome),1)]; W = Z;
d = [delta Ti ones(n,1) x21];
[out,std] = rpat0(Y,X,Z,W,patt,d,2)
```

**Model** For further details see "A Random Pattern-Mixture Model for Longitudinal Data with Dropouts".

$$y_{ij} = X_{ij1}\alpha_1 + Z_{ij}\beta_{ij} + W_{ij}u_i + e_{ij}$$

$$R_{ij} = \mathbf{x}_{ij2}^T\alpha_2 + b^T u_i + \varepsilon_{ij}$$

where  $\beta_{ij} \sim N(\mathbf{0}, \Sigma_\beta)$ ,  $u_i \sim N(0, \Sigma_u)$ ,  $e_{ij} \sim N(0, \sigma_e^2)$ , and  $\varepsilon_{ij} \sim N(0, s^2)$ . Note:  $y_{ijk}$  is the  $k$ -th measurement on the  $j$ -th subject within the  $i$ -th group/pattern. Also,  $i = 1 \dots m$ ,  $j = 1 \dots n_i$ , and  $k = 1 \dots n_{ij}$ . The parameter estimates  $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{b}, \hat{\Sigma}_\beta, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{s}^2)$  are estimated using the EM algorithm.

## surv

**Purpose** Performs joint modelling of longitudinal and survival data via a common frailty.

**Syntax** `[out,stderr] = surv(data,X,Z,patt,drop);`  
`[out,stderr] = surv(data,X,Z,patt,drop,maxit,tol,ind);`

**Description** `[out,stderr] = surv(data,X,Z,patt,drop);` for longitudinal response data given in `data`, with corresponding fixed effects and subject level random effects design matrices `X` and `Z`, pattern/grouping information given in `patt` and survival information given in `drop`, returns the parameter estimates for the model in `out` and their corresponding standard errors in `stderr`.

The `data = [i Y]`, where `i` is the subject number and `Y` is the response variable, contains the responses for each subject stacked on top of one another. For each row in `data`, the corresponding fixed effects and subject level random effects design vectors are given in the same row in `X` and `Z`, respectively. The grouping/pattern information for each subject is given in `patt`. This must be ordered in the same order as the subjects appear in `data`. The information for determining the survival (hazard) model is given in `drop = [delta Ti x2]`, where `delta` indicates if the subject informatively dropped out of the study (1) or was censored (0), `Ti` is the time at which this happened, and `x2` contains the fixed effect design vector for each subject's survival time (one row per subject).

`out` contains the parameter estimates for  $\alpha_1$  (`out.alpha1`),  $\Sigma_\beta$  (`out.SigBeta`),  $\sigma_u^2$  (`out.sig2u`),  $\sigma_e^2$  (`out.sig2e`),  $\alpha_2$  (`out.alpha2`),  $b$  (`out.b`), and the cumulative baseline hazard  $\Lambda_0$  (`out.Lambda0`). The estimates are found using the EM and Newton-Raphson algorithms.

The corresponding standard errors are given in `stderr` and are calculated using the method of Louis (1982).

The initial parameter values should be stored in the MATLAB database file `survinit.mat`. If this file doesn't exist, then initial values will be set to ones and identity matrices.

`[out,stderr] = surv(data,X,Z,patt,drop,maxit)` optionally specifies the maximum number of iterations (default is 100), and `tol` specifies the tolerance used for determining convergence (default is 1e-5). `ind` optionally specifies the output to be displayed; 0 = estimates at each iteration (default), 1 = iteration number only, 2 = iteration number plus final estimates for different values of `tol`, 3 = no output.

**Examples** First, create a MATLAB database file containing the initial values.

```
SigBeta = 0.4684; sig2u = 0.0744; sig2e = 0.8665;
alpha1 = [6.5233; 0.03713; 0.00706; 0.1376;];
alpha2 = [-0.0204; 0.0403]; b = -0.07418;
save survinit alpha1 alpha2 SigBeta sig2u sig2e b;
```

Set up the data as needed for the analysis with a random subject level intercept. Then run the analysis.

```
Y = [ids outcome];
X = [ones(length(outcome),1) x11 x12 x13];
Z = [ones(length(outcome),1)];
d = [delta Ti ones(n,1) x21];
[out,std] = surv(Y,X,Z,patt,d)
```

**Model** For further details see "Joint Modeling of Longitudinal and Survival Data via a Common Frailty".

$$y_{ij} = X_{ij1}\alpha_1 + u_i + \varepsilon_{ij}^*$$

$$\lambda_{ij}(t|\mathbf{x}_{ij2}, u_i) = \lambda_0(t)e^{x_{ij2}^T\alpha_2 + bu_i}$$

where

$$\varepsilon_{ij}^* = \mathbf{Z}_{ij}\boldsymbol{\beta}_{ij} + \varepsilon_{ij}$$

$$u_i \sim N(0, \sigma_u^2)$$

$$\boldsymbol{\beta}_{ij} \sim N(0, \Sigma_\beta)$$

$$\varepsilon_{ij} \sim N(0, \sigma_e^2)$$

and  $\lambda_0(t)$  is the baseline hazard function. Also,  $t_{ij} = \min(v_{ij}, c_{ij})$ , the minimum of the survival and censoring times. Note:  $y_{ijk}$  is the  $k$ -th measurement on the  $j$ -th subject within the  $i$ -th group/pattern. Also,  $i = 1 \dots m$ ,  $j = 1 \dots n_i$ , and  $k = 1 \dots n_{ij}$ .

**See Also** `rpat`